The Vocabulary of MEDICAL ENGLISH

A CORPUS-BASED STUDY

Renáta Panocová

The Vocabulary of Medical English

The Vocabulary of Medical English:

A Corpus-based Study

By

Renáta Panocová

Cambridge
Scholars
Publishing

The Vocabulary of Medical English: A Corpus-based Study

By Renáta Panocová

This book first published 2017

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-9578-4
ISBN (13): 978-1-4438-9578-1

Table of Contents

LIST OF FIGURES

Figure 2-1 The relationship between general vocabulary, specialized vocabulary, terms, and terms in the narrow sense based on ten Hacken (2010, 2015)

LIST OF TABLES

ACKNOWLEDGEMENTS

This monograph is based on my habilitation thesis submitted and defended in 2016 at the Faculty of Arts, Pavol Jozef Šafárik University in Košice, Slovakia. I would like to express my sincere thanks to the members of the habilitation committee, Prof. PhDr. Pavel Stekauer, DrSc. (chair), Assoc. Prof. Dr. José Igor Prieto-Arranz, and doc. PhDr. Edita Kominarecová, PhD, and to the reviewers of my habilitation thesis, Univ.-Prof. Dr. Pius ten Hacken, dr. hab. Maria Bloch-Trojnar, and doc. PhDr. František Šimon, CSc. The constructive comments made in the reviews and during the discussion as well as the suggestions some of them gave me afterwards were instrumental in the revision leading to the current version of the text. All remaining errors are of course my own.

CHAPTER ONE

INTRODUCTION: DEFINING MEDICAL ENGLISH

This monograph explores the vocabulary of medical English from a corpus-based perspective. In investigating medical corpora, I will highlight the question of the characterization of medical vocabulary in English. One of the central issues I address is how to design a methodology appropriate for the purpose of description of medical English. This contrasts with the pedagogical perspective, which uses medical corpora for compiling word frequency lists used as a basis for developing teaching materials. The difference in aims of the two perspectives, pedagogical and descriptive, points to the need for a different methodology to be applied in research of the phenomenon referred to as *medical language* or *medical English* (ME).

Language is an important tool in professional communication in medicine. The history of medicine clearly points to Latin as a dominant language in the field throughout the middle ages and the early modern era, when it was the main international language not only in medicine, but also in religion and philosophy (Fischbach, 1993: 94). Even today, the influence of Latin in medical language should not be ignored. Several textbooks in medicine and medical terminological dictionaries in a number of different languages take Latin as a basis, for instance Vojteková's (2015a) trilingual dictionary of anatomical terms in Latin, Slovak, and Polish. A discussion of the importance of Latin in current medical terminology is given in Vojteková (2015b).

From the 17th century onwards a new tendency for the use of national languages such as German, French, and English in medical writings emerged (Ferguson, 2013: 282). The relatively equal status of German, French, and English changed in the second half of the 20th century, resulting in English taking over the most prominent role in medical texts. A piece of evidence comes from Maher (1986), who reports that in 1980 72.2% of the articles in the *Index Medicus* database were published in English. A similar tendency is observed by Giannoni (2008) who reports that more than 99% of medical journal papers by Italian authors are in

English. Gunnarsson's (2009) findings confirm this tendency in Scandinavian countries.

In this context it is interesting to report some results from a search in Google. A Google search for *medical English* gives 1 030 000 000 hits in half a second.[1] The top hits include medical English online exercises and games, medical English worksheets and teaching resources, exercises for doctors and their patients, reading and listening exercises for medical workers, offers for courses of medical English. Many of these webpages promise improvement of communication in a medical environment by mastering ME, not only for native speakers of English but also for doctors, nurses, and other health-care workers with a native language other than English. This also demonstrates the importance of ME at the international level.

The prominent role of English in medicine raises at least two important questions; the first addresses the nature of ME and its definition and the second concentrates on the position of ME in relation to general English. Two main perspectives have been adopted in determining the notion of ME. Firstly, ME can be defined in terms of the distinction from other language variants and common or general language, as in Lankamp (1989). The second perspective considers ME as a sublanguage.

A central hypothesis in Lankamp's (1989) research is that ME is so distinct from general language "that it would have to be acquired or learnt (two interchangeable terms in here) by language users with only general language knowledge" (1989: 14). Then he raises the question in what sense ME differs from general language or other language variants. In his study, Lankamp (1989: 14) focuses on "the investigation of the ways in which written English medical language differs on the various linguistic levels of analysis (discourse, syntax, semantics, lexicon and morphology) from other English written language variations". Following Hudson (1980) and Picht and Draskau (1985), Lankamp (1989: 20) views ME in terms of register and language for special purposes. On this basis Lankamp (1989: 20-22) distinguishes the five dimensions of variation given in (1).

(1) a. medical specialism
b. manner of transmission of medical language
c. relations between participants in medical exchange
d. communicative purpose
e. national language

[1] Retrieved 29 January, 2016

The dimensions in (1a-c) are in line with Picht & Draskau's (1985) discussion of terminology and specialized language, the remaining two are added by Lankamp (1989: 22). The dimension in (1a) highlights the importance of linguistic differences among different medical specializations or professional groups. These are similar to linguistic differences between ME and other language variations. In (1b) it is emphasized that important differences are expected between spoken medical language and written medical language. For instance, medical jargon or slang used in spoken medical context will not occur in research articles in respectable medical scientific journals. The label *tenor* is sometimes used to refer to the dimension in (1c). It indicates that linguistic properties may differ depending on the roles of participants, e.g. doctor-doctor conversation, doctor-patient conversation, doctor-nurse conversation, equipment manufacturer-doctor, etc. (Lankamp, 1989: 21). An additional dimension in (1d) highlights "the function of a language for special purposes-type register to communicate information of a specialist nature at any level of complexity in the most economic, precise and unambiguous terms possible, i.e. as efficiently as possible, especially in the expert-to-expert tenor" (Lankamp, 1989: 22; cf. also Sager et al. 1980: 290-291). In this dimension the role of terminology based on the need for precise, and preferably non-synonymous language items to label relevant concepts is crucial. The central point of (1e) is the fact that "medical language is differentiated according to specific national languages expressing international medical concepts. In this dimension, medical language is differentiated in medical Dutch, medical English, medical French etc." (Lankamp, 1989: 22).

Lankamp (1989: 23) suggests that defining ME as a type of register is fully compatible with the lexical competence of a language user in the psycholinguistic model he presents in his book. However, it should be noted that English for Specific Purposes (ESP) and register are not one and the same phenomenon. Biber and Conrad (2009: 3) in their book *Register, genre, and style* explain that "ESP focuses on description of the language used in registers/genres from a particular profession or academic discipline, e.g. biochemistry or physical therapy". It is important to emphasize that Biber and Conrad (2009: 2) use the terms *register*, *genre*, and *style* to refer to three different perspectives on text varieties.

It is understood that "the register perspective combines an analysis of linguistic characteristics that are common in a text variety with analysis of the situation of use of the variety" (Biber and Conrad, 2009: 2). This means that linguistic features such as pronouns and verbs are functional, and, therefore their use depends on situational context and a type of communicative purpose.

The genre and style perspectives each concentrate on similarities and differences with respect to the register perspective. The property shared by the genre perspective and the register perspective is that the purposes and situational context of a text variety can be immediately identified. By contrast, linguistic analysis concentrates on the conventional structures used to build a text within the variety, for example, the conventional way in which a letter begins and ends (Biber and Conrad, 2009: 2).

Finally, the style perspective is similar to the register perspective in its linguistic focus. This means that linguistic features occurring in a particular variety are analysed. The crucial difference from the register perspective is "that the use of these features is not functionally motivated by the situational context; rather, style features reflect aesthetic preferences, associated with particular authors or historical periods" (Biber and Conrad, 2009: 2).

In the history of the development of ESP, Hutchinson and Waters (1987: 13) identify an initial stage in which, they say, "the analysis had been of the surface forms of the language" in the form of register analysis, that is, the study at the sentence level of the use of language in different communicative settings, such as the language used by nurses, airplane mechanics, and bank tellers. In this stage, the teaching of reading received minimal attention. It was at the next stage of development that reading pedagogy in ESP took major steps forward: "Whereas in the first stage of its development, ESP had focused on language at the sentence level, the second phase of development shifted attention to the level above the sentence, as ESP became closely involved with the emerging field of discourse or rhetorical analysis" (Hutchinson and Waters, 1987: 10). A more detailed discussion of the role of reading in ESP from various perspectives can be found in Hirvela (2013). To sum up, ME may be approached from the register perspective, but a register perspective leads to a much more fine-grained division. ME is not a register, but a range of registers. ME covers for instance, research papers, doctor-patient conversation, and Patient Information Leaflets with instructions on the use of pharmaceuticals. In line with Biber and Conrad (2009: 32) "there is no single "right" level for a register analysis". Obviously, even in the register of medical research articles more specific subregisters can be identified, e.g. there may be some degree of variation between research articles in the domain of psychiatry and cardiology. The same applies to doctor-patient communication in different situational contexts such as medical examination, surgery or disclosure that a patient suffers from a lethal disease. As Biber and Conrad (2009: 33) emphasize, differences between registers can be viewed as a continuum of variation.

Ten Hacken and Panocová (2015: 2-3) note that it is not common to ask whether someone speaks medical English as opposed to the question whether someone speaks English. Similar to Lankamp (1989: 22) they point to the fact that medical language is language-specific and medical English differs from medical Slovak or medical Dutch. They also raise the question of the relationship between English medical language and general English, but also of the relationship with Dutch and Slovak medical language (ten Hacken and Panocová, 2015: 3).

An alternative is to approach medical language from the perspective of sublanguages (Harris, 1968; Kittredge, 1987; Lehrberger, 2014). This means that English medical language is taken to be a sublanguage of English. In the first perspective, ME is like a register of English. In the second one, it is considered as a subset. In either perspective, the vocabulary of a particular area of study or professional use, for instance medicine, is an example of specialized vocabulary.

Harris (1968) introduced the notion of *sublanguage* in linguistics "by analogy with *subsystem* used in mathematics" (Lehrberger, 2014: 20). A sublanguage is viewed as a theoretical construct. Lehrberger (2014: 20) points out that whereas in mathematics "a subsystem can be readily defined in terms of restrictions on the sets and operations of the system of which it is a part", in a natural language and its sublanguage "the relation between part and whole is not so clear-cut". This is in line with Kittredge's (1982: 110) observation that "[i]n considering which samples of specialized language can be regarded as representing "genuine" sublanguages we are immediately faced with the lack of an empirically adequate definition of the term" and there is a need for more precise delimitation criteria. Kittredge (1982: 110) claims that "the closure property proposed by Harris (1968) is not in itself sufficient to resolve this question" and the main reason is that "[i]f a sublanguage can be *any* subset of sentences which is closed under the transformational operations, this definition could identify a very large number of linguistic subsets as sublanguages". In mathematics, *closure* refers to the property that if a particular operation is applied to members of a set, the result will always again be a member of that set. Kittredge (1982: 110) understands the closure property as a necessary condition. This means that even if we have a set of sentences which can be considered as a sublanguage, "we must include in it all sentences generated from the candidate set by means of transformational operations of negation, question formation, clefting, conjunction, etc." (Kittredge, 1982: 110). It is important to add that Harris (1968) points out that a linguistic subsystem can be closed only under some, but not all of the operations.

Another type of closure is vocabulary closure. This has been investigated by McEnery and Wilson (2001) and Temnikova (2013). They examine relationships between types and tokens in a corpus of the genre. If a genre shows closure properties, the number of types stops growing after some number of tokens has been processed. On the other hand, if it does not exhibit closure, then the number of types will continue to rise continually as the number of tokens increases (Temnikova, 2013: 72).

Kittredge (1983: 49) repeatedly emphasizes that although sublanguages have been investigated in a number of ways and perspectives, "there is no widely accepted definition of the term", but there is an agreement about certain factors that are usually present in a subset of a particular natural language and are essential for semantic processing. The factors mentioned by Kittredge (1983: 49) are presented in (2).

(2) a. restricted domain of reference
b. restricted purpose and orientation
c. restricted mode of communication
d. community of participants sharing specialized knowledge

In (2a) the main point is that linguistic expressions refer to a set of objects and relations and their number is relatively small. In (2b) it is emphasized that there are clearly identifiable relationships among participants. The same applies to the goals of the exchange. The factor in (2c) indicates that there are differences not only between spoken and written communication, but "there may be constraints on the form because of "bandwidth" limitations (e.g. telegraphic style)" Kittredge (1983: 49). In medicine one could think of prescriptions. The last factor in (2d) suggests it is easier to determine properties of a sublanguage if a community of users who share it can be identified. This is important in order to determine characteristic patterns of usage which contribute to a complete characterization of a sublanguage as a linguistic system.

If we compare the lists of features in (1) and (2) we can see that some of the features are shared by the register perspective and sublanguage perspectives. It is obvious that taken together they must overlap to a certain extent. A striking difference is that (1e) is not explicitly mapped in (2). What the two perspectives share is that a specialized language is central in a more in-depth research. The main difference between the two perspectives is that register is about the use of competence whereas sublanguage concentrates on a subset of a language.

Generally, it is well-known that language users on advanced and proficient levels must have implicit knowledge about register, word meaning, and lexical and grammatical patterns, because otherwise it would not be possible to write and speak appropriately (Nesi, 2013: 451). It is also obvious that language users have intuitions about language, but it is questionable whether and when they are reliable or misleading. According to Sinclair (1991: 4) "human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use [it]". This may be seen as a sufficiently strong argument for the use of corpora and especially specialized corpora in the research into specialized languages, e.g. medical language. At present, large corpora for English are available for online search of a number of linguistic features, for instance the Corpus of Contemporary American English (COCA), the British National Corpus (BNC), etc. There is a strong tendency for other languages to compile similar national corpora, e.g. the Russian National Corpus, the Slovak National Corpus, the National Corpus of Polish.

Tognini-Bonelli (2001) differentiates between *corpus-based* research and *corpus-driven* research. Corpus-based research makes it possible to verify intuitions a researcher has about language use, whereas corpus-driven research uses corpus data to formulate relevant observations and generalizations about language use. Although it is always necessary to start from a theoretically informed research question, the present research assigns corpora a much more central role than as only a tool to test intuitions. Therefore it should be considered as corpus-driven in Tognini-Bonelli's sense.

Against this background, it is possible to raise the most relevant questions that guide the research in the remaining chapters:

- How can the structure of medical vocabulary in English be determined on the basis of a specialized corpus?
- How does the choice of a particular perspective (pedagogical versus characterizing/descriptive) influence the methodology of corpus-based research?
- How does the text type influence the structure of medical vocabulary?
- How does the choice of a corpus influence the results?

The monograph is divided into four chapters and a conclusion. After this introduction, chapter 2 presents an overview of previous efforts to characterize and determine medical English. Defining the key notions of *lemma/lexeme*, *word family*, *specialized vocabulary*, and *terminology* and the relationships between them are central issues which are addressed. Then, an overview of methods relevant for identifying ESP vocabulary with an emphasis on medical English is given. The role of corpora and specialized corpora in determining the vocabulary of medical English is discussed in detail. Word lists of academic vocabulary by Coxhead (2000) and Gardner and Davies (2013) and of medical vocabulary by Wang et al. (2008) are described. All these word lists are based on specialized corpora of academic texts and medical research journal papers. The methodologies applied in these word lists are compared and critically evaluated.

In chapter 3 I argue that the methodology used in a pedagogical approach, which results in medical word lists, is neither sufficient nor adequate if the main aim is to characterize or describe medical vocabulary and modifications of methodology are suggested. First, the chapter explains why it is reasonable to use The Corpus of Contemporary American English (COCA) to find answers to the above mentioned research questions. Then, the chapter concentrates on the description of the medical subcorpus ACAD: Medicine in COCA. It discusses how the structure of the medical corpus influences the characterization of medical vocabulary. Finally, an overview of the procedure applied to arrive at the characterization of medical vocabulary is presented. It explains why it is better to approach medical vocabulary from the perspective of a cline or continuum based on two dimensions: absolute and relative frequency. Determining the threshold values for each of these dimensions is a crucial decision. It also demonstrates what effect the different threshold values might have on the structure or description of medical vocabulary in English. The chapter concludes by presenting a model of medical vocabulary as a two-dimensional continuum based on the interaction of absolute frequency and relative frequency.

Chapter 4 compares the results based on the subcorpus of medicine in COCA with an alternative corpus of medical texts, a specially compiled corpus for illustrative purposes. This medical corpus is based on the Wikipedia corpus, which was made available as a supplement to COCA in 2015. The Wikipedia corpus is based on the full text of the English version of the Wikipedia at a particular point in time and it contains 4.4 million Wikipedia articles with 1.9 billion words. The Wikipedia articles on medical topics represent a different type of medical text to medical journal articles. The chapter compares the results based on the two specialized corpora and evaluates their usefulness with respect to the characterization of medical vocabulary in English.

Finally, the conclusion summarizes the most relevant findings of the previous chapters and indicates their significance. On one hand, it is argued that the perspective to ME adopted here contributes to a better understanding of language use in medical communication. On the other, lines of further research are outlined.

CHAPTER TWO

DETERMINING THE VOCABULARY OF MEDICAL ENGLISH

A central question in any subfield of English for Specific Purposes (ESP) is how it relates to the lexicon. Johns (2013: 23) points out that this issue has been discussed since the early years (1961-1982) of the history of the ESP research by the *Washington School*. The main representatives of this school, John Lackstrom, Larry Selinker, and Louis P. Trimble, made the relationships of the grammar and lexicon of English for science and technology with the authors' rhetorical purposes central in their research (see e.g. Lackstrom et al. 1972). Since then, it has continued to be in the focus of ESP research. Although in most ESP research, the main aim is pedagogical, with emphasis on an accurate definition of which vocabulary ESP learners need for their professional communication, the approach also triggers interesting theoretical questions about the lexicon. These are the main focus of this chapter.

The terminology that is used to refer to ESP vocabulary includes the terms *specialized*, *technical*, *sub-technical*, and *semi-technical vocabulary* (Coxhead, 2013: 141). The term *sub-technical* vocabulary is used by Cowan (1974). Farrell (1990) prefers the term *semi-technical vocabulary*. According to Coxhead "such terms usually refer to the vocabulary of a particular area of study or professional use" (2013: 141). This shows the importance of defining the key notions of *general vocabulary*, *specialized vocabulary*, and *terms in the narrow sense* and the relationships between them. The definition of these terms is addressed in section 2.1. Section 2.2 gives an overview of methods relevant for identifying ESP vocabulary with an emphasis on medical English. It discusses in detail the role of corpora in determining the vocabulary of medical English. In 2.3 the importance of corpus-based methods applied in identifying ESP vocabulary is emphasized and Coxhead's (2000) Academic Word List (AWL) is described in more detail. In section 2.4 I will present the Medical Academic Word List (MAWL) by Wang, Liang and Ge (2008). A more recent contribution, the New Academic Vocabulary List (AVL) by Gardner and Davies (2013), is discussed in 2.5.

2.1 Defining specialized vocabulary

The vocabulary of medical English clearly belongs to a specialized professional area. This means that non-professionals might not have knowledge of medical vocabulary or at least of the specialized senses of vocabulary items relevant in a medical professional environment. On the other hand, there is a certain degree of overlap with general vocabulary. This raises crucial questions about specialized vocabulary and its relationship to words and terms.

Coxhead (2013: 141) emphasizes that “the range of a word is important in ESP. That is, a specialized word would have a narrow range of use within a particular subject area”. She distinguishes three types of specialized words: words of Greek or Latin origin, highly technical words, and words used also in general language (Coxhead, 2013: 141). These three different types of specialized words are exemplified in (1).

(1) a. malleus
b. trocar
c. jacket

The first type of specialized words based on Coxhead (2013) includes words with Greek or Latin elements. It is exemplified in (1a). The specialized word (1a) is of Latin origin and means hammer in the sense of ‘the largest ossicle of the three auditory ossicles’ (Stedman, 1997). The word in (1b) is a highly technical word and it represents the second class of specialized words distinguished by Coxhead (2013). The meaning of (1b) is ‘an instrument for withdrawing fluid from a cavity, or for use in paracentesis’ (Stedman, 1997). A corpus search confirms that (1b) is a highly specialized technical word, the Corpus of Contemporary American English (COCA) gives 12 occurrences, and the British National Corpus (BNC) only 1.[1] This also means that only experts are likely to store the meaning of (1b) in their mental lexicon. In (1c) we see the third type in line with Coxhead (2013), a word which is used in much narrower senses in medicine than in general English. For (1c), Stedman (1997) gives two senses typical of medicine. In one sense it may mean ‘a fixed bandage applied around the body in order to immobilize the spine’ (Stedman, 1997)

[1] Corpus results were retrieved on 30 July, 2015 from COCA.

whereas in dentistry, it means ‘an artificial crown composed of fired porcelain or acrylic resin’ (Stedman, 1997).

Especially the third type of specialized vocabulary, exemplified in (1c) has been the main object of a number of research studies such as those by Crawford Camiciottoli (2007), Nation (2008), etc. The top ten word list in business studies by Crawford Camiciottoli (2007) includes *price*, *work*, and *market*, which are frequently used also in common contexts of general language use. In medical vocabulary, Nation (2008) reports that *neck* and *by-pass* occur frequently. However, they are also frequent in general lexis but in different senses, e.g. a *city by-pass* or a *bottleneck*. According to Coxhead (2013: 151), the question of polysemy of ESP and its vocabulary is a challenging issue in a pedagogical perspective. In her view, “new technical meaning requires […] learners to build their knowledge of both the concept of a word and its meaning” (Coxhead, 2013: 151).

Many ESP researchers identified another essential problematic question related to specialized vocabulary. Specialized vocabulary is dynamic and develops rapidly. It is important that the fast progress in specialized vocabulary development is reflected in teaching material. This is the main reason why, for instance, Crawford Camiciottoli (2007) questions the correspondence of specialized vocabulary between professional texts and university level texts.

A different view of specialized words combining lexicographic and terminological perspectives can be found in ten Hacken (2008, 2010, 2015). He discusses the relationship between *general vocabulary*, *specialized vocabulary*, and *terms*. Ten Hacken’s approach is based on a theory of prototypes (e.g. Labov, 1973) and preference rules formulated by Jackendoff (1983). Labov’s experiment with the concept of *cup* is a classical demonstration of the fact that a judgement whether a particular object is a cup or not is prototype-based. The informants had a stronger tendency to reject the label cup for an object which was further removed from the prototype. Scalar conditions and preference rules determine the distance from the prototype. Ten Hacken (2010: 917) gives the height-width relation as an example of a scalar condition in the case of cup and the presence of a handle exemplifies a preference rule. It is obvious that preference rules interact with scalar conditions in the sense that if the object has a handle, “it can be further removed from the prototypical height-width relation and still be judged a cup” (ten Hacken, 2010: 917).

Let me now turn to the consequences of the assumption of prototypes for the distinction between general words, specialized words, and terms in ten Hacken’s perspective. Both general words and specialized words are based on prototypes. The difference between the two is that the latter is a label for expressions “used only in specialized language” (ten Hacken, 2010: 918). The example in (1c) is a case in point. The meaning of (1c) in general language is distinct from its specialized meaning in medicine. This also means that specialized words “are in the mental lexicon of a much smaller group of speakers” (ten Hacken, 2015: 6) as opposed to general lexis. It is reasonable to assume that (1c) in the sense of ‘an outer garment for the upper part of the body’ must be stored in the mental lexicon of most speakers of English. However, retrieving its specialized meaning of ‘a fixed bandage applied around the body in order to immobilize the spine’ requires a specialized context, familiar to a much smaller number of speakers.

Ten Hacken (2008, 2010) observes that two conditions can be used for defining terms, specialization and the precise delimitation of the extension. He emphasizes the different nature of these conditions. Specialization represents a scalar condition, whereas “having a precisely delimited extension produces a dichotomy” (ten Hacken, 2010: 917). To put it differently, with the former we can decide where the cutoff point is, but for the latter we have to select one or another. According to ten Hacken (2010: 917), only expressions with precisely delimited meanings can be labelled as terms in the narrow sense. A direct consequence of the fact that the conditions for specialized words and terms are independent is that the overlap of the two categories is possible without triggerring any problems.

The overlap of the two categories in certain contexts, specialized vocabulary items and terms, brings us to another crucial distinction ten Hacken (2010, 2015) makes; the difference between *terms* and a subset of terms he labels *terms in the narrow sense*. Only the latter can be made distinct from specialized vocabulary in the sense of terminological definition proper. This is illustrated in (2).

(2) **trocar** – an instrument for withdrawing fluid from a cavity, or for use in paracentesis; it consists of a metal tube (cannula) into which fits an obturator with a sharp three-cornered tip, which is withdrawn after the instrument has been pushed into the cavity; the name t. is usually applied to the obturator alone, the entire instrument being designated t. and cannula. (Stedman, 1997)

First, the definition in (2) may seem as an example of a classical terminological definition. It classifies the instrument within a particular category, or a class of objects and specifies its typical properties. However, the definition in (2) does not determine precise boundaries consisting of necessary and sufficient conditions relevant for (1b). The concept remains prototype-based and there is no need to impose clear-cut boundaries in order to determine whether a particular instrument is a trocar as in (1b) or not. Thus, it may be argued that (2) represents a well-formed lexicographic definition taken from a specialized medical dictionary. In (2), 'an instrument' fulfils the function of the hyperonym. A detailed description of the relevant parts specifies material, shape, and purpose. This information is an example of scalar conditions. *Usually* in the final part of the definition in (2) indicates a preference rule. It allows a user to select whether he wishes to refer to the instrument as a whole or only to a specific component. Although *trocar* is used only in specialized contexts, which makes it distinct from any general vocabulary item, similarly to natural concepts it is based on a prototype. a consequence is that *trocar* in (1b) is an example of a specialized vocabulary item and a term, but not of what ten Hacken (2010, 2015) labels as a term in the narrow sense.

Ten Hacken (2015: 7) demonstrates that "the distinction between terms (in the narrow sense) and specialized vocabulary is determined by the need to resolve conflicts. Unless there is such a need, we can continue to use prototypes, which correspond to the natural state of concepts." The usefulness of the difference between terms in the narrow sense and specialized vocabulary items arises in contexts where it is necessary to adopt clear boundaries of the concept in contrast to a continuum. Ten Hacken (2015) identifies two such contexts, legal disputes and scientific theories. The example in (3) illustrates the former.

(3) The term "drug" means

> (A) articles recognized in the official United States Pharmacopoeia, official Homoeopathic Pharmacopoeia of the United States, or official National Formulary, or any supplement to any of them; and
>
> (B) articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease in man or other animals; and
>
> (C) articles (other than food) intended to affect the structure or any function of the body of man or other animals; and
>
> (D) articles intended for use as a component of any article specified in clause (A), (B), or (C). A food or dietary supplement for which a claim, subject to sections 343 (r)(1)(B) and 343 (r)(3) of this title or sections 343 (r)(1)(B) and 343 (r)(5)(D) of this title, is made in accordance with the requirements of section 343 (r) of this title is

> not a drug solely because the label or the labeling contains such a claim. A food, dietary ingredient, or dietary supplement for which a truthful and not misleading statement is made in accordance with section 343 (r)(6) of this title is not a drug under clause (C) solely because the label or the labeling contains such a statement.

The source of the definition in (3) is the *Code of Laws of the United States of America*, commonly abbreviated to U.S. Code.[2] This document is the official compilation and codification of the general and permanent federal statutes of the United States.[3]

The definition in (3) is an example of a terminological definition. It delimits the precise boundaries of what is and what is not a *drug* by making necessary and sufficient conditons explicit. This means that *drug* as defined in (3) is a term in the narrow sense. On the basis of (3) it is possible to make a distinction between, for instance, a drug and a dietary supplement. Such a distinction is relevant in legal contexts and provides guidance in resolving legal disputes. An example of a relatively current issue is the case of Cholestin. The precise categorization of the product as a dietary supplement or drug was the main issue in a federal court case (cf. Havel, 1999; Heber et al., 1999).

Another type of context illustrating the need for a more precise terminological definition required for terms in the narrow sense is scientific theory. Ten Hacken (2010: 920) points out that terms in empirical science classify entities in the real world. This explains why theories in empirical science need a certain degree of precision for claims to be testable. In a medical context, this may be illustrated with different types of incisions. In (4), two definitions of *incision* are given.

(4) a. The action of cutting into something; esp. into some part of the body in surgery. (OED, 2015)

[2] Available at available at http://uscode.house.gov/, accessed on 2 August, 2015

[3] According to Wikipedia (United States Code, 2 August, 2015), the main edition is published every six years by the Office of the Law Revision Counsel of the House of Representatives, and cumulative supplements are published annually. The U.S. Code is organized in 52 titles, Food and Drugs can be found under title 21. The basic structure of the titles includes sections. The sections are numbered sequentially across the entire title without regard to the previously-mentioned divisions of titles. Frequently, the sections are further structured into subsections, paragraphs, subparagraphs, clauses, subclauses, items, and subitems. (3) is Title 21 › Chapter 9 › Subchapter II › § 321 (g).

b. a cut; a surgical wound; a division of the soft parts made with a knife. (Stedman, 1997)

The definitions in (4) can be described as examples of lexicographic definitions. In (4a), the OED gives a general definition of a vocabulary item that is likely to be used in common everyday situations. There is also a good reason to assume this item is stored in the mental lexicon of a large number of individual speakers of English. The definition in (4b) is slightly more specific, it indicates a degree of specialization of this vocabulary item. A precise delimitation of the boundaries in the sense of necessary and sufficient conditions is not given. Thus, the concepts described in (4a) and (4b) are prototype-based. This implies that *incision* is another example of the two overlaps. First, the general vocabulary item overlaps with the specialized word, and second, the specialized word with the term. However, neither (4a) nor (4b) can be considered as a definition of a term in the narrow sense.

In the medical theoretical literature, a number of different types of incisions are distinguished and sets of conditions specifying precisely when a particular incision is the best with respect to healing. The so-called *MacFee incision* is a good example to discuss. It is a type of incision used for neck dissection (Werner and Davies, 2004). Detailed descriptions and discussions of the MacFee incision can be found in books on theoretical medicine examining clinical judgement and reasoning. *Metastases in Head and Neck Cancer* by Werner and Davies (2004) is an example of such a book. It summarizes the types of health problems related to head and neck cancer, explains and describes a range of methods for their treatment and evaluates them with respect to a set of criteria. With the incisions used for neck dissection, nine evaluation criteria for the selection of a particular type are listed. These include the tendency of necrosis of the detached skin parts, the planned extent of the tumor intervention, the primary defect coverage in cases of more extended skin resections, the blood supply of the flaps, the overview of the entire operation field, the additional performance of tracheotomy, the possible excision of existing scars, the potential for avoiding skin incision when mucosal incisions suffice, and the possibility of an extension of the incision if additional cervical lymph node regions must be dissected (Werner and Davies, 2004).

Then, the illustrations of the incisions are presented. The illustration of the MacFee incision shows the procedural details. Individual steps represent necessary and sufficient conditions combined in the definition of the concept. The choice of the name is less relevant. This means that in contrast to *incision*, *MacFee incision* is an example of a scientific term or in ten Hacken's terminology, a term in the narrow sense. The term *MacFee incision* has precisely delimited boundaries. The illustrations make it possible to determine the type of incision immediately and unambiguously.

The illustrations are followed by a summary of their advantages and in some cases by a comparison with competing incisions. a detailed description of the advantages of the MacFee incision is given in (5).

(5) "The so-called MacFee incision probably has the best chance of healing because this type of incision addresses the blood supply of the neck […]. It leads to very good esthetic results as long as the incisions are performed along skin lines, especially in pre-formed creases. Furthermore, this type of incision protects the carotid artery. The operative procedure is more difficult to perform in patients with short necks. Additionally, exposure of the operative field is often impaired so that intensive retraction by the assistant is required. The MacFee incision is preferred for patients suffering from a peripheral vascular disease or for patients who have undergone prior radiotherapy […]. It is often used in young patients undergoing neck dissection for thyroid cancer." (Werner and Davies, 2004, references deleted)

The description of the incision in (5) shows at least two important facts. First, it delimits which factors for the selection of a particular type of incision are essential for the MacFee incision, e.g. blood supply and the overview of the operation field. Second, it demonstrates that the type of incision, i.e. the concept identified by this name, includes a number of benefits for a patient. This becomes a part of theorizing about the best practice in a particular context. The overview of the positive outcomes is based on a reasonable amount of empirical data, their collection, and evaluation. There may be discussion about advantages and disadvantages influenced by different empirical data, but this does not mean that the scientific term has fuzzy boundaries.

It is interesting to compare Coxhead's understanding of specialized vocabulary with ten Hacken's interpretation. An essential similarity is immediately obvious. Both linguists agree on the fact that specialized vocabulary requires a specialized context. Coxhead (2013) uses this label to cover general vocabulary items with a narrowed meaning, words of Greek and Latin origin, and highly technical words. These types were exemplified in (1). This typology is sufficient in the context of ESP for pedagogical purposes. The questions related to the representation of specialized vocabulary in the lexicon require a more fine-grained theoretical approach. This is what we find in ten Hacken's approach to general vocabulary, specialized vocabulary, terms in general, and terms in the narrow sense. An overview is presented in Figure 2-1.

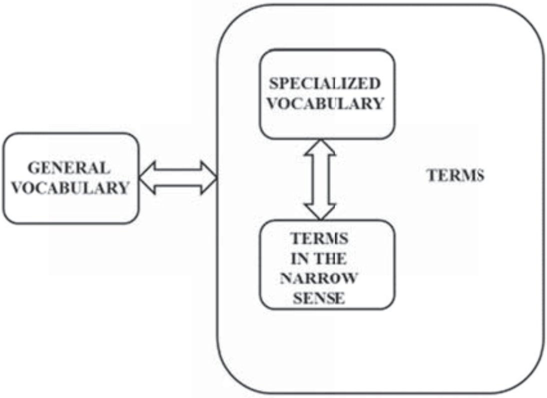


Figure 2-1 The relationship between general vocabulary, specialized vocabulary, terms, and terms in the narrow sense based on ten Hacken (2010, 2015)

Figure 2-1 summarizes ten Hacken's (2010, 2015) perspective of general words, terms, specialized words, and terms. Figure 2-1 has four boxes representing general vocabulary, specialized vocabulary, and terms in the narrow sense. The last two, specialized vocabulary and terms in the narrow sense are covered under terms. They are connected by two-headed arrows indicating that some items may move from one category to another depending on the context. Having a prototype-based character is the property that general words and specialized vocabulary have in common, but not terms in the narrow sense. There may be cases in which a general word is used with a special meaning in the specialized context and this property is highlighted when a vocabulary item is classified as a specialized word. It is reasonable to assume that specialized vocabulary items are stored in a much smaller number of speakers than general lexis. Specialized vocabulary and terms in the narrow sense may overlap as discussed for (1c) above. For specialized and general vocabulary, lexicographic definitions similar to the one discussed in (2) are sufficient. However, when a need for a precise delimitation arises, the nature of the definition changes. A proper terminological definition may be necessary in legal or scientific contexts. Thus, the distinction between specialized words and terms in the narrow sense is useful only when a need for clear boundaries arises. Fuzzy boundaries vs precisely delimited boundaries are then crucial for the distinction between the two. Otherwise, the scalar nature of terms (in a broad sense) is reasonable and sufficient.

2.2 Three methods for delimiting specialized vocabulary

In ESP, determining domain-specific vocabulary has always been a main focus of research. Coxhead (2013: 142-147) mentions three fundamental methods used to delimit specialized vocabulary: consulting experts and technical dictionaries, Chung and Nation's (2003) four-step scale, and the use of corpora. Let us consider each of them in more detail.

The expertise of specialists in a particular field is one of the possible sources of identifying specialized vocabulary. Although Schmitt (2010) considers this method useful, he comments on difficulties that may be encountered. Problems may arise, for instance, when consulting several experts whose opinions differ. In addition, unaided judgements about level of specialization are often vague.

Chung and Nation (2004) used technical dictionaries to identify specialized vocabulary items. The main criterion they applied was relatively simple, if a word has a main entry or sub-entry in a technical dictionary, it was considered sufficient to label such a word a technical word (Chung and Nation, 2004: 255). Their research focused on anatomical technical words covered by two medical dictionaries. In their study they compared a dictionary approach to three competing ways: the rating scale approach (see below for more detailed description), the clues-based approach, and the computer-based approach. They found out that the average percentage of terms and non-terms identified correctly is 79.8%. This placed the dictionary-based approach on the third position after the rating scale with 100% and the computer-based approach with 82.7%. Chung and Nation (2004) point to two main disadvantages of dictionary use in identifying technical words. First, they discovered how data from the two dictionaries can differ from each other and that the choice of the best dictionary is critical (Chung and Nation, 2004: 261). Second, the selection of words to be included in a dictionary depends on an intuitive judgement of one person or a group of specialists, which means "that there could be little consistency of decision making between dictionaries" (Chung and Nation, 2004: 256). The general problem with a dictionary approach is that dictionary making requires a lot of decision-making, including the size of the dictionary, or the position of the word in a main entry or sub-entry, dependent on lexicographic policy.

Chung and Nation (2003) developed and tested a four-step rating scale to identify technical words. This method is noteworthy here especially because the technical words were determined in anatomy texts, i.e. a clearly medical domain. Step 1 and step 2 in the scale cover non-technical words. Step 1 includes words unrelated to anatomy, e.g. *the*, *is*, *between*, *it*, *by*, *12*, *adjacent*, *amounts*, *common*, *commonly*. Words minimally related to the field of anatomy in the sense in that they describe the positions, movements, or features of the body fall into Step 2 and some examples are *superior*, *part*, *forms*, *pairs*, *structures*, *surrounds*, *supports*, *associated*. Words that belong to Step 3 and Step 4 are considered technical. Step 3 is a class of words closely related to anatomy, which means that they refer to parts, structures or functions of the body, such as the regions of the body and systems of the body. Some words may also be used in general language, others may have some restrictions of usage depending on the subject field. Some examples are: *chest*, *trunk*, *neck*, *abdomen*, *ribs*, *breast*, *cage*, *cavity*, *shoulder*, *girdle*, *skin*, *muscles*, *wall*, *heart*, *lungs*, *organs*. Step 4 is for words specific to the field of anatomy. It is reasonable to assume that such words are not used or not commonly used in general language, for instance, *thorax*, *sternum*, *costal*, *vertebrae*, *pectoral*, *fascia*, *trachea*, *mammary*, *periosteum*, *hematopoietic*. Perhaps the most striking finding is that almost one in every three words (31.2%) in the anatomy text they analysed is a technical term (Chung and Nation, 2003). Earlier research by Coxhead (1998), or Nation (2001) reported a significantly lower proportion, with technical vocabulary accounting for only around 5% of the running words of a text.

Coxhead (2013: 142) points out that “developing new scales means new ways of classifying technical words, so different results from different studies have to be accompanied by a clear understanding of how principles of selection and classification have taken place”. However, it is obvious that the four-step rating scale represents rather a cline with fuzzy boundaries between the steps. Similar to consulting experts and technical dictionaries, this method involves a large amount of subjective decision-making. Technical dictionaries represent a different source of knowledge, which may be used as an instrument for classification, but this is not their intended purpose. It should be noted that there is an important difference between a rating scale method and dictionary use. When using a rating scale, we do the analysis ourselves and it depends on our decision. When using a dictionary, we already use the analysis presented in the dictionary. Such an analysis results from a lexicographer’s decision, presumably based on some kind of rating scale. The lexicographer presents the results of such an analysis in a dictionary.

Corpora represent an important tool in research of ESP vocabulary. The main advantage of corpus-based studies is that they “allow for large-scale investigations of words in context” (Coxhead, 2013: 144). On the basis of their accessibility, Nesi (2013: 452) distinguishes ESP corpora with restricted access and ESP corpora in the public domain. Examples of the former are the *TOEFL 2000 Spoken and Written Academic Language Corpus* (T2K-SWAL), the *Cambridge and Nottingham Business English Corpus* (CANBEC), or Coxhead’s Academic Corpus. The data from T2K-SWAL are discussed in a number of research papers, for instance, Biber et al. (2002), Biber, Conrad, and Cortes (2004), Biber (2006), and Csomay (2005, 2006). The CANBEC data were the main source in the research performed by McCarthy and Handford (2004) and Handford (2010).

An example of a specialized ESP corpus with open access is the one offered by the *Professional English Research Consortium* (PERC). Nesi (2013: 452) considers it the largest corpus available at present for the language of science and technology. The PERC Corpus is a 17-million-word corpus of copyright-cleared English academic journal texts.[4] The texts are taken from approximately 170 subdomains and are classified into the following 22 domains: agriculture, biology, chemistry, civil engineering, computer science, construction and building technology, earth science, electrical and electronic engineering, engineering, environmental sciences, fisheries, food science, forestry, general science, materials science, mathematics, medicine, metallurgy and metallurgical engineering, nuclear science and technology, oceanography, physics, telecommunication, and media: academic journals. These domains can be accessed separately as sub-corpora. The selection of journal texts is based on the data obtained from the Journal Citation Reports (JCR). JCR presents quantifiable statistical data for an objective and systematic approach to determining the relative importance of journals within their subject categories. At the PERC website, it is mentioned that in 2001, the Science Edition of the JCR contained about 5,700 journals. It uses an indicator called “Impact Factor,” which provides a way to evaluate or compare a journal’s relative importance as perceived by others in the same field. Employing these data, the journals with the top 20% impact factor in each field were selected for inclusion in the PERC Corpus. JCR classifications were also used to define the subject fields. In addition to these selection criteria, all journals

[4] Information about PERC is available at https://scn.jkn21.com/~percinfo/eng_sub1.html#, retrieved 14 February, 2016.

are monolingual (English). The texts published in distinguished scientific journals represent a good example of professional writing in English. The information about the regional variety of English (British English, American English, etc.) in which the research paper is written is also included. PERC is a synchronic corpus because it covers academic texts published between 1995 and 2002. The PERC corpus was compiled with the intention to be used for research in the field of Professional English. However, the fact that it stopped adding more recent research papers than 2002 makes it slightly disadvantaged for current research. Another disadvantage for my research is that the PERC corpus does not indicate the size of the medical corpus or of any other domain included in it.

Large general English corpora, the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), include subdivisions of academic specialized texts and represent a good source of data in ESP research. An interesting type of open access corpora is a compilation of texts where English is used as a second language. It should be emphasized that this type of a corpus is not equal to learner corpora. Granger (2002) discusses in detail methodologies associated with computer learner corpora and differences between native and L2 learners of a language. The Vienna-Oxford International Corpus of English (VOICE) and the English as a Lingua Franca in Academic Settings (ELFA) include recordings of experienced non-native speakers of English in academic contexts. The former was compiled at the University of Vienna, whereas the latter resulted from a research project at the University of Tampere and the University of Helsinki in Finland. Each of these corpora contains approximately 1 million words of transcribed speech.

Corpora, including many of those mentioned above, are central in vocabulary delimitation research in ESP. Current technologies made it possible to develop a number of corpus-based academic word lists such as the Academic Word List (Coxhead, 2000), the New Academic Vocabulary List by Gardner and Davies (2013), and a number specialized word lists, e.g. Business word lists (Nelson, 2000; Konstantakis, 2007), a pilot science-specific EAP word list (Coxhead and Hirsh, 2007), a Medical Academic Word List (Wang, Liang and Ge, 2008), Ward's English list of Basic Engineering Words (Ward, 2009).

2.3 Coxhead's Academic Word List

The Academic Word List (AWL) compiled by Coxhead (2000) is a well-known and widely used list, which resulted from a corpus-based approach.

The development of the AWL was motivated by the need to identify the academic vocabulary that could be used in designing materials for language courses and supplementary materials for individual and independent study.

The size of the corpus is an important methodological criterion. Coxhead's corpus includes 3.5 million running words. Coxhead (2000: 217) points out that "[t]he decision about size was based on an arbitrary criterion relating to the number of occurrences necessary to qualify a word for inclusion in the word list: If the corpus contained at least 100 occurrences of a word family, allowing on average at least 25 occurrences in each of the four sections of the corpus, the word was included." We will return to the notion of word family and the structure of the corpus below. The decision about the number of occurrences in a corpus of 3.5 million words was based on the findings of Francis & Kučera (1982). Their data, based on the Brown corpus, suggest that a corpus of around 3.5 million words would be needed to identify 100 occurrences of a word family.

The criteria for word selection were crucial in designing the AWL. Coxhead (2000) used the definition of *word* and *word family* proposed by Bauer and Nation (1993). Their delimitation of a word family takes into account the importance for vocabulary teaching. From the perspective of reading, Bauer and Nation (1993: 253) define a word family as consisting of "a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately". This definition is a guiding principle in setting up six levels of inflection and affixation given in (6).

(6) a. develop
develops
developed
developing
b. develop, develops, developed, developing
c. developable, undevelopable, developer(s), undeveloped
d. development(s), developmental, developmentally
e. developmentwise, semideveloped, antidevelopment
f. redevelop, predevelopment

Level 1 is illustrated in (6a). a characteristic property of this level is that every different form represents a different word. Bauer and Nation (1993: 258) point out that the level is not significant from the perspective of readers recognizing that the items in (6a) belong to the same lexeme. The level is important in distinguishing cases of polysemy, e.g. *ring* 'call' vs *ring* ‘an item of personal wear’, or zero inflectional ending, e.g. *hit*PRESENT TENSE and *hit*PAST TENSE.

In (6b), however, all listed items have the same base but different inflections and belong to the same word family at Level 2. The inflectional categories Bauer and Nation (1993: 258) consider relevant for Level 2 include plural, third person singular present tense, past tense, past participle, *-ing*, comparative, superlative, and possessive. Bauer and Nation (1993) discuss two main problems with infectional affixes. First, there is no general agreement in morphological theory about determining the inflectional affixes in English. For instance, Beard (1982) excludes the category of plural. Jensen (1990) argues for counting comparative and superlative, whereas Mugdan (1989) does not. Zwicky (1992) does not include possessive. The dictinction between derviation and inflection is a much-debated issue in the field of morphology cf. ten Hacken (2014), Štekauer (2015). A second problem is posed by word forms which are in particular contexts inflectional, but in others their inflectional status is less clear, e.g. *He is shooting clay-pigeons* vs *Clay-pigeon shooting is an expensive pastime*. Bauer and Nation (1993) made a decision to treat all -*ing* and *-ed* items as inflectional for the purposes of setting up the six levels.

The examples in (6c) represent Level 3. This level includes the most frequent and regular derivational affixes *-able*, *-er*, *-ish*, *-less*, *-ly, -ness*, *-th*, *-y*, *non-*, *un-*.[5] The selection of the affixes was based on the criteria listed in (7).

(7) a. Frequency
b. Productivity
c. Predictability
d. Regularity of the written form of the base
e. Regularity of the spoken form of the base
f. Regularity of the spelling of the affix
g. Regularity of the spoken form of the affix
h. Regularity of function

All criteria listed in (7) were applied strictly at Level 3, but at subsequent levels they were gradually relaxed. The criterion in (7a) takes into account the number of words (types) in which a particular affix occurs. In line with Bauer (1983, 1988) there is a correlation between the frequency of affixes

[5] Some of these affixes are ambiguous. Thus, *-th* in the sense used at Level 3 is the one that forms ordinal numbers such as *sixth*, *seventh*. The *-th* suffix such as in *warmth* is placed at Level 6.

and the degree of generalization. Frequent affixes tend to be highly generalized. In (7b) the probability that an affix will be used to create new words is considered important. Bauer and Nation (1993: 256) point out that the affixes placed at lower levels, (6a), are highly productive, but less likely to be given separate dictionary entries. This increases the need to recognize such words as related to their bases. Bauer and Nation (1993: 256) exemplify this with *-ly* and *-ness* which are often used to create words that have not been met before. For (7c) it is important to note that the meaning of an affix is more predictable when the word class of the base to which it attaches is determined, e.g. in case of the suffix *-s* whether its meaning is plural or third person singular present tense. The application of (7d) shows that at earlier levels the recognition of the base is not affected orthographically, e.g. *green+ish* whereas at higher levels this may not necessarily be the case, e.g. *sacrilegious*. Similarly, (7e) is applied to the spoken form. A phonologically unaffected base is expected at lower levels. Higher levels may include bases that are not potential free forms, e.g. *permeable*. Criterion (7f) explains a gradual decline of predictability of unchanged orthographic forms from easily recognizable lower levels to forms that do not always seem related orthographically. This may be illustrated by, for instance, *in-*, *im-*, *il-*, and *ir-*. This contrasts with the prefix *pre-* with one form. The problem in the case of *pre-* is not allomorphy, but recognizing it. It is likely to be mis-analysed in words such as *prescribe*, *present*, *prevent*, etc. The same is true for (7g), but it applies to phonological form. The last criterion in (7h) refers to the property of affixes to attach to the bases of a specific word class with the predictable and regular word class of the output, e.g. the suffix *-ess* is a nominal suffix which always attaches to nouns.

These eight criteria are essential for Level 4 exemplified in (6d). The frequency parameter, or generalization, plays a more important role than productivity, and orthography is prioritized to phonology. Level 4 includes the suffixes *-al*, *-ation*, *-ess*, *-ful*, *-ism*, *-ist*, *-ity*, *-ize*, *-ment*, *-ous*, *in-*. Interestingly, (6d) includes the item *developmentally* with the suffix *-ly* categorized as Level 4.

The examples in (6e) illustrate Level 5. This level covers regular, but infrequent affixes e.g. *-age*, *-dom*, *-ship*, *-wise*, *anti-*. The complete list includes 50 affixes (Bauer and Nation, 1993: 261). The affixes at Level 5 always select free bases.

Level 6 is presented by the instances in (6f). Frequent but irregular affixes are placed there. This class of affixes may cause segmentation problems. The main reason is orthographic allomorphy in the bases of the words they attach to, for example, *describe→description*, *diagram→diagrammatic*.

The suffixes classified at this level *-able*, *-ee*, *-ic*, *-ify*, *-ion*, *-ist*, *-ition*, *-ive*, *-th*, *-y*, *pre-*, *re-*. It is interesting to see that *re-* belongs to this level although it is much more productive than, for instance, *-th*. Bauer and Nation (1993: 279) explain that the problem is to recognize this suffix correctly because "there are so many lexicalized instances with *re-* and cases that are likely to be mis-analyzed, that trying to use it might be counter-productive". Their examples of words that are likely to be wrongly analysed include *rebut*, *recap*, *recommend*, *record*, *recover*, *recur*, *redoubt*, etc. Some of the affixes were listed at earlier levels, but Bauer and Nation (1993: 261) assume that earlier levels cover more transparent cases whereas Level 6 treats more opaque ones.

In addition to the six levels listed in (6), Bauer and Nation (1993: 262) distinguish Level 7, where they place neoclassical roots and affixes. These occur as bound roots, e.g *embolism*, or in neoclassical compounds, e.g. *geography*. Apart from common combining forms, this level includes frequent prefixes, e.g. *ab-*, *ad-*, *com-*, *de-*, *dis-*, *ex-*, and *sub-*.

The main aim for the development of this multilevel scale was the recognition of written words. In line with Bauer and Nation (1993: 257) the levels should be taken as steps along a continuum, or a cline and they have purely practical rather than theoretical value. It should be noted that the levels include affixation only, although transparent compounds might be added to one of the later levels.

On the basis of Bauer and Nation (1993), Coxhead (2000: 218) defines a word family as a stem plus all closely related affixed forms as defined by Levels 1-6 of Bauer and Nation's (1993) scale. Only affixes that can be added to free stems are included. This means that, for instance, *specify* and *special* are not placed in the same word family because *spec* cannot stand alone as a free form (Coxhead, 2000: 218). Interestingly, she states that Bauer and Nation's "Level 6 definition of affix includes all inflections and the most frequent, productive, and regular prefixes and suffixes" (Coxhead, 2000: 218). However, Bauer and Nation (1993: 261) describe Level 6 as the level with "frequent but irregular affixes". It may be that Coxhead intended to refer to Level 3 with "the most frequent and regular derivational affixes" (Bauer and Nation, 1993: 258).

Coxhead (2000: 218-219) formulates six research questions about academic vocabulary. They are listed in (8).

(8) 1. Which lexical items occur frequently and uniformly across a wide range of academic material but are not among the first 2,000 words of English as given in the General Service List (GSL) (West, 1953)?

2. Do the lexical items occur with different frequencies in arts, commerce, law, and science texts?
3. What percentage of the words in the Academic Corpus does the AWL cover?
4. Do the lexical items identified occur frequently in an independent collection of academic texts?
5. How frequently do the words in the AWL occur in non-academic texts?
6. How does the AWL compare with the University Word List UWL (Xue & Nation, 1984)?

The first two questions in (8) frame the description of the AWL and the remaining four concentrate on the assessment of the AWL.

A crucial step in the process is the corpus design. The corpus (Coxhead's Academic Corpus) contains articles from academic journals, edited academic journal articles available online, university textbooks or course books, and texts from several corpora of the Learned and Scientific section of the Wellington Corpus of Written English (Bauer, 1993), the Learned and Scientific section of the Brown Corpus (Francis & Kučera, 1982), the Learned and Scientific section of the Lancaster-Oslo/Bergen (LOB) Corpus (Johansson, 1978), and the MicroConcord academic corpus (Murison-Bowie, 1993).

The texts were collected in electronic form and the word count was determined after the bibliography had been removed. The texts were classified into three categories depending on their length. The category of short texts included texts of 2,000-5,000 running words, medium texts ranged from 5,000 to 10,000 running words, and long texts had more than 10,000 running words. The corpus consisted of four subcorpora: arts, commerce, law, and science, each containing an approximately equal number of running words. Each subcorpus was subdivided into seven subject areas. To maintain a balance of long and short texts, the four main sections (and, within each section, the seven subject areas) each contained approximately equal numbers of texts of each length.

Words in the corpus were processed by the corpus analysis program *Range* (Heatley & Nation, 1996). The selection of items for the AWL was based on three criteria: specialized occurrence, range, and frequency. Specialized occurrence meant that the word families had to be outside the first 2,000 most frequently occurring words of English, as represented by West's (1953) GSL in order to be included. Range was determined by the occurrence of a member of a word family at least 10 times in each of the four main sections of the corpus and at least once in 15 or more of the 28 subject areas. Last, but not least, word families were selected only if the members of the word family occurred at least 100 times in Coxhead's Academic Corpus.

The results of Coxhead's research show that the answer to her first question in (8) is a list of 570 word families in the AWL. All included vocabulary items are beyond the first 2,000 in West's (1953) GSL. Coxhead divides them into ten sublists according to frequency. Sublists 1-9 have 60 words each, sublist 10 only 30. Among the most frequent words in the AWL (in Sublist 1) are, for instance, *assess*, *assume*, *concept*, *constitute*, *define*, and *estimate*. The second question in (8) concerns the differences in occurrence of words across disciplines. The findings reveal that "the list appears to be slightly advantageous for commerce students, as it covers 12.0% of the commerce subcorpus. The coverage of arts and of law is very similar (9.3% and 9.4%, respectively), and the coverage of science is the lowest among the four disciplines (9.1%)" (Coxhead, 2000: 222). The data indicate that the AWL covers 10% of the total tokens in the Coxhead's Academic Corpus. Interestingly, recent research by Billuroğlu and Neufeld (2005) confirmed that although the GSL needs revision, the headwords in the list still provide approximately 80% text coverage in written English.

Coxhead (2000) compared the AWL with a second independent corpus she compiled on the basis of the same criteria and sources to select texts and dividing them into the same four disciplines. This corpus covers 678,000 tokens (82,000 in arts, 53,000 in commerce, 143,000 in law, and 400,000 in science). In the second corpus, all 570 word families in the AWL are attested, but interestingly, Coxhead (2000: 224) found that the AWL's coverage of a second independent corpus of academic texts is 8.5%.

Then, Coxhead (2000) compiled a collection of 3,763,733 running words of fiction texts. The collection consisted of 50 texts from Project Gutenberg's (http://www.gutenberg.net). The comparison of the AWL with non-academic fiction texts reveals that the AWL accounts for approximately 1.4% of the tokens. This is much lower than the 10% in the Academic corpus.

Finally, Coxhead (2000) compared the AWL with the University Word List (UWL) developed by Xue and Nation (1984). The UWL is a list of vocabulary items that characterize academic texts. It includes 808 words, which are divided into 11 levels. This list is designed to be a list of specialized vocabulary for students who intend to study in an English-language university. The overlap between the AWL and the UWL represents 51% with 435 word families found in both. This means that 401 word families occurred only in the UWL and 135 word families occurred only in the AWL.[6] This finding is the answer to the last question in (8). According to Nesi (2013: 452), the AWL has become a staple resource used in EAP materials design (cf. also Schmitt and Schmitt, 2005, 2011) and it is widely cited across many scientific disciplines. For our purposes, however, the AWL is of limited direct use because it does not include medicine among its domains.

2.4 The Medical Academic Word List by Wang, Liang and Ge (2008)

Research focused on the academic vocabulary specific to one discipline is based on the underlying assumption that the academic vocabulary in a single scientific field may have unique properties. Academic vocabulary in computer science was investigated by Lam (2001). A specialized Student Engineering English Corpus (SEEC) with 2 million running words was compiled by Mudraya (2006). Chen and Ge (2007) considered the distribution of the AWL word families in medical research papers. Their findings suggested the need to establish a medical academic word list. Wang et al. (2008) report on the development of such a Medical Academic Word List (MAWL).

The first step by Wang et al. (2008) was to compile a corpus of medical research articles. The size of the corpus was 1,093,011 running words. This is approximately one third of the size of the Academic Corpus developed by Coxhead, but the medical corpus is more homogeneous. The medical research papers were collected from the ScienceDirect Online database. The papers were selected from journals covering 32 medical subfields such as anesthesiology and pain medicine, cardiology and cardiovascular medicine, nephrology, dentristry, dermatology, hematology, oncology, etc. In addition to these traditional medical subdisciplines, also such areas as health informatics or public health and health policy are included. The research articles were compiled from volumes published between 2000 and 2006 and were written by native speakers, or at least the first author must have been a native speaker and affiliated with an institution in an English-speaking country. Wang et al. (2008) designed a three-round procedure to select the research papers for the corpus. In the first round, they chose three journals from each of the 32 medical subdisciplines by the method of stratified random sampling. This resulted

[6] UWL seems to use word families as the basic unit as well. For pairs such as eliminate and elimination only one item occurs in the list.

in a total number of 96 journals. In the second round, one issue out of each of the 96 journals was chosen at random. Then, in the third round, the articles were evaluated on the basis of three criteria, native speaker authorship, length between 2000 and 12000 words, and a conventionalized Introduction-Method-Result-Discussion structure. Only papers that met all three criteria were included in the corpus. Bibliographies, charts, and diagrams were removed from the texts in order to ensure the texts can be processed by corpus management software. Similarly to Coxhead (2000), the definition of a word family by Bauer and Nation (1993) was used in data processing. This means that *demonstrate*, *demonstrates*, *demonstrating*, *demonstrated*, and *demonstration* were labelled and counted as one item.

Following Coxhead (2000), three criteria, specialized occurrence, range and frequency of a word family, were taken to be relevant in the development of the MAWL by Wang et al. (2008). The criteria by Wang et al. (2008) were slightly adjusted to make them more appropriate for the structure of their corpus. Specialized occurrence was identical with Coxhead (2000). This means that the word families had to be outside the first 2000 most frequent words of English, included in West's GSL (1953). For range, 50% of the subcorpora was set as a criterion for inclusion of a word family in the MAWL. This gave a different number of subject areas. In Coxhead (2000: 221), members of a word family had to occur at least 10 times in each of the four main sections of the corpus and at least once in 15 or more of the 28 subject areas whereas in Wang et al. (2008) it was at least once in 16 or more of the 32 subject areas. In both studies, range was prioritized to frequency. In practice, this means that high frequency of a word family was not sufficient for inclusion in the word list if the word family was covered in less than 50% of the subject areas in the corpus (15 in Coxhead, 16 in Wang et al.). Borderline cases between technical vocabulary and academic vocabulary were consulted with two professors of English for Medical Purposes (EMP). Unfortunately, Wang et al. (2008) do not give any examples of the items that were submitted to the EMP specialists.

Wang et al. (2008) found 31 275 word families in their corpus. When the frequency criterion was applied, the number of word families reduced to 3345. Then, the GSL families were excluded, leaving 1446 word families. Out of these, 650 word families occurred in 16 or more subdisciplines of the corpus. Consultations with two EMP experts excluded 27 more families. Although Wang et al. do not make it explicit on what basis these borderline word families were excluded, it may be assumed that the experts considered them to be highly specialized terms such as *pathogenesis*, *cytokine*, *necrosis*, *stent*, etc. (Wang et al., 2008:

448). The final number of word families in the MAWL was 623. Wang et al. list all these word families in their Appendix (Wang et al., 2008: 452-457). It is interesting that more than 78% of the word families were found in 20 or more out of 32 subdisciplines and nearly 17% occurred in all 32. A significant correlation between the range and frequency of the word families was observed.

A comparison with the AWL reveals a marked difference between the two word lists. Only 342 MAWL word families, representing 54.9%, overlap with the AWL.[7] Wang et al. interpret this difference as undermining "the usefulness of general academic word lists across different disciplines" (Wang et al., 2008: 451). Hyland and Tse (2007) also argue for a more restricted, discipline-based vocabulary. By contrast, Coxhead (2013: 147) argues that the overlap between MAWL and AWL is a consequence of the fact that Wang et al. (2008) used the GSL as a common core instead of the AWL. Table 2-1 compares the top 15 words in MAWL and AWL.

Frequency number in MAWL	MAWL top headwords	Sublist number in AWL
1	cell	not included
2	data	1
3	muscular	not included
4	significant	1
5	clinic	not included
6	analyze	1
7	respond	1
8	factor	1
9	method	1
10	protein	not included
11	tissue	not included
12	dose	not included
13	gene	not included
14	previous	2
15	demonstrate	3

Table 2-1 The distribution of top 15 headwords in the MAWL and AWL

[7] In analysing the two lists, I only found 337 overlapping word families. As Wang et. al (2008) do not specify which word families overlap, I could not verify the cause of this difference.

The data in Table 2-1 show that eight headwords in the MAWL and AWL overlap, which is just over 50%. This roughly corresponds with the overlap result of 54.9% by Wang et al. (2008) for the total number of word families in both word lists. Interestingly, a majority of the overlapping headwords are placed in Sublist 1 in the AWL. Sublist 1 covers the most frequent words in the list, and Sublist 10 covers the least frequent. The distribution of MAWL across AWL sublists is given in Figure 2-2.

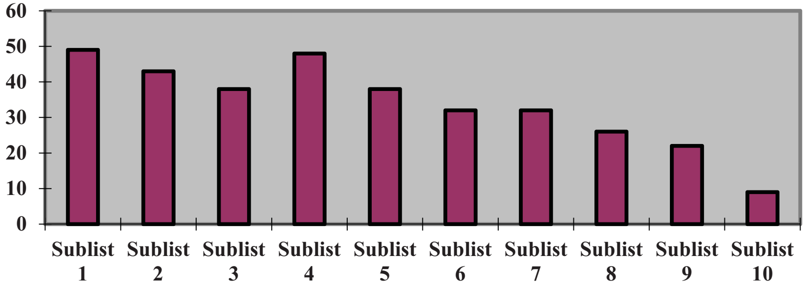


Figure 2-2 Distribution of MAWL across AWL ten sublists

In Figure 2-2 we can see how many items from each sublist of AWL also occur in MAWL. My analysis shows that the total number of overlapping words is 337, which corresponds to the more than 50% overlap given above. We might expect a slow downward curve. However, it is interrupted by the unexpectedly high proportion in sublist 4. For sublist 10, which is only half the size of the other sublists, the column should be doubled for proper comparison.

Coxhead (2013: 147) views the MAWL as "an example of early specialization", which allows learners to focus directly on their subject field of specialization rather than on general academic vocabulary. The headwords not found in the AWL are better classified as specialized vocabulary items and terms in the narrow sense in ten Hacken's (2010, 2015) typology, but also in Coxhead's (2013) categorization. This raises important methodological questions about the structure of medical vocabulary items across academic word lists.

Another interesting observation about the MAWL is that it includes *symptom* (number 81) and *syndrome* (number 211), but not *disease*. The reason is that *disease* occurs among the first 2000 GSL vocabulary items (number 1156) and in line with Wang et al.'s methodology the overlapping words were excluded. It is not obvious whether the AWL does not list *disease* for the same reason or because medicine is not a field which was included in the corpus. However, the notions of *symptom*, *syndrome*, and *disease* and relationships among them are eminently relevant in medicine. OED (2015) explains *syndrome* as ‘a concurrence of several symptoms in a disease; a set of such concurrent symptoms’ (OED, 2015). It is only when the mechanism linking symptoms and this cause is understood and explained sufficiently, that the corresponding condition is described as a disease. OED (2015) defines *disease* as ‘any one of the various kinds of such conditions; a species of disorder or ailment, exhibiting special symptoms or affecting a special organ’. This means that understanding the relationship between the notions of *symptom*, *syndrome*, and *disease* in medicine is important. Panocová and ten Hacken’s (2017) claim that symptoms are descriptive terms that can be used to classify observations and even if their names are standardized, their concepts remain based on prototypes and they are referred to as specialized vocabulary in ten Hacken’s typology described in section 2.1. The search results of *syndrome*, *symptom*, and *disease* in the COCA corpus are summarized in Table 2-2.

Word	Frequency in COCA
disease	15160
symptom	1266
syndrome	2195

Table 2-2 Frequency of occurrence in COCA

The frequency in Table 2-2 is based on a search restricted to the Academic section in the COCA corpus. *Disease* is found nearly 12 times more frequently than *symptom*, and almost 7 times more often than *syndrome*. It is interesting to see how these three words collocate together. The frequency of collocates for *disease* and *syndrome*, with a 5-item distance on the right and left, is 32, for *disease* and *symptom* 20 whereas for *symptom* and *syndrome* it is only 5. Some example sentences illustrating the context are given in (9).

(9) a. This definition, and every other definition, of autism is a description of **symptoms**. As such, autism is recognized as a **syndrome**, not a **disease** in the traditional sense of the word.[8]

b. Normal individuals free from any evident **symptom** of the **disease** were taken as controls.[9]

c. **Symptom** management for this **syndrome** is an important focus for nurses and healthcare professionals.[10]

The examples in (9) indicate that these three words often co-occur in the context. Therefore, it seems reasonable to assume all of them should be included in a proper description of medical vocabulary. Such examples cast doubt on the use of the GSL in discipline-based academic word lists.

2.5 A New Academic Word List by Gardner and Davies (2013)

Rapid progress in computer technologies used in corpus linguistics makes it possible to improve the existing academic word lists or design new ones. An example is the research by Gardner and Davies (2013) resulting in the development of a new Academic Vocabulary List (AVL). Although Coxhead's AWL became accepted and has been widely used as a standard for more than ten years, Gardner and Davies (2013: 3) address two main methodological concerns about it, repeatedly pointed out in the literature, the use of word families and the reliance on the GSL in the AWL.

Coxhead's use of word families was based on Bauer and Nation (1993), presented in section 2.2.2. Coxhead (2000) uses word families in the sense of a stem, i.e. a headword and all its inflected forms, and all transparent derivations. Gardner and Davies (2013: 3) find this decision problematic because the notion of word family does not consider the

[8] The source of this example in COCA is Shriver, Mark D., Allen, Keith D., Mathews, Judith R. (1999), 'Effective assessment of the shared and unique characteristics of children with autism.', *School Psychology Review*, vol. 28, Issue 4, p. 538.

[9] The source of this example in COCA is Sharma, Ritu; Singh, Balwant; Mahajan, Mridula; Kant, Ravi (2007), 'Age and Sex: Important Determinants In Affecting The Levels Of Serum Apolipoprotein B And A1 In Indian Population.', *Internet Journal of Cardiology*, vol. 4, Issue 1, p. 2.

[10] The source of this example in the COCA is Taggart, Helen M.; Arslanian, Christine L.; Bae, Sejong; Singh, Karan (2003), 'Effects of T'ai Chi Exercise on Fibromyalgia Symptoms and Health-Related Quality of Life.', *Orthopaedic Nursing*, vol. 22, Issue 5, p.353.

syntactic category of the word form. This makes it impossible to separate *proceeds* with the meaning 'continues' and with the meaning 'profits', as they are in the same word family. Gardner and Davies (2013) propose to solve this problem by counting lemmas in the sense of inflectional relationships only, instead of word families. Their suggestion is based on research by Nippold and Sun (2008), who found that morphologically complex derived words are attained later. However, their research focused on the fifth grade school children whose primary language spoken at home was English (Nippold and Sun, 2008: 367). First language acquisition does not concern specialized word list. Therefore this argument is not so strong.

Another problem Gardner and Davies (2013: 4) see is that the AWL was designed on top of the GSL (West, 1953). They compared the AWL word families with the top 4000 lemmas of COCA and their data give evidence that "the AWL is largely a subset of the high-frequency words of English and should therefore not be thought of as an appendage to the GSL, and the GSL, as a whole, is no longer an accurate reflection of high-frequency English" (Gardner and Davies, 2013: 5).

Gardner and Davies (2013: 5) also report on the reverse of the problem described above. Words with a high frequency but with a clearly academic meaning were excluded from the AWL. This may be exemplified by *company*, *interest*, *market*, *account*. Similar to the example *disease* discussed in section 2.2.2, these words were not considered for the AWL because they occurred among first 2000 GSL words. For a more detailed discussion see Neufeld et al. (2011); and Nagy and Townsend (2012). On the basis of the latest research, Gardner and Davies (2013: 8) formulate five design criteria for desirable properties of a new list of the common academic core vocabulary listed in (10).

(10) 1. The new list must initially be determined by using lemmas, not word families.
2. The new list must be based on a large and representative corpus of academic English, covering many important academic disciplines.
3. The new list must be statistically derived (using both frequency and dispersion statistics) from a large and balanced corpus consisting of both academic and non-academic materials.
4. The academic materials in the larger corpus, as well as the non-academic materials to which it will be compared, must represent contemporary English, not dated materials from 20 to 100 years ago.

5. The new list must be tested against both academic and non-academic corpora, or corpus-derived lists, to determine its validity and reliability as a list of core academic words.

The properties listed in (10) were used as design criteria in the development of the new Academic Vocabulary List by Gardner and Davies (2013: 8). An important step preceding the development of the AVL was corpus design. Gardner and Davies (2013) compiled a much larger corpus of academic texts. In contrast to Coxhead's 3.5 million word corpus, their corpus included 120 million words from academic journals from nine disciplines including medicine. All the words were tagged for grammatical parts of speech. The CLAWS 7 tagger from Lancaster University was used for tagging.[11] This means that lemmas rather than word families could be counted. Tagging allowed for "distinguishing between the verb used in *he used a rake* and the adjective used in *they bought a used car*" (Gardner and Davies, 2013: 9).

In order to establish the AVL, Gardner and Davies (2013: 9-12) applied four main criteria: ratio, range, dispersion, and discipline measure. Ratio is determined by the condition that the frequency of the word (lemma) was at least 50% higher in the academic corpus than in the non-academic portion of COCA (per million words). In contrast to Coxhead (2000), Gardner and Davies (2013) use ratio instead of the absolute frequency (at least 100 times in the Academic corpus) used by Coxhead. Range means that the word (lemma) had to be present with at least 20% of the expected frequency in at least seven of the nine academic disciplines. These included:

- Education,
- Humanities,
- History,
- Social Science,
- Philosophy, religion, and psychology,
- Law and political science,
- Science and technology,
- Medicine and health,
- Business and finance.

[11] An earlier version of the tagger is described by Garside (1987). CLAWS 7 refers to the tagset used. It is listed at http://ucrel.lancs.ac.uk/claws/

This again differs from Coxhead's use of range, which did not take into account the expected frequency measure. Coxhead's (2000: 221) range criterion required a member of a word family to occur at least 10 times in each of the four main sections of her corpus and in 15 or more of the 28 subject areas.

Dispersion is a criterion not applied in the development of AWL or MAWL. This feature "shows how 'evenly' a word is spread across the corpus, and it varies from 0.01 (the word only occurs in an extremely small part of the corpus) to 1.00 (perfectly even dispersion in all parts of the corpus)" (Gardner and Davies, 2013: 12). The authors consider the Range criterion secondary to Dispersion. They illustrate it with the example that if a word is found at the 20% expected level in seven of nine disciplines, but in, for instance, Science and Medicine it is much more frequent than in the remaining five, this level corresponds to a dispersion value below 0.80. Therefore, this word would be excluded from the list (Gardner and Davis, 2013: 12). The reasoning is that it is less evenly spread across the corpus. Gardner and Davies (2013: 12) admit that there is no research-recommended score for Dispersion, but their research results revealed that 0.80 did well in eliminating highly specialized words like *taxonomy*, *sect*, *microcosm*, etc. while keeping core academic words like *detect*, *relational*, *coercion* with Dispersion between 0.80 and 0.84.

The last criterion states that "the word cannot occur more than three times the expected frequency (per million words) in any of the nine disciplines. For example, *student* occurs in Education about 6.8 times the expected frequency (taking into account the size of the Education discipline); because this is above 3.0, the word was excluded from the academic core" (Gardner and Davies, 2013: 12). Then they compared the AVL with Coxhead's AWL word families. Their results indicate that the AVL "has nearly twice the coverage as the AWL" (Gardner and Davies, 2013: 19). This means that in a new corpus, AVL covers twice as many words (tokens) as AWL.

The AVL represents the state of the art of academic word lists at present and the main aim of its use should be in English for academic purposes.[12] One main difference with the AWL and MAWL is that Gardner and Davies reject the use of word families and prioritize lemmas. However, because the AWL has been an influential and widely used word list, a comparison with word families was also performed. The authors of the AVL do not deny the importance of discipline-based academic word

[12] The full AVL by Gardner and Davies (2013) is freely accessible at http://www.academicwords.info/ in the lemma format and the word family format.

lists. On the contrary, Gardner and Davies (2013: 21) express the hope that the AVL will be used for “research of English academic vocabulary in its many contexts”, which also includes medicine and medical English vocabulary. Building on the insights gained from the compilation of these lists, I will proceed to investigate the best method for compiling a list characterizing medical vocabulary in the next chapters.

CHAPTER THREE

METHODOLOGICAL CONSIDERATIONS

In the discussion of Coxhead's (2000) AWL, Wang et al.'s (2008) MAWL, and Gardner and Davies's (2013) AVL in chapter 2, three main problematic issues arose, namely the use of word families, the use of West's (1953) GSL, and the structure of the corpus. It is obvious that all of them are methodological issues. The problems are even more prominent when the main aim is to characterize medical vocabulary in English rather than just producing a word list for teaching. The former is the aim of this corpus-based study of medical vocabulary in English.

This chapter aims to demonstrate that if the main purpose is to characterize or describe medical vocabulary, the methodology used in pedagogical approach, which results in medical word lists, is not optimal, and to outline a more adequate alternative methodology. It is organized in three sections. Section 3.1 explains and justifies the use of the Corpus of Contemporary American English (COCA), one of the largest corpora available at present. It also shows how the methodological problems mentioned above can be solved. Section 3.2 concentrates on the medical subcorpus ACAD: Medicine in COCA. It discusses how the structure of the medical corpus influences the characterization of medical vocabulary. Section 3.3 gives an overview of the procedure applied here to arrive at the characterization of medical vocabulary. It explains why it is better to approach medical vocabulary from the perspective of a cline or continuum based on two dimensions: absolute and relative frequency. Determining the threshold values for each of these dimensions is a crucial criterion. This section shows what effect the different threshold values might have on the structure or description of medical vocabulary. Finally, section 3.4 summarizes the main results.

3.1 The Corpus of Contemporary American English (COCA)

Three methodological problems concerning corpus design as a basis for AWL or MAWL were identified in chapter 2: the use of word families, the use of the GSL, and the structure of the corpus. Let us consider each of them in turn.

The first problem is visible when we consider the words in MAWL that do not occur in AWL. Whereas AWL contains many words that have a large word family and refer to general concepts used in academic reasoning, MAWL also has more specific words, which refer to concepts of medical reality, e.g *cell*, *dose*, *tissue*, *liver*. This casts doubt on the usefulness of word families in compiling specialized vocabulary lists. They work very differently for this type of words than for the general academic words (e.g. *demonstrate*) we find in AWL. Whereas for AWL, the full extent of word families is listed in an appendix, there is no such information available for MAWL. Another disadvantage of word families is that they combine elements from different word classes so that the distinction between word classes is lost (Gardner and Davies, 2013). For instance, for *dose*, the frequency values for the noun and verb are combined. However, in describing medical vocabulary, we are interested in the difference between the values for the nominal and verbal readings of *dose*. This suggests that for characterizing medical vocabulary, lexemes are a better unit than word families.

The second problem concerns the gaps in the selected vocabulary. An example is *disease*, discussed in more detail in section 2.4. This word is not found in MAWL because it occurs among the first 2000 GSL vocabulary items, but it was demonstrated that *disease* is a relevant word in medical vocabulary. Here I will illustrate this issue with another example, *blood*. Similarly to *disease*, it does not appear in AWL and MAWL because GSL lists it (number 784). However, the example in (1) shows that in medical texts *blood* tends to occur with *cell*, which is the most frequent headword in the MAWL.

(1) Phagocyte is a type of white **blood cell** that eats bacteria and waste.[1]

Finally, evaluation of *red* **blood cell** velocity would have been yet another method for evaluating immediate changes in blood *flow*.[2]

[1] The source of this example in COCA is Helman, Amanda L.; Calhoon, Mary Beth; Kern, Lee (2015) 'Improving Science Vocabulary of High School English Language Learners With Reading Disabilities', *Learning Disability Quarterly*, 38 (1): 40-52.

[2] The source of this example in COCA is Galdyn, Izabela; Swanson, Edward; Gordon, Chad; Kwiecien, Grzegorz; Bena, James; Siemionow, Maria; Zins, James

The examples in (1) are from COCA. It is interesting to see how *blood* and *cell* collocate in this corpus. We also see the collocation with *red* and *flow*. The frequency of collocates for these two words with a 1-item distance on the right and left is 264. The search for the most frequent collocates for *blood* with a 1-item distance on either side shows that the top eight collocates with frequencies ranging from 6925 to 818 are *pressure*, *high*, *vessels*, *sugar*, *flow*, *cells*, *test*, *red*. Except *cells*, none of these collocates can be found in MAWL. The reason is again that these words are among the top 2000 items in the GSL. Another interesting observation about the example in (1) is that in certain contexts not only *blood* and *cell* co-occur, but they also combine with other top collocates, *red* and *flow*. However, Biber, Conrad, and Reppen (1998: 265) point out that "frequency information alone may present a biased measure of the strength of associations between words". In order to determine more accurately how strong the association between two words is, the *mutual information index* or *mutual information score* is "the most commonly used measure of collocation" (Biber and Jones, 2009: 1298). As Biber, Conrad, and Reppen (1998: 265) state "it focuses on the likelihood of two words appearing together within a particular span of words". This means that the mutual information index (MI) does not show how frequent individual words are. It does not show either how frequent the combination of two words is. It indicates how strong the tendency of the two words to occur together is. The search for collocates in COCA gives the information about the MI scores.[3] The MI scores for the most frequent collocates of *blood* are presented in Table 3-1.

(2015) 'Microcirculatory effect of topical vapocoolants', *Plastic Surgery*, 23 (2): 71-76.

[3] In COCA, Mutual Information is calculated as follows:
MI = log ((AB * sizeCorpus) / (A * B * span)) / log (2). Suppose we are calculating the MI for the collocate *color* near *purple* in BYU-BNC.
A = frequency of node word (e.g. *purple*): 1262
B = frequency of collocate (e.g. *color*): 115
AB = frequency of collocate near the node word (e.g. *color* near *purple*): 24
sizeCorpus= size of corpus (# words; in this case the BNC): 96,263,399
span = span of words (e.g. 3 to left and 3 to right of node word): 6
log (2) is literally the log10 of the number 2: .30103
MI = 11.37 = log ((24 * 96,263,399) / (1262 * 115 * 6)) / .30103, available at http://corpus.byu.edu/mutualInformation.asp , retrieved 24 January, 2016

Collocation	Frequency	All	MI
blood pressure	6925	56529	8.78
high blood	2094	244551	4.94
blood vessels	1722	7386	9.70
blood sugar	1542	32602	7.40
blood flow	1297	23891	7.60
blood cells	1008	28710	6.97
blood test	866	81730	5.24
red blood	818	105304	4.80

Table 3-1 Top eight collocates for *blood* in COCA

Table 3-1 gives the eight most frequent collocates for blood. The Frequency gives the total frequency of the given expression in the corpus, e.g. *blood pressure* occurs 6925 times. The column All gives the information about the frequency of the collocate in the corpus, regardless of context. For example, *pressure* occurs 56529 times in the corpus. The frequency of *blood* is 58628. The last column gives the MI score as given by COCA. MI scores around 3.0 or higher indicate 'a semantic bonding' between the two words.[4] All collocates in Table 3-1 have an MI index above 3.0 which COCA gives as a threshold for significance. This suggests that the relationships among these words are relevant in medicine. If MI is 1.0 then the distribution is random. If the two items are semantically and syntactically compatible, then higher values are expected.

It is obvious that AWL and MAWL use GSL as an exclusion list. Gardner & Davies (2013) object to the use of GSL, because it is an old list. However, if we want to avoid such gaps, any exclusion list will be problematic. A much better measure is relative frequency. In this method, words are selected when their frequency in the specialized corpus is significantly higher than in a general language corpus. Gardner and Davies (2013) also argue for the use of relative frequency as an alternative. In AVL, they offer users the relative frequency of a word across academic disciplines as a type of information (Gardner & Davies, 2013: 21).

[4] This information is available at http://corpus.byu.edu/coca/help/display_table_simple_e.asp, retrieved 9 February, 2016.

Finally, it is worth taking a critical look at the structure of the corpora. Section 2.2.2 describes the methodology Coxhead (2000) used in compiling a highly structured corpus. This structure was essential to exclude biased frequencies. This may be important for AWL, but in a characterization of medical language, we will in any case have more names of specialized concepts that appear in medical reality. This suggests a different approach. The subcorpora have the effect of eliminating words that are characteristic of a small range of subdomains. It is questionable whether this effect is desirable in a characterization perspective. A larger, but still balanced corpus is likely to give a better characterization. Moreover, Coxhead (2000) and Wang et al. (2008) stipulate threshold values without arguing for them or showing what the effect of different values would be. I argue that thresholds should be determined on the basis of the analysis of the effects they have.

In view of these observations, I propose a new methodology for compiling a list of medical vocabulary that can be used to characterize medical English. It should be based on lexemes rather than word families as units, relative frequency rather than an exclusion list and a less strict compartmentalization of the corpus. An alternative to a highly structured corpus is using dispersion as a measure, which also excludes too much influence of individual texts. Dispersion is a measure which tells us how ‘evenly’ a word is spread across the corpus. Its value varies from 0.01, which means that the word only occurs in an extremely small part of the corpus, to 1.00 which indicates a perfectly even distribution in all parts of the corpus (Gardner and Davies, 2013: 12).

Obviously, the way a corpus is compiled and processed is a central issue. Compiling a corpus and preparing it for linguistic research is a complex task influenced by at least three important factors. First, it depends on the available software for tagging and querying. This determines the syntactic and semantic variation that can be retrieved. Hyland and Tse (2007) give examples of words that may have different significance in different disciplines. For instance, *process*$_{\mathrm{N}}$ is more frequent than *process*$_{\mathrm{V}}$ in science and engineering but not in social sciences (Hyland and Tse, 2009: 117). Second, it depends on the size and structure of the relevant corpus or corpora. Last but not least, it is determined by the central research questions related to the characterization of medical vocabulary in English. These factors played a crucial role in taking the decision whether to use an existing large corpus with advanced functionalities that help solve meaning-variation problems or design a new one.

The Corpus of Contemporary American English (COCA) is one of the largest corpora of English available at present.[5] The corpus was created by Mark Davies, Professor of Corpus Linguistics at Brigham Young University and its popularity among professional and non-professional users is increasing. COCA has more than 520 million words in 220,225 texts and is balanced in the sense that it is equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. At the same time it is balanced in the sense that it includes 20 million words for each year from 1990-2015. The corpus is regularly updated by adding an annual portion as a supplement. The COCA web page indicates that the most recent texts are from December 2015. Table 3-2 illustrates word counts for each of the five genres of spoken, fiction, popular magazines, newspapers, and academic journals.

genre	number of words
spoken	109,391,643
fiction	104,900,827
popular magazines	110,110,637
newspapers	105,963,844
academic journals	103,421,981
TOTAL	**533,788,932**

Table 3-2 Distribution of word counts across genres in COCA from 1990-2015[6]

In Table 3-2 we can see what an even distribution across five genres means in practice. It means that word counts in each genre are approximately the same. The subcorpus spoken includes transcripts of unscripted conversation from more than 150 different TV and radio programs. Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts occur in fiction. The subcorpus popular magazines contains texts from nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). The subcorpus newspapers covers ten newspapers from across the US, including *USA Today*, *New York Times*, *Atlanta Journal-Constitution*, *San Francisco*

[5] All facts about COCA in this section are available at http://corpus.byu.edu/coca, information retrieved 13 January, 2016.

[6] Data available at http://corpus.byu.edu/coca , retrieved 13 January, 2016.

Chronicle, etc. In most cases, there is a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc. Finally, the subcorpus academic journals is compiled from nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.).

Copyright policy is important to mention because it influences the text collection process in COCA. The majority of the texts are copyrighted and COCA uses them under the so-called *Fair Use* agreement. The COCA Fair Use policy is based on four criteria. First of all, the corpus is for academic use only, commercial use is prohibited. Then, the COCA users do not have access to full versions of copyrighted texts, for instance to newspapers articles, research journal articles or short stories. COCA allows registered users to investigate the collection of texts via a web interface. In most cases, the users can see outputs of their search in tables with relevant information, e.g. frequency. The function tool *Keyword in Context* (KWIC) makes it possible to display small portions of the original text, normally only a few words on the right and left of the word a user searched for. COCA users must be logged on and depending on their status they can do a limited number of searches per day. The number of queries per day ranges from 20 for an unregistered user, to 400 for a university professor or graduate student in the field of language or linguistics. These two criteria, a limited number of searches per day and small portions of the original text make it almost impossible to reconstruct even such a small part as a paragraph of the original text. This means that there is virtually no effect on the potential market. The final criterion is associated with the nature of the copyrighted work. Although around 20% of creative work is included in the corpus, e.g. short stories and small portions of novels, the remaining 80% of the copyrighted materials are transcripts of TV shows, and articles from newspapers, magazines, and academic journals. They have a reduced commercial value.

The COCA interface has functionalities which allow users to search for exact words or phrases, lemmas, part of speech (PoS), and collocates. Another important type of information provided by the corpus is frequency. It is possible to compare the frequency of words, phrases, and grammatical constructions by genre and over the time. An example of the former is e.g. a comparison between spoken, fiction and academic, or even between sub-genres, e.g. newspaper editorials or sports magazines. The frequency over time makes it possible to compare different years from 1990 to present. I will not present more details of these functionalities, because with the exception of frequency the others were beyond the main focus of my research.

Let us turn now to word frequency data that can be retrieved from the COCA corpus. COCA offers a number of different formats available for the 5,000-60,000 word lists. The COCA word list with genre frequency included and comprising 60,000 words was central for my research aimed at the characterization of medical vocabulary. The list is available in Excel format, it is a large file that can be printed and copied. A sample illustrating the information available in this wordlist is given in Table 3-3.

rank	lemma / word	PoS	disp	totFreq	spok	fic	mag	news	acad	M1	M2	N1	N2	A1	A2
25083	piglet	n	0.88	239	20	97	54	46	22	10	2	3	3	0	2
25088	woodsman	n	0.70	300	10	176	77	12	25	1	2	1	3	2	0
25090	candied	j	0.87	242	17	49	102	73	1	0	1	2	1	0	0
25093	metacognitive	j	0.69	306	0	0	0	0	306	0	0	0	0	0	0
25107	industry-wide	j	0.89	236	16	2	64	109	45	19	10	2	1	10	6
25108	health-food	j	0.85	246	10	19	154	55	8	6	4	7	1	0	2
25110	posterior	n	0.88	240	6	30	36	27	139	0	5	4	0	0	99

Table 3-3 A sample from the word list with genre frequency in COCA[7]

Table 3-3 gives an overview of what kind of information is available in the word frequency data in COCA. The column *rank* gives the position in the COCA frequency list based on the total corpus. The next column includes

[7] Due to space constraints, in this sample only six of the 40+ sub-genres are shown (M1: MAG-Financial; M2: MAG-Science/Tech; N1: NEWS-Sports; N2: NEWS-Editorial; A1: ACAD-Law/PolSci; A2: ACAD-Medicine). The green shading for the five main genres highlights those words whose frequency in that genre is at least double what would otherwise be expected (based on genre size). This table and explanatory notes are available at http://www.wordfrequency.info/sample.asp, retrieved 13 January, 2016.

a lemma (word).[8] The abbreviation *PoS* refers to part of speech. The part of speech tagger CLAWS was used for determining the word class. The most often used labels are *n* (noun), *v* (verb), *j* (adjective) and *r* (adverb).[9] According to information given on the COCA website, the COCA team spent 2-3 months manually correcting the part of speech tags. There were a number of difficult cases such as -*ing* words, for example, differentiating between -*ing* nouns and verbs: *learning*, *meeting*, *thinking*, *beginning*, *living*, *teaching*, *reading*, *feeling*, etc., or verbs and adjectives: *leading*, *following*, *growing*, *changing*, *developing*, *missing*, *supporting*, etc. The COCA part of speech tagging is very good, although the COCA team admits it may not be perfect. No percentage of accuracy is given on their website.[10] Compared to MAWL and AWL, this methodology is much more accurate and advanced.

Dispersion statistics shows how evenly a word (lemma) is distributed among all parts of a corpus. It ranges from 0.01 to 1.00. The minimum shows that the word can be found only in a very small part of the corpus, whereas the maximum points to an even dispersion across all parts of the corpus. The highest dispersion score in Table 3-3 is for *industry-wide*. It means that it can be found fairly evenly in all five major genres: spoken, fiction, popular magazines, newspapers, and academic and also their subdomains.

Total frequency shows how many times a word occurs in the whole corpus. The remaining columns represent frequencies in each genre and subdomain. For instance, *posterior* is most frequent in the academic subcorpus, but it does not occur in M1: MAG-Financial, N2: NEWS-Editorial, and A1: ACAD-Law/PolSci. In contrast, its frequency in A2: ACAD-Medicine is 99. The green shading of the frequency 139 in the column academic means that *posterior* occurs at least twice as many times in this genre as would be expected if the distributions were even.

All things considered, COCA has many advantages over a newly created corpus. The most relevant properties for my research can be summarized as follows:

[8] Lemma is used in the sense of inflectional relationships only, e.g. for the verb $\text{WORK}_{\text{V,}}$ the forms *work*, *works*, *worked*, and *working* are not listed separately.

[9] The full list of part of speech tags is available at http://ucrel.lancs.ac.uk/claws7tags.html , accessed 13 January, 2016. The full list of part of speech tags used in this monograph is available in Appendix 1.

[10] Information available at http://www.wordfrequency.info/100k_faq.asp, retrieved 9 February, 2016

1. COCA makes use of lexemes, not of word families.
2. COCA is the largest and most balanced corpus available at present.
3. The texts are recent and new texts are added regularly. What makes COCA unique among corpora of English is that it offers roughly the same genre balance from year to year.
4. COCA gives frequency data based on accurate PoS tagging for the texts in the corpus.
5. Frequency data are available not only for the whole COCA, but also across the five main genres and their subdomains, including medicine.
6. COCA follows the copyright Fair Use policy.

On the basis of these properties I decided that COCA is currently the best corpus of English for my purpose.

3.2 The medical subcorpus in COCA

From the perspective of the characterization of medical vocabulary it is essential that the selection of COCA includes a medical subcorpus ACAD: Medicine. This is one of the subdomains in the genre academic. The academic corpus is divided into 10 subdomains given in Table 3-4.

The total number of words in the academic corpus is 103,421,981. This is the most recent update in December 2015 available for download at the COCA website. The distribution over the time is presented in Table 3-5.

In Table 3-5 we can see that the corpus academic in COCA is extended by around 4 million words annually. The table demonstrates a thoughtfully planned process of corpus building, year by year from 1990 up to the present and in roughly the same proportion each year. This compiling strategy is also a major advantage of COCA. In contrast, another well-known, large corpus, The British National Corpus (BNC), covers a number of genres, but it is static and, as it was compiled in the early 1990s, by now slightly outdated. On the COCA website, we also find comparisons to other corpora, e.g. the Oxford English Corpus (OEC). This is a corpus which is more recent, up to 2006, but the genre balance is more varied from year to year. This means that it is not always possible to determine if the changes in the corpus indicate changes or developments in the language or reflect only changes in the composition of the corpus. In contrast to COCA and BNC, OEC is not freely available, its use is restricted to researchers working on projects for Oxford University Press.

subdomains of the genre academic	number of words (in millions)
ACAD: History	12.64
ACAD: Education	14.11
ACAD: General (Journals)	0.76
ACAD: Geog/SocSci	16.75
ACAD: Law/PolSci	8.99
ACAD: Humanities	13.68
ACAD: Phil/Rel	7.00
ACAD: Sci/Tech	14.87
ACAD: Medicine	8.93
ACAD: Misc	4.70

Table 3-4 Number of words in the ten subdomains of the genre academic in COCA 1990-2015[11]

YEAR	# WORDS in ACAD	YEAR	# WORDS in ACAD	YEAR	# WORDS in ACAD
1990	3,943,968	**2000**	4,053,691	**2010**	3,816,420
1991	4,011,142	**2001**	3,924,911	**2011**	4,064,535
1992	3,988,593	**2002**	4,014,495	**2012**	4,300,876
1993	4,109,914	**2003**	4,007,927	**2013**	3,467,083
1994	4,008,481	**2004**	3,974,453	**2014**	3,383,971
1995	3,978,437	**2005**	3,890,318	**2015**	3,523,931
1996	4,070,075	**2006**	4,028,620	---------	------------
1997	4,378,426	**2007**	4,267,452	-------	------------
1998	4,070,949	**2008**	4,015,545	--------	------------
1999	3,983,704	**2009**	4,144,064	---------	------------

Table 3-5 Distribution of number of words added to ACADEMIC in COCA from 1990-2015[12]

From the COCA website an Excel file can be downloaded with a listing of the 220,225 texts in the 520+ million word corpus. The file contains six sheets. The sheet [TOTALS] has the overall totals by each of the five genres (spoken, fiction, magazine, newspaper, and academic) in each year,

[11] Available at http://corpus.byu.edu/full-text/, retrieved January 13, 2016.

[12] A full version of the table with all five genres is available at http://corpus.byu.edu/coca/ , retrieved 13 January, 2016.

1990-2015. The remaining sheets display a listing of all of the texts in each of the five genres.

On closer inspection, the listings provide a detailed description of the structure of the subcorpus ACAD: Medicine. It shows that the texts are taken from 51 scientific medical journals from different medical disciplines, published between 1990 and 2015, e.g. *American Journal of Public Health*, *ENT: Ear, Nose & Throat Journal*, *Emerging Infectious Diseases*, *Canadian Journal of Plastic Surgery*, *Internet Journal of Cardiology*, *Internet Journal of Gastroenterology*, etc. The selection of journals was determined by their online availability and copyright criteria. In contrast to MAWL the strict subdivision into particular subdomains was not decisive for ACAD: Medicine. The listing is produced in a systematic way including information about the identification number of a text, word count, year, domain, journal, title of the research article, publication information, and subject of the text. Depending on the purpose of the search the data in the file can be categorized in a number of different ways. This makes it the largest corpus of recent medical texts available at present. The corpus compiled for MAWL was eight times smaller, 1,093,011 running words. As mentioned above, another important advantage of ACAD: Medicine is its balanced structure over time. The design of this corpus required a team of specialists, specialized sofware, and a considerable amount of time. Therefore, ACAD: Medicine in COCA is the best possible choice, especially for a linguist who needs a relevant medical corpus for linguistic analysis, but does not have such a specialized team for designing a new corpus of medical texts.

3.3 ACAD: Medicine in COCA and the characterization of medical vocabulary

Word frequency data for ACAD: Medicine in COCA are the starting point for my analysis of medical vocabulary. As mentioned in section 3.2, different formats are available for the 5,000-60,000 word lists. In this study the 60,000 frequency word list with frequencies in COCA including the frequency counts in the five main genres and 40 subdomains or subcorpora was used. These data were used to search for answers to the following research questions:

- How can the total frequencies for the whole COCA be compared with the frequencies in ACAD: Medicine?

- How can the results of the comparison of these frequencies be interpreted in terms of characterization of medical vocabulary in English?
- What do the results of the comparison of the frequencies tell us about medical vocabulary in English?

In Table 3-3 it was illustrated what information is given in the frequency word list. For this analysis, rank, lemma, PoS, total frequency and frequency in ACAD: Medicine were the most relevant. The first step was to compare the frequencies for the whole corpus with the frequencies of the words occurring in ACAD: Medicine. The word list shows that 27,166 lexemes out of the 60,000 in the full list occur in ACAD: Medicine.

There is a great difference in size of the two corpora on which these two lists are based. In order to make these frequency data comparable we need to turn the raw frequencies into normalized frequencies to 10,000 words.[13] Turning raw to normalized frequencies was performed separately for the general COCA corpus and for ACAD: Medicine. An example of some words with their raw and normalized frequencies in both corpora is given in Table 3-6.

Word (lemma)/rank	PoS	Total freq. COCA	Norm. freq. COCA	Raw freq. ACAD: Med.	Norm. freq. ACAD: Med.
the (1)	a	22995858	616.7468	308224	682.3573
because (79)	c	459382	12.32058	3976	8.802211
study (240)	n	183613	4.924484	15246	33.75214
toxicologist (26851)	n	210	0.005632	4	0.008855
rhinitis (31340)	n	186	0.004989	178	0.394063

Table 3-6 An example of raw and normalized frequencies for five words in general COCA and ACAD: Medicine

[13] The reference number of a common base varies across corpus studies from 1,000 to 1,000,000. McEnery, Xiao and Tono (2006: 53) explain that "the common base for normalization must be comparable to the sizes of the corpora (or corpus segments) under consideration". They give as an example that comparing two corpora one with 10,000,000 words and the other with 90,000,000 words using the reference number 1,000 words would be inappropriate "as the results obtained on an irrationally enlarged or reduced common base are distorted" (McEnery, Xiao and Tono, 2006: 53).

Table 3-6 presents five randomly selected words. Their rank in COCA varies greatly, from the highest position to more than 25,000 ranks lower. There are three nouns, an article and a conjunction. The raw frequency of the definite article *the* in the general COCA corpus is considerably higher than its raw frequency in ACAD: Medicine. However, it is striking that the comparison of the normalized frequencies indicates it is proportionally more frequent in ACAD: Medicine. This may be explained by the fact that the definite article is typically used before names of organs, e.g. *the lungs*, *the spleen*, *the liver*, etc. and names of the systems of the body, e.g. *the circulatory system*, *the lymphatic system*, *the respiratory system*, etc. Then there are a number of other contexts requiring the use of definite article, especially the second mention. For instance, *If **injections** are not working well, it is important to document precisely the effect of **the injections** (percentage of improvement, duration of benefit, side effects).*[14] Corpus-based research also shows that nominal phrases dominate in medical texts. Segura-Bedmar et al.'s (2010: 95) results show that noun phrases are most frequent in their medical corpus. Their findings also indicate that nominal phrase anaphora is prevalent in biomedical texts and that "these phrases consist of the definite article *the*, possessives *its*, *their*, demonstratives *this*, *these*, *those*, distributives *both*, *such*, *each*, *either*, *neither* and indefinites *other*, *another*, *all* followed by a generic term for drugs (such as *antibiotic*, *medicine*, *medication* etc)" (Segura-Bedmar et al., 2010: 96).

It is interesting to observe that *because* ranks among the top 100 words in general COCA. MAWL does not include function words although the COCA frequency list shows they play an important role especially from the perspective of syntax. In medicine, it is relevant to explain phenomena in terms of cause and effect relationships. The difference between the raw frequencies and the normalized frequencies shows that both counts are lower in the subcorpus ACAD: Medicine. However, the difference between normalized counts is much less striking. This may indicate that causality in medical texts is expressed by other means.

The word *study*$_{\mathrm{N}}$ is an example of a word which occurs more frequently in the general corpus, but where the normalized counts indicate its frequency is notably higher in the specialized subcorpus. The subcorpus ACAD: Medicine contains research articles, which may also explain the relative frequency of the noun *study*.

Rhinitis and *toxicologist* represent terms with very low frequencies as opposed to the examples described above. The methodology applied in

[14] Retrieved 16, January, 2016, available at http://www.dystonia.org.uk/index.php/about-dystonia/treatments/botulinum-toxin-injections/when-botulinum-toxin-is-not-working

MAWL compilation aims at excluding terms in the narrow sense due to their low frequency across subdisciplines. Such terms are too domain-specific to meet the principle of range, i.e. “members of a word family had to occur at least in 16 or more of the 32 subject areas” (Wang et al., 2008: 447). In COCA, there are probably two main reasons why they are included in the frequency word list. First of all, the COCA word list contains 60,000 words whereas MAWL has only 623 word families. Second, the dispersion value for *toxicologist* is 0,882 and for *rhinitis* 0,623 (dispersion values are not included in Table 3-6). In the case of *rhinitis* almost all occurrences in COCA are in the medical subcorpus. The normalized frequencies are in all these examples less than 0.5, which points to their rare use also in narrowly specialized contexts.

After producing normalized frequencies for the general COCA corpus and the ACAD: Medicine subcorpus, the relative frequency was calculated. The term *relative frequency* may be used in different senses in corpus linguistics. Therefore it is worth explaining its different possible uses. These are illustrated in (2).

(2) a. “a *normalised frequency* (or relative frequency)…answers the question ‘how often might we assume we will see the word per *x* words of running text?’ Normalised frequencies (*nf*) are calculated as follows, using a *base of normalisation*:
nf = (number of examples of the word in the whole corpus ÷ size of corpus) x (base of normalisation).” [McEnery and Hardie (2011: 49)]
b. “..[w]hat is needed are the *observed relative frequencies*, which are typically normalized and reported as frequencies per 1,000 or 1,000,000 words. ….It is therefore important to bear in mind that one can only compare corpus frequencies or use them to make statements about what is more frequent when the frequencies have been normalized. Second, relative frequencies can be used to compare different corpora with each other just by computing the *relative frequency ratio*, the quotient of the relative frequencies of a word in both corpora (Damerau 1993). ” [Gries (2010: 271-272).]

McEnery and Hardie (2011: 49) in (2a) make the terms *normalized frequency* and *relative frequency* synonymous. Interestingly, Crawford and Csomay (2016) do not use the term *relative frequency* in their recent textbook on corpus linguistics at all, but they use the same formula as in (2a) to calculate normalized frequency. In (2b) we can see that *observed* *relative frequency* is contrasted with *relative frequency ratio*. The former is fully compatible with the use of *normalized frequency* by McEnery and Hardie (2011) in (2a) and Crawford and Csomay (2016). The latter is dependent on the former, but they are not identical. Gries (2010: 272) exemplifies it with the frequencies of *give* in spoken and written corpora. After normalizing frequencies in both corpora, the relative frequency ratio between the two corpora is 1.37. It was calculated by taking first the frequency of occurrence of *give* in the spoken corpus multiplied by 1,000,000 and divide it by the total number of words in the spoken corpus, which yields a normalized frequency score 465.75. The same procedure was applied to calculate the normalized frequency in the written corpus, resulting in a score of 339.96. In order to calculate the relative frequency ratio, the normalized frequency score in the spoken corpus was divided by the normalized frequency in the written corpus, yielding 1.37.[15]

In my analysis I will use the term *relative frequency* in the sense of relative frequency ratio in (2b) by Damerau (1993) and Gries (2010). Table 3-6 presented a sample of the normalization process in the general COCA corpus and the ACAD: Medicine subcorpus. The next step in my analysis was to calculate the relative frequency on the basis of the normalized counts in the two corpora. The relative frequency is calculated by taking the normalized frequency of the medical corpus and dividing it by the normalized frequency of the general corpus. This is illustrated in Table 3-7.

Word (lemma)/rank	PoS	Norm. freq. COCA	Norm. freq. ACAD: Med.	Relative freq. Med./COCA
odynophagia (57981)	n	0.000617	0.050918	82.54449
patient (572)	n	2.034344	54.88763	26.98051
mortality (4706)	n	0.163467	2.525987	15.45255
need (132)	v	7.829885	7.903394	1.009388
tonight (911)	r	1.360198	0.002214	0.001628

Table 3-7 A sample of relative frequency scores for five words occurring in general COCA and ACAD: Medicine

[15] This is based on the following calculations: *give* spoken: 297x1000000÷637682=465.75; *give* written: 144x1000000÷ 423581=339.96; 465.75÷339.96=1.37. (Gries, 2010: 271)

The words in Table 3-7 were randomly selected and sorted by the relative frequency score. The highest value is assigned to *odynophagia*, the lowest by *tonight*. These relative frequency scores can be seen as the values close to the ends of a continuum from highly specialized medical vocabulary to general vocabulary that may appear in medical texts in the corpus, but not to the extent to which the words are typical of medical texts. Relative frequency is a measure of typicality of a word in medical vocabulary. If the relative frequency values is close to 1, for instance for *need*, the frequency in general COCA and ACAD: Medicine is approximately the same, the normalized frequencies in the two corpora are very similar. If the relative frequency is higher, the word is more frequent in medicine than in the general corpus, e.g. *patient*, *mortality*. The extreme value for *odynophagia* confirms that it is a highly specialized medical term, but low normalized frequencies also suggest that it may not be necessarily frequent even in medical texts. As ACAD: Medicine is a subcorpus of COCA, all occurrences in the ACAD: Medicine subcorpus are also occurrences in COCA. Therefore, a value over 80 shows that (almost) all occurrences in COCA are in the ACAD: Medicine subcorpus. Higher relative frequency measures are not possible in this setting. The very low relative frequency value for *tonight* clearly shows that although the word occurs in ACAD: Medicine, it is not typical in medical vocabulary. At this point, the question arises how to determine a threshold value indicating when words are frequent enough to be considered characteristic of medical vocabulary.

In addition, we should not forget that the relative frequency is only one of the two crucial dimensions in characterizing medical vocabulary. The other continuum is the absolute frequency in the medical corpus. This is used to produce a threshold of words frequent enough in the medical corpus. The best value depends on the size of the corpus. Therefore, it is only possible to characterize medical vocabulary accurately when the relative frequency and absolute frequency are both taken into account.

Whereas the relative frequency can be calculated for any word, my aim was to identify only those words that have at least a certain absolute frequency in the medical corpus. Otherwise, rare words which are more frequent in ACAD: Medicine than in general COCA would also be included. This can be illustrated by the example of the word *exchangeable*. Its normalized frequency in ACAD: Medicine is 0.004428 whereas in general COCA it is 0.001314. The relative frequency value is fairly high, 3.37. However, the absolute frequency values are very low, only 2 in ACAD: Medicine and 49 in general COCA. This indicates that the absolute values are not sufficiently high to draw conclusions or generalizations.

In order to implement this idea, I first sorted the frequency word list by absolute frequency (total frequency) in the medical corpus, then selected the range within a threshold, e.g. 100,000, 10,000; 100, 90, etc. and re-sorted this by relative frequency. Table 3-8 shows possible threshold values and the number of words falling into each category.

Threshold level	Absolute frequency	Number of words
1.	308,224-107,344	6
2.	61,162-21,625	13
3.	18,329-10,103	14
4.	9819-9365	6
5.	8948-8175	4
6.	7855-7133	12
7.	6959-6014	11
8.	5925-5043	25
9.	4902-4006	22
10.	3983-3000	54
11.	2995-2003	129
12.	1995-1001	327
13.	999-1	26,543

Table 3-8 Threshold levels with word counts based on absolute frequency in ACAD: Medicine

Table 3-8 illustrates 13 threshold-determined classes and the number of words in each. The differences among the threshold levels from the perspective of word counts are remarkable. The highest threshold level includes only 6 words whereas the lowest covers 26,543 words. Starting from the threshold level ranked nine, the word counts gradually increase with a sharp increase in the final class. Taking into account that the total number of words (types) in ACAD: Medicine is 27,166, only 623 words are distributed among the first 12 threshold levels and the remaining 26,543 fall into the last threshold range. This means that the top 12 threshold levels represent only 2.29% of words as opposed to 97.71% on threshold level 13. Obviously, it is worth investigating what words are on some of the top 12 levels. Threshold level 1 is given in Table 3-9.

Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
the	1	a	308,224	1.106382
of	2	i	197,595	1.511305
be	3	v	188,494	1.18527
and	4	c	159,292	1.169854
a	5	a	107,607	0.836898
in	6	i	107,344	1.213356

Table 3-9 Absolute frequency in ACAD: Medicine and relative frequency ACAD: Medicine/COCA for threshold level 1

The top six words in Table 3-9 represent threshold level 1, with the highest absolute frequency in ACAD: Medicine. Interestingly, the rank indicates that these words are the top six also in the general COCA corpus. A closer look at the relative frequencies shows that they are all around 1.0, which suggests that their frequencies in general COCA and ACAD: Medicine are roughly the same. All these words are function words. This finding is entirely in line with the results in the frequency word lists based on the Cambridge International Corpus (CIC) by Carter (2012). Carter (2012: 103) emphasizes that "the function words dominate the top frequencies of both lists, and indeed, one of the defining criteria of function words is their frequency". Gardner (2013: 13) confirms that function words "tend to be high frequency in all types of communication". Function words also clearly dominate threshold levels 2, 3, 4, and 5 as can be seen in Table 3-10.

Threshold level	Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
2	to	7	t	61162	0.76194
2	for	13	i	55434	1.331599
2	to	9	i	53226	1.089812
2	with	16	i	49595	1.456549
2	have	8	v	39519	0.724322
2	that	12	c	37715	0.868724
2	or	32	c	27448	1.568969
2	by	30	i	27327	1.45068
2	on	17	i	26381	0.835574
2	patient	572	n	24793	26.98051
2	this	20	d	23610	0.984299
2	from	26	i	22106	1.063597
2	not	28	x	21625	1.09104

Threshold level	Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
3	at	22	i	18329	0.816882
3	it	10	p	15907	0.324492
3	study	240	n	15246	6.853945
3	as	49	i	14973	1.427913
3	use	92	v	13517	2.531168
3	health	344	n	12840	8.08704
3	as	33	c	12667	0.772611
3	these	82	d	12571	1.934441
3	their	36	a	12568	0.917369
3	can	37	v	11759	0.908932
3	do	18	v	11047	0.337979
3	we	24	p	10407	0.452123
3	which	58	d	10388	1.1993
3	may	119	v	10103	2.458534
4	they	21	p	9819	0.416737
4	child	115	n	9645	2.278981
4	than	73	c	9622	1.311325
4	all	43	d	9589	0.845158
4	group	163	n	9474	3.247969
4	also	87	r	9365	1.589169
5	who	38	p	8948	0.693832
5	other	75	j	8767	1.268219
5	case	186	n	8253	3.256388
5	between	140	i	8175	2.452636

Table 3-10 Absolute frequency in ACAD: Medicine and relative frequency ACAD: Medicine/COCA for the threshold levels 2-5

In Table 3-10 we can see that 30 words out of 37 covered by the threshold levels 2-5 are function words. For instance *to* counts as two words, because it has a different PoS, *to* with rank 7 is infinitival *to* whereas *to* with rank 9 is a preposition. Only the seven words in light grey shadowing are content words. The relative frequency scores of the function words vary. For instance, the lowest relative frequency 0.324492 in Table 3-10 is for the pronoun *it* from threshold level 3, which means it is much more typical of general vocabulary. Its high rank of 10 in general COCA tells us it is one of the top frequent words in COCA. Other words much more typical of general vocabulary than medicine are *do*, *we*, *they*, and *who*. It is not surprising that the first person pronoun is included. The research of specific features of academic writing showed that "first person pronouns help writers create a sense of newsworthiness and novelty about their work" (Livnat, 2012: 94). For more information about the use of pronouns in academic writing, cf. Myers (1992), Hyland (2002), Harwood (2005). It is interesting to observe that the relative frequency of *they* in Table 3-10 is less than 1.0 whereas for *these* it is nearly 2.0. This fact may be connected to a style of presentation typical of medical text, which prefers the more specific demonstrative determiner *these*, which combines with a noun, to *they*.

Then, there are function words with the relative frequency around 1 such as *to*, *this*, *from*, *not*, *their*, *can*, *which*, and *other*. This means that their frequency of occurrence in general COCA and ACAD: Medicine is very similar. Only a few function words, *or*, *these*, *may*, *also*, *between*, have relative frequency scores higher than 1.5. By contrast, all content words in Table 3-8 have relative frequencies higher than 2.0. For *patient*, *study*, and *health* it is considerably higher. This indicates these words are more typical of medical vocabulary. Another piece of evidence is their ranking in the general COCA corpus, it is much lower than for the function words in Table 3-8. For example, *patient* ranks 572 in the general COCA corpus whereas other words in threshold class 2 take up ranks between 7 and 32.

Gardner (2013: 54) argues for separating function words from content words in the top frequency lists in core vocabulary lists. He gives two main reasons for this decision from the pedagogical perspective. The first reason is connected with the so-called learning burden, which is different for function and content words. The latter require more attention to meaning and form. The second is that word lists can be misleading when a high portion is taken up by function words because they "do not impact meaning (thus comprehension) in the same way that content words do" (Gardner, 2013: 54). From the learner's point of view, separating function words from content words seems reasonable. However, from the point of view of characterizing medical vocabulary, an a *priori* elimination of function words would result in a distorted description by omitting an important part of vocabulary.

From threshold level 6 down, the number of content words increases significantly. Table 3-8 shows that threshold level 13 is the largest one. This means it is worth exploring it in more detail to find out how useful it can be in describing medical vocabulary. This threshold level can be subdivided into a number of sublevels. This is illustrated in Table 3-11.

Sublevels of threshold level 13	Number of words
1. 999-900	75
2. 899-800	84
3. 799-700	102
4. 699-600	150
5. 599-500	182
6. 499-400	239
7. 398-300	370
8. 299-200	677
9. 199-100	1423
10. 99-1	23,242

Table 3-11 Threshold level 13 with sublevels based on absolute frequency in ACAD: Medicine

Table 3-11 demonstrates that when threshold level 13 is divided into 10 levels, the word count increases with each sublevel. Similar to Table 3-8, the last sublevel is the largest one. In a corpus analysis, Zipf's law (originally formulated by Zipf 1935, 1949) might be used to approximate many types of data studied. The law states that the rank order of a word based on frequency values in a corpus is inversely related to its absolute frequency value. This means that the most frequent word occurs roughly twice as often as the second most frequent word and three times as often as the third most frequent word, etc. a study on applications of Zipf's law by Powers (1998: 151) claims that "the most frequent 150 words typically account for around half the words of a corpus, although this figure varies significantly with the size of the corpus, the size of the lexicon, the genre, register and medium of communication and the linguistic complexity of the text". In Table 3-9 we could see that short words also tend to be among the most frequent words. Li (1992) proved that words created by a random combination of letters are in line with Zipf's law. Li (1992) found that in his text generated by random combination of letters, the frequency distribution of word length was exponential. This means that words of length 1 occurred more than words of length 2, etc. The results suggested that frequency declined exponentially with word length, short words tend to occur frequently as opposed to long words. Li (1992: 1845) argues that "Zipf's law is not a deep law in natural language" as Zipf distributions are based on randomly-generated texts with no linguistic structure.

A characteristic feature of the thresholds 1-8 is that the absolute frequency in the subcorpus ACAD: Medicine decreases gradually with only a few words showing identical values. This is shown in Table 3-12.

Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
basis	1311	n	999	2.62367
refer	1157	v	996	2.30189
observation	2117	n	995	4.511743
protocol	4526	n	989	12.80773
third	584	m	987	1.158268
success	777	n	987	1.542669
expect	405	v	985	0.815379
direct	1483	j	985	3.00856
only	328	j	984	0.661917
communication	1254	n	984	2.448048
combination	1818	n	983	3.885702
duration	5486	n	982	17.06499
medication	3168	n	981	7.890105
error	2212	n	979	4.68796
consist	2204	v	976	4.666286

Table 3-12 Top 15 words in sublevel 1 (999-900) of threshold level 13

Table 3-12 shows one word with the absolute frequency value 999, one with 996, one with 995, two with 987, etc. The absolute frequency scores are decreasing gradually. Table 3-12 illustrates this for the top 15 words in sublevel 1 of threshold level 13. These include nouns (n), verbs (v), adjectives (j), and a numeral (m). Only in two cases, *expect* and *only*, is the relative frequency less than 1, which means these words are more typical of general vocabulary. All the remaining words have relative frequency scores (often considerably) higher than 1. This suggests these words are typical of medical vocabulary. The relative frequency indicates the degree of typicality.

A comparison of the words in Table 3-12 with MAWL shows that only four words overlap: *protocol*, *duration*, *error*, and *consist*. In MAWL, they are all placed among the top 200 headwords. The relative frequency in COCA for these words indicate that *protocol* and *duration* are more typical of medical vocabulary than *error* and *consist*, although the latter pair's relative frequencies still comfortably exceed 1.0. This raises a question about the threshold level of relative frequency relevant for typical medical vocabulary. In section 3.1 we saw that for the MI values given in

COCA, the reference point 3.0 or higher is assumed to indicate ‘a semantic bonding’ between the two words. Therefore it may be reasonable to consider this value also as a threshold for relative frequency values showing which words are more typical of medical vocabulary. This results in the four words in MAWL plus *observation*, *direct*, *combination* and *medication* to be included in medical vocabulary.

Let us now turn to somewhat less frequent words. Sublevel 5 of threshold level 13 is exemplified in Table 3-13.

Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
write	228	v	599	0.290356
terms	1289	i	598	1.5236
themselves	449	p	596	0.553473
above	894	i	596	1.071306
primarily	2147	r	596	2.793987
possibility	1197	n	594	1.456624
advantage	1305	n	594	1.600295
supply	1499	n	594	1.885897
extension	3211	n	594	4.876323
city	290	n	593	0.354135
concept	1159	n	592	1.223616
operate	1313	v	591	1.599783
experiment	2000	n	590	2.550471
perspective	1416	n	589	1.621488
temporal	6970	j	589	14.33334

Table 3-13 Top 15 words in sublevel 5 (599-500) of threshold level 13

Table 3-13 covers nouns (n), verbs (v), an adjective (j), an adverb (r), a pronoun (p), and prepositions (i). *Terms* is labelled as a preposition presumably because it occurs in *in terms of*. The relative frequency values for *write*, *themselves*, and *city* suggest that although they occur in medical texts, they are more typical of general than medical vocabulary. The relative frequency for *above* points to very similar frequencies in the general COCA corpus and the subcorpus ACAD: Medicine. Taking into account the rank in the general COCA corpus, it is obvious that the words more typical of medical texts tend to rank lower from the top, e.g. *experiment* (2000) than more general vocabulary items, e.g. *write* (228).

Compared to Table 3-12, we can see that in Table 3-13 there hardly are any gaps in the decreasing absolute frequency levels and often there is more than one word with the same absolute frequency, e.g. *themselves*, *above*, *primarily*, or *possibility*, *advantage*, *supply*, and *extension*. It is interesting to address the words with the relative frequency scores between 1.5-5.0, which include *advantage*, *supply*, *extension*, *operate*, *experiment*, *primarily*, *terms*, and *perspective*. Obviously at least *operate* and *experiment* may be immediately associated with a medical context. At the same time, the relative frequency values for these words indicate a weak point of relative frequency, which may be illustrated by the fact that the relative frequency of *experiment* is lower than *extension*. Intuitively, *experiment* is more characteristic for medical vocabulary than *extension*. The reverse order maybe because *experiment* is also frequent in other specialized domains, e.g. MAG: Sci/Tech, NEWS: Misc. This illustrates how the relative frequency is inherently dependent on the coverage of other domains used in the general corpus as a benchmark.

Let us now turn to sublevel 9 of threshold level 13 illustrated in Table 3-14.

Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
morning	357	n	199	0.137029
kitchen	1052	n	199	0.399863
interested	1284	j	199	0.541266
philosophy	2266	n	199	1.011849
capable	2408	j	199	1.129658
comfort	2945	n	199	1.45469
radical	2946	j	199	1.44306
ethics	3252	n	199	1.628468
conduct	3718	n	199	2.011062
fraction	4763	n	199	2.953318
resulting	5472	j	199	3.541689
dietary	6453	j	199	4.720217
stent	22318	n	199	40.459
cutaneous	28656	j	199	67.8775
violent	2171	j	198	0.996452

Table 3-14 Top 15 words in the sublevel 9 (199-100) of threshold level 13

In contrast to Tables 3-12 and 3-13, Table 3-14 shows that at this level the frequency bands tend to be larger, they include more than ten words, e.g. 14 words with the absolute frequency value 199. Similarly to Tables 3-12 and 3-13, a few words with a relative frequency smaller than 1 are included. It is expected that words such as *morning*, *kitchen*, *interested*, and *violent* are more typical of other parts of the corpus than medicine. The highest relative frequency values in Table 3-14 are for the words *cutaneous* and *stent*. These are specialized medical terms and their occurrence in COCA is restricted to the academic genre, especially in medicine. The total frequency of *cutaneous* in COCA is 242, so that the absolute frequency value of 199 in Table 3-14 means that it occurs mostly (over 80% of tokens) in medicine.

With a more fine-grained division of sublevels it is possible to show how the absolute frequency measure and the relative frequency measure interact. Table 3-15 gives the words with an absolute frequency of 100 in the medical corpus, sorted by relative frequency.

Table 3-15 illustrates that when the words in the absolute frequency band 100 are sorted by the relative frequency, the top words are specialized medical terms. High relative frequencies indicate that words such as *thyroidectomy* or *rotavirus* occur almost exclusively in the subcorpus ACAD: Medicine. Only five words in this frequency band, *occasional*, *firm*, *vast*, *religious*, and *character* are less frequent in medical vocabulary than expected on the basis of their overall frequency in COCA. This raises the question about the role of words with relative frequency values of less than 1 in medical vocabulary. The answer may not be straightforward. In Tables 3-12, 3-13, and 3-14 we saw that each threshold level includes vocabulary items less typical of medical texts. Their portion in the samples illustrated in the tables above is very similar. Kittredge (1982: 111) points out that "most articles in scientific journals have some degree of "contamination" from the general language". He gives an example of the word *dress* used in a research paper on subatomic particles, which clearly demonstrates that "most dynamic scientific sublanguages are constantly borrowing terms form the standard language, particularly when new concepts are being introduced and analogies are needed" (Kittredge, 1982: 110). Because *dress* is quite frequent in other domains, it is easy to miss this word on the basis of its relative frequency. Excluding words with a rather low relative frequency may therefore also exclude genuinely medical words. Therefore it seems safer to accept such words as part of medical vocabulary on the basis of their absolute frequency, even if their relative frequency is lower.

Word (lemma)	Rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
thyroidectomy	41094	n	100	80.14028
rotavirus	43623	n	100	65.5115
anthropometric	35114	j	100	63.49576
microbiology	21969	n	100	22.99289
fiber-optic	14871	j	100	10.12816
sclerosis	14298	n	100	9.643048
evacuation	8373	n	100	3.508053
informant	7217	n	100	2.750566
staff	6175	v	100	2.030615
nutrient	4896	n	100	1.479027
shared	4042	j	100	1.109022
occasional	3535	j	100	0.971798
firm	3269	j	100	0.875897
vast	1975	j	100	0.442717
religious	885	j	100	0.171247
character	786	n	100	0.157787

Table 3-15 Frequency band 100 in sublevel 9 (199-100) of threshold level 13 sorted by relative frequency ACAD: Medicine/general COCA

Another question arises concerning the role of threshold levels. There is no simple correlation between level of specialization and threshold levels. General words occur in all threshold levels and medical words in all but the highest ones. Another key factor is that frequency bands tend to be larger at lower thresholds, e.g. 199, and 100. Therefore it seems reasonable to investigate the situation with lower threshold levels. Table 3-16 shows the situation at the threshold level 99.

Word (lemma)	Rank	PoS	Absol. Freq. ACAD:Med.	Rel. Freq. Med./COCA
dysplasia	29578	n	99	43.93497
midline	28196	n	99	39.66944
cytoplasm	27938	n	99	38.91383
public-health	20601	j	99	19.27336
quadriceps	20495	n	99	19.04873
dizziness	15359	n	99	11.02821
diabetic	14852	j	99	10.21488
irrespective	15039	i	99	10.08877
masking	14348	n	99	9.202595
approximate	12165	v	99	6.919479
subset	12269	n	99	6.57434
for-profit	11826	j	99	5.900292
positioning	10984	n	99	5.883301
faucet	10980	n	99	5.84961
analog	10217	n	99	5.146036
solving	9621	n	99	4.236343
unchanged	8907	j	99	4.184283
notify	6193	v	99	2.300649
zero	5442	n	99	1.814366
contract	5039	v	99	1.61564
cure	4630	v	99	1.450462
project	3626	v	99	0.990054
sponsor	3476	v	99	0.914595
inquiry	3389	n	99	0.861833
modest	3120	j	99	0.803609
disagree	2695	v	99	0.631572
massive	2006	j	99	0.440605
cheese	2116	n	99	0.43217
else	440	r	99	0.089015

Table 3-16 Frequency band 99 in the sublevel 10 (99-1) of threshold level 13 sorted by relative frequency ACAD: Medicine/general COCA

Table 3-16 illustrates that the frequency band 99 covering 29 items is larger than the frequency band 100 in Table 3-14. The ordering based on the relative frequency confirms that the highest relative frequencies are for the terms in the narrow sense, e.g. *dysplasia*, and specialized terms, e.g. *dizziness*. The degree of specialization decreases with lower relative frequencies. The words more typical of general vocabulary are placed towards the other end of a continuum. Their relative frequency values are smaller than 1. This distribution pattern occurs across all threshold levels and frequency bands. It seems it would be arbitrary to exclude, for instance, the frequency band 99, because this would mean we would lose words such as *dysplasia*, *cytoplasm*, or *diabetic*, which certainly are typical examples of medical vocabulary. Taking into account how relative frequency interacts with absolute frequency, it is obvious that not so typical medical words have fairly high relative frequency values, e.g. *masking*, *for-profit*, or *unchanged* even if their absolute frequency values are low. Therefore it seems interesting to compare the data if the relative frequency values are constant as opposed to the absolute frequency values. An example is given in Table 3-17.

Word (lemma)	**Rank**	**PoS**	**Rel. Freq. Med./COCA**	**Absol. Freq. ACAD:Med.**
patient	572	n	26.98051	24793
fungal	12857	j	26.89503	405
MRI	17082	n	26.55453	212
cleft	25175	j	26.56604	84
antifungal	27275	j	26.70557	66
nonsteroidal	31109	j	26.91668	45
occipital	34030	j	26.82696	39
dilated	34286	j	27.01456	36
patella	35354	n	27.01456	36
elastin	38212	n	26.53216	27
crashworthiness	42849	n	26.66822	21
instrumented	43561	j	26.9857	17
age-matched	51078	j	26.77119	12
teacher-rated	53329	j	26.70557	11
spectrophotometer	54632	n	26.70557	11
colloid	54791	n	26.70557	11
reabsorption	50087	n	26.62726	10
transversely	50319	r	26.62726	10
nasal	51167	n	26.62726	10
wait-list	58990	j	26.53216	9
heterotopic	59001	j	26.53216	9

Table 3-17 An example of medical vocabulary sorted by relative frequency ACAD: Medicine/general COCA (frequency band between 26.5-27.02)

Table 3-17 suggests that there might be a correlation between the absolute frequency and relative frequency. It seems the lower the threshold in absolute frequency, the higher the threshold which must be adopted for the relative frequency. Perhaps the most striking exception is *patient* with the highest absolute frequency, then *fungal* and the abbreviation *MRI* follow. The remaining words belong to the final sublevel in threshold level 13 in Table 3-11. The lower-ranked items in Table 3-17 have a very low absolute frequency, which makes them less typical of medical vocabulary despite their high relative frequency, for instance, *age-matched*, or *teacher-rated*. It seems therefore that the measure of relative frequency alone is not sufficient.

3.4 Conclusion

The main aim of this chapter was to investigate the structure of medical vocabulary on the basis of the general COCA corpus and its subcorpus ACAD: Medicine. The results suggest it is better to view medical vocabulary in English as a cline than as a dichotomy with a clear-cut boundary.

The continuum approach to vocabulary acquisition is a generally accepted perspective, especially in second language acquisition research. The frequency factor is crucial in evaluating receptive and productive vocabulary knowledge from a pedagogical perspective. It is relevant in understanding the language learning process based on the interaction between receptive and productive dimensions, cf. Melka (1997). However, for Laufer and Goldstein (2004) and Schmitt (2010) it is not so obvious how to determine “the threshold at which receptive knowledge becomes productive” (Pignot-Shahov, 2012: 38). For Meara (1990), the receptive and productive dimensions are distinct. Meara (1990: 153) argues that the receptive dimension is qualitatively different from the productive one. Therefore, at least in this respect, Meara’s position seems rather oriented towards a dichotomy than a cline.

Words with a high frequency of occurrence tend to be shorter and “are not semantically restrained like words with a lower frequency” (Pignot-Shahov, 2012: 41). Pignot-Shahov (2012: 41) explains that such high-frequency words “can be used in a variety of contexts because there is no connotation or collocation attached to them”. According to Nation (2010) low-frequency words account for 5% of academic texts and they include proper nouns, and specialized or technical words. Frequency word lists inspired a number of research studies to verify the so-called frequency hypothesis based on the assumption that more frequent words are mastered by learners before less frequent words (Palmberg, 1987; Laufer and Goldstein, 2004; Milton, 2009; Nation, 2010; Schmitt, 2010).

From the perspective of the characterization of medical vocabulary, receptive and productive dimensions are not key factors, whereas frequency plays an important role. Absolute and relative frequency can be combined in a two-dimensional model representing medical vocabulary as given in Figure 3-1.

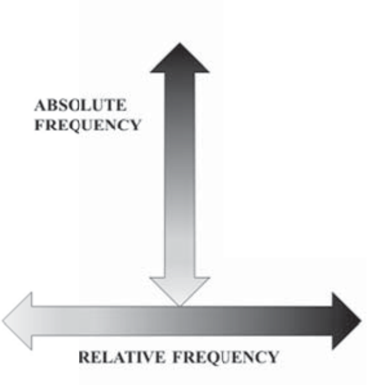


Figure 3-1 A model of medical vocabulary as a two-dimensional continuum based on absolute frequency and relative frequency

The model of medical vocabulary in Figure 3-1 is based on the interaction of two frequency dimensions: absolute frequency and relative frequency. It was demonstrated that relative frequency is a measure which indicates a degree of typicality of a word in medical vocabulary. The continuum is between two ends from general vocabulary to highly specialized vocabulary and terms in the narrow sense. The data show that specialized words tend to have much higher relative frequency scores than words less typical of medical vocabulary. a key question was how to set a threshold value to determine whether or not words are frequent enough to be considered typical of medical vocabulary. This was the point where the dimension based on the absolute frequency must be taken into account. The absolute frequency continuum was used to produce a threshold of words frequent enough in the medical corpus. Investigating the threshold levels in more detail revealed that especially with the frequency bands with a larger number of words it is only possible to characterize medical vocabulary accurately when the relative frequency and absolute frequency are both taken into account. Panocová (2016) argues that a measure based on the interaction of absolute frequency and relative frequency provides a better tool for identifying medical vocabulary than previously used measures. The data also demonstrate that all frequency bands contain words ranging from rather untypical of medical vocabulary to very typical of medical vocabulary. Therefore it is not possible to determine a threshold value without excluding words that are typical of medical vocabulary on the basis of their relative frequency scores. This was shown in Tables 3-15 and 3-16. This suggests that medical vocabulary is best characterized as a two-dimensional continuum where the two dimensions interact and depend on each other. The data indicate that the choice for a particular threshold is always to some extent arbitrary. The combination of the two values suggests that lower absolute frequency requires higher relative frequency. If the relative frequency value and the absolute frequency value are interconnected, as suggested by the discussion of the tables, the line separating medical vocabulary from general vocabulary in Figure 3-1 will become a curve. This raises the questions of what the precise parameters of the curve depend on and how it is possible to determine their values.

A starting point in the search for the answer to the above-mentioned question was inspired by the so-called Edmundsonian paradigm (Mani, 2001) in automatic text summarization. Edmundson (1969) laid down the foundations of work on automatic extraction. Extraction is based on the selection of units of texts which contain relevant information and their evaluation. The corpus in his experiment included 200 scientific papers. The main aims were to identify relevant features of the text, have a computer program to recognize them and weight them. Edmundson (1969) used four basic methods: the Cue method, the Key method, the Title method, and the Location method.

The Cue method was based on the assumption that hedging words such as *significant*, *impossible* or *hardly* determine the relevance of the sentence. The analysis was based on three statistical parameters: frequency (frequency of occurrence in the corpus), dispersion (number of scientific papers in which the word occurred) and selection ratio (ratio of number of occurrences in extractor-selected sentences to number of occurrences in all sentences of the corpus). The sum of the cue weights of the constituent words in each sentence represents the final cue weight for each sentence.

The Key method focused on identifying the document-specific key words. The words were sorted according to their frequencies starting from the highest frequency values downward. Non-cue words with frequencies above the threshold were labelled as key words. The threshold was not fixed but fractional. Edmundson (1969: 272) explains this means that “key words were chosen from a given percent of the total number of words in the document, and their key weights were taken to be their frequency of occurrence over all words in the document". The key weight of a sentence was then the sum of the key weights of individual words.

The Title method produced a title glossary covering all non-Null words in the title, subtitle, and headings of each scientific paper with positive weights. Then the title weight for each sentence was calculated as a sum of the title weights of its constituent words.

In contrast to the three methods described above, the Location method does not use a word, but a sentence as a characteristic feature. This method was based on two assumptions. First, headings such as Introduction, or Conclusion were manually created and sentences under these headings were assigned positive weights. Second, sentences in the first paragraphs and last paragraphs of a section were given positive weights.

In the final evaluation, these four methods were combined and their relative weights were parameterized in the linear function in (3):

(3) a. $a_1C + a_2K + a_3T + a_4L$ (Edmundson, 1969: 273)
b. W = a1 * C + a2 * K + a3 * T + a4 * L (Sarkar, 2014: 162)
c. $W(s) = \alpha C(s) + \beta K(s) + \gamma T(s) + \delta L(s)$ (Mani, 2001: 45)

In (3a) we can see the formula as presented in the original paper by Edmundson (1969). The formula played a key role in the development of summarization approaches. As a classic method it tends to be included in the history of approaches to automatic text summarization. It is referred to as Edmundsonian Methods by Sarkar (2014) or the Edmundsonian Paradigm by Mani (2001). The differences between the formulae in (3a), (3b), and (3c) are merely in formal expressions, the intended meaning is identical. Therefore I will describe in more detail the formula in (3c). In (3c) W represents the overall weight of a sentence *s*, C, K, T, and L are the values for the four calculation methods explained above and α, β, γ, and δ are their parameters i.e. positive integers. It is an empirical research question how to determine the optimal values of α, β, γ, and δ. These values regulate the weights of the four methods or features. Although the precise values of these parameters are not given, Edmundson's evaluation of the data set suggested that a method based on a combination of Cue, Title and Location values proved to be the most effective whereas the Key method used in isolation was the worst.

It seems that a similar method of relative weights among the dimensions in the model of medical vocabulary given in Figure 3-1 can prove useful. The model in Figure 3-1 suggests that connecting the points with the same degree of 'medicalness' as found in the balance between the horizontal axis of relative frequency and the vertical axis of absolute frequency should be represented by a curve. This assumption is based on a tendency for lower absolute frequency values to combine with higher relative frequency values. Obviously, this raises the question of how (i.e. by which formula) absolute frequency and relative frequency are related. If the search for an answer is driven by a similar approach as given in (3) we can hypothesize about the parameters which should be covered by a formula. The outcome of the formula in (3c) is the overall weight of a sentence *s*. This corresponds to the degree of typicality of medical vocabulary. The parameters of absolute frequency and relative frequency are crucial. The relation between them cannot be a simple addition as in (3). The reason is that such a simple addition would result in a linear function. However, medical vocabulary seems better represented by a curve where the weight of the relative frequency depends on the absolute frequency. The lower the absolute frequence, the higher the relative frequency has to be in order to attain a similar degree of 'medicalness'. This suggests that the formula must include a division. Determining the points that are intuitively correct must also be taken into account in determining the formula. We can use parameters such as α, β, γ, and δ in (3c) to ensure that the curve crosses these points. Determining the optimal values of the multiplier or multipliers would then delimit the function of the curve indicated in Figure 3-1.

Another measure which may be taken as an inspiration in searching for appropriate expression of a degree of medicalness is the Body Mass Index (BMI), used in medicine and health care to express the sense of a healthy weight. It originated in the late 19th century, when the Belgian anthropologist Adolphe Quetelet (1798-1874) proposed a measure of body shape (Quetelet, 1869; Jelliffe and Jelliffe, 1979). Approximately a century later, the BMI started to be used in studies of the link between body adiposity and disease (Bedogni, Tiribelli, and Bellentani, 2005: 1). At present, the BMI formula in (4) is widely recognized and used not only among medical specialists.

(4) $\text{BMI} = \text{W}/\text{H}^2$

The formula in (4) defines BMI as a ratio of the body weight (W) in kilograms and the square of the height (H) in meters. The resulting score is then classified among five categories: underweight (18.5 and less), normal weight (18.5-25.0), overweight (25.0-30.0), obese (30.0-40.0), and extremely obese (over 40.0). The formula in (4) gives a curve when we keep the BMI value constant and consider the relation between weight and height.

Some aspects of the BMI may be viewed as a good basis for determining the degree of typicality of medical vocabulary. The ‘medicalness’ would be the output of the formula corresponding to BMI. The formula is based on the division of the two variables, weight and height. The degree of typicality of medical vocabulary is also based on two measures, absolute and relative frequency. Because the resulting function is expected to be a curve, division must be involved. However, a difference is that in determining ‘medicalness’, we would expect an inverted division because a lower value of absolute frequency requires a higher relative frequency value to obtain the same degree of medicalness.

In conclusion, it seems that a method of relative weights among different dimensions might be an effective way of delimiting the degree of typicality of medical vocabulary. Absolute frequency and relative frequency will play a role in the formula. The formulae in (3) and (4) are examples of how in other fields formulae have been proposed and used that can inspire the search for a formula of ‘medicalness’ of vocabulary. Elaborating such a formula will have to be left for future research.

Chapter Four

An Alternative Corpus: The Wikipedia-Based Medical Corpus

The corpus-based investigation of the medical subcorpus of COCA in chapter 3 indicates that medical vocabulary is best represented as a continuum. The continuum is based on two frequency dimensions: *relative frequency* and *absolute frequency*. It was shown that these two dimensions work together, but they have different roles. The relative frequency produces a continuum from words that are very typical of medical vocabulary to words that are comparatively rare in medical text. The other continuum is the absolute frequency in the medical corpus. This is used to apply a threshold of words frequent enough in the medical corpus. The best value for this threshold depends on the size of the corpus and influences the measure of relative frequency.

The aim of this chapter is to compare the results based on the subcorpus of medicine in COCA with an alternative corpus of medical texts, a specially compiled corpus for illustrative purposes. The illustrative medical corpus was compiled on the basis of the Wikipedia corpus, which was made available on the BYU website also offering COCA in 2015. It is based on the full text of the English version of the Wikipedia in 2012. The corpus covers 4.4 million Wikipedia articles with 1.9 billion words. The Wikipedia articles on medical topics represent a different type of medical text to medical journal articles. This feature, a different text type, makes the corpus based on the Wikipedia web pages related to medicine appropriate as an alternative corpus. The comparison of two corpora, the medical corpus of COCA and the specially compiled Wikipedia medical corpus, raises the following research questions:

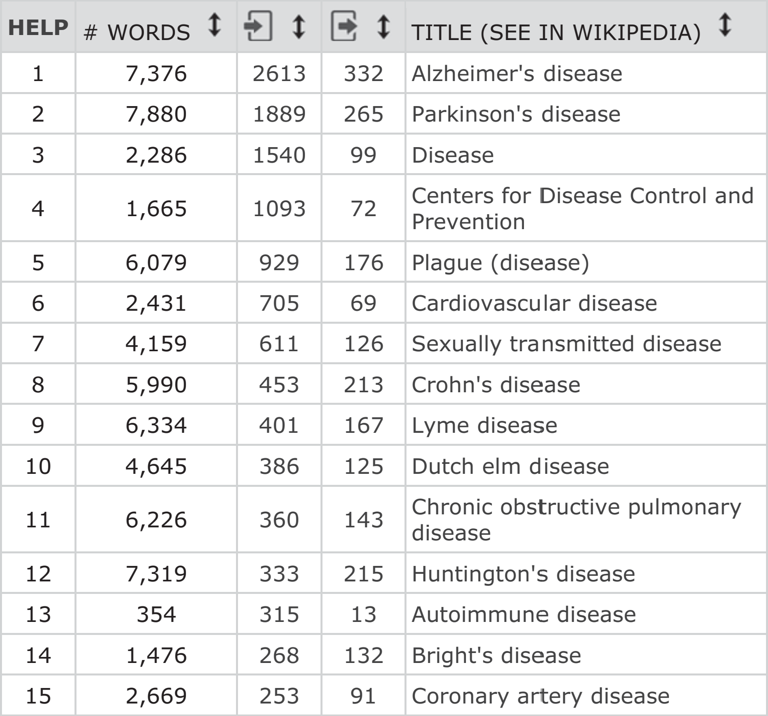
- Do the two corpora lead to very different frequency distributions?
- What are the differences in the results obtained from the analysis of the two corpora and how can they be interpreted and explained?
- How does the text type influence the structure of medical vocabulary?

Answers to these research questions are presented and discussed across the four sections of this chapter. Section 4.1 gives a detailed description of the methodology used to compile the illustrative Wikipedia-based medical corpus. Section 4.2 discusses the frequency distribution of nouns in the illustrative Wiki corpus and compares it with the results from COCA. Similarly, in section 4.3 the frequency distribution of verbs, adjectives and adverbs is addressed. The results are compared with the COCA medical corpus. Threshold values, their role and influence on the structure of medical vocabulary are discussed in detail. Finally, section 4.4 evaluates the usefulness of the two corpora of medical vocabulary based on different text types and summarizes the most relevant conclusions.

4.1 Methodology of an illustrative medical corpus design

The research question determines the criteria for compiling a relevant corpus. The key research question underlying the illustrative medical corpus is *How does the text type influence the structure of medical vocabulary?* This is a specific research question which requires compiling a specialized corpus. Specialized corpora are usually significantly smaller than large general corpora such as COCA or BNC. In line with Crawford and Csomay (2016: 79), specialized corpora "are designed to investigate a restricted set of questions, and therefore, are less likely a representative of language use in general terms". A definition of a specialized corpus by Hunston (2002: 14) is frequently referred to by books on corpus linguistics; it emphasizes that a specialized corpus "is used to investigate a particular type of language. Researchers often collect their own specialised corpora to reflect the kind of language they want to investigate. There is no limit to the degree of specialisation involved, but the parameters are set to limit the kind of the texts included". Hunston (2002: 14) gives an example of a corpus restricted to a particular time, e.g. a century, or a social context, e.g. newspaper articles about the European Union.

In corpus linguistics, it is also frequently emphasized that "a well designed corpus includes texts that address the research question(s) of the study" (Crawford and Csomay, 2016: 79). The Wikipedia corpus available in COCA meets this selection criterion. Wikipedia articles represent a different text type from scientific journal articles. The most obvious differences can be observed in the structure and length of the articles. Scientific journal papers follow a conventionalized Introduction-Method-Result-Discussion structure whereas the Wikipedia articles have a somewhat less uniform structure. There are certain topic-dependent patterns, but they are less rigorously imposed than for scientific journal papers. Table 4-1 illustrates how the length of the Wikipedia articles may vary.



HELP	# WORDS ↕	⇥ ↕	⇨ ↕	TITLE (SEE IN WIKIPEDIA) ↕
1	7,376	2613	332	Alzheimer's disease
2	7,880	1889	265	Parkinson's disease
3	2,286	1540	99	Disease
4	1,665	1093	72	Centers for Disease Control and Prevention
5	6,079	929	176	Plague (disease)
6	2,431	705	69	Cardiovascular disease
7	4,159	611	126	Sexually transmitted disease
8	5,990	453	213	Crohn's disease
9	6,334	401	167	Lyme disease
10	4,645	386	125	Dutch elm disease
11	6,226	360	143	Chronic obstructive pulmonary disease
12	7,319	333	215	Huntington's disease
13	354	315	13	Autoimmune disease
14	1,476	268	132	Bright's disease
15	2,669	253	91	Coronary artery disease

Table 4-1 The top 15 examples of the Wikipedia articles from the Wikipedia corpus in COCA based on the keyword *disease*

Table 4-1 shows that among the articles selected, the article about Parkinson’s disease in line 2 has the highest number of words, nearly 8,000. In contrast, the Wikipedia web page about Autoimmune disease contains less than 400 words. In between we can see articles of less than 2,000 words, such as in line 4 or 14, or exceeding 4,000 words, for instance the article about Sexually transmitted disease in line 7, or Dutch elm disease in line 10. Similar differences are not expected in scientific medical journal articles. The third column in Table 4-1 gives the number of articles that link to this article. This measure is used to express the centrality of the article. It is used here to sort the articles. The fourth column gives the number of articles that the article in the last column links to.

Before elaborating on the methodology used for the Wikipedia medical corpus design, it is worth exploring characteristic features of Wikipedia articles in more detail. As is well known, the word *Wikipedia* is a blend of the Hawaiian word *wiki* meaning ‘quick’, at present used to refer to a technology for creating collaborative websites, and the word *encyclopedia*.[1] Wikipedia is written by volunteer contributors. Users can contribute under their real identity, using a pseudonym, or anonymously. This also means the contributors are not necessarily experts or specialists in the field, although this possibility cannot be excluded either. This is another crucial difference concerning the text type. The Wiki pages can be edited, revised, and modified by anyone with Internet access. Wikipedia is based on five fundamental principles, labelled as five ‘pillars’. They are given in (1).[2]

(1) a. Wikipedia is an encyclopedia.
 b. Wikipedia is written from a neutral point of view.
 c. Wikipedia is free content that anyone can use, edit, and distribute.
 d. Editors should treat each other with respect and civility.
 e. Wikipedia has no firm rules.

The main principles in (1) aim to explain best practices and standards the users should adopt and follow, although it is frequently emphasized that Wikipedia does not employ hard-and-fast rules and using common sense is recommended when applying any principles or guidelines. The pillar in (1a) explains that the main ambition of Wikipedia is to serve as a source of encyclopedic information and knowledge, a kind of a reference source or compendium. The neutral perspective in (1b) means that the purpose of Wikipedia is to present information. There may be contexts where one well-recognized point is generally accepted, or contexts with multiple perspectives. In the latter, it is important to describe all perspectives accurately, avoiding the ‘best view’ presentation. The main point given in (1c) is that no contributor owns the Wikipedia article and therefore it may be freely revised and redistributed. The contributors or editors should respect copyright laws and avoid plagiarism. The message in (1d) is to stay away from personal attacks and disruptive practices. The last pillar in (1e) emphasizes that the guidelines and policies are subject to change over

[1] Available at https://en.wikipedia.org/wiki/Wikipedia:About, information retrieved 29 December, 2015.

[2] Available at https://en.wikipedia.org/wiki/, information retrieved 29 December, 2015.

time. It is recommended to be bold but not reckless in updating Wikipedia articles. Wikipedia saves all version of the articles, which makes it relatively easy to correct mistakes by restoring an earlier version.

Wikipedia was launched in 2001 and since then it has grown and turned into one of the largest reference websources with 374 million unique visitors monthly as of September 2015.[3] Wikipedia attracts more than 70,000 active contributors who work on more than 35,000,000 articles in 290 languages. As of 2015, there are 5,042,122 articles in English.[4] Wikipedia has become a dynamic tool with a collective output of tens of thousands of edits and thousands of new articles every day.

On the other hand, the rules and principles which made Wikipedia grow rapidly often raise doubts and criticism of the reliability of Wikipedia as a source of information. On the basis of studies investigating the quality of Wikipedia, He (2012: 52) identified four reasons undermining its trustworthiness: authority, responsibility, persistence of content, and the absence of reputation gain. The authority problem is directly linked with the open-edit system. It raises a crucial epistemological question whether we can trust an article written by an anonymous author or authors. The scandal in 2005 with the Wikipedia article about the U.S. journalist John Seigenthaler containing the false information that he had been a suspect in the assassination of President Kennedy indicates that the Wikipedia articles "are not rigorously fact-checked and reviewed like a traditional encyclopedia, they may contain errors and outright falsehoods that have been overlooked" (Anderson, 2011: 85) or indeed consciously introduced. More, similar scandals appeared later. According to Sanger (2009) and Wray (2009), the fact that non-specialists can edit, or even delete information discourages experts, specialists, and professionals from active participation. Worldwide, in many colleges and universities, this resulted in the exclusion of Wikipedia from the set of reliable sources to be used in research. For instance, in 2007, "the history department of Middlebury College in Vermont banned the use of Wikipedia as a source of research papers" (Anderson, 2011: 85). On the other hand, Tapscott and Williams (2006) point to the fact that Wikipedia is a remarkable example of mass collaboration, an important internet phenomenon. Without such a resource, the earlier Nupedia project, carried out by scholars and experts only and subject to a long peer-reviewing process (He, 2012: 46), did not manage to

[3] "Report card". Wikimedia. Retrieved September 3, 2015. Available at https://en.wikipedia.org/wiki/Wikipedia:About#cite_note-1

[4] Available at https://en.wikipedia.org/wiki/Wikipedia:About#cite_note-1, retrieved 29 December, 2015.

achieve a sufficiently wide coverage, although it is now regarded as the precursor of Wikipedia.

Responsibility, persistence of content, and the lack of reputation gain are the other factors with a negative impact on Wikipedia's credibility. Similarly to authority, these are also associated with the open-edit policy of Wikipedia. Even if individual edits and editors can be traced, the editors cannot be legally held responsible for their revisions (Sanger, 2009; He, 2016). Tollefsen (2009) uses the term 'doxastic instability' to label the fact that the accuracy of information in the Wikipedia may change; there is no mechanism to guarantee the persistence of correct information. Wikipedia is run by the non-profit Wikimedia Foundation. This also means that contributors "cannot obtain any reputation from their works, nor the property of their work" (He, 2012: 53). In line with Wray (2009: 38) "an invisible hand cannot ensure quality in Wikipedia".

Despite ongoing criticism of credibility and quality of the Wikipedia, it is a fact that it ranks among the top ten most visited websites globally and in many countries. This is illustrated in Table 4-2.

Wikipedia.org ranking	Country
6	Italy
6	Sweden
7	United States of America
7	Austria
8	Canada
8	Hong Kong
8	India
9	Australia
9	Kuwait
10	United Kingdom
10	Mexico
7	global

Table 4-2 Ranking of Wikipedia among most visited sites globally and by country[5]

[5] Ranking by Alexa Traffic Rank available at http://www.alexa.com/topsites, retrieved 30 December, 2015. The statistics carried out by Alexa Traffic Rank providing web analytics services since 1996. Alexa Traffic gives details of their calculations. According to their web site, the sites in the top sites lists are ordered by their one month Alexa Traffic results. The one month rank is calculated using a combination of average daily visitors and pageviews over the past month. The site

Table 4-2 demonstrates that Wikipedia is a powerful tool used by large numbers of Internet users all over the world. The significant popularity of Wikipedia does not mean that it has become reliable without any reservations. Persisting doubts about the quality and reliability of Wikipedia resulted in an attempt to launch a competing web encyclopedia Citizendium. The project Citizendium started in 2006. Larry Sanger, who previously worked in Wikipedia, is the founder of Citizendium and was its first editor-in-chief. Citizendium has strict editing rules and contributors are not anonymous, their real names are required. According to Anderson (2011: 92) "Sanger started Citizendium in the hopes of creating a more reliable online encyclopedia". However, the current information on Citizendium is that they have 16,891 articles at different stages of development, of which 160 have expert-approved citable versions.[6] These numbers support Anderson's view (2011: 92) that "Sanger's project has a long way to go before it can rival Wikipedia".

The significant, measurable attractiveness of Wikipedia may have been crucial in the decision to include its articles into COCA. Two more factors may also have played a role in this decision. First, Wikipedica is free of copyright and second, it cumulates a large amount of text. The facts presented above indicate that the Wikipedia corpus represents a separate text type and therefore it is appropriate for designing an illustrative alternative corpus of medical vocabulary.

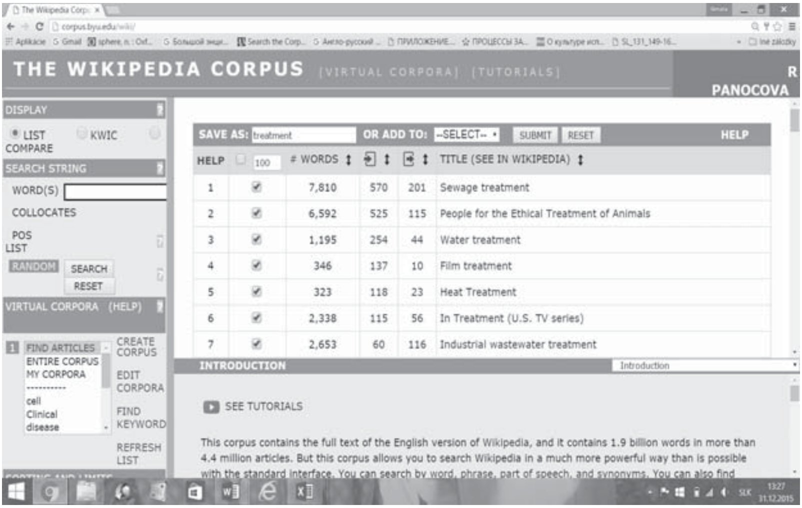
I decided to compile a separate corpus based on the Wikipedia, which I called WIMECO. In order to compile this corpus, I had to find selection criteria for the articles to be included. The COCA interface allows users to create their own virtual corpora from Wikipedia for any topic. The size of the virtual corpora is limited, each of these corpora can contain up to 1,000,000 words in up to 1,000 articles. Then, virtual corpora can be searched in the same way as any of the other corpora from BYU, including the ability to find keywords. Three options for building a virtual corpus are available: based on words in the title, for instance *medicine*, based on words in the text of articles, and a combination of words that appear in the title and words that appear in the content. Once the virtual corpus has been created, it can be edited, articles can be added, deleted, or moved to another corpus.

with the highest combination of visitors and pageviews is ranked #1. Information retrieved 30 December, 2015 from http://www.alexa.com/topsites.

[6] Available at http://en.citizendium.org/wiki/Welcome_to_Citizendium , retrieved 30 December, 2015.

The key question in compiling my illustrative medical corpus WIMECO was how to select relevant seed words. The results based on the analysis in the subcorpus Academic Medicine in COCA were my starting point. The Academic Medicine COCA word list was sorted by frequency. The aim was to select the top ten most frequent medical content words to build up 10 subcorpora of 100 articles each, which would be merged into WIMECO. Although it is possible to choose between 100, 300, and 1,000 Wikipedia articles (pages), 100 seems optimal. The reason is that depending on the title or text word, there are not always 300 or 1,000 articles available. One hundred pages ensures a better balance between subcorpora containing a maximum of 1,000,000 words in the final WIMECO.

The main criteria in the selection process were frequency, being a content word, and having an immediate link with medicine. The top fifteen most frequent words in the Academic Medicine in COCA are all function words: *the*, *of*, *be*, *and*, *a*, *in*, *to*, *for*, *with*, *have*, *that*, *or*, *by, on*. Obviously, these would not be helpful in searching Wikipedia articles about medicine. Therefore the top ten most frequent content words seemed a better alternative. However, the frequency list for content words starts with *study*, *use*, *child*, *group*, *case*, *system*, *time*, *result*, *figure*, *treatment*. Although frequency is an important criterion, it is itself not sufficient. The words above are used in a number of contexts and are polysemous. An example is given in Figure 4-1.



SAVE AS: treatment

HELP	☐ 100	# WORDS			TITLE (SEE IN WIKIPEDIA)
1	☑	7,810	570	201	Sewage treatment
2	☑	6,592	525	115	People for the Ethical Treatment of Animals
3	☑	1,195	254	44	Water treatment
4	☑	346	137	10	Film treatment
5	☑	323	118	23	Heat Treatment
6	☑	2,338	115	56	In Treatment (U.S. TV series)
7	☑	2,653	60	116	Industrial wastewater treatment

This corpus contains the full text of the English version of Wikipedia, and it contains 1.9 billion words in more than 4.4 million articles. But this corpus allows you to search Wikipedia in a much more powerful way than is possible with the standard interface. You can search by word, phrase, part of speech, and synonyms. You can also find

Figure 4-1 Example of a Wikipedia corpus with title word *treatment*

Figure 4-1 illustrates the search results in the Wikipedia Corpus with the title word *treatment*. The top seven Wikipedia articles include different domains such as environmental protection in lines 1, 3, and 7, movie industry in 4 and 6, animal rights in 2, and a rock music album in line 5. It is interesting that no result is related to medicine. Despite the possibility to delete articles and add new ones this would complicate the compilation procedure. The risk was that even the maximum number of 1,000 pages would not yield enough relevant results. This explains why frequent, but too general and polysemous words were also excluded. Word class (PoS) was not a selection criterion, but it is included in the table as an information category. The final list of lemmas selected for building up the illustrative medical corpus is given in Table 4-3.

Lemma	PoS (part of speech)	Frequency ACAD: Medicine	Rank in COCA
patient	n	26381	572
health	n	12840	344
disease	n	6058	775
hospital	n	5508	647
clinical	j	4422	2600
cell	n	4006	896
medical	j	3942	646
tissue	n	3467	2704
infection	n	2925	2708
symptom	n	2914	2263

Table 4-3 List of top ten medical, most frequent lemmas in ACAD: Medicine in COCA

The first column in Table 4-3 lists the seed words to be used for creating virtual corpora in the Wikipedia corpus. Their word class is marked in the second column. Frequency in the COCA subcorpus of academic medical texts determined which words were included, after function words and general academic words had been excluded. The last column provides information about the rank of the selected words (lemmas) in the general corpus. This shows that the selected words are also among the first 3,000 most frequent words in general COCA. The ranking in COCA is available for 60,000 lemmas in general and specialized subcorpora.

The next step was to create ten virtual subcorpora, each based on one of the seed words in Table 4-3. The combination of a word in the title and text of the Wikipedia article was used. The search aimed at 100 pages. The Wikipedia corpus in COCA makes it possible to view the web page with the article immediately after clicking the title. Then each page was inspected in order to exclude irrelevant ones. For instance, the 100-page virtual corpus based on the word *health* included the article about the music band called *Health*. In other cases the title word occurred in the title of a movie, for example, *The English Patient* (1996), *Dr. Patient* (2009), *The Infection* (2004), *The Cell* (2000), a mobile game or computer game, e.g. *The 7th Guest: Infection*, a TV series (soap opera), e.g. *General Hospital: Night Shift*, the title of a book, e.g. *Scar Tissue*, the name of a song and a compilation album, *Symptom of the Universe: The Original Black Sabbath 1970-1978* released in 2002, or it referred to a name of a company, *Irving Tissue Company Limited*. All such articles were deleted from the virtual subcorpora. Table 4-4 gives an overview of the final versions of ten virtual subcorpora.

Virtual corpus name	# Articles	# Words
CELL	73	151,174
CLINICAL	100	78,689
DISEASE	100	206,507
HEALTH	99	207,055
HOSPITAL	100	126,371
INFECTION	92	55,849
MEDICAL	98	165,133
PATIENT	97	83,983
SYMPTOM	15	5,484
TISSUE	88	50,152
TOTAL WIMECO	830	1,111,735

Table 4-4 Size of virtual subcorpora based on the Wikipedia Corpus in COCA

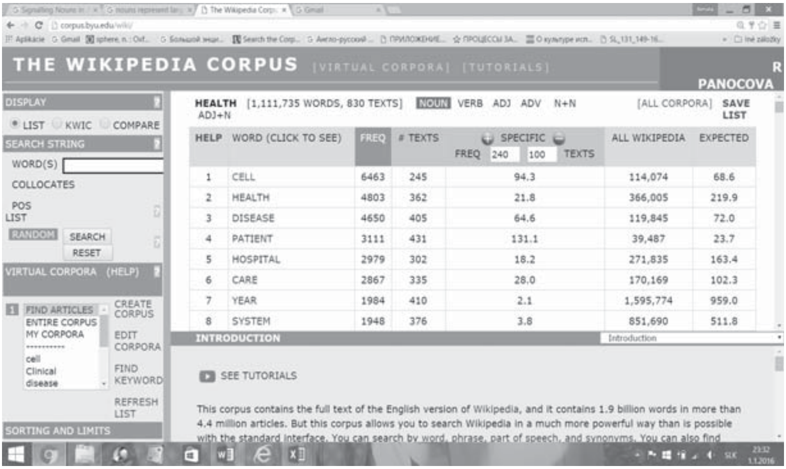
In Table 4-4 we can see the number of the Wikipedia articles included in each subcorpus. The size of each virtual subcorpus varies depending on the number of words in the articles. For three seed words, the full range of 100 articles was reached. In other cases there were not enough articles matching the relevant criteria. The only case where a significantly smaller number of articles was found is *symptom.* This confirms the hypothesis that aiming for 100 articles is a good starting point for collecting the data. Then all subcorpora were merged into a final corpus to produce WIMECO. It is interesting to examine in detail the total numbers. The final number of articles in WIMECO is 830, which is 30 articles less than the sum of the numbers in the lines above. A similar difference can be observed in the total number of words, there is a difference 18,662 words. This can be explained by the COCA corpus tool, which automatically deletes duplicated articles when the same article is selected in different subcorpora.

The WIMECO size of more than a million words seems sufficient for an illustrative specialized corpus. WIMECO is slightly larger than the corpus for MAWL compiled by Wang et al. (2008), which has 1,093,011 running words. In general, the question of the ideal or minimal size of a specialized corpus remains unsolved. Flowerdew (2004: 26) suggests that "in fact there is no optimum size, but what is of paramount importance is that the size of the specialized corpus must be closely matched with the features under investigation". Some authorities emphasize that smaller, specialized corpora concentrating on business or medical language can often be much more useful than large general language corpora (e.g. Tribble, 1997; Bernardini and Gavioli, 1999; Gavioli, 2005). Flowerdew (2004: 21) gives a corpus size between 1,000,000-5,000,000 words as one of the parameters for optimal size of specialized corpus. This indicates that my illustrative medical corpus was designed in line with these recommendations and meets standard criteria for a specialized corpus.

4.2 Distribution of nouns in WIMECO

The results of the analyses in chapter 3 demonstrate that nouns represent the largest word class in the ACAD: Medicine subcorpus. Nouns cover 45.3% out of the total number of 27,166 lemmas occurring in the medical subcorpus. This is in line with the figure of 44.9% of nouns found in the whole COCA corpus. Therefore we will first consider nouns.

A function tool available in COCA extracts the most frequent words, commonly referred to as *keywords*, from a virtual corpus based on the Wikipedia corpus. It is understood that keywords are the words occurring in a text with a higher frequency than their expected frequency "to an extent which is statistically significant" (Wynne, 2008: 730). The keywords are computed for a text or a set of texts in comparison to a reference corpus in order to obtain keywords that characterize a text or a set of texts. For a more detailed account of keywords in corpus linguistics cf. Lüdeling (2008) and Culpeper and Demmen (2015). The options are to search for the keyword nouns, keyword verbs, keyword adjectives, keyword adverbs, keyword noun+noun, or keyword adjective+noun combinations. The total number of keywords is 500 for each word class. a printscreen illustrating noun keywords in WIMECO is presented in Figure 4-2.



HELP	WORD (CLICK TO SEE)	FREQ	# TEXTS	SPECIFIC FREQ 240 100 TEXTS	ALL WIKIPEDIA	EXPECTED
1	CELL	6463	245	94.3	114,074	68.6
2	HEALTH	4803	362	21.8	366,005	219.9
3	DISEASE	4650	405	64.6	119,845	72.0
4	PATIENT	3111	431	131.1	39,487	23.7
5	HOSPITAL	2979	302	18.2	271,835	163.4
6	CARE	2867	335	28.0	170,169	102.3
7	YEAR	1984	410	2.1	1,595,774	959.0
8	SYSTEM	1948	376	3.8	851,690	511.8

This corpus contains the full text of the English version of Wikipedia, and it contains 1.9 billion words in more than 4.4 million articles. But this corpus allows you to search Wikipedia in a much more powerful way than is possible with the standard interface. You can search by word, phrase, part of speech, and synonyms. You can also find

Figure 4-2 Example of a list of top eight keyword nouns in WIMECO

The output list in Figure 4-2 includes the information about frequency [FREQ]. This value gives us the number of times the word was used in the corpus. Another important information # TEXTS refers to the number of articles in which the word was used. The total number of texts is 830. In certain contexts, this may be crucial for excluding a word despite a high value for [FREQ]. It ensures that the word does not occur many times in only a small number of articles.

The ALL WIKIPEDIA column provides information about the number of times the word was used in all articles in all 4.4 million articles in Wikipedia. For example, *cell* was used a total of 114,074 times in all articles. The data in the last column EXPECTED give expected number of occurrences in the virtual corpus are based on the total number of words by a virtual corpus if it were used at the same rate as in all articles. For instance, *health* is expected to occur 219.9 times providing it was used at the identical rate as in all other Wikipedia articles.

When we have the keyword nouns, we can sort them by the most frequent words or by how specific the words are to the virtual corpus. As the list in Fig. 4-2 shows, words such as year and system are frequent in WIMECO, but less frequent than would be expected on the basis of their general frequency. But if the feature [SPECIFIC] is selected, then more corpus-specific words will be placed higher in the list. Both values [FREQ] and [SPECIFIC] to the corpus can be used as a sorting criterion by clicking on these buttons. This will change the minimum frequency for the words and the minimum number of texts in which the word must occur in the boxes below [SPECIFIC]. These values can also be changed manually, i.e. overwriting. The resulting numbers show how much more frequent the word is in the custom-made corpus, based on the total frequency of that word in all of the Wikipedia and the size of the custom-made corpus. It is interesting to observe how these two values affect the list of keyword nouns in the illustrative medical corpus. This is exemplified in Table 4-5.

Top 20 nouns in WIMECO sorted by FREQ	Top 20 nouns in WIMECO sorted by SPECIFIC
cell	health-care
health	symptom
disease	patient
patient	cell
hospital	infection
care	professional
year	disease
system	tissue
treatment	medication
research	diagnosis
infection	disorder
service	physician
tissue	provider
study	care
blood	therapy
program	health
time	muscle
symptom	treatment
case	nurse
people	hospital

Table 4-5 Comparison of the top 20 nouns in WIMECO sorted by the value FREQ and SPECIFIC

Table 4-5 displays the differences in the top 20 keywords between the list created on the basis of the value [FREQ] and the list sorted by [SPECIFIC]. Overlapping words are highlighted. It is striking that they are a minority and occur in quite different positions in the two columns. For instance, *hospital* moved to the last place among the top 20 keywords after selecting the value [SPECIFIC]. This sort criterion resulted in the inclusion of the more corpus-specific words *health-care*, *professional*, *medication*, *diagnosis*, *disorder*, *physician*, *therapy*, *muscle*, and *nurse* in the top 20 keywords in contrast to more general keyword nouns such as *year*, *program*, *time*, or *people* in the list sorted by frequency. This indicates the importance of the value [SPECIFIC]. The default for the minimum frequency for the corpus-specific words was 240 and for the minimum number of texts in which the word must occur it was 100. Although it is possible to adjust the minimum frequency and the minimum number of texts in which the word must occur, the modifications did not give different results in the top 20 keywords. The list of nouns based on [SPECIFIC] with these parameters includes 168 items and all of them overlap with the list based on the value [FREQ].

Another interesting observation about the comparison in Table 4-5 is that eight out of nine overlapping keywords, *cell*, *disease*, *health*, *hospital*, *infection*, *patient*, *symptom*, *tissue* (listed alphabetically here), are among the seed words used to compile illustrative virtual corpus WIMECO. The only seed words that are missing, but listed in Table 4-3, are *medical* and *clinical*. As these are adjectives, they cannot occur among the overlapping keyword nouns. This may raise objections against the methodological decision about how to compile WIMECO because of an apparent bias. However, there are several reasons which explain and justify the methodology used. A first, highly practical reason is that designing a virtual corpus based on the Wikipedia corpus in COCA is only possible via keywords in the title, main body of the article, or a combination of the two.

A second consideration is based on quantifying the influence of the individual words. In order to address this, I produced a control corpus without the part based on the seed word *health*. Also in this case, the resulting list of 500 keyword nouns included *health*. Rank, frequency and other features slightly differed for the values given in Figure 4-2. For the control corpus, the keyword noun *health* ranked 8 with absolute frequency 1438 as opposed to rank 4 and absolute frequency 4803. It may be assumed that similar results would be obtained for a subcorpus based on each of the remaining seed words. This also confirms that a selection process based on the most frequent, content, medical keywords from the subcorpus ACAD: Medicine in COCA is not overly arbitrary, as it is based on relevant criteria.

Third, despite the fact that these keywords appear among the top nouns in the word lists in ACAD: Medicine, WIMECO sorted by the value [FREQ], and WIMECO sorted by [SPECIFIC], their positions differ. For example, the keyword *cell* ranked 6 in the top 10 lemmas selected from ACAD: Medicine (see Table 4-3) whereas in WIMECO it moved to higher positions. It appears on the top in WIMECO sorted by the value [FREQ] and ranks fourth in WIMECO sorted by [SPECIFIC] (see Table 4-5). These differences in frequency-based ranking in the two corpora suggest that these nouns are indeed representative of medical vocabulary. The results also indicate how a different text type and different frequency values affect the distribution of keywords in medical vocabulary.

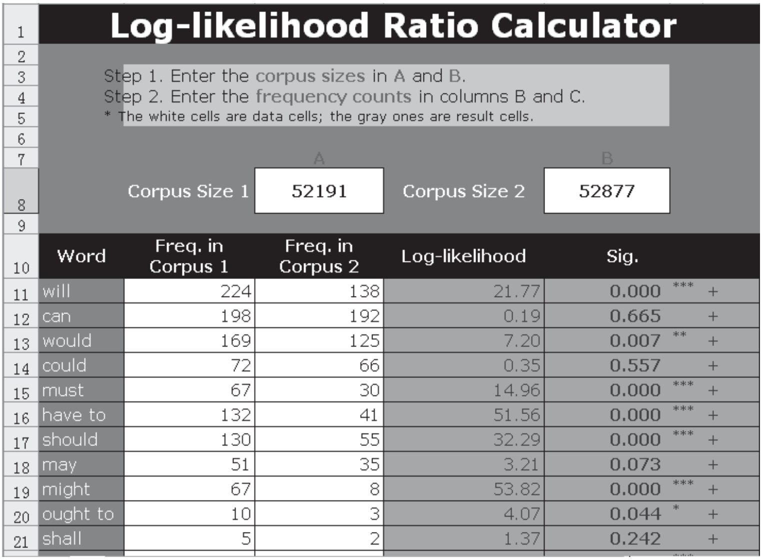
In case we deal with shared or overlapping words across the two types of text, it seems worth exploring differences in their frequencies. Our two corpora WIMECO and ACAD: Medicine in COCA are not of equal size. This means it is necessary to normalize the counts. Normalized counts then serve as a basis for comparison. The results for the eight overlapping keywords listed above are given in Table 4-6.

Keyword shared noun	Frequency WIMECO	Frequency ACAD: Medicine in COCA
cell	6463 (58.13)	3983 (6.06)
disease	4650 (41.82)	6014 (9.15)
health	4803 (43.20)	12840 (19.53)
hospital	2979 (26.79)	5508 (8.38)
infection	1460 (13.13)	2925 (4.45)
patient	3111 (27.98)	24793 (37.72)
symptom	1133 (10.19)	2914 (4.43)
tissue	1257 (11.30)	3429 (5.21)

Table 4-6 Comparison of frequencies of eight overlapping nouns in WIMECO and ACAD: Medicine in COCA

The frequency columns in Table 4-6 include two numbers. The top count indicates raw frequency, i.e. how many times a word occurs in the whole corpus. The count in brackets has been normalized to 10,000 words. For a more detailed explanation of the selection of this reference number see section 3.3. In order to convert raw frequency counts to normalized counts, we divide the raw frequency of a particular word by the total number of words in the corpus. Then, we multiply the result by a reference number, in our study it is 10,000. This means that the final number tells us how frequently a word occurs per 10,000 words. For instance, *cell* in WIMECO occurs 58.13 times per 10,000 words and *cell* in ACAD: Medicine in COCA occurs 6.06 times per 10,000 words. The difference between the normalized frequencies is much more remarkable than the difference between the raw frequencies. Interestingly, even though raw frequency counts of *disease*, *health*, *hospital*, *symptom*, *tissue* are lower in WIMECO, their normalized counts are higher. The most striking contrast between raw frequencies is in the word *patient*. However, in this case, the normalized frequencies do not differ as greatly.

This raises the question about the relevant comparison of these quantitative data derived from two corpora. McEnery and Wilson (2001: 81) note "that the use of quantification in corpus linguistics typically goes well beyond simple counting: many sophisticated statistical techniques are used……to show with some degree of certainty that differences between texts, genres, languages and so on are real ones and not simply a fluke of the sampling procedure". Frequencies represent a type of descriptive statistics. This means that they are useful in summarizing a dataset (McEnery, Xiao and Tono, 2006: 54). However, if we are interested in generalizations about how typical particular features or patterns are, we need different statistical procedures (Crawford and Csomay, 2016: 94). Inferential statistical measures are used to determine whether the results are statistically significant. Significance tests are "typically used to formulate or test hypotheses" (McEnery, Xiao and Tono, 2006: 54). Therefore, the next step in my research was to test whether the differences in frequency values for nouns in WIMECO and Acad: Medicine are statistically significant. According to McEnery, Xiao and Tono (2006: 55), the most commonly used techniques for testing statistical significance are the *Chi-square test*, also called *Pearson chi-square test*, and the *Log-likelihood (LL) test*, also called the *log-like chi-square* or *G-square test*. McEnery, Xiao and Tono (2006: 55) prefer the LL test for statistic significance because it "does not assume the data are normally distributed". In order to determine a measure of likelihood termed the *significance* or the *p* value, another value *degree of freedom* (d.f.) is necessary. It is calculated by multiplying the number of rows less 1 with the number of columns less 1 in a frequency (contingency) table. The LL critical values with 1 d.f. are 3.84 ($p < 0.05$), 95 percentile; 6.63 ($p < 0.01$), 99 percentile; 10.83 ($p < 0.001$), 99.9 percentile; and 15.13 ($p < 0.0001$), 99.99 percentile. McEnery, Xiao and Tono (2006: 56) point out that there are many web-based LL calculators. In addition, the LL test is a part of standard statistical packages, such as the statistical package for social sciences SPSS. In my research, I used the LL calculator by Xu available online.[7] Figure 4-4 illustrates sample calculations.



Log-likelihood Ratio Calculator

Step 1. Enter the corpus sizes in A and B.
Step 2. Enter the frequency counts in columns B and C.
* The white cells are data cells; the gray ones are result cells.

Corpus Size 1 (A): 52191 Corpus Size 2 (B): 52877

Word	Freq. in Corpus 1	Freq. in Corpus 2	Log-likelihood	Sig.		
will	224	138	21.77	0.000	***	+
can	198	192	0.19	0.665		+
would	169	125	7.20	0.007	**	+
could	72	66	0.35	0.557		+
must	67	30	14.96	0.000	***	+
have to	132	41	51.56	0.000	***	+
should	130	55	32.29	0.000	***	+
may	51	35	3.21	0.073		+
might	67	8	53.82	0.000	***	+
ought to	10	3	4.07	0.044	*	+
shall	5	2	1.37	0.242		+

Figure 4-3 Example of calculations by Xu's Log-likelihood Ratio Calculator

The sample counts in Figure 4-3 demonstrate that the higher the LL value, the more significant is the difference between the two frequency scores. For instance, the LL value for *will* is higher than the critical value 15.13. This means it is significant at $p < 0.0001$. Or in other words, it shows we

[7] Created by Jiain Xu, available at https://www.academia.edu/6050773/Log_likelihood_calculation_Excel_spreadsheet, retrieved 3 January, 2016

can be more than 99.99% confident that the difference is not due to chance.

I uploaded the data from WIMECO and ACAD: Medicine in COCA to this application and the LL values and significance values for the 500 keyword nouns identified in WIMECO were computed. Figure 4-4 summarizes the results.

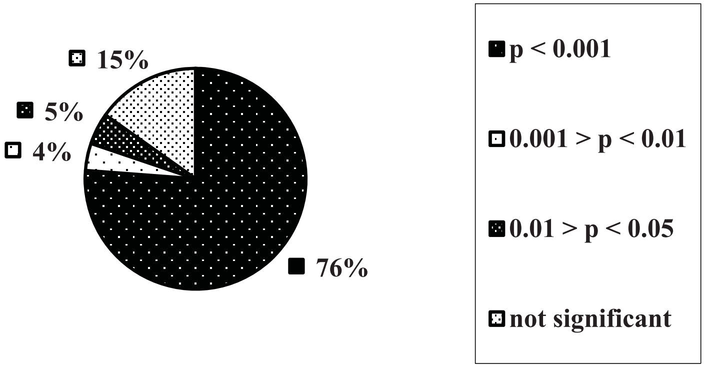


Figure 4-4 LL values and statistical significance of 500 keyword nouns

The percentage in Figure 4-4 shows that 85% values of the total number of keyword nouns have a significantly different frequency in the two corpora. The largest class, 76.2%, includes nouns with statistically significant values for the two frequency scores at $p < 0.001$. The LL values in this class are higher than the critical value of 10.83. An additional 4% of nouns have a difference in frequency that is significant at $p < 0.01$. A further 4.6% of the values are statistically significant at $p < 0.05$, which means they are higher than the critical value of 3.84. For the remaining 15.2%, the difference between the two frequencies from different corpora may well be due to chance. For them, $p > 0.05$ so that they are not statistically significant.

Let us explore the individual classes in more detail. We will inspect for which keyword nouns the frequency scores are highly significant. The largest class significant at $p < 0.001$ covers 381 keyword nouns. The top 20 with the highest significance are given in Table 4-7.

keyword noun	LL value	significance value	+/-
cell	12346,05168	0,0000	+
disease	5249,637211	0,0000	+
stem	2398,204392	0,0000	+
hospital	2239,096963	0,0000	+
health	1924,628837	0,0000	+
care	1621,759684	0,0000	+
insurance	1578,106197	0,0000	+
country	1053,491634	0,0000	+
infection	979,1833862	0,0000	+
protein	892,4764954	0,0000	+
million	817,4355053	0,0000	+
research	755,0097511	0,0000	+
study	743,2770544	0,0000	-
subject	740,7928914	0,0000	-
medicine	713,7370664	0,0000	+
blood	711,5692376	0,0000	+
doctor	707,4595856	0,0000	+
gene	694,3414046	0,0000	+
molecule	694,0533455	0,0000	+
campus	658,8017860	0,0000	+

Table 4-7 Top 20 keyword nouns with $p < 0.001$

The keyword nouns in Table 4-7 are sorted by their LL value. The LL value and the significance value *p* are correlated. However, the significance values are rounded, which results in the same value in this column. The LL value is more discriminating because it shows differences. If we compare the LL values for *cell* and *campus*, the difference is a factor 19. This means that with the LL scores it is possible to see a complete continuum of medical keyword nouns. On one end of the continuum, shown in Figure 4-7, we can see nouns typical of the WIMECO corpus as opposed to the ACAD: Medicine in COCA. The top ten words in Table 4-7 include five seed words used for compiling WIMECO, *cell*, *disease*, *hospital*, *health*, and *infection*. This fact may be viewed as a sign of bias, although I demonstrated above that even if the seed word *health* was excluded from corpus, it still occurred among the 500 keyword nouns.

The + or - in the last column indicate which of the two corpora each keyword is characteristic of. Table 4-7 indicates that all but two nouns, *study* and *subject*, are typical of WIMECO as opposed to the other corpus. These two nouns are therefore typical of ACAD: Medicine, which is the corpus comprising journal research papers. On the other end of the continuum there are nouns equally important in both corpora. With the keywords nouns in Table 4-7 we are 99.99% confident that the difference in frequency found in WIMECO and ACAD: Medicine in COCA is statistically significant. This indicates that the two corpora give different perspectives of what is medical vocabulary.

As most of the keywords in Table 4-7 have a + value, Table 4-8 presents the top twenty keyword nouns typical of ACAD: Medicine as opposed to WIMECO.

keyword noun	LL value	significance value	+/-
study	743,2770544	0,0000	-
subject	740,7928914	0,0000	-
sample	542,3486026	0,0000	-
analysis	508,7700859	0,0000	-
water	492,1165284	0,0000	-
difference	479,399467	0,0000	-
behavior	455,0454037	0,0000	-
child	447,4158437	0,0000	-
group	420,5646067	0,0000	-
value	353,0209373	0,0000	-
finding	348,2500777	0,0000	-
score	336,7467924	0,0000	-
result	314,9913955	0,0000	-
food	281,2358226	0,0000	-
patient	266,554134	0,0000	-
control	234,268026	0,0000	-
intervention	194,814667	0,0000	-
pressure	190,8175094	0,0000	-
force	188,0301916	0,0000	-
frequency	175,0003534	0,0000	-

Table 4-8 Top 20 keyword nouns with p < 0.001 typical of ACAD: Medicine as opposed to WIMECO

The data in Table 4-8 also show one end of the continuum of nouns typical of medical vocabulary in the ACAD: Medicine corpus as opposed to the WIMECO corpus. It is interesting to see that it includes nouns that are indeed characteristic for research papers, for instance, *study*, *subject*, *sample*, *analysis*, *difference*, *finding*, *result*, *control*, *frequency*. The WIMECO corpus consists of the Wikipedia medical articles whereas the ACAD: Medicine contains exclusively papers from medical research journals. It is obvious that the keyword nouns in Table 4-7 and Table 4-8 demonstrate that this aspect of the corpus has a significant influence on what actually emerges as medical vocabulary.

4.3 Distribution of verbs, adjectives, and adverbs in WIMECO

The analysis in section 4.2. only concerned nouns. This section concentrates on investigating whether a similar statistical significance can be observed across other word classes. First, the keyword verbs are discussed. Then, a description of the results for adjectives and adverbs follows.

Similarly to WIMECO nouns, WIMECO verbs were first compared with verbs in ACAD: Medicine in COCA. The result showed that all 500 WIMECO verbs can also be found in ACAD: Medicine in COCA, but one. The verb *adipose* occurs in WIMECO (frequency 85 in 12 texts), but is not listed among the 27,166 lemmas in the subcorpus ACAD: Medicine. The next step in the analysis was to perform the LL test, in order to find out whether the differences between the two corpora are statistically significant. The online LL test calculator, described in detail in section 4.2, was used to calculate LL values and their statistical significance values for keyword verbs. Table 4-9 presents the top 20 keyword verbs typical of WIMECO as opposed to ACAD: Medicine.

keyword verb	**LL value**	**significance value**	**+/-**
call	1136,212493	0,0000	+
found	700,6725607	0,0000	+
cause	643,1095175	0,0000	+
open	575,15082	0,0000	+
know	565,3731393	0,0000	+
form	498,1697718	0,0000	+
name	497,6749714	0,0000	+

keyword verb	LL value	significance value	+/-
publish	481,4858467	0,0000	+
include	442,7867935	0,0000	+
bind	400,5339204	0,0000	+
become	390,9674125	0,0000	+
pass	378,9913224	0,0000	+
regulate	302,1571074	0,0000	+
create	290,0650592	0,0000	+
kill	289,6632249	0,0000	+
lead	285,9281924	0,0000	+
establish	261,2983069	0,0000	+
announce	260,1273297	0,0000	+
cover	248,425646	0,0000	+
appoint	244,3954623	0,0000	+

Table 4-9 Top 20 keyword verbs with p < 0.001 typical of WIMECO as opposed to ACAD: Medicine

It is interesting to compare the LL values for nouns in Table 4-7 with the LL values for verbs in Table 4-9. Generally, the LL values for the verbs are less extreme than for the nouns. I will come back to this in section 4.4. In Table 4-9 we can see that many of the top twenty keyword verbs typical of WIMECO in contrast to the medical subcorpus of COCA are verbs that might easily be found in general contexts. In (2) some of the verbs in Table 4-9 are illustrated in their context in WIMECO.

(2) a. Cells are the smallest unit of life that can replicate independently, and are often **called** the building blocks of life.
b. In February 2009, the Phillip T. and Susan M. Ragon Institute of immunology was **founded** to bolster research into creating vaccines and other therapies for acquired immune system conditions.
c. It is **caused** by the Bluetongue virus (BTV).
d. This discovery is expected to **open** the avenue to new treatments in the coming years.

The examples in (2), taken from Wikipedia articles in WIMECO, illustrate contexts of the top four verbs that characterize this corpus in contrast to

ACAD: Medicine. They suggest that the general nature of the texts in WIMECO is more popular than strictly scientific. The keyword verbs typical of the ACAD: Medicine are given in Table 4-10.

keyword verb	LL value	significance value	+/-
report	290,2435026	0,0000	-
indicate	229,0421406	0,0000	-
show	190,3734905	0,0000	-
compare	187,5809283	0,0000	-
reveal	148,7208352	0,0000	-
calculate	134,5009087	0,0000	-
note	123,8919257	0,0000	-
need	117,1144379	0,0000	-
determine	114,7792082	0,0000	-
assess	114,6674818	0,0000	-
obtain	114,5993765	0,0000	-
evaluate	108,8249139	0,0000	-
ask	108,4507502	0,0000	-
measure	107,8733315	0,0000	-
suggest	107,5932611	0,0000	-
feel	96,1967938	0,0000	-
examine	93,95908468	0,0000	-
present	92,2561225	0,0000	-
observe	84,37381729	0,0000	-
demonstrate	77,11616676	0,0000	-

Table 4-10 Top 20 keyword verbs with p < 0.001 typical of ACAD: Medicine as opposed to WIMECO

In Table 4-10, it is straightforward that the top keyword verbs characterizing ACAD: Medicine as opposed to WIMECO are more typical of research papers. This is illustrated by the context senteces from ACAD: Medicine in (3).

(3) a. Females were more likely than males to **report** that they were currently on a diet (17% vs. 6%).

b. The higher prevalence detected in controls in 2008 could **indicate** a seasonal infection pattern because specimens were collected in a single weekend….

c. Overall, the linear regression relationship failed to **show** any association in the adjusted model (Table 2), and the likelihood ratio…

d. Age-adjusted incidence rates allowed us to **compare** the incidence rate for a standardized population during the study period and were performed by…..

The example sentences in (3) from ACAD: Medicine clearly demonstrate that the top four keyword verbs are typically used to present research findings. These keyword verbs may be considered as markers of an appropriate writing style for research papers. Examples of the use in medical contexts of other verbs from Table 4-10, such as *reveal*, *determine*, *assess*, *obtain*, *evaluate*, *measure* would be stylistically similar and therefore typical of research-oriented writing. In a sense, verbs are less medical than nouns and the contrast we see is more one of journalistic versus academic style.

Let us now turn to exploring the situation with the LL test results for keyword adjectives. The list of keyword adjectives in WIMECO includes five adjectives, namely, *coeliac*, *vitro*,[8] *pluripotent*, *eukaryotic*, and *multicellular*, which do not occur in ACAD: Medicine. Table 4-11 presents the top 20 keyword adjectives most representative of WIMECO.

Table 4-11 shows a cline of adjectives typical of the medical corpus based on Wikipedia articles as opposed to ACAD: Medicine. It may be assumed that these lemmas characterize medical practice rather than medical research. Similar to Table 4-7, in Table 4-11 we can also see that the top two key adjectives *medical* and *clinical* are seed words used to compile WIMECO. This slightly biased situation may be explained in a similar way as in the keyword nouns above. Another interesting observation is that the keyword adjective *red* is among the top 20 adjectives typical of WIMECO in comparison with ACAD: Medicine. Among the keyword nouns in Table 4-7, *cell* also ranked as the top lemma characteristic for WIMECO. Taking into account that *red cell* or *red blood cell* are certainly appropriate for the context of medical practice, it confirms that WIMECO is more representative of health care language than academic research. Similar to the observations in Table 4-10, we can see that Table 4-11 includes the adjective *new*, which is more characteristic of journalistic style. *Private* and *federal* might often refer to aspects of the health care system.

[8] *Vitro* might seem dubious as an adjective. Its absolute frequency in WIMECO is 75, occurring in 34 texts. In all cases it was found in the combination *in vitro*.

keyword adjective	LL value	significance value	+/-
medical	3869,075787	0,0000	+
clinical	746,6672532	0,0000	+
immune	642,8011713	0,0000	+
embryonic	579,3874201	0,0000	+
new	521,4042478	0,0000	+
genetic	517,6766558	0,0000	+
private	492,468782	0,0000	+
military	309,7168242	0,0000	+
other	292,6939181	0,0000	+
dendritic	282,1598478	0,0000	+
federal	262,400657	0,0000	+
mental	259,3760359	0,0000	+
modern	253,2992358	0,0000	+
electronic	251,9909773	0,0000	+
main	251,2761352	0,0000	+
red	249,4448174	0,0000	+
biomedical	244,9463588	0,0000	+
human	236,7011266	0,0000	+
nonprofit	228,3474435	0,0000	+
pharmaceutical	227,0404133	0,0000	+

Table 4-11 Top 20 keyword adjectives with $p < 0.001$ typical of WIMECO as opposed to ACAD: Medicine

Table 4-12 gives the top 20 keyword adjectives typical of ACAD: Medicine as opposed to WIMECO.

Table 4-12 includes adjectives typically used in reporting scientific research, for instance, *significant*, *consistent*, *total*, *previous*, *statistical*, *average*, *relative*, *normal*. The adjective *environmental* is on top. This might also be explained by the fact that the sources of medical research articles in ACAD: Medicine in COCA include journals such as *Journal of Environmental Health* and *Environmental Health Perspectives*. In general it is very difficult to avoid bias in a corpus entirely, so that the bias we found in WIMECO due to the seed words used is not categorically worse than the bias in COCA. The adjectives *left*, *right*, and *middle* are used in medical contexts to refer to the orientation in the body. Interestingly, the corpus data show that *left* may also be used in a different sense in combinations such as if *left untreated*. In the content of Table 4-11, I discussed the use of the adjective *red* and Table 4-12 lists another adjective describing colour, *black*. For medical contexts, COCA gives combinations such as *Black Death* as a synonym of plague, *black lung disease*, etc.

keyword adjective	LL value	significance value	+/-
environmental	771,7889869	0,0000	-
significant	363,4908341	0,0000	-
left	227,3368015	0,0000	-
black	200,4167243	0,0000	-
right	192,0036629	0,0000	-
middle	184,4497376	0,0000	-
positive	156,2658245	0,0000	-
consistent	147,955733	0,0000	-
social	136,6600878	0,0000	-
total	128,4877816	0,0000	-
previous	124,4373713	0,0000	-
statistical	117,775128	0,0000	-
low	109,6902146	0,0000	-
lateral	107,0317435	0,0000	-
soft	105,5267692	0,0000	-
average	95,56120655	0,0000	-
relative	77,0667286	0,0000	-
external	75,60244132	0,0000	-
normal	75,30920042	0,0000	-
great	72,17218333	0,0000	-

Table 4-12 Top 20 keyword adjectives with p < 0.001 typical of ACAD: Medicine as opposed to WIMECO

Adverbs are the last word class to be discussed. A comparison of WIMECO and ACAD: Medicine revealed that 14 WIMECO keyword adverbs were not found in ACAD: Medicine. The top 20 keyword adverbs characteristic for WIMECO as opposed to ACAD: Medicine are presented in Table 4-13.

keyword adverb	LL value	significance value	+/-
also	851,911334	0,0000	+
usually	309,8719854	0,0000	+
often	261,8785508	0,0000	+
formerly	195,7336839	0,0000	+
now	172,0879493	0,0000	+
officially	167,4295239	0,0000	+
sometimes	162,2191487	0,0000	+
eventually	158,9686161	0,0000	+
typically	158,9401628	0,0000	+
over	154,1351828	0,0000	+
currently	149,6611402	0,0000	+
originally	146,0440797	0,0000	+
generally	141,5649138	0,0000	+
worldwide	131,3528743	0,0000	+
commonly	129,1014252	0,0000	+
around	104,7663173	0,0000	+
publicly	101,4060821	0,0000	+
normally	98,62556593	0,0000	+
sexually	91,44688872	0,0000	+
mostly	90,1625968	0,0000	+

Table 4-13 Top 20 keyword adverbs with p < 0.001 typical of WIMECO as opposed to ACAD: Medicine

It is interesting that the top three adverbs in Table 4-13 are not included among the top 3000 lemmas of AVL by Gardner and Davies (2013). However, *also*, *usually*, and *often* can be found in general COCA and they all rank among the first 1,000 lemmas. For two of them, *also* and *often*, the absolute frequency value in the subcorpus ACADEMIC is the highest of the subcorpora. For *usually*, the absolute frequency in ACADEMIC is the second highest after POPULAR MAGAZINES. The LL values for all items in Table 4-13 indicate that these adverbs are typical of the WIMECO corpus as opposed to the ACAD: Medicine corpus. The reverse is illustrated in Table 4-14.

keyword adverb	LL value	significance value	+/-
however	439,7514	0,0000	-
least	379,2271372	0,0000	-
significantly	276,3928209	0,0000	-
in	269,8562917	0,0000	-
statistically	183,8965656	0,0000	-
respectively	84,02606186	0,0000	-
general	80,95052772	0,0000	-
therefore	80,62432125	0,0000	-
how	75,49986677	0,0000	-
additionally	66,36281752	0,0000	-
that	64,81557358	0,0000	-
here	64,30348267	0,0000	-
thus	56,72588864	0,0000	-
nevertheless	56,23697469	0,0000	-
overall	55,21227066	0,0000	-
unfortunately	51,54588912	0,0000	-
finally	49,51518706	0,0000	-
no	44,83114263	0,0000	-
similarly	44,57527585	0,0000	-
clearly	41,30983976	0,0000	-

Table 4-14 Top 20 keyword adverbs with $p < 0.001$ typical of ACAD: Medicine as opposed to WIMECO

Similar to Table 4-8 and Table 4-10, keyword adverbs in Table 4-14 also indicate that the two different medical corpora give different perspectives on what is medical vocabulary in English. It may seem striking that Table 4-14 includes *that* and *in* as adverbs. *That* can be used not only as a pronoun, determiner, and a conjunction, but also as an adverb. In such a case it is typically used before an adjective or adverb, for example, *that early.* Although *in* usually functions as preposition, in certain contexts it may be used as an adverb, in contexts without a following noun, for instance, *regulates what moves in and out (selectively permeable)*.

Two characteristic classes of adverbs can be identified in Table 4-14. A first class is typically used in academic reasoning. It includes items such as *therefore*, *thus*, *however*, *additionally*, *nevertheless*. The second class is connected to the use of statistics as part of an argumentation line. It is represented by the adverbs such as *significantly*, *statistically*, *respectively*, *overall*. Both classes are typical of ACAD: Medicine as opposed to WIMECO. This gives further confirmation that these adverbs are more characteristic of a writing style associated with reporting on research also in medicine, whereas the adverbs in Table 4-13 are more typical of medical practice and texts more accessible to a non-expert readership.

4.4 How useful are the two medical corpora?

This chapter aimed at the comparison of the results based on the subcorpus of medicine in COCA with an alternative corpus of medical texts, a specially compiled corpus based on Wikipedia articles named WIMECO. The comparison of two corpora, the medical subcorpus of COCA ACAD: Medicine and the WIMECO corpus, was carried out to answer the following research questions:

- Do the two corpora lead to very different frequency distributions?
- What are the differences in the results obtained from the analysis of the two corpora and how can they be interpreted and explained?
- How does the text type influence the structure of medical vocabulary?

The first question concerns the occurrence of differences in frequency distribution between WIMECO and ACAD: Medicine in COCA. The two corpora differ in size. The function tool available in COCA for the virtual WIMECO corpus based on Wikipedia makes it possible to create a frequency list for four major word classes: nouns, verbs, adjectives and adverbs. Each list includes 500 items. First, it was necessary to compare WIMECO nouns, verbs, adjectives and adverbs with the same word classes in ACAD: Medicine in COCA. The comparison revealed that for nouns, all 500 noun keywords in WIMECO also occur in ACAD: Medicine in COCA. For the other word classes, slight differences were found. One verb occurred in WIMECO, but not in ACAD: Medicine; five adjectives and fourteen adverbs were also found only in WIMECO.

Then a statistical analysis based on the LL test was carried out. The results confirmed that the differences between frequencies for individual keywords in WIMECO and ACAD: Medicine are statistically significant, for many words, even with $p < 0.0001$.

An interesting phenomenon is that the general levels of the LL values are different per syntactic category and also per corpus. This is obvious from the comparison of the data presented in Tables 4-7 to 4-14. For convenience, Table 4-15 gives a summary of the ranges of the LL values per syntactic category and per corpus for the top (most significant) 20 words in each word class.

syntactic category	range of LL values WIMECO	range of LL values ACAD: Medicine
nouns	12346.05168-658.8017860	743.2770544-175.0003534
verbs	1136.212493-244.3954623	290.2435026-77.11616676
adjectives	3869.075787-227.0404133	771.7889869-72.17218333
adverbs	851.911334-90.1625968	439.7514-41.30983976

Table 4-15 Range of LL values of the 20 most significant keywords per syntactic category in WIMECO and ACAD: Medicine

From Table 4-15 it is obvious that nouns show the highest LL values in both medical corpora. Interestingly, in both corpora nouns are followed by adjectives, then verbs and adverbs. In general, the LL values are higher for WIMECO than for ACAD: Medicine. This can be explained because the words for which the values are calculated are the keywords of WIMECO.

These findings provide further evidence that medical vocabulary is best viewed as a continuum. However, this continuum differs from the continuum based on the interaction of absolute and relative frequencies presented in section 3.4. The continuum based on the LL values may be seen as having the words typical of the medical corpus WIMECO on one end and the words characteristic of ACAD: Medicine on the other end.

At the same time, on one end of a continuum on a different dimension, there are nouns, verbs, adjectives and adverbs typical of WIMECO as opposed to the ACAD: Medicine in COCA. On the other end of that continuum there are nouns, verbs, adjectives and adverbs equally important in both corpora. As these were not discussed so far, in Table 4-16 I list the keyword nouns with the lowest LL values.

Table 4-16 shows nouns which occur close to the middle point of the continuum where we find words approximately equally important in WIMECO and ACAD: Medicine. The LL values are much lower than in the other tables presented in this chapter. The difference between the nouns in the two medical corpora is not statistically significant. This means that the difference between their frequencies in WIMECO and ACAD: Medicine is very likely due to chance. Figure 4-5 illustrates how these results map on the other continuum.

keyword noun	LL value	significance value	+/-
lack	0,000010723	0,9974	-
skin	0,000219754	0,9882	-
amount	0,000242735	0,9876	-
project	0,002087167	0,9636	-
effort	0,021302693	0,884	-
evidence	0,025072512	0,8742	+
implementation	0,041480834	0,8386	-
action	0,043206377	0,8353	+
tool	0,073928073	0,7857	+
nature	0,077677615	0,7805	-
technique	0,085594203	0,7699	-
prevalence	0,097949216	0,7543	+
component	0,16230723	0,687	-
device	0,209358173	0,6473	+
director	0,298549256	0,5848	+
involvement	0,304709848	0,5809	-
use	0,310436676	0,5774	+
business	0,3221111	0,5703	-
man	0,366900612	0,5447	+
surgeon	0,408029246	0,523	-

Table 4-16 Top 20 keyword nouns with the lowest LL values

In Figure 4-5 we can see the combined continuum of medical vocabulary represented as a U-shaped curve. The dotted line shows that the continuum can be divided into the part where nouns, verbs, adjectives and adverbs typical of WIMECO are placed as opposed to the part more characteristic of ACAD: Medicine. On the bottom middle point of the curve we find the words that are of similar importance in both medical corpora. The highest LL values are on the top of each side of the U-curve and the lowest on the bottom. What the figures from Fig. 4-4 show is that the common core, i.e. the words that belong to medical vocabulary according to both corpora, is very small. Only 15% of the WIMECO keywords are not significantly more frequent in one of the two corpora.

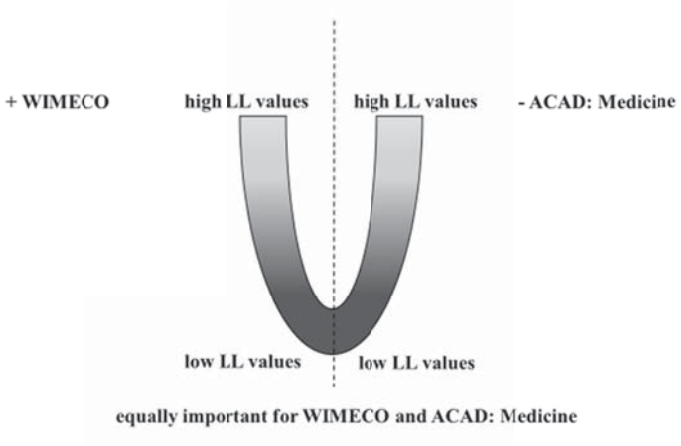


Figure 4-5 LL Continuum of medical vocabulary based on the LL values in WIMECO and ACADEMIC: Medicine

The issue of the influence of a different text type is the main subject of investigation in this chapter. In this study, Wikipedia medical articles and scientific journal articles compiled in ACAD: Medicine in COCA were selected to represent different text types. The analysis of the LL test results clearly indicate that ACAD: Medicine contains vocabulary typical of medical research in contrast to the practice of the health care system often characteristic of WIMECO. These findings provide evidence that the nature of the corpus has a highly significant influence on what emerges as medical vocabulary.

The structure of the corpus plays an important role, especially the fact to which medical specialization a text belongs. We can expect that a corpus based on texts from one domain, for instance, cardiology, is likely to give a slightly different overview of medical vocabulary than in a corpus based on the texts in neurology. In addition, the text type plays such an important role in determining the outcome that it is hard to imagine a single list covering two text types equally well. Therefore it is less attractive to try to define medical vocabulary for all contexts, because it will not be adequate for each individual domain.

CONCLUSION

This monograph investigated the vocabulary of medical English from a corpus-based perspective. The question of the characterization of medical vocabulary in English was the main aim of my research. This perspective of characterization of medical English contrasts with the pedagogical perspective which uses medical corpora for compiling word frequency lists used as a basis for teaching materials. Whereas previous work can be seen as belonging to applied linguistics, the present study is rather part of empirical linguistics. The monograph focused primarily on four central research questions:

- How can medical vocabulary in English be determined on the basis of a specialized corpus?
- How does the choice of a particular perspective (pedagogical versus characterizing/descriptive) influence the methodology of corpus-based research?
- How does the text type influence the structure of medical vocabulary?
- How does the choice of a corpus influence the results?

As a background to the study of these questions, chapter 2 considered specialized word lists developed earlier. There are three prominent methodological issues in which the empirical approach has different requirements from Coxhead's (2000) Academic Word List (AWL). They concern the use of word families, the use of West's (1953) GSL, and the structure of the corpus.

In approaching the research questions from an empirical perspective, I first explained the advantages of the use of the COCA as a general corpus of English and especially its subcorpus of medical English texts ACAD: Medicine for the purpose of the characterization of medical vocabulary. A first advantage is that COCA makes use of lexemes, not of word families. This solves some methodological problems of AWL and MAWL, which did not make a distinction between, for instance, the use of *dose* as a noun and as a verb. At present, COCA is the largest and most balanced freely available corpus of English, with a size of more than 520 million words. The subcorpus of the genre ACADEMIC includes more than 103 million words and the medical section ACAD: Medicine has nearly 9 million words. What makes COCA exceptional among other corpora of English is that it is extended on an annual basis and offers roughly the same genre balance from year to year. The final advantage is that frequency data are available not only for the whole COCA, but also across the five main genres and their subdomains, including medicine. The main source of the medical subcorpus were research papers published in scientific medical journals.

The use of West's (1953) GSL has two disadvantages, it is more than sixty years old and it is a kind of exclusion list. My aim was to replace the use of any potentially problematic exclusion list by a more sophisticated measure based on relative frequency. In my analysis, the first step in the comparison of COCA and ACAD: Medicine was to convert raw frequencies in COCA and raw frequencies in the subcorpus ACAD: Medicine to normalized frequencies to 10,000 words. This is a regular procedure which makes it possible to compare two corpora of different sizes. The results for normalized counts demonstrated that the general COCA and ACAD: Medicine include highly specialized medical terms such as *toxicologist* and *rhinitis*. In MAWL, such words are systematically and intentionally excluded by the methodology, which rejects terms in the narrow sense if they display a low frequency across subdisciplines.

The next step in my analysis was to calculate the relative frequency on the basis of the normalized counts in the two corpora. The relative frequency was calculated by taking the normalized frequency in the medical corpus and dividing it by the normalized frequency in the general corpus. Relative frequency is a measure of the typicality of a word in medical texts. If the relative frequency value is close to 1, the frequency in general COCA and ACAD: Medicine is approximately the same, i.e. the normalized frequencies in the two corpora are very similar. If the relative frequency is higher, the word is more frequent in medicine than in the general corpus. An extreme value, for instance 82.54449 for *odynophagia*, confirmed that this was a highly specialized medical term, but the low normalized frequency in ACAD: Medicine (0.050918) also suggests that such words may not be necessarily frequent even in medical texts.

The crucial question is how to determine a threshold value indicating when words are frequent enough to be considered characteristic of medical vocabulary. At this point, it is important to add that the relative frequency is only one of the two crucial dimensions in characterizing medical vocabulary. The other continuum is the absolute frequency in the medical corpus. This can be used to produce a threshold of words frequent enough in the medical corpus. The best value depends on the size of the corpus.

The claim made here is that it is only possible to characterize medical vocabulary accurately when the relative frequency and absolute frequency are both taken into account.

Whereas the relative frequency can be calculated for any word, my aim was to identify only those words that have at least a minimum absolute frequency in the medical corpus. Otherwise, rare words which are more frequent in ACAD: Medicine than in general COCA would also be included. This was illustrated by the example of *exchangeable*. Its normalized frequency in ACAD: Medicine is 0.004428 whereas in general COCA it is 0.001314. As a result, the relative frequency value is quite high, 3.37. However, the absolute frequency values are very low, only 2 in ACAD: Medicine and 49 in general COCA. In this case, the absolute values are not high enough to draw the conclusion that *exchangeable* is a typical word of medical vocabulary. Therefore, I first sorted the frequency word list by absolute frequency (total frequency) in the medical corpus, then selected the range within a threshold, e.g. 100,000, 10,000; 100, 90, etc. and re-sorted this by relative frequency. This yielded 13 threshold levels with a different number of words for each level. Level 13 included such a large proportion of the words that I divided it into ten sublevels.

On the basis of the results combining absolute and relative freqeuncy, a two-dimensional model representing medical vocabulary was proposed in section 3.3. The results indicate that relative frequency is a measure for the degree of typicality of a word in medical vocabulary. Then, the continuum is between two ends from general vocabulary with lower relative frequency values to highly specialized vocabulary and terms in the narrow sense with much higher relative frequency values.

A weak point of the use of relative frequency is that it is a measure that is inherently dependent on the coverage of other domains used in the general corpus as a benchmark. For instance, the relative frequency of *experiment* is lower than that of *extension*. Intuitively, *experiment* is more characteristic for medical vocabulary than *extension*. The reverse order can be explained because *experiment* is more frequent in other specialized domains in COCA, e.g. MAG: Sci/Tech, NEWS: Misc. To some extent, this effect can be countered by making use of absolute frequency.

The absolute frequency continuum was used to set a threshold of words frequent enough in the medical corpus. The data also confirmed that all frequency bands cover words ranging from less typical of medical vocabulary to more typical of medical vocabulary. This suggests that medical vocabulary is best characterized as a two-dimensional continuum where the two dimensions interact and depend on each other. The data also indicate that the choice for any particular threshold is always to some extent arbitrary. The combination of the two values suggests that lower absolute frequency requires higher relative frequency. This indicates that the relative frequency value and the absolute frequency value should be interconnected. a direct consequence is that the line separating medical vocabulary from general vocabulary in the continuum model given in Figure 3-1 will become a curve.

This raises the question of whether the relation between absolute frequency and relative frequency can be expressed by a formula. A point of departure in answering this question was the Edmundson's approach to text summarization (1969) based on a combination of four measures and their relative weights which were parameterized in a linear function. As mentioned above, the model in Figure 3-1 suggests that connecting the points with the same degree of 'medicalness' as found in the balance between the horizontal axis of relative frequency and the vertical axis of absolute frequency should be represented by a curve. This brings us to hypothesizing about the parameters which a corresponding formula should include. The outcome of the formula should correspond to the degree of typicality of medical vocabulary. The threshold of the relative frequency depends on the absolute frequency. This suggests that the formula must in any case include a division. Determining the points that are intuitively correct must also be taken into account in determining the formula. We can use parameters such as α, β, γ, and δ in Edmundson's (1969) formula. Including these parameters would ensure that the curve crosses these points. Determining the optimal values of the multiplier or multipliers would then delimit the function of the curve suggested in connection with Figure 3-1.

Another source of inspiration is the formula used for the BMI. Here we find two variables rather than four. Moreover, there is a division in the formula, which results in a curve. However, the relationship between the two variables in the BMI is different from what we need for our index of 'medicalness'. Whereas in the BMI, a higher weight can be compensated by a greater height to yield the same value, for the same degree of 'medicalness', a higher absolute frequency requires a lower relative frequency.

In order to answer the last two research questions I compared the results based on the subcorpus of medicine in COCA with an alternative corpus of medical texts. This illustrative medical corpus was compiled on the basis of the Wikipedia corpus, which was made available on the COCA website in 2015. It is based on the full text of the English version of the Wikipedia. The corpus covers 4.4 million Wikipedia articles with 1.9 billion words. The Wikipedia articles on medical topics represent a different type of medical text to medical journal articles. Especially this feature, a different text type, makes the corpus based on the Wikipedia web pages related to medicine interesting as an alternative corpus for the comparison of data in the two medical corpora. While it can be expected that the results are different, quantifying the influence of the text type is important in future attempts to determine medical vocabulary on an empirical basis.

The COCA interface allows users to create their own virtual corpora from Wikipedia for any topic. The size of the virtual corpora is limited, each of these corpora can contain up to 1,000,000 words in up to 1,000 articles. Once the virtual corpus has been created, it can be edited, articles can be added, deleted, or moved to another corpus.

In designing my Wikipedia medical corpus (WIMECO), determining the criteria for the choice of articles played an important role. The results based on the analysis of the subcorpus ACAD: Medicine in COCA were my starting point in selecting seed words. I selected the top ten most frequent medical content words to compile ten subcorpora of 100 articles each. All these were merged into the final WIMECO, which contains 1,111,735 words. A corpus analysis function offered by COCA makes it possible to produce a list of 500 keywords for different syntactic categories. First, it was necessary to compare WIMECO nouns, verbs, adjectives and adverbs with the same word classes in ACAD: Medicine in COCA. The comparison showed that for nouns, all 500 noun keywords in WIMECO also occur in ACAD: Medicine in COCA. For the other word classes, slight differences were found, one verb occurred in WIMECO, but not in ACAD: Medicine, five adjectives and fourteen adverbs were also found only in WIMECO.

Then a statistical comparison based on the Log-likelihood (LL) test was carried out. The results confirmed that the differences between frequencies for individual keywords in WIMECO and ACAD: Medicine are in most cases statistically significant. These findings reconfirm that medical vocabulary is best viewed as a continuum, but in a different sense from the one described above (see Figure 4-5). This continuum is based on the LL values and + and – values. These values result in producing a U-shaped curve with the highest LL values at the top and the lowest values at the bottom. The values + and – divide the continuum into two sides of a parabola. On the + end of the continuum there are nouns, verbs, adjectives and adverbs typical of WIMECO as opposed to the ones typical of ACAD: Medicine in COCA on the – end. On the bottom of the continuum curve, there are nouns, verbs, adjectives and adverbs equally important in both corpora.

The issue of the influence of a different text type was one of the main aims of my analysis. In this study, Wikipedia medical articles and scientific journal articles compiled in ACAD: Medicine in COCA were selected to represent different text types. The LL test results clearly indicate that ACAD: Medicine contains vocabulary typical of medical research whereas in WIMECO, health care practice is more prominently represented. In 85% of the vocabulary identified as keywords in WIMECO, the difference in frequency with the values for ACAD: Medicine was statistically significant. These findings might be seen as evidence to what extent the nature of the corpus has a significant influence on what emerges as medical vocabulary.

In conclusion, two main findings emerge from this research. The first is the curve based on the interaction of relative frequency and absolute frequency. The data indicate that determining a particular threshold is always to some extent arbitrary, but they show that a lower absolute frequency requires a higher relative frequency. Although no precise formula can be given at the moment, the method of relative weights among different dimensions might prove to be an efficient way of how to delimit the degree of typicality of medical vocabulary. The second finding is that the selection of a corpus plays a crucial role in determining what is found as medical vocabulary. In elaborating the empirical approach to identifying medical vocabulary, future research should address the question of the representativity of different text types again.

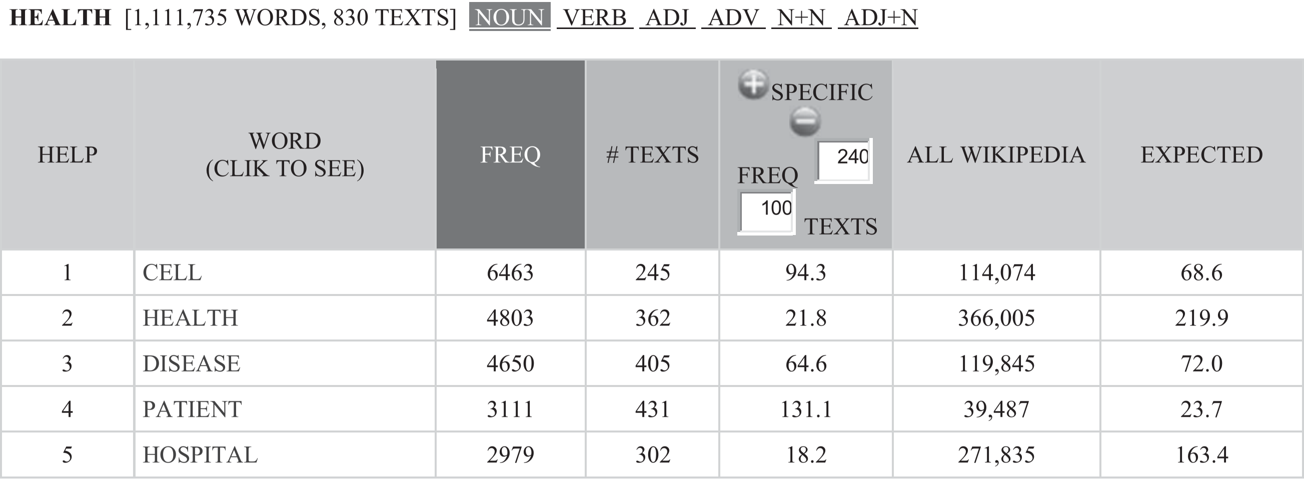
Appendix 1

List of Part of Speech Tags (PoS)

a	article
c	conjunction
d	determiner
e	existential *there*
i	preposition
j	adjective
m	numeral
n	noun
p	pronoun
r	adverb
t	infinitival *to*
u	interjection
v	verb
x	negative

APPENDIX 2

LIST OF KEYWORD NOUNS IN WIMECO



HEALTH [1,111,735 WORDS, 830 TEXTS] NOUN VERB ADJ ADV N+N ADJ+N

HELP	WORD (CLIK TO SEE)	FREQ	# TEXTS	SPECIFIC FREQ 240 100 TEXTS	ALL WIKIPEDIA	EXPECTED
1	CELL	6463	245	94.3	114,074	68.6
2	HEALTH	4803	362	21.8	366,005	219.9
3	DISEASE	4650	405	64.6	119,845	72.0
4	PATIENT	3111	431	131.1	39,487	23.7
5	HOSPITAL	2979	302	18.2	271,835	163.4

6	CARE	2867	335	28.0	170,169	102.3
7	YEAR	1984	410	2.1	1,595,774	959.0
8	SYSTEM	1948	376	3.8	851,690	511.8
9	TREATMENT	1650	369	19.0	144,792	87.0
10	RESEARCH	1609	352	5.1	526,321	316.3
11	INFECTION	1460	196	84.1	28,876	17.4
12	SERVICE	1351	253	2.9	771,355	463.5
13	TISSUE	1257	197	63.7	32,841	19.7
14	STUDY	1252	288	8.5	246,347	148.0
15	BLOOD	1246	222	13.2	156,628	94.1
16	PROGRAM	1240	245	4.2	491,408	295.3
17	TIME	1152	365	0.9	2,234,484	1,342.8
18	SYMPTOM	1133	179	447.9	4,209	2.5
19	CASE	1076	302	4.9	367,133	220.6
20	PEOPLE	1072	286	1.3	1,362,579	818.8
21	TYPE	975	289	5.0	326,976	196.5
22	USE	940	307	2.8	550,803	331.0

23	SCHOOL	931	152	0.7	2,211,136	1,328.8
24	NUMBER	902	325	1.4	1,093,924	657.4
25	RISK	896	205	15.9	93,833	56.4
26	COUNTRY	895	225	2.3	639,759	384.5
27	BODY	868	290	4.0	365,540	219.7
28	PHYSICIAN	854	220	30.2	46,984	28.2
29	CONDITION	817	283	13.9	97,949	58.9
30	GROUP	811	289	1.1	1,185,575	712.5
31	CHILD	810	201	5.3	255,737	153.7
32	LEVEL	809	239	3.8	358,095	215.2
33	AREA	799	309	1.1	1,211,544	728.1
34	INSURANCE	792	71	16.2	81,365	48.9
35	PART	791	349	1.0	1,359,856	817.2
36	STEM	786	57	42.6	30,700	18.4
37	DRUG	771	208	12.2	105,551	63.4
38	INFORMATION	750	204	3.0	417,675	251.0
39	DATA	740	139	4.2	291,796	175.4

40	STATE	734	230	0.8	1,511,279	908.2
41	TRIAL	718	116	8.6	139,222	83.7
42	PROTEIN	711	127	16.6	71,371	42.9
43	MEDICINE	707	222	9.0	131,194	78.8
44	FACTOR	695	210	14.3	81,044	48.7
45	TEST	676	168	5.9	189,745	114.0
46	RESULT	675	273	3.4	329,621	198.1
47	ORGANIZATION	664	198	3.7	300,060	180.3
48	PRACTICE	659	222	4.9	225,074	135.3
49	DISORDER	653	165	30.8	35,321	21.2
50	INDIVIDUAL	648	206	16.3	66,261	39.8
51	PROCESS	644	245	3.5	302,446	181.8
52	MILLION	639	186	2.1	504,723	303.3
53	DEVELOPMENT	631	268	1.7	623,408	374.6
54	CENTER	629	215	1.4	754,732	453.5
55	THERAPY	625	188	23.1	44,964	27.0
56	MEMBER	623	239	1.2	846,154	508.5

57	GOVERNMENT	622	161	1.1	955,093	574.0
58	CANCER	617	190	9.3	110,691	66.5
59	STUDENT	613	111	3.8	267,084	160.5
60	EFFECT	611	212	5.7	178,841	107.5
61	DOCTOR	597	183	6.7	149,316	89.7
62	POPULATION	584	211	1.0	970,134	583.0
63	RATE	578	204	5.1	188,847	113.5
64	FACILITY	577	176	7.8	123,515	74.2
65	COST	575	140	7.3	130,366	78.3
66	PERSON	571	193	3.7	256,838	154.3
67	DIAGNOSIS	565	180	40.0	23,528	14.1
68	FORM	561	233	2.4	390,646	234.8
69	DEATH	557	181	1.3	697,789	419.3
70	HEALTH-CARE	548	147	709.7	1,285	0.8
71	FUNCTION	546	220	6.2	147,323	88.5
72	FAMILY	544	211	0.7	1,219,805	733.0
73	TRAINING	542	156	2.6	348,004	209.1

74	EDUCATION	536	167	1.4	637,824	383.3
75	SITE	534	212	2.1	430,013	258.4
76	LIFE	528	210	0.8	1,130,634	679.4
77	PROBLEM	524	218	6.2	140,856	84.6
78	COMMUNITY	520	172	1.4	610,931	367.1
79	QUALITY	511	174	5.4	156,896	94.3
80	SAFETY	511	117	6.1	138,806	83.4
81	GENE	508	101	10.7	79,150	47.6
82	CHANGE	499	213	4.1	203,348	122.2
83	DEGREE	484	118	3.2	254,248	152.8
84	BUILDING	482	134	1.3	621,033	373.2
85	CAUSE	476	183	9.0	87,685	52.7
86	ROLE	468	210	1.6	474,661	285.2
87	PROVIDER	466	113	29.1	26,646	16.0
88	ACTIVITY	465	216	5.6	138,724	83.4
89	VIRUS	465	96	20.2	38,241	23.0
90	TERM	454	227	2.3	335,286	201.5

91	WORK	449	206	0.9	860,996	517.4
92	HEART	446	139	3.4	221,044	132.8
93	WOMAN	445	157	2.9	251,122	150.9
94	PLAN	445	131	3.7	198,821	119.5
95	METHOD	443	185	5.0	148,677	89.3
96	MANAGEMENT	434	155	2.3	319,663	192.1
97	SURGERY	433	156	11.5	62,766	37.7
98	MEMBRANE	419	81	31.5	22,106	13.3
99	RESPONSE	417	158	5.9	117,215	70.4
100	PROCEDURE	411	146	15.3	44,794	26.9
101	WORLD	410	199	0.4	1,816,394	1,091.5
102	MUSCLE	409	111	20.6	33,075	19.9
103	TUMOR	409	72	47.4	14,359	8.6
104	EMERGENCY	408	120	8.2	82,656	49.7
105	COMPANY	405	123	0.6	1,082,599	650.6
106	DAY	402	189	0.8	861,080	517.5
107	MODEL	400	132	2.1	314,122	188.8

108	ILLNESS	396	142	14.8	44,465	26.7
109	UNIT	394	134	2.5	261,163	156.9
110	POLICY	394	113	3.0	216,719	130.2
111	ANIMAL	391	108	6.0	109,108	65.6
112	LAW	389	114	1.0	626,110	376.3
113	STANDARD	388	136	6.4	100,350	60.3
114	FIELD	386	177	1.3	503,941	302.8
115	BONE	382	99	14.6	43,623	26.2
116	RECORD	381	88	1.3	506,756	304.5
117	INSTITUTION	372	151	6.5	94,764	56.9
118	ISSUE	370	174	3.3	188,889	113.5
119	BRAIN	368	112	7.6	80,891	48.6
120	SYNDROME	362	111	14.0	43,033	25.9
121	AGE	360	157	0.7	852,682	512.4
122	ADULT	353	136	6.4	91,192	54.8
123	CONTROL	352	167	1.6	377,589	226.9
124	WAY	347	205	1.1	544,134	327.0

125	INJURY	344	126	5.7	99,864	60.0
126	MONTH	343	153	3.3	170,809	102.6
127	MEDICATION	337	106	54.8	10,238	6.2
128	DEVICE	336	81	6.1	91,070	54.7
129	MOLECULE	332	82	35.1	15,753	9.5
130	LABORATORY	331	156	8.2	66,986	40.3
131	BENEFIT	331	131	12.5	44,084	26.5
132	PERIOD	329	157	1.1	504,047	302.9
133	LIVER	327	85	27.2	19,999	12.0
134	VACCINE	326	66	51.4	10,548	6.3
135	BACTERIA	325	72	22.3	24,299	14.6
136	PROFESSIONAL	324	128	69.6	7,741	4.7
137	STRUCTURE	323	154	2.3	238,713	143.5
138	REPORT	323	144	2.8	191,301	115.0
139	NURSE	323	126	18.9	28,470	17.1
140	NAME	320	189	0.4	1,202,029	722.3
141	STAGE	320	127	1.6	341,227	205.1

142	EVIDENCE	316	147	2.9	183,802	110.5
143	RECEPTOR	316	73	24.5	21,449	12.9
144	WORKER	314	93	13.4	38,938	23.4
145	STAFF	313	147	2.1	243,487	146.3
146	HUMAN	312	132	15.5	33,578	20.2
147	AGENCY	309	104	3.7	139,351	83.7
148	SUPPORT	308	145	1.7	308,853	185.6
149	GROWTH	308	132	3.3	153,538	92.3
150	WALL	305	61	3.3	155,086	93.2
151	HISTORY	302	160	0.4	1,258,359	756.2
152	SKIN	300	112	7.4	67,434	40.5
153	TECHNIQUE	298	130	6.8	73,396	44.1
154	EVENT	298	130	1.3	378,740	227.6
155	SERIES	297	133	0.4	1,292,661	776.8
156	PAIN	296	103	8.5	57,733	34.7
157	CULTURE	296	88	1.8	274,996	165.3
158	LOSS	294	133	2.4	202,967	122.0

159	DEPARTMENT	288	137	0.9	515,485	309.8
160	OTHERS	284	181	1.2	409,343	246.0
161	PLACE	284	168	0.6	794,095	477.2
162	ORGAN	284	133	6.7	70,647	42.5
163	SURFACE	282	99	2.7	176,969	106.3
164	NEED	280	161	4.8	97,534	58.6
165	UNIVERSITY	279	86	0.3	1,602,335	962.9
166	REGION	278	140	0.9	541,873	325.6
167	COURSE	274	109	2.1	219,740	132.1
168	TECHNOLOGY	272	122	1.4	314,956	189.3
169	ANTIBODY	272	73	93.0	4,866	2.9
170	OFFICER	271	60	1.7	267,652	160.8
171	FUNDING	269	94	4.8	94,005	56.5
172	ABILITY	267	138	3.0	149,928	90.1
173	LUNG	267	94	20.8	21,329	12.8
174	LINE	267	94	0.6	793,493	476.8
175	SIGN	266	99	6.0	73,597	44.2

176	SCIENCE	265	114	1.1	413,488	248.5
177	CLASS	265	106	1.0	453,710	272.7
178	RESEARCHER	262	115	17.9	24,344	14.6
179	CENTURY	262	112	0.6	773,866	465.0
180	OUTCOME	261	120	14.5	29,912	18.0
181	ANTIGEN	260	48	96.8	4,470	2.7
182	SOURCE	259	143	1.9	228,001	137.0
183	ANALYSIS	259	108	3.0	141,961	85.3
184	WEEK	257	119	1.5	289,742	174.1
185	ACCESS	256	117	2.2	192,927	115.9
186	CLINIC	252	113	15.7	26,709	16.1
187	AGENT	250	105	3.4	121,755	73.2
188	MAN	247	100	0.7	573,149	344.4
189	CAMPUS	247	77	2.3	179,412	107.8
190	COVERAGE	247	47	6.0	68,056	40.9
191	BILL	247	34	1.6	249,363	149.9
192	APPLICATION	245	120	3.4	118,992	71.5

193	ORGANISM	245	67	33.1	12,327	7.4
194	APPROACH	244	124	3.3	122,332	73.5
195	ACID	243	78	5.9	69,031	41.5
196	PRODUCT	242	116	2.7	148,989	89.5
197	MEASURE	242	113	7.7	52,390	31.5
198	REQUIREMENT	241	100	12.1	33,270	20.0
199	FOOD	241	99	1.6	245,252	147.4
200	INDUSTRY	239	98	1.4	284,744	171.1
201	EXAMINATION	239	91	9.8	40,604	24.4
202	TRANSPLANT	238	75	56.7	6,983	4.2
203	PRESSURE	236	106	3.0	130,770	78.6
204	HOME	235	121	0.6	645,325	387.8
205	ANTIBIOTIC	232	57	102.6	3,762	2.3
206	AMOUNT	231	123	2.7	143,440	86.2
207	REGULATION	231	103	9.1	42,324	25.4
208	HOUR	231	102	3.8	101,397	60.9
209	DAMAGE	231	85	3.5	111,390	66.9

210	JOURNAL	230	98	1.9	203,075	122.0
211	ASSESSMENT	230	91	8.8	43,261	26.0
212	PRACTITIONER	230	87	37.8	10,127	6.1
213	POINT	229	116	0.8	472,072	283.7
214	ENVIRONMENT	228	126	2.6	146,087	87.8
215	SPECIALTY	228	85	18.4	20,599	12.4
216	KIDNEY	227	72	27.8	13,581	8.2
217	EFFORT	226	126	3.3	112,726	67.7
218	BED	226	91	7.6	49,425	29.7
219	CONTACT	226	78	4.9	76,473	46.0
220	IMAGE	226	75	2.0	184,203	110.7
221	PHASE	226	70	3.6	103,562	62.2
222	MOUSE	226	52	10.3	36,637	22.0
223	BEHAVIOR	224	92	4.9	75,414	45.3
224	COMPONENT	222	129	6.9	53,853	32.4
225	RESOURCE	222	116	9.3	39,544	23.8
226	IMAGING	219	53	22.6	16,135	9.7

227	TEAM	218	106	0.3	1,396,036	838.9
228	LOCATION	216	136	1.3	275,641	165.6
229	TESTING	216	92	7.4	48,750	29.3
230	PRODUCTION	215	107	0.7	530,634	318.9
231	RANGE	214	139	1.2	296,478	178.2
232	HOST	214	79	2.9	120,767	72.6
233	OUTBREAK	214	47	9.3	38,487	23.1
234	TRACT	213	78	19.4	18,256	11.0
235	VARIETY	212	149	1.8	195,130	117.3
236	PREVENTION	212	115	11.2	31,395	18.9
237	MECHANISM	212	109	7.3	48,300	29.0
238	ROOM	211	92	1.6	218,309	131.2
239	REVIEW	210	100	1.9	187,046	112.4
240	PERCENT	209	86	2.6	133,497	80.2
241	THEORY	209	74	1.6	217,276	130.6
242	PLANT	209	47	1.7	203,433	122.3
243	SECTION	208	114	1.1	311,689	187.3

244	IMPROVEMENT	208	103	7.6	45,314	27.2
245	END	207	126	0.4	786,620	472.7
246	NETWORK	207	104	1.0	338,486	203.4
247	EXPOSURE	207	79	7.8	44,001	26.4
248	NUCLEUS	207	44	19.5	17,674	10.6
249	SCIENTIST	206	105	6.5	52,833	31.7
250	ACTION	204	114	1.2	289,641	174.1
251	BIRTH	204	102	2.6	130,182	78.2
252	PURPOSE	201	120	2.8	119,942	72.1
253	WATER	200	76	0.6	588,134	353.4
254	EXAMPLE	198	124	2.1	157,491	94.6
255	KNOWLEDGE	198	101	2.3	143,254	86.1
256	EYE	198	85	3.3	99,769	60.0
257	OFFICE	197	103	0.7	492,918	296.2
258	PROVISION	195	83	9.8	32,962	19.8
259	FAILURE	194	89	3.4	93,815	56.4
260	RIGHT	192	94	1.6	205,935	123.8

261	PROJECT	192	92	0.8	416,840	250.5
262	DIVISION	192	66	0.5	643,172	386.5
263	MATERIAL	191	102	1.6	193,223	116.1
264	SUBJECT	191	91	2.5	128,637	77.3
265	PATHOGEN	191	69	71.7	4,432	2.7
266	TRANSMISSION	191	63	5.3	59,729	35.9
267	PRESCRIPTION	191	48	41.5	7,655	4.6
268	RELATIONSHIP	190	96	1.6	193,566	116.3
269	NERVE	190	63	13.9	22,789	13.7
270	SPECIES	189	74	0.5	585,820	352.0
271	SIDE	188	118	0.6	544,879	327.4
272	PUBLIC	188	108	1.7	182,979	110.0
273	INTEREST	188	107	1.3	248,748	149.5
274	EXPERIENCE	188	103	2.0	153,207	92.1
275	BOARD	188	77	0.8	376,693	226.4
276	BILLION	187	81	3.1	99,570	59.8
277	MOVEMENT	186	96	1.1	281,969	169.4

278	INCREASE	186	95	3.7	82,856	49.8
279	ERROR	186	45	6.7	45,891	27.6
280	NATION	185	89	1.9	160,781	96.6
281	FEVER	185	75	11.5	26,768	16.1
282	DIFFERENCE	184	102	3.8	79,772	47.9
283	SIZE	184	95	1.0	296,934	178.4
284	DECISION	183	91	1.7	184,400	110.8
285	CYCLE	183	65	3.9	78,310	47.1
286	DOSE	183	53	24.4	12,462	7.5
287	GAME	183	23	0.3	1,102,818	662.7
288	FUND	182	70	3.5	87,349	52.5
289	PRESENCE	181	106	2.5	121,826	73.2
290	PREGNANCY	181	60	14.7	20,424	12.3
291	MARROW	181	54	56.6	5,321	3.2
292	RESPONSIBILITY	180	94	4.0	75,708	45.5
293	DIABETES	180	86	20.2	14,824	8.9
294	EMPLOYEE	180	64	8.2	36,537	22.0

295	AUTHORITY	179	98	1.8	169,349	101.8
296	OUTPATIENT	179	96	75.8	3,931	2.4
297	COMPLICATION	179	80	102.8	2,898	1.7
298	SIGNAL	178	57	3.2	91,173	54.8
299	NEURON	178	45	72.3	4,097	2.5
300	STATUS	177	108	1.4	203,660	122.4
301	CONCERN	177	103	5.7	51,346	30.9
302	SETTING	177	95	5.8	51,103	30.7
303	GRADUATE	177	63	3.5	83,957	50.5
304	GOAL	176	99	1.3	232,802	139.9
305	ORDER	176	83	0.6	458,075	275.3
306	MORTALITY	176	75	18.7	15,670	9.4
307	VESSEL	175	65	4.8	60,325	36.3
308	INCOME	175	45	1.1	266,669	160.3
309	FLUID	174	86	9.5	30,456	18.3
310	ARTERY	174	35	22.9	12,663	7.6
311	FINDING	173	88	25.3	11,388	6.8

312	CITY	173	86	0.2	1,831,383	1,100.5
313	INTERVENTION	173	72	8.8	32,738	19.7
314	REFORM	173	40	3.5	81,624	49.1
315	OPERATION	172	118	1.2	235,035	141.2
316	LACK	172	101	2.4	117,628	70.7
317	MUTATION	172	52	25.6	11,181	6.7
318	PSYCHOLOGIST	172	33	20.8	13,768	8.3
319	ADMINISTRATION	171	94	1.3	215,792	129.7
320	PATHWAY	171	57	19.3	14,774	8.9
321	HAND	170	98	1.3	220,048	132.2
322	PERFORMANCE	167	70	0.8	364,029	218.8
323	PLAGUE	167	14	17.1	16,244	9.8
324	IMPACT	166	104	2.3	119,927	72.1
325	SPECIALIST	166	89	7.3	37,694	22.7
326	ARTICLE	166	88	1.4	191,713	115.2
327	MOTHER	166	68	0.8	360,965	216.9
328	FEATURE	164	107	1.8	155,269	93.3

329	EXPANSION	164	69	2.6	105,765	63.6
330	ENZYME	164	66	7.0	38,807	23.3
331	RULE	164	65	1.8	151,186	90.9
332	EMPLOYER	164	48	13.3	20,557	12.4
333	ASSOCIATION	163	93	0.6	482,200	289.8
334	COMPLEX	162	67	2.6	105,685	63.5
335	DELIVERY	161	90	5.5	48,629	29.2
336	VERSION	161	72	0.5	556,056	334.2
337	GUIDELINE	160	78	111.9	2,379	1.4
338	SKILL	159	72	7.8	33,860	20.3
339	SPACE	158	88	0.8	330,000	198.3
340	TARGET	158	88	2.9	90,749	54.5
341	RESIDENT	158	84	8.2	32,209	19.4
342	COLLEGE	158	55	0.3	889,591	534.6
343	AMBULANCE	158	28	14.7	17,863	10.7
344	PAPER	157	82	1.7	157,781	94.8
345	SUBSTANCE	156	76	9.2	28,120	16.9

346	BASIS	155	110	1.8	146,795	88.2
347	PAYMENT	155	47	7.0	37,104	22.3
348	SHIP	155	29	0.7	380,921	228.9
349	COMBINATION	154	113	3.1	82,598	49.6
350	MAJORITY	153	97	1.3	196,633	118.2
351	CATEGORY	153	96	2.3	112,214	67.4
352	MARKET	153	59	1.0	247,541	148.8
353	ACTIVATION	153	52	16.2	15,689	9.4
354	INFLAMMATION	152	63	35.2	7,179	4.3
355	CHARACTER	152	46	0.6	404,150	242.9
356	PROPERTY	151	87	1.1	222,889	133.9
357	TEACHING	151	86	3.9	64,184	38.6
358	PARENT	151	64	5.3	47,346	28.5
359	EXCHANGE	151	47	1.9	134,010	80.5
360	POSITION	150	100	0.6	447,192	268.7
361	SAMPLE	150	74	5.1	49,016	29.5
362	CRITERIA	150	68	7.4	33,959	20.4

363	REASON	149	110	2.1	117,995	70.9
364	PNEUMONIA	149	49	25.7	9,665	5.8
365	SECTOR	149	47	2.2	111,480	67.0
366	OPTION	148	82	4.1	60,245	36.2
367	SET	148	78	1.0	237,464	142.7
368	CHAIN	147	56	2.8	86,898	52.2
369	FACULTY	147	48	2.0	121,628	73.1
370	LIST	145	98	0.5	502,357	301.9
371	COMMUNICATION	145	71	2.4	100,655	60.5
372	LAYER	145	56	5.2	46,144	27.7
373	CHROMOSOME	144	36	24.4	9,834	5.9
374	SURGEON	143	76	8.5	28,013	16.8
375	POWER	143	76	0.4	589,172	354.1
376	PROFESSION	143	59	8.0	29,561	17.8
377	PSYCHOLOGY	143	36	4.2	56,980	34.2
378	OXYGEN	143	34	6.9	34,702	20.9
379	DESIGN	142	67	0.6	420,445	252.7

380	ENTITY	142	58	6.5	36,319	21.8
381	VALUE	141	90	1.4	169,947	102.1
382	TOOL	141	82	4.6	51,110	30.7
383	PLASMA	141	51	12.4	18,996	11.4
384	DISCOVERY	140	78	2.9	79,861	48.0
385	HORMONE	139	48	26.4	8,776	5.3
386	ENTRY	138	66	2.3	101,135	60.8
387	AIR	138	59	0.3	664,705	399.4
388	DONOR	138	50	17.2	13,322	8.0
389	PATTERN	137	88	3.4	67,931	40.8
390	INVESTIGATION	136	89	2.5	91,866	55.2
391	REACTION	136	81	2.7	83,556	50.2
392	INPATIENT	136	78	98.0	2,310	1.4
393	TAX	136	39	1.8	126,141	75.8
394	EXPRESSION	135	55	3.5	64,444	38.7
395	EMBRYO	135	29	42.9	5,233	3.1
396	SITUATION	134	87	1.9	117,518	70.6

397	STRATEGY	134	74	3.1	72,699	43.7
398	DEFINITION	132	71	3.0	72,971	43.9
399	MISSION	132	68	1.0	229,590	138.0
400	LEGISLATION	132	61	3.1	70,869	42.6
401	ONSET	132	51	15.9	13,845	8.3
402	QUESTION	131	83	2.2	101,127	60.8
403	PERSONNEL	131	62	1.3	163,538	98.3
404	EXAM	131	38	11.2	19,485	11.7
405	POTENTIAL	130	66	4.3	50,414	30.3
406	SURVIVAL	130	62	6.0	35,974	21.6
407	WEIGHT	129	71	1.9	111,298	66.9
408	STRESS	129	63	5.7	37,637	22.6
409	CONSUMER	129	44	4.4	48,747	29.3
410	RADIATION	129	41	5.6	38,035	22.9
411	INTERACTION	128	74	5.6	38,051	22.9
412	FAT	128	29	15.6	13,645	8.2
413	FORMATION	127	82	1.5	140,185	84.2

414	DIRECTOR	127	74	0.4	542,541	326.0
415	CONSTRUCTION	127	69	0.6	355,131	213.4
416	FOCUS	126	86	2.4	87,274	52.4
417	RESISTANCE	126	61	2.3	90,216	54.2
418	TRANSPLANTATION	126	61	49.7	4,216	2.5
419	FREQUENCY	126	60	3.2	64,934	39.0
420	STRAIN	126	47	11.7	17,937	10.8
421	REDUCTION	125	71	4.4	47,570	28.6
422	PROTECTION	125	69	1.5	139,079	83.6
423	INVOLVEMENT	125	68	3.5	59,086	35.5
424	ATTACK	125	62	0.9	220,544	132.5
425	FACT	124	95	0.9	224,970	135.2
426	PATHOLOGY	124	69	20.1	10,265	6.2
427	NATURE	123	96	1.0	206,374	124.0
428	ORIGIN	123	81	1.5	132,698	79.7
429	INFANT	123	65	8.9	23,103	13.9
430	BUSINESS	123	64	0.4	515,663	309.9

431	MOTOR	123	41	2.3	90,168	54.2
432	STATION	122	48	0.2	872,859	524.5
433	SOCIETY	121	62	0.4	469,906	282.4
434	HEAD	120	71	0.5	385,914	231.9
435	INCIDENCE	120	69	19.8	10,092	6.1
436	SPREAD	120	56	8.7	22,970	13.8
437	IMMUNITY	120	37	12.8	15,569	9.4
438	MEMORY	120	34	1.6	126,685	76.1
439	STEP	119	82	3.1	64,848	39.0
440	COLLECTION	119	62	0.7	266,762	160.3
441	EPISODE	119	48	0.5	424,813	255.3
442	CAPACITY	118	84	1.4	144,051	86.6
443	PRINCIPLE	118	80	3.4	57,198	34.4
444	VISIT	118	64	3.4	57,998	34.9
445	COMPUTER	118	55	1.0	205,190	123.3
446	PREVALENCE	118	53	22.3	8,815	5.3
447	DIET	118	42	5.6	35,295	21.2

448	MANDATE	118	25	8.2	23,938	14.4
449	CONE	118	8	13.1	15,029	9.0
450	ASPECT	117	84	4.9	40,124	24.1
451	FORCE	117	66	0.4	469,514	282.1
452	INITIATIVE	117	61	3.2	60,139	36.1
453	DISABILITY	117	61	10.4	18,654	11.2
454	WARD	116	57	2.0	96,280	57.9
455	CLINICIAN	116	54	122.9	1,571	0.9
456	DEFICIENCY	115	62	18.1	10,551	6.3
457	PARTNER	115	61	1.8	105,746	63.5
458	CORE	115	55	1.9	99,486	59.8
459	CLASSIFICATION	115	50	2.4	80,048	48.1
460	SCALE	115	47	2.0	94,242	56.6
461	SOFTWARE	115	41	1.1	172,221	103.5
462	RESIDENCY	114	43	13.6	13,967	8.4
463	ELM	114	1	17.6	10,755	6.5
464	BUDGET	113	53	2.3	80,572	48.4

465	SECURITY	113	43	0.8	236,231	142.0
466	TRANSFER	113	41	2.2	86,677	52.1
467	ADMISSION	112	50	6.6	28,440	17.1
468	EPIDEMIC	112	45	16.4	11,398	6.8
469	CERTIFICATION	112	41	6.1	30,528	18.3
470	PUBLICATION	111	77	1.5	121,037	72.7
471	EQUIPMENT	111	59	1.2	151,560	91.1
472	SCORE	111	41	1.0	185,783	111.6
473	FIBER	111	33	10.6	17,425	10.5
474	INSURER	111	31	90.3	2,045	1.2
475	WORD	110	71	0.8	219,898	132.1
476	CHARACTERISTIC	110	71	9.5	19,244	11.6
477	LEADER	110	70	0.7	280,354	168.5
478	GENERATION	110	60	1.5	124,260	74.7
479	PROTOCOL	110	52	5.0	36,808	22.1
480	CHARITY	110	46	2.8	65,287	39.2
481	COMMITTEE	110	39	0.6	331,197	199.0

482	FOUNDATION	109	74	0.8	239,780	144.1
483	SOLUTION	109	63	2.5	73,669	44.3
484	HOUSE	109	53	0.2	971,871	584.0
485	IMPLEMENTATION	109	53	3.6	50,430	30.3
486	VICTIM	109	52	4.2	43,573	26.2
487	STORAGE	109	50	2.5	73,457	44.1
488	CORD	109	45	12.6	14,369	8.6
489	URINE	109	38	21.0	8,624	5.2
490	BOWEL	109	24	51.5	3,525	2.1
491	CONTRAST	108	75	2.7	67,797	40.7
492	CONCEPT	108	74	1.3	141,869	85.3
493	EVALUATION	108	67	5.4	32,978	19.8
494	TITLE	108	61	0.3	539,305	324.1
495	TRANSPORT	108	46	1.1	161,780	97.2
496	SUCCESS	107	73	0.7	268,485	161.3
497	RECOMMENDATION	107	69	10.5	16,980	10.2
498	GENETICS	107	42	11.8	15,079	9.1

499	IDEA	106	68	1.1	153,427	92.2
500	VISION	106	54	2.2	79,458	47.7

APPENDIX 3

LL- TEST RESULTS FOR KEYWORD NOUNS IN WIMECO AS OPPOSED TO ACAD: MEDICINE

Log-likelihood Ratio Calculator

Step 1. Enter the **corpus sizes** in **A** and **B**.
Step 2. Enter the **frequency counts** in columns B and C.
* The white cells are data cells; the gray ones are result cells.

	A		B
Corpus Size 1	**1111735**	**Corpus Size 2**	**6571682**

Word	Freq. in Corpus 1	Freq. in Corpus 2	Log - likelihood	Sig.		
cell	6463	3983	12346,05	0	***	+
health	4803	12840	1924,63	0	***	+
disease	4650	6014	5249,64	0	***	+
patient	3111	24793	266,55	0	***	-
hospital	2979	5508	2239,1	0	***	+
care	2867	6396	1621,76	0	***	+
year	1984	7287	321,37	0	***	+
system	1948	7814	216,45	0	***	+
treatment	1650	7589	80,81	0	***	+
research	1609	3981	755,01	0	***	+
infection	1460	2925	979,18	0	***	+
service	1351	5765	108,52	0	***	+
tissue	1257	3429	481,91	0	***	+

study	1252	15246	743,28	**0**	***	-
blood	1246	2764	711,57	**0**	***	+
program	1240	5557	73,24	**0**	***	+
time	1152	7750	17,32	**0**	***	-
symptom	1133	2914	492,32	**0**	***	+
case	1076	8253	69,04	**0**	***	-
people	1072	3281	310,6	**0**	***	+
type	975	3408	189,02	**0**	***	+
use	940	5448	0,31	**0,577**		+
school	931	3521	134,28	**0**	***	+
number	902	4806	8	**0,005**	**	+
risk	896	5511	1,23	**0,268**		-
country	895	1107	1053,49	**0**	***	+
body	868	2115	419,39	**0**	***	+
physician	854	2654	237,42	**0**	***	+
condition	817	3000	132,46	**0**	***	+
group	811	9474	420,56	**0**	***	-
child	810	9645	447,42	**0**	***	-
level	809	6860	103,58	**0**	***	-
area	799	5061	3,35	**0,067**		-
insurance	792	449	1578,11	**0**	***	+
part	791	1962	369,38	**0**	***	+
stem	786	112	2398,2	**0**	***	+
drug	771	2067	307,06	**0**	***	+
information	750	4742	2,98	**0,084**		-
data	740	6907	157,72	**0**	***	-
state	734	2507	153,78	**0**	***	+
trial	718	2210	204,9	**0**	***	+
protein	711	818	892,48	**0**	***	+
medicine	707	1022	713,74	**0**	***	+
factor	695	4520	5,62	**0,018**	*	-

test	676	4363	4,62	**0,032**	*	-
result	675	7645	314,99	**0**	***	-
organization	664	2174	158,89	**0**	***	+
practice	659	2785	55,92	**0**	***	+
disorder	653	1491	354,97	**0**	***	+
individual	648	2497	86,39	**0**	***	+
process	644	3191	16,04	**0**	***	+
million	639	719	817,44	**0**	***	+
development	631	2447	81,83	**0**	***	+
center	629	1518	309,45	**0**	***	+
therapy	625	2826	34,69	**0**	***	+
member	623	1966	165,98	**0**	***	+
government	622	923	610,54	**0**	***	+
cancer	617	1837	192,12	**0**	***	+
student	613	3946	3,94	**0,047**	*	-
effect	611	5560	114,99	**0**	***	-
doctor	597	733	707,46	**0**	***	+
population	584	2728	25,32	**0**	***	+
rate	578	4751	61,04	**0**	***	-
facility	577	1695	185,85	**0**	***	+
cost	575	2042	105,46	**0**	***	+
person	571	2037	103,04	**0**	***	+
diagnosis	565	2558	31,07	**0**	***	+
form	561	1580	200,01	**0**	***	+
death	557	1803	137,9	**0**	***	+
health-care	548	2230	56,76	**0**	***	+
function	546	2146	66,7	**0**	***	+
family	544	3678	8,84	**0,003**	**	-
training	542	2952	3,01	**0,083**		+
education	536	2285	43,3	**0**	***	+
site	534	2948	2,08	**0,149**		+

life	528	1763	118,95	**0**	***	+
problem	524	5169	143,48	**0**	***	-
community	520	2562	13,77	**0**	***	+
quality	511	2695	5,45	**0,02**	*	+
safety	511	2165	42,75	**0**	***	+
gene	508	527	694,34	**0**	***	+
change	499	3839	32,81	**0**	***	-
degree	484	2102	35,02	**0**	***	+
building	482	546	613,1	**0**	***	+
cause	476	1415	148,8	**0**	***	+
role	468	1879	51,77	**0**	***	+
provider	466	1138	224,21	**0**	***	+
activity	465	3048	4,42	**0,036**	*	-
virus	465	684	460,78	**0**	***	+
term	454	660	455,58	**0**	***	+
work	449	2553	0,57	**0,45**		+
heart	446	902	294,93	**0**	***	+
woman	445	2442	2,05	**0,153**		+
plan	445	1303	144,49	**0**	***	+
method	443	4006	80,77	**0**	***	-
management	434	3360	29,98	**0**	***	-
surgery	433	2853	4,55	**0,033**	*	-
membrane	419	845	278,26	**0**	***	+
response	417	3798	78,83	**0**	***	-
procedure	411	3378	43,38	**0**	***	-
world	410	759	307,65	**0**	***	+
muscle	409	2769	6,78	**0,009**	**	-
tumor	409	2875	11,22	**0,001**	***	-
emergency	408	1068	170,96	**0**	***	+
company	405	637	372,54	**0**	***	+
day	402	3885	100,61	**0**	***	-

model	400	3086	26,89	**0**	***	-
illness	396	1111	142,46	**0**	***	+
unit	394	1677	32,11	**0**	***	+
policy	394	1296	92,96	**0**	***	+
animal	391	728	291,1	**0**	***	+
law	389	807	247,46	**0**	***	+
standard	388	1408	65,78	**0**	***	+
field	386	1151	119,72	**0**	***	+
bone	382	2847	19,19	**0**	***	-
record	381	922	186,49	**0**	***	+
institution	372	699	273,49	**0**	***	+
issue	370	2303	0,86	**0,354**		-
brain	368	349	538,42	**0**	***	+
syndrome	362	1122	101,38	**0**	***	+
age	360	3726	120,23	**0**	***	-
adult	353	1592	19,95	**0**	***	+
control	352	4447	234,27	**0**	***	-
way	347	1811	4,39	**0,036**	*	+
injury	344	2625	21,17	**0**	***	-
month	343	3080	60,32	**0**	***	-
medication	337	981	111,02	**0**	***	+
device	336	1933	0,21	**0,647**		+
molecule	332	170	694,05	**0**	***	+
laboratory	331	919	121,96	**0**	***	+
benefit	331	1434	24,31	**0**	***	+
period	329	2725	36,48	**0**	***	-
liver	327	547	279,62	**0**	***	+
vaccine	326	629	230,93	**0**	***	+
bacteria	325	1030	85,54	**0**	***	+
professional	324	1341	30,78	**0**	***	+
structure	323	1268	39,67	**0**	***	+

report	323	2615	30,86	**0**	***	-
nurse	323	2863	53,06	**0**	***	-
name	320	322	447,87	**0**	***	+
stage	320	1275	36,71	**0**	***	+
evidence	316	1850	0,03	**0,874**		+
receptor	316	259	511,25	**0**	***	+
worker	314	1612	5,06	**0,024**	*	+
staff	313	2223	9,68	**0,002**	**	-
human	312	391	362,83	**0**	***	+
agency	309	1577	5,39	**0,02**	*	+
support	308	2219	11,13	**0,001**	***	-
growth	308	1184	41,49	**0**	***	+
wall	305	638	191,54	**0**	***	+
history	302	1964	2,44	**0,118**		-
skin	300	1775	0	**0,988**		-
technique	298	1794	0,09	**0,77**		-
event	298	1945	2,6	**0,107**		-
series	297	858	99,68	**0**	***	+
pain	296	2770	63,88	**0**	***	-
culture	296	1050	54,51	**0**	***	+
loss	294	2304	22,29	**0**	***	-
department	288	1574	1,47	**0,225**		+
others	284	1222	21,73	**0**	***	+
place	284	902	74,3	**0**	***	+
organ	284	352	333,66	**0**	***	+
surface	282	2007	8,93	**0,003**	**	-
need	280	2598	58	**0**	***	-
university	279	392	290,15	**0**	***	+
region	278	1499	1,94	**0,164**		+
course	274	1287	11,35	**0,001**	***	+
technology	272	1200	17,83	**0**	***	+
antibody	272	679	125,45	**0**	***	+

officer	271	370	290,16	**0** ***	+
funding	269	433	240,87	**0** ***	+
ability	267	1401	3,1	**0,078**	+
lung	267	755	94,31	**0** ***	+
line	267	1012	38,16	**0** ***	+
sign	266	818	76,1	**0** ***	+
science	265	391	261,73	**0** ***	+
class	265	857	65,79	**0** ***	+
researcher	262	769	84,57	**0** ***	+
century	262	216	422,27	**0** ***	+
outcome	261	2495	62,19	**0** ***	-
antigen	260	345	286,35	**0** ***	+
source	259	1944	13,85	**0** ***	-
analysis	259	5043	508,77	**0** ***	-
week	257	2298	44,21	**0** ***	-
access	256	768	78,18	**0** ***	+
clinic	252	930	40,07	**0** ***	+
agent	250	904	42,95	**0** ***	+
man	247	1400	0,37	**0,545**	+
campus	247	64	658,8	**0** ***	+
coverage	247	368	241,4	**0** ***	+
bill	247	176	435,56	**0** ***	+
application	245	1061	18,04	**0** ***	+
organism	245	506	156,93	**0** ***	+
approach	244	2006	25,8	**0** ***	-
acid	243	565	127,95	**0** ***	+
product	242	1762	9,74	**0,002** **	-
measure	242	2359	62,93	**0** ***	-
requirement	241	915	34,21	**0** ***	+
food	241	3736	281,24	**0** ***	-
industry	239	938	29,39	**0** ***	+

examination	239	1797	12,98	**0**	***	-
transplant	238	148	452,51	**0**	***	+
pressure	236	3186	190,82	**0**	***	-
home	235	1852	18,57	**0**	***	-
antibiotic	232	963	21,71	**0**	***	+
amount	231	1367	0	**0,988**		-
regulation	231	653	81,65	**0**	***	+
hour	231	1782	15,52	**0**	***	-
damage	231	611	94,7	**0**	***	+
journal	230	246	306,8	**0**	***	+
assessment	230	2290	65,53	**0**	***	-
practitioner	230	662	77,88	**0**	***	+
point	229	1934	28,62	**0**	***	-
environment	228	1750	14,71	**0**	***	-
specialty	228	176	383,18	**0**	***	+
kidney	227	293	256,76	**0**	***	+
effort	226	1350	0,02	**0,884**		-
bed	226	516	122,87	**0**	***	+
contact	226	1157	3,78	**0,052**		+
image	226	771	47,53	**0**	***	+
phase	226	1399	0,42	**0,518**		-
mouse	226	235	308,33	**0**	***	+
behavior	224	4432	455,05	**0**	***	-
component	222	1351	0,16	**0,687**		-
resource	222	1405	0,91	**0,341**		-
imaging	219	497	120,32	**0**	***	+
team	218	1445	2,56	**0,11**		-
location	216	1130	2,63	**0,105**		+
testing	216	1393	1,45	**0,229**		-
production	215	563	90,02	**0**	***	+
range	214	2092	56,34	**0**	***	-
host	214	217	297,74	**0**	***	+

outbreak	214	1199	0,51	**0,473**		+
tract	213	540	95,28	**0**	***	+
variety	212	746	40,16	**0**	***	+
prevention	212	1112	2,48	**0,116**		+
mechanism	212	924	15,01	**0**	***	+
room	211	867	20,81	**0**	***	+
review	210	1490	6,42	**0,011**	*	-
percentage	209	1020	6,1	**0,013**	*	+
theory	209	423	138,06	**0**	***	+
plant	209	497	105,67	**0**	***	+
section	208	1760	26,3	**0**	***	-
improvement	208	1548	10,31	**0,001**	**	-
end	207	1142	0,82	**0,365**		+
network	207	585	73,21	**0**	***	+
exposure	207	2445	111,35	**0**	***	-
nucleus	207	260	240,22	**0**	***	+
scientist	206	179	320,58	**0**	***	+
action	204	1187	0,04	**0,835**		+
birth	204	464	111,64	**0**	***	+
purpose	201	1120	0,59	**0,444**		+
water	200	4350	492,12	**0**	***	-
example	198	906	10,08	**0,001**	**	+
knowledge	198	1647	22,55	**0**	***	-
eye	198	466	102,01	**0**	***	+
office	197	823	17,81	**0**	***	+
provision	195	343	156,56	**0**	***	+
failure	194	1498	13,13	**0**	***	-
right	192	662	38,99	**0**	***	+
project	192	1139	0	**0,964**		-
division	192	158	309,82	**0**	***	+
material	191	1916	56,16	**0**	***	-
subject	191	5316	740,79	**0**	***	-

pathogen	191	588	54,48	**0**	***	+
transmission	191	551	64,32	**0**	***	+
prescription	191	195	264,35	**0**	***	+
relationship	190	2067	76,77	**0**	***	-
nerve	190	1987	66,1	**0**	***	-
species	189	398	117,46	**0**	***	+
side	188	1242	2,07	**0,15**		-
public	188	807	14,59	**0**	***	+
interest	188	906	6,21	**0,013**	*	+
experience	188	1629	27,27	**0**	***	-
board	188	789	16,58	**0**	***	+
billion	187	129	335,96	**0**	***	+
movement	186	1033	0,6	**0,437**		+
increase	186	1846	52,28	**0**	***	-
error	186	979	2,05	**0,152**		+
nation	185	300	164,22	**0**	***	+
fever	185	435	95,47	**0**	***	+
difference	184	4121	479,4	**0**	***	-
size	184	1460	15,28	**0**	***	-
decision	183	1218	2,31	**0,129**		-
cycle	183	662	31,39	**0**	***	+
dose	183	1170	0,99	**0,319**		-
game	183	158	286,03	**0**	***	+
fund	182	247	196,04	**0**	***	+
presence	181	1451	16,16	**0**	***	-
pregnancy	181	435	89,74	**0**	***	+
marrow	181	106	354,89	**0**	***	+
responsibility	180	628	35,12	**0**	***	+
diabetes	180	329	137,41	**0**	***	+
employee	180	1383	11,71	**0,001**	***	-
authority	179	637	32,59	**0**	***	+
outpatient	179	353	122,87	**0**	***	+

complication	179	1488	20,31	**0**	***	-
signal	178	810	9,44	**0,002**	**	+
neuron	178	197	230,88	**0**	***	+
status	177	1412	15,27	**0**	***	-
concern	177	1208	3,3	**0,069**		-
setting	177	1413	15,35	**0**	***	-
graduate	177	273	166,48	**0**	***	+
goal	176	1509	24,04	**0**	***	-
order	176	554	47,22	**0**	***	+
mortality	176	1141	1,33	**0,249**		-
vessel	175	388	100,03	**0**	***	+
income	175	344	120,69	**0**	***	+
fluid	174	553	45,43	**0**	***	+
artery	174	641	27,86	**0**	***	+
finding	173	3408	348,25	**0**	***	-
city	173	593	35,82	**0**	***	+
intervention	173	2644	194,81	**0**	***	-
reform	173	225	194,27	**0**	***	+
operation	172	1122	1,48	**0,223**		-
lack	172	1017	0	**0,997**		-
mutation	172	373	101,98	**0**	***	+
psychologist	172	69	397,94	**0**	***	+
administration	171	820	5,91	**0,015**	*	+
pathway	171	311	131,41	**0**	***	+
hand	170	1315	11,67	**0,001**	***	-
performance	167	2012	95,82	**0**	***	-
plague	167	13	556,37	**0**	***	+
impact	166	1303	12,74	**0**	***	-
specialist	166	535	41,63	**0**	***	+
article	166	1476	27,72	**0**	***	-
mother	166	543	39,83	**0**	***	+

feature	164	1141	3,96	**0,046**	*	-
expansion	164	130	271,08	**0**	***	+
enzyme	164	232	169,35	**0**	***	+
rule	164	533	40,12	**0**	***	+
employer	164	379	87,29	**0**	***	+
association	163	1070	1,59	**0,207**		-
complex	162	209	183,32	**0**	***	+
delivery	161	674	14,39	**0**	***	+
version	161	327	105,79	**0**	***	+
guideline	160	829	2,26	**0,133**		+
skill	159	1807	75,09	**0**	***	-
space	158	760	5,31	**0,021**	*	+
target	158	687	11,35	**0,001**	***	+
resident	158	761	5,25	**0,022**	*	+
college	158	641	16,62	**0**	***	+
ambulance	158	28	462,03	**0**	***	+
paper	157	1005	0,88	**0,349**		-
substance	156	769	4,11	**0,043**	*	+
basis	155	999	1,02	**0,311**		-
payment	155	273	124,23	**0**	***	+
ship	155	81	321,03	**0**	***	+
combination	154	983	0,8	**0,371**		-
majority	153	696	8,13	**0,004**	**	+
category	153	1112	6,05	**0,014**	*	-
market	153	413	60,03	**0**	***	+
activation	153	260	128,32	**0**	***	+
inflammation	152	339	86,02	**0**	***	+
character	152	100	280,4	**0**	***	+
property	151	721	5,43	**0,02**	*	+
teaching	151	208	160,24	**0**	***	+
parent	151	2122	136,49	**0**	***	-
exchange	151	255	127,62	**0**	***	+

position	150	1907	101,75	**0**	***	-
sample	150	4011	542,35	**0**	***	-
criteria	150	1494	42,8	**0**	***	-
reason	149	1192	13,12	**0**	***	-
pneumonia	149	310	94,36	**0**	***	+
sector	149	332	84,46	**0**	***	+
option	148	788	1,33	**0,249**		+
set	148	808	0,78	**0,378**		+
chain	147	266	113,73	**0**	***	+
faculty	147	319	87,06	**0**	***	+
list	145	665	7,26	**0,007**	**	+
communication	145	984	2,49	**0,114**		-
layer	145	369	64,37	**0**	***	+
chromosome	144	78	293,29	**0**	***	+
surgeon	143	895	0,41	**0,523**		-
power	143	1202	17,37	**0**	***	-
profession	143	487	30,25	**0**	***	+
psychology	143	58	329,47	**0**	***	+
oxygen	143	551	19,07	**0**	***	+
design	142	1723	83,35	**0**	***	-
entity	142	283	95,98	**0**	***	+
value	141	3094	353,02	**0**	***	-
tool	141	813	0,07	**0,786**		+
plasma	141	488	28,27	**0**	***	+
discovery	140	107	236,73	**0**	***	+
hormone	139	205	137,35	**0**	***	+
entry	138	356	59,6	**0**	***	+
air	138	1657	78,32	**0**	***	-
donor	138	235	115,44	**0**	***	+
pattern	137	1559	64,99	**0**	***	-
investigation	136	1252	27,11	**0**	***	-
reaction	136	934	2,77	**0,096**		-

inpatient	136	361	55,33	**0**	***	+
tax	136	142	184,94	**0**	***	+
expression	135	424	36,45	**0**	***	+
embryo	135	87	251,85	**0**	***	+
situation	134	1243	27,72	**0**	***	-
strategy	134	1838	113,3	**0**	***	-
definition	132	584	8,5	**0,004**	**	+
mission	132	200	126,64	**0**	***	+
legislation	132	255	93,34	**0**	***	+
onset	132	544	12,82	**0**	***	+
question	131	1876	124,59	**0**	***	-
personnel	131	621	5,02	**0,025**	*	+
exam	131	109	209,86	**0**	***	+
potential	130	663	2,29	**0,13**		+
survival	130	816	0,41	**0,522**		-
weight	129	1301	38,78	**0**	***	-
stress	129	1666	91,72	**0**	***	-
consumer	129	576	7,81	**0,005**	**	+
radiation	129	615	4,7	**0,03**	*	+
interaction	128	1053	13,59	**0**	***	-
fat	128	142	165,7	**0**	***	+
formation	127	499	15,54	**0**	***	+
director	127	712	0,3	**0,585**		+
construction	127	310	61,16	**0**	***	+
focus	126	692	0,57	**0,452**		+
resistance	126	816	0,93	**0,335**		-
transplantation	126	217	103,92	**0**	***	+
frequency	126	2099	175	**0**	***	-
strain	126	620	3,38	**0,066**		+
reduction	125	1621	89,97	**0**	***	-
protection	125	1243	35,42	**0**	***	-
involvement	125	779	0,3	**0,581**		-

attack	125	377	37,65	**0**	***	+
fact	124	1122	22,68	**0**	***	-
pathology	124	390	33,35	**0**	***	+
nature	123	747	0,08	**0,78**		-
origin	123	337	46,7	**0**	***	+
infant	123	472	16,69	**0**	***	+
business	123	768	0,32	**0,57**		-
motor	123	478	15,8	**0**	***	+
station	122	237	85,59	**0**	***	+
society	121	611	2,42	**0,12**		+
head	120	1586	91,31	**0**	***	-
incidence	120	1317	50,07	**0**	***	-
spread	120	306	53,06	**0**	***	+
immunity	120	92	202,53	**0**	***	+
memory	120	214	94,64	**0**	***	+
step	119	1103	24,53	**0**	***	-
collection	119	849	3,87	**0,049**	*	-
episode	119	533	7,06	**0,008**	**	+
capacity	118	631	0,98	**0,323**		+
principle	118	471	13,42	**0**	***	+
visit	118	826	3,08	**0,079**		-
computer	118	894	6,84	**0,009**	**	-
prevalence	118	676	0,1	**0,754**		+
diet	118	318	46,47	**0**	***	+
mandate	118	102	184,29	**0**	***	+
cone	118	33	307,97	**0**	***	+
aspect	117	795	2,05	**0,152**		-
force	117	2077	188,03	**0**	***	-
initiative	117	392	26,08	**0**	***	+
disability	117	796	2,09	**0,149**		-
ward	116	98	183,97	**0**	***	+
clinician	116	577	2,77	**0,096**		+

deficiency	115	334	38,1	**0**	***	+
partner	115	338	36,99	**0**	***	+
core	115	289	52,35	**0**	***	+
classification	115	487	9,65	**0,002**	**	+
scale	115	1575	96,82	**0**	***	-
software	115	483	10,1	**0,001**	**	+
residency	114	82	199,92	**0**	***	+
elm	114	2	421,17	**0**	***	+
budget	113	354	30,73	**0**	***	+
security	113	288	50,01	**0**	***	+
transfer	113	476	9,76	**0,002**	**	+
admission	112	602	0,84	**0,361**		+
epidemic	112	126	143,3	**0**	***	+
certification	112	129	140,45	**0**	***	+
publication	111	268	54,56	**0**	***	+
equipment	111	932	13,39	**0**	***	-
score	111	2694	336,75	**0**	***	-
fiber	111	552	2,66	**0,103**		+
insurer	111	68	212,7	**0**	***	+
word	110	1217	47,25	**0**	***	-
characteristic	110	1618	112,28	**0**	***	-
leader	110	457	10,24	**0,001**	**	+
generation	110	263	55,06	**0**	***	+
protocol	110	989	19,47	**0**	***	-
charity	110	168	104,61	**0**	***	+
committee	110	448	11,35	**0,001**	***	+
foundation	109	123	139,1	**0**	***	+
solution	109	887	10,78	**0,001**	**	-
house	109	255	56,77	**0**	***	+
implementation	109	658	0,04	**0,839**		-
victim	109	400	17,71	**0**	***	+

storage	109	412	15,76	**0**	***	+
cord	109	329	32,76	**0**	***	+
urine	109	272	50,31	**0**	***	+
bowel	109	155	111,95	**0**	***	+
contrast	108	714	1,21	**0,272**		-
concept	108	592	0,51	**0,475**		+
evaluation	108	1851	160,25	**0**	***	-
title	108	120	139,63	**0**	***	+
transport	108	213	74,13	**0**	***	+
success	107	987	21,54	**0**	***	-
recommendation	107	686	0,62	**0,43**		-
genetics	107	40	254,1	**0**	***	+
idea	106	500	4,24	**0,039**	*	+
vision	106	354	23,87	**0**	***	+

BIBLIOGRAPHY

Anderson, Jennifer Joline (2011), *Wikipedia: The Company and Its Founders.* Minnesota: ABDO Publishing Company.

Bauer, Laurie (1983), *English Word-Formation*. Cambridge: Cambridge University Press.

—. (1988), *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Bauer, Laurie, & Nation, Paul (1993), 'Word families', *International Journal of Lexicography*, 6, 253-279.

Beard, Robert (1982), 'The plural as a lexical derivation', *Glossa* 16: 133-148.

Bedogni, Giorgio, Tiribelli, Claudio and Bellentani, Stefano (2005), 'Body Mass Index: From Quetelet to Evidence-Based Medicine', in Ferrera, Linda A. (ed.), *Body Mass Index: New Research*, New York: Nova Science Publishers, Inc., pp. 1-12.

Bernardini, Silvia and Gavioli, Laura (1999), 'L'analisi di piccoli e grandi corpora', in Haarman (Ed.), *Ricerche linguistiche: Strumenti e riflessioni metodologiche.* Pescara: Libreria dell'Universita Editrice. pp. 83-109.

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

Biber, Douglas, Conrad, Susan, Reppen, Randi (1998), *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Pat and Helt, Marie. (2002), 'Speaking and writing in the university: A multi-dimensional comparison', *TESOL Quarterly* 36: 9-48.

Biber, Douglas, Conrad, Susan and Cortes, Viviana (2004) '*If you look at…*: Lexical bundles in university teaching and textbooks', *Applied Linguistics* 25: 371-405.

Biber, Douglas and Jones, K. James (2009), 'Quantitative methods in corpus linguistics', in Lüdeling Anke and Kytö, Merja (eds.), *Corpus Linguistics, An International Handbook, Volume 2*, Berlin: Mouton de Gruyter, pp. 1287-1304.

Biber, Douglas and Susan Conrad (2009), *Register, Genre, and Style*. New York, NY: Cambridge University Press.

Billuroğlu, Ali & Neufeld, Steven (2005), 'The Bare Necessities in Lexis: a new perspective on vocabulary profiling.', available at http://lextutor.ca/vp/BNL_Rationale.doc

Carter, Ronald (2012), *Vocabulary: Applied Linguistic Perspectives.* New York: Routledge.

Code of Laws of the United States of America, U. S. code, available at http://uscode.house.gov/

Chen, Qui and Ge, Guang-Chun (2007), 'A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs)', *English for Specific Purposes* 26: 502–14.

Chung, Teresa Mihwa and Nation, Paul (2003), 'Technical vocabulary in specialised texts', *Reading in a Foreign Language* 15: 103-116.

Chung, Teresa Mihwa and Nation, Paul (2004), 'Identifying technical vocabulary', *System*, 32: 251-263.

Cowan, J. Ronayne (1974), 'Lexical and syntactic research for the design of EFL reading materials', *TESOL Quarterly* 8(4): 389-399.

Coxhead, Averil (1998), *The development and evaluation of a new academic word list.* Unpublished MA thesis, Victoria University of Wellington, New Zealand.

—. (2000), 'A new academic word list', *TESOL Quarterly* 34: 213–38.

—. (2013), 'Vocabulary and ESP', in Brian Paltridge and Sue Starfield (eds.), *The Handbook of English for Specific Purposes*, Chichester, UK: John Wiley & Sons, Ltd, pp. 141-158.

Coxhead, Averil and Hirsh, David (2007), 'A pilot science word list for EAP', *Revue Française de linguistique appliqueé* xii (2): 65-78.

Crawford Camiciottoli, B. (2007), *The Language of Business Studies Lectures*. Amsterdam: John Benjamins.

Crawford, William J. and Csomay, Eniko (2016), *Doing Corpus Linguistics*. New Yourk: Routledge.

Csomay, Eniko (2005), 'Linguistic variation within university classroom talk: A corpus-based perspective', *Linguistics and education* 15/3: 243-274.

—. (2006), 'Academic talk in American university classrooms: Crossing the boundaries of oral – literate discourse', *Journal of English for academic purposes*. 5: 117-135.

Culpeper, Jonathan and Demmen, Jane (2015), 'Keywords', in Douglas Biber and Randi Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics*. pp. 90-105. [Online]. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. Available from: Cambridge Books Online

<http://dx.doi.org/10.1017/CBO9781139764377.006> [Accessed 21 February 2016].

Davies, Mark (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Damerau, Fred J. (1993), 'Generating and evaluating domain-oriented multi-word terms from texts', *Information Processing & Management* 29:433-447.

Edmundson, Harold P. (1969), 'New Methods in Automatic Extracting', *Journal of the Association for Computing Machiner*y, 16 (2): 264-285.

Farrell, Paul (1990), *Vocabulary in ESP: a lexical analysis of the English of electronics and a study of semitechnical vocabulary*. CLCS Occasional Paper No. 25 Trinity College.

Ferguson, Gibson (2013), 'English for Medical Purposes.', in: Brian Paltridge and Sue Starfield (eds.), *The Handbook of English for Specific Purposes,* pp. 243-262.

Fischbach, Henry (1993), 'Translation, the Great Pollinator of Science', in Wright, Sue Ellen & Wright Jr., Leland D. (eds.), *Scientific and Technical Translation*, Amsterdam/Philadelphia: John Benjamins, pp. 89-100.

Flowerdew, Lynne (2004), 'The argument for using English specialized corpora', in Connor, Ulla and Thomas A. Upton (eds.) *Discourse in the Professions: Perspectives from corpus linguistics*. Amsterdam/ Philadelphia: Benjamins, pp. 11-36.

Francis, W. Nelson and Kučera, Henry (1982), *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Gardner, Dee (2013), *Exploring Vocabulary: Language in Action*. New York: Routledge.

Gardner, Dee and Davies, Mark (2013), 'A new academic vocabulary list', *Applied Linguistics*, 35: 1- 24.

Garside, Roger (1987). 'The CLAWS Word-tagging System'. In: Garside, Roger; Leech, Geoffrey and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman, pp. 30-41.

Gavioli, Laura (2005), *Exploring Corpora for ESP learning*. . Amsterdam/Philadelphia: John Benjamins.

Giannoni, Davide (2008), 'Medical writing at the periphery: The case of Italian journal editorials', *Journal of English for Academic Purposes* 7: 97-107.

Granger, Sylviane (2002), 'A Bird's-eye View of Computer Learner Corpus Research', in: Granger Sylviane; Hung Joseph; Petch-Tyson

Stephanie, *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (Language Learning and Language Teaching; 6), Amsterdam & Philadelphia: John Benjamins, pp. 3-33. http://hdl.handle.net/2078.1/75823

Gries, Stefan Th. (2010), 'Useful statistics for corpus linguistics', In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, Frankfurt am Main: Peter Lang, pp. 269-291.

Gunnarsson, Britt-Louise (2009), *Professional Discourse*, London: Continuum.

ten Hacken, Pius (2008). 'Prototypes and Discreteness in Terminology', in Bernal, Elisenda; DeCesaris, Janet (eds.). *Proceedings of the XIII Euralex International Congress*. IULAUPF. Barcelona, pp. 979-987.

—. (2010), 'The Tension between Definition and Reality in Terminology', in Dykstra, Anne & Schoonheim, Tanneke (eds.), *Proceedings of the XIV Euralex International Congress.* Ljouwert: Fryske Akademy / Afuk, pp. 915-927.

—. (2014), 'Delineating Derivation and Inflection', in Lieber, Rochelle and Pavol Štekauer (eds.), *The Oxford handbook of derivational morphology*, Oxford: Oxford University Press, pp. 10-25.

—. (2015), 'Terms and Specialized Vocabulary: Taming the Prototypes', in Hendrik J. Kockaert, Frieda Steurs (eds.), *Handbook of Terminology*, *Volume 1*, Amsterdam/Philadelphia: John Benjamins, pp. 3–13.

ten Hacken, Pius and Panocová, Renáta (2015), 'Introduction: Medical Language, Word Formation and Transparency', in ten Hacken Pius and Renáta Panocová (eds.), *Word Formation and Transparency in Medical English*, Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 1-12.

Handford, Michael (2010), 'What can a corpus tell us about specialist genres?', in: A. O'Keeffe & M. McCarthy (eds.), *The Routledge handbook of corpus linguistics*. London: Routledge, pp. 255-269.

Harris, Zellig (1968), *Mathematical Structures of Language*, New York: Interscience.

Harwood, Nigel (2005), '*Nowhere has anyone attempted...In this article I aim to do just that.* A corpus-based study of self-promotional I & WE in academic writing across four disciplines'. *Journal of Pragmatics* 37: 1207-1231.

—. (2007), 'Political scientists on the functions of personal pronouns in their writing: an interview-based study of *I and we*. ', *Text & Talk* 27(1): 27-54.

Havel, Richard J. (1999), ‘Dietary supplement or drug? The case of Cholestin’ (editorial), *The American Journal of Clinical Nutrition*, 69: 175-176.

He, Zeyi (2012), *Digital By-Product Data in Web 2.0: Exploring Mass Collaboration of Wikipedia.* New Castle upon Tyne: Cambridge Scholars Publishing.

Heatley, A., & Nation, Paul (1996), *Range [Computer software]*. Wellington, New Zealand: Victoria University of Wellington. (Available from http://www.vuw.ac.nz/ lals

Heber, David, Yip, Ian, Ashley, Judith M., Elashoff, David A., Elashoff, Robert M., Go, Vay Liang W. (1999), ‘Cholesterol-lowering effects of a proprietary Chinese red-yeast-rice dietary supplement’, *The American Journal of Clinical Nutrition*; 69:231-236.

Hirvela, Alan (2013), ‘ESP and Reading’, in Brian Paltridge and Sue Starfield (eds.), *The Handbook of English for Specific Purposes*, Chichester, UK: John Wiley & Sons, Ltd, pp. 100-115.

Hudson, Richard A. (1980), *Sociolinguistics,* Cambridge: Cambridge University Press.

Hunston, Susan (2002), *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hutchinson, Tom and Waters, Alan (1987). English for specific purposes *: a learning-centred approach.* Cambridge: Cambridge University Press.

Hyland, Ken (2002), ‘Authority and invisibility: authorial identity in academic writing. ’, *Journal of Pragmatics,* 34 (8): 1091-1112.

Hyland, Ken, & Tse, Polly (2007), ‘Is There an "Academic Vocabulary?"’, *TESOL Quarterly*, 41(2): 235-253.

Hyland, Ken, & Tse, Polly (2009), ‘Academic Lexis and Disciplinary Practice: Corpus Evidence for Specifity’, *International Journal of English Studies*, 9(2): 111-129.

Jackendoff, Ray (1983), *Semantics and Cognition*. Cambridge (Mass.): MIT Press.

Jelliffe, Derrick B. and Jelliffe, Eleanore F. Patrice (1979), ‘Undeappreciated pioneers. Quetelet: man and index.’ *American journal of clinical nutrition* 32(12): 2519-2521.

Jensen, John. T. (1990), *Morphology*. Amsterdam/Philadelphia: John Benjamins.

Johansson, Stig (1978), *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo, Norway: University of Oslo, Department of English.

Johns, Ann M. (2013), ‘The History of English for Specific Purposes Research’, in Brian Paltridge and Sue Starfield (eds.), *The Handbook* *of English for Specific Purposes*, Chichester, UK: John Wiley & Sons, Ltd, pp. 22-49.

Kittredge, Richard I. (1982), 'Variation and Homogeneity of Sublanguages', in Kittredge, Richard and Lehrberger, John, *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin, New York: de Gruyter, pp. 107-137.

—. (1983), 'Semantic Processing of Texts in Restricted Sublanguages', *Comp. and Maths with Applications* 9.1: 45-58.

—. (1987), 'The significance of sublanguage for automatic translation', in Nirenburg, Sergei (ed.), *Machine Translation: Theoretical and Methodological Issues*, Cambridge: Cambridge University Press, pp. 59-67.

Konstantakis, Nikolaos (2007), 'Creating a business word list for teaching English', *Estudios de lingüística inglesa aplicada*, 7: 79-102.

Labov, William (1973). 'The Boundaries of Words and Their Meanings'. In Bailey, C.-J. N.; Shuy, R.W. (eds.). *New Ways of Analyzing Variation in English*. Washington DC: Georgetown University Press. 340-373. Repr. in Aarts, Bas; Denison, David; Keizer, Evelien; Popova, Gergana (eds.) (2004). *Fuzzy Grammar: A Reader*. Oxford: Oxford University Press, pp. 67-89.

Lackstrom, John, Selinker, Larry and Louis P. Trimble (1972), 'Grammar and Technical English', *English Teaching Forum*: X, 5.

Lam, Jacqueline Kam-mei (2001), *A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography* (Research reports, Vol. 3). Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Lankamp, Robert Eduard (1989), *A study on the effect of terminology on l2 reading comprehension: should specialist terms in medical texts be avoided?*. Amsterdam: Rodopi.

Laufer, Batia, & Goldstein, Zahava. (2004), 'Testing vocabulary knowledge: size, strength and computer adaptiveness', *Language Learning 54*, 399-436.

Lehrberger, John (2014), 'Sublanguage Analysis', in Grishman, Ralph and Kittredge, Richard (eds.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, New York, London: Psychology Press. First published 1986, pp. 19-38.

Li, Wentian (1992), 'Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution', *IEEE Transactions on Information Theory* **38** (6): 1842–1845. doi:10.1109/18.165464

Livnat, Zohar (2012), *Dialogue, Science and Academic Writing.* Amsterdam: John Benjamins.

Lüdeling, Anke (2008), *Corpus Linguistics: An International Handbook, Volume 1,* Berlin: Mouton de Gruyter.

Maher, John (1986), 'The development of English as an international language of medicine', *Applied Linguistics* 7: 206-218.

Mani, Inderjeet (2001), *Automatic Summarization*, Amsterdam: Benjamins.

McCarthy, Michael and Handford, Michael (2004), 'Invisible to us': A preliminary corpus-based study of spoken business English', in Connor, U. and Upton, T. (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics*, Amsterdam: John Benjamins, pp. 167 –201.

McEnery, Tony and Wilson, Andrew (2001), *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

McEnery, Tony; Xiao, Richard, and Tono, Yukio (2006), *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

McEnery, Tony and Andrew Hardie (2011), *Corpus Linguistics*. Cambridge: Cambridge University Press.

Meara, Paul (1990), 'A note on passive vocabulary', *Second Language Research 6*, 150-154.

Melka, Francine (1997), 'Receptive vs. productive aspects of vocabulary'. In Norbert Schmitt and Michael McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press, pp. 84-102.

Milton, James (2009). *Measuring Second Language Vocabulary Acquisition*. Clevedon: Multilingual Matters.

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

Mugdan, Joachim (1989), 'Review of Bauer 1988', in Geert E. Booij and Jaap van Marle (eds.),*Yearbook of Morphology* 2: 175-183.

Murison-Bowie, Simon (1993), *MicroConcord manual*. Oxford: Oxford University Press.

Myers, Greg (1992), 'In this paper we report: speech acts and scientific facts. ', *Journal of Pragmatics*, 17: 295-313.

Nagy, William and Townsend, Dianna (2012), 'Words as tools: Learning academic vocabulary as language acquisition,' *Reading Research Quarterly* 47: 91–108.

Nation, I.S.P. (Paul) (2001), *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

—. (2008), *Teaching Vocabulary: Strategies and Techniques*. Boston: Heinle Cengage Learning.

—. (2010). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nelson, Mike (2000), *A Corpus-based Study of Business English and Business English Teaching Materials*. Unpublished PhD Dissertation. Manchester: University of Manchester. Available at http://users.utu.fi/micnel/thesis.html

Nesi, Hilary (2013), 'ESP and Corpus Studies', in Brian Paltridge and Sue Starfield (eds.), *The Handbook of English for Specific Purposes*, Chichester, UK: John Wiley & Sons, Ltd, pp. 451-473.

Neufeld, Steven, Hancioğlu, Nilgün and Eldridge, John (2011), 'Beware the range in RANGE, and the academic in AWL', *System* 39: 533–8.

Nippold, Marylin A. and Sun, Lei (2008), 'Knowledge of morphologically complex words: a developmental study of older children and young adolescents', *Language, Speech, and Hearing Services in Schools* 39: 365–73.

OED (2014, 2015), *Oxford English Dictionary*, Third edition, edited by John Simpson, www.oed.com.

Palmberg, Rolf (1987). Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition 9*, 201-220.

Panocová, Renáta (2016), 'A Descriptive Approach to Medical English Vocabulary', in: *Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity*. Tbilisi: Ivane Javakhishvili Tbilisi University Press, pp. 529-540.

Panocová, Renáta and ten Hacken, Pius (2017), 'Naming Symptoms, Syndromes, and Diseases', in Calderón-Tichy, Marietta, Heuberger, Reinhard, and Chamson, Emil (eds.), *Health and Language*. Vienna: Peter Lang, pp. 225-234.

Pignot-Shahov, Virginie (2012), 'Measuring L2 Receptive and Productive Vocabulary Knowledge', University of Reading: *Language Studies Working Papers* 4, 37-45.

Picht, Heribert and Draskau, Jennifer (1985), *Terminology: An introduction.* Guildford: University of Surrey.

Powers, David M. W. (1998), 'Applications and Explanations Of Zipf's Law', *Association For Computational Linguistics*: 151–160.

Quetelet, Adolphe (1869), *Physique sociale ou essai sur le développement des facultés de l'homme*. Brussels: C. Muquardt.

Sager, Juan C., Dungworth, David and McDonald, Peter F. (1980), *English Special Languages*. Wiesbaden: Brandstetter.

Sanger, Lawrence Mark (2009), 'The Fate of Expertise after WIKIPEDIA', *Episteme* 6: 52-73.

Sarkar, Kamal (2014), 'Multilingual Summarization Approaches', in *Computational Linguistics: Concepts, Methodologies, Tools, and Applications*, Information Resources Management USA: IGI Global, pp. 158-177.

Segura-Bedmar, Isabel; Crespo, Mario and Cesar de Pablo-Sánchez (2010), 'Score-based Approach for Anaphora Resolution in Drug-Drug Interactions Documents', in Helmut Horacek, Elisabeth Metais, Rafael Munoz (eds.), *Natural Language Processing and Information Systems: 14th International Conference on Applications of Natural Language to Information Systems , NLDB 2009, Saarbrücken, Germany, June 24-26, 2009. Revised Papers*. Berlin/Heidelberg: Springer Verlag.

Schmitt, Diane and Schmitt, Norbert (2005), *Focus on Vocabulary: Mastering the Academic Word List*. Longman Press.

Schmitt, Diane and Schmitt, Norbert (2011, 2nd edition). *Focus on Vocabulary: Mastering the Academic Word List*. Pearson Longman Press.

Schmitt, Norbert (2010), *Researching Vocabulary. A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.

Sinclair, John (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stedman (1997), *Stedman's Concise Medical Dictionary for the Health Professions*, ed. John H. Dirckx, Baltimore: Williams & Wilkins.

Štekauer, Pavol (2015), 'The delimitation of derivation and inflection', in Müller, Peter O.; Ohnheiser, Ingeborg; Olsen, Susan & Rainer, Franz (eds.) (2015-16), *Word-Formation: An International Handbook of the Languages of Europe*, Berlin: Mouton De Gruyter, Vol. 1, pp. 218-235.

Tapscott, Don and Williams, Anthony D. (2006), *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio.

Temnikova, Irina, P. (2013), 'Recognizing sublanguages in scientific journal articles through closure properties', in the *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing* (BioNLP 2013), Sofia, Bulgaria, August 4-9 2013, pp.72–79.

Tognini-Bonelli, Elena (2001), *Corpus Linguistics at Work*, Amsterdam: John Benjamins.

Tollefsen, Deborah Perron (2009), 'Wikipedia and the Epistemology of Testimony', *Episteme: Special Issue on the Epistemology of the Mass Collaboration* 6 (1), 8-24.

Tribble, Christopher (1997), 'Improvising corpora for ELT: quick and dirty ways of developing corpora for language teaching', in Barbara Lewandowska-Tomaszczyk and Patrick James Melia, *Practical* *applications in language corpora*. Lodz: Lodz University Press. 106-117.

Vojteková, Marta (2015a), *Latinsko-slovensko-poľský slovník anatomických termínov*, Prešov: Vydavateľstvo Prešovskej univerzity.

—. (2015b), 'Koncepcia latinsko-slovensko-poľského slovníka anatomických termínov', in: Olchowa, Gabriela and Balowski Mieczysław (eds.), *Jezyki slowianskie w procesie przemian*, Banská Bystrica: Univerzita Mateja Bela v Banskej Bystrici, pp. 63-71.

Wang, Jing, Liang, Shao-lan, & Ge, Guang-chun. (2008), 'Establishment of a Medical Academic Word List', *English for Specific Purposes*, 27: 442-458.

Ward, Jeremy (2009), 'A basic engineering English word list for less proficient foundation engineering undergraduates', *English for Specific Purpose*, 28: 170-182.

Werner, Jochen A. and Davies, Kim (2004), *Metastases in Head and Neck Cancer*. Berlin, Heidelberg: Springer Science & Business Media. This publication is available in an electronic form without page numbers.

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

Wray, K. Brad (2009), 'The Epistemic Cultures of Science and WIKIPEDIA: A Comparison', *Episteme*, 6: 38-51.

Wynne, Martin (2008), 'Searching and concordancing', in Lüdeling Anke and Kytö, Merja (eds.), *Corpus Linguistics, An International Handbook, Volume 1*, Berlin: Mouton de Gruyter, pp. 706-737.

Xue, Guo-yi and Nation, Paul (1984), 'A university word list', *Language Learning and Communication*, 3: 215-229.

Zipf, George, K. (1935), *The Psychobiology of Language*. Boston: Houghton-Mifflin.

—. (1949), *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Zwicky, Arnold M. (1992), 'Clitics', in William Bright (ed.), *International Encyclopaedia of Linguistics*. Oxford: Oxford University Press.

Author Index

SUBJECT INDEX