A Descriptive Approach to Medical English Vocabulary

Renáta Panocová
Pavol Jozef Šafárik University in Košice
e-mail: renata.panocova@upjs.sk

Abstract

This paper presents research into the characterization of medical vocabulary in English. It aims to develop an optimal methodological approach to the characterization of medical vocabulary in English. It is based on the analysis of data from the medical subcorpus of the Corpus of Contemporary American English (COCA). Earlier corpus-based research into medical vocabulary was carried out mainly from a pedagogical perspective and resulted in medical word lists. In those approaches, all criteria are based on absolute frequencies. It would not be sufficient to replace absolute frequency with relative frequency, because a minimal degree of absolute frequency is also necessary. What I show is that the threshold to be set for the absolute frequency interacts with the relative frequency. Therefore a measure based on the interaction of absolute frequency and relative frequency is shown to provide a better tool for identifying medical vocabulary than previously used measures.

Keywords: relative frequency; absolute frequency; corpus; language for specific purposes (LSP)

Language is an important tool in professional communication in medicine. The history of medicine clearly points to Latin as a dominant language in medicine especially since the middle ages. This status has changed in the 20th century, especially towards the end, resulting in English taking over the most prominent role in medical texts. In this paper I explore the optimal methodology for characterizing English medical vocabulary or medical English (ME). First, I discuss the role of a corpus-based research in specialized languages including ME (section 1). Then I contrast this perspective with a descriptive approach to ME and I argue that each perspective requires a different methodology, although both may include corpus data (section 2). On this basis, I conclude that there are good arguments for developing a specific methodology appropriate for characterizing medical vocabulary (section 3) and I outline its principal steps (section 4). Finally, the main findings are summarised in the conclusion (section 5).

1 The Role of Corpora in Identifying Medical English

Corpora represent an important tool in research of the vocabulary of *English for Specific Purposes (ESP)*. This obviously includes English used in medical domains. The first initiative in the vocabulary delimitation in corpus-based research into ESP was Coxhead's Academic Word List (AWL) (Coxhead, 2000). Then, on this basis a number of specialized word lists were produced, including Wang et al.'s (2008) Medical Academic Word List (MAWL). The development of these academic word lists illustrates the significant role of corpora in identifying specialized vocabulary.

The development of 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. 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."

A crucial step in the process is corpus design. Coxhead's Academic Corpus contains articles from academic journals, edited academic journal articles available online, university textbooks or course books, and texts from several previously compiled corpora. The texts were collected in electronic form and the word count was determined after the bibliography had been removed. The texts were classified into four categories depending on their length. The corpus consisted of four subcorpora: arts, commerce, law, and science, each of them further subdivided into seven domain-specific corpora of 125,000 words each. Interestingly, the corpus does not include medicine. Words in the corpus were processed by the corpus analysis program Range (Heatley & Nation, 1996). This is a dedicated package by means of which complex queries can be answered very quickly.

The selection criteria for words are essential in the compilation of 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". On the basis of Bauer and Nation (1993), Coxhead (2000: 218) defines a word family as a stem plus all closely related affixed forms. 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).

The selection of the items for AWL was based on three criteria: specialized occurrence, frequency, and range. Specialized occurrence means that the word families had to be outside the first 2,000 most frequently occurring words of English, as represented by West's (1953) General Service List (GSL) in order to be included. As for frequency, a word family was considered relevant only if its members occurred at least 100 times in the Academic Corpus. 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 in 15 or more of the 28 subject areas. This eliminates words that are typical of only specific domains. As a result, Coxhead's AWL has 570 word families. On the basis of their frequency, they are divided into 10 sublists.

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. Wang et al. (2008) aimed at the development of a Medical Academic Word List (MAWL). Their first step was to compile a corpus of medical research articles. The size of their corpus was 1 093 011 running words. This is approximately one third of the Academic Corpus developed by Coxhead but the domain is much 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, etc. The research articles were selected from journal volumes published in the period 2000 to 2006 and all were written by native speakers. 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.

Similar to Coxhead (2000), the definition of a word family by Bauer and Nation (1993) was used in data processing. Coxhead's (2000) three criteria, specialized occurrence, range and frequency of a word family, were taken to be relevant in the development of MAWL. Word families with at least one

member in GSL were excluded, which meant that *blood* or *disease* were deleted from the list. The final number of word families in MAWL was 623. Fifty-four per cent of MAWL word families overlapped with Coxhead's AWL. Wang et al. interpret this difference as undermining "the usefulness of general academic word lists across different disciplines" (Wang et al., 2008: 451). Coxhead (2013: 147) suggests that the overlap between MAWL and AWL results from the fact that Wang et al. (2008) used GSL as a common core instead of AWL.

Both AWL and MAWL represent word lists and were designed to be used primarily in language teaching. The idea of word lists of specialized language is compatible with language learner's needs (Felber, 1984; Sager et al. 1980). It should be noted, however, that language learners are not the only target group of speakers who need ME. The learner may be an expert or a non-specialist. Also native speakers of English may need it, especially if they are not domain experts. Among non-specialists, translators represent a large group of users. If the target group of speakers of ME is more heterogenous, as this suggests, their needs may be reflected in the choice of methodology.

2 Does a Different Approach to Medical English Need a Different Methodology?

The comparison of AWL and MAWL raises at least three issues that are problematic when it is our aim to characterize medical vocabulary. They concern the use of word families, the use of the GSL, and the structure of the corpus.

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 do not mark the word class (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. In line with Bauer et al. (2013: 9), lexemes "are tied to particular inflectional paradigms (each lexeme is realized by a set of word-forms)".

The second problem concerns the gaps in the selected vocabulary. An example is *disease,* which is not found in MAWL. The reason is that *disease* occurs among the first 2000 GSL vocabulary items (number 1156) and, in line with Wang et al.'s methodology, it was excluded. AWL does not list *disease* either. This may be for the same reason or because medicine is not a field which was included in the corpus. As opposed to AWL, MAWL does include *symptom* (number 81) and *syndrome* (number 211). However, the example in (1) shows that the notions of *symptom*, *syndrome*, and *disease* and relationships among them are relevant in medicine.

(1) 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.

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

A syndrome is often explained in terms of symptoms, e.g. 'a concurrence of several symptoms in a disease; a set of such concurrent symptoms' (OED, 2015). Only when the mechanism of interrelation between symptoms and cause is understood and explained sufficiently, the corresponding condition is described as a disease. The example in (1a) indicates that these three words often co-occur in the same context. Therefore, it seems reasonable to assume that all of them should be included in a proper description of medical vocabulary. The example in (1) suggests that by excluding *disease*, MAWL does not give a full, coherent description of the medical vocabulary of English.

To sum up, both 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 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.

Finally, it is worth taking a critical look at the structure of the corpora. Coxhead (2000) compiled a highly structured corpus and used the structure 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. Coxhead (2000) and Wang et al. (2008) stipulate threshold values without arguing for them or showing what the effect of different values would be. It would be preferable to determine thresholds 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.

3 Frequency in the COCA Corpus

A medical corpus plays a crucial role in the characterization of medical vocabulary. This means that also the way a corpus is compiled and processed is central. The decision whether to use an existing corpus, which already solves some of the methodological issues described above, or design a new medical corpus was essential at the beginning of my research. Given the fact that compiling a new medical corpus is time-consuming and requires a well-trained team, I turned to already existing large corpora available online.

The Corpus of Contemporary American English (COCA) includes a subcorpus of academic texts labelled ACAD: Medicine. At present, COCA is one of the largest corpora of English.[1] 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 five main genres of 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 genre of academic journals

[1] Details about the design of COCA in this section were taken from at http://corpus.byu.edu/coca , information retrieved 13 January, 2016.

covers a separate section of medical research articles ACAD: Medicine. The remaining academic subdomains are history, education, general journals, geography, law and politics, humanities, philosophy and religion, science and technology and miscellaneous. The size of ACAD: Medicine is eight times bigger than that of the medical corpus for MAWL. This makes ACAD: Medicine the largest available corpus of medical research articles. ACAD: Medicine is part of COCA and therefore it shares relevant properties which solve some of the methodological concerns described above. First of all, COCA makes use of lexemes, not word families. COCA gives frequency data based on accurate part of speech (PoS) tagging for the texts in the corpus. Frequency data are available not only for the whole COCA, but also across main genres including a subdomain of medicine. Another advantage of COCA is that the texts it comprises are regularly added and they are recent. In addition, COCA offers approximately the same genre balance from year to year. From the COCA website an Excel file can be downloaded with a complete listing of the texts in the more than 520 million word corpus. The listings of the texts included in ACAD: Medicine provide a detailed description of the structure of this medical corpus. The listing gives the information that the texts are collected from more than 50 scientific medical journals. The articles cover different medical disciplines. All research papers were published between 1990 and 2015. It is important to mention that journals were selected on the basis of their availability in electronic form and copyright criteria. If compared to MAWL, the subdivision into particular subdomains did not play an essential role in ACAD: Medicine. The listing of medical journal articles is systematic. It includes information about the word count, year, domain, journal title of the paper, publication details, and identification label number of a text. All these advantages make ACAD: Medicine the best possible selection of the medical corpus for linguistic analysis.

Let us turn now to word frequency data based on the COCA corpus. The COCA word list with genre frequency 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 of 54 MB that can be printed and edited. These data served as a basis in search for answers to these main 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?

4 Analysis of Frequency Data in COCA

The first step in my analysis was to compare the frequencies for the whole corpus with the frequencies of the words found in ACAD: Medicine. The comparison of the full word list based on COCA with ACAD: Medicine reveals that 27,166 lexemes out of the 60,000 lexemes in the full list can also be found in ACAD: Medicine. Due to a great difference in size of the full COCA and ACAD: Medicine corpora, it was necessary to make these frequency data comparable. This was done by turning raw frequencies to normalized frequencies per 10,000 words. This was calculated separately for the general COCA corpus and for ACAD: Medicine. After producing normalized frequencies for the general COCA corpus and the ACAD: Medicine subcorpus, the relative frequency was calculated. Following Damerau (1993) and Gries (2010) I use the term *relative frequency* in the sense of relative frequency ratio, the quotient of the relative frequencies of a word in both corpora (Gries, 2010: 271-272). This means that in my analysis the relative frequency was calculated by taking the normalized frequency of the medical corpus ACAD: Medicine and dividing it by the normalized frequency of the general corpus COCA. This is illustrated in Table 1.

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 1: Relative frequency scores for five randomly selected words from ACAD: Medicine and general COCA.

Table 1 shows five randomly selected words occurring in COCA and in ACAD: Medicine with their relative frequency values. The word *odynophagia* has the highest relative frequency value in contrast to the lowest relative frequency value of the word *tonight*. These values represent the end points of a continuum ranging from highly specialized medical vocabulary items to general vocabulary items found in medical part of the corpus, but certainly not to the extent that they can be considered typical of medical texts. Relative frequency is a measure of typicality of word in the vocabulary of medical English. The values are interpreted in the following way. If the relative frequency value is close to 1, for example for *need*, it means that its frequency in general COCA and ACAD: Medicine is roughly the same. This is confirmed by very similar scores of normalized frequencies. 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 the 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 minimal 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. For instance, the word *exchangeable* shows a relative frequency value of 3.37. If we look at its normalized frequencies, in COCA it is 0.001344 and in ACAD: Medicine 0.004428. This suggests that the occurrence of *exchangeable* in both corpora is very low. In fact, the absolute frequency in ACAD: Medicine is 2 whereas in general COCA it is 49. It is obvious that such low absolute frequency values are not sufficient to arrive at relevant conclusions. This clearly indicates that medical vocabulary can be accurately described only when both frequency measures, relative frequency and absolute frequency, are taken into account.

This was performed by first sorting the frequency word list by absolute frequency in the medical corpus ACAD: Medicine. Then, I selected the range within a particular threshold, for example, 100,000; 10,000; 100; etc. and resorted this by relative frequency. I explored 13 threshold levels given in Table 2.

Threshold level	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 2: Threshold levels with word counts based on absolute frequency in ACAD: Medicine and relative frequency in ACAD: Medicine/COCA.

It is interesting to observe that the differences among the threshold levels from the perspective of word counts are remarkable. The highest threshold level (level 1) includes only 6 words whereas the lowest (level 13) covers 26,543 words. Starting from the ninth threshold level, the word counts gradually increase with a sharp increase in the final class. Taking into account that the total number of words in ACAD: Medicine is 27,166, only 623 words are distributed among the first 12 threshold values 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. Let us first turn to threshold level 1, which includes only 6 words with the highest absolute frequency in ACAD: Medicine given in Table 3.

Word (lemma)	COCA 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	1.185270
and	4	c	159,292	1.169854
a	5	a	107,607	0.836898
in	6	i	107,344	1.213356

Table 3: Threshold level 1 – absolute frequency in ACAD: Medicine and relative frequency in ACAD: Medicine/COCA.

The top six words in Table 3 representing threshold level 1 also have the highest absolute frequency in the medical corpus. It is interesting to see that at the same time, these words are the most frequent in general COCA. This is not surprising when we inspect their relative frequency values. These are in all six cases around 1.0 which suggests that their frequencies in ACAD: Medicine and general COCA are approximately the same. The observation that all words in Table 2 are function words is fully 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".

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.

Our data show that function words dominate threshold levels 1, 2, 3, 4, and 5. From threshold level 6 down, the number of content words is increasing. As mentioned above, threshold level 13 is the largest one in terms of number of words. Therefore it seems reasonable to subdivide this threshold level into sublevels, as presented in Table 4.

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 4: Threshold level 13 with sublevels based on absolute frequency in ACAD: Medicine.

Table 4 clearly shows that when threshold level 13 is divided into 10 sublevels, the word count increases with each sublevel. As it was described above, the last level is the largest one again. A characteristic feature of the thresholds 1-6 is that the absolute frequency in the subcorpus ACAD: Medicine decreases gradually with only a few words showing identical values. In sublevel 9, the frequency bands tend to be larger, they include more than ten words, e.g. 14 words with the absolute frequency value 199. With a more fine-grained division of sublevels we can demonstrate how the absolute frequency measure interacts with the relative frequency measure. Table 5 lists the words with an absolute frequency 100 in the medical corpus. The words are sorted by relative frequency.

Word (lemma)	COCA rank	PoS	Absol. freq. ACAD:Med.	Rel. freq. Med./COCA
thyroidectomy	41094	n	100	80.14028
rotavirus	43623	n	100	65.51150
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 5: Frequency band 100 in sublevel 9 (199-100) of threshold level 13 sorted by relative frequency in ACAD: Medicine/general COCA .

Table 5 illustrates that when the words in the absolute frequency band 100 are re-sorted by relative frequency, the top words are specialized medical terms. High relative frequencies indicate the 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 the words with relative frequency values less than 1 in medical vocabulary. The answer may not be straightforward. The data show that each threshold level includes vocabulary items less typical of medical texts with a very similar proportion. 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). This also suggests that it would be wrong to exclude them from the characterization of medical vocabulary altogether, because their absolute frequency shows they occur in medical texts. They may be less relevant for medicine, but they occur fairly frequently. Therefore it seems better to assume that such more general words somehow constitute a part of medical vocabulary.

Another question arises concerning the role of threshold levels. With general words the threshold levels are not helpful because they are included in all threshold levels. 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 6 shows the situation at frequency band 99.

Word (lemma)	COCA 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	6.574340
for-profit	11826	j	99	5.900292
positioning	10984	n	99	5.883301
faucet	10980	n	99	5.849610
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	1.615640
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	0.432170
else	440	r	99	0.089015

Table 6: Frequency band 99 in sublevel 10 (99-1) of threshold level 13 sorted by relative frequency in ACAD: Medicine/general COCA.

Table 6 illustrates that the frequency band 99 covering 29 items is larger than the frequency band 100

in Table 4. 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 part 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 7.

Word (lemma)	COCA 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	26.98570	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 7: An example of medical vocabulary sorted by relative frequency ACAD: Medicine/general COCA (frequency band between 26.5-27.02).

Table 7 suggests that there might be a significant 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 lower ranked items in Table 7 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 neither absolute frequency nor relative frequency alone is sufficient in delimiting medical vocabulary.

5 Conclusion

The main aim of this paper was to explore medical vocabulary on the basis of the general COCA corpus and its subcorpus ACAD: Medicine. The results suggest it is more reasonable to view medical vocabulary in English as a cline rather than a dichotomy with clear-cut boundaries. From the perspective of the characterization of medical vocabulary, absolute and relative frequency can be combined in a two-dimensional model representing medical vocabulary. 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 when words are frequent enough or not 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 frequency bands with a larger number of words it is only possible to characterize medical vocabulary accurately when relative frequency and absolute frequency are both taken into account. The data also demonstrate that all frequency bands contain words ranging from less typical of medical vocabulary to more 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. 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. For a particular absolute frequency we find a continuum of relative frequency scores. This continuum shifts when we take a lower absolute frequency. Therefore if we want to posit a threshold, we should refer to both absolute frequency and relative frequency.

6 References

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

Bauer, Laurie, Rochelle Lieber, and Ingo Plague (2013). *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

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

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

Coxhead, Averil (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.

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

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

Felber, Helmut (1984). 'Language and the professions: The role of special language in communication', *Taal en beroep*, 19:17-30.

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.

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.

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

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.

Sager, J.C., Dungworth, D., & McDonald, P. F. (1980). *English Special Languages*. Wiesbaden: Brandstetter.

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

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