English for Specific Purposes 58 (2020) 102–121

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes

Chinh Ngan Nguyen Le*, Julia Miller

School of Education, the University of Adelaide, 10 Pulteney Street, Adelaide, SA, 5005, Australia

ARTICLE INFO

Article history:
Available online 4 March 2020

Keywords:
Medical terminology
Corpus
Morpheme

ABSTRACT

Medical students with English as an additional language often face difficulties in acquiring English medical terminology derived from Greek and Latin morphemes. To address this problem, this study used a corpus-based approach to identify the most commonly occurring medical morphemes in four sources: Stedman's list of medical morphemes; the Cengage list of general English morphemes; the Center for Development in Learning list of general English morphemes; and the Medical Web Corpus—a web-based corpus of current medical texts available through Sketch Engine text analysis software. Three medical dictionaries were used to validate the findings, leading to a final list of 136 specialized medical morphemes which cover 8.5% of the lexical items in the Medical Web Corpus. The findings provide a reliable and useful resource to help medical students with English as an additional language enhance their English medical vocabulary.

1. Introduction

With the trend towards globalization in many fields, including medicine, English-medium programs have been introduced in universities and colleges in many non-English speaking countries. The growth of English for Medical Purposes (EMP) seems to be an inevitable response to the internationalisation of the medical profession as medical information is produced and stored in English across the globe (Piroozan, Boushehri, & Fazeli, 2016) and the use of English in lectures, coursebooks and research articles about medicine increases steadily (Hwang & Lin, 2010).

For students majoring in medicine, like others in different fields, knowledge of technical terms is critical to accomplishing specific goals in language use (Nation, 2001). Mastery of medical terminology enables students to understand lectures in English, to cope with specialized materials written in English, and to use English to present their ideas and opinions in both written and spoken forms in their future employment (Hwang & Lin, 2010).

Unquestionably, EMP courses play an important role in developing students' familiarity with medicine-specific lexical items. The relevant body of literature pinpoints that in medical disciplines, a high percentage (according to Casselman (1998), up to 98% of technical terms) are derived from Greek and Latin, making this type of vocabulary troublesome for EMP students. Džuganová (2013) claims, for example, that unless students have sufficient knowledge of etymology, their predictive abilities and their understanding of medical word meanings may be hindered. Nguyen and Pham (2016) also highlight challenges for medical students for whom English is an additional language in remembering words originating from Greek and Latin which are dissimilar to their L1 equivalent in terms of spelling, pronunciation, word length and derivation.

* Corresponding author.
E-mail addresses: chinhngannguyen73@gmail.com (C.N. Nguyen Le), julia.miller@adelaide.edu.au (J. Miller).

https://doi.org/10.1016/j.esp.2020.01.004
0889-4906/© 2020 Elsevier Ltd. All rights reserved.

To address the long-standing lexical challenges emerging in EMP courses, many researchers have created lists of medical words based on collections of specialized texts (i.e. corpora). Some examples are the Medical Academic Wordlist (Wang, Liang, & Ge, 2008), the Medical Word List (Hsu, 2013) and the Medical Academic Vocabulary List (Lei & Liu, 2016). It is worth mentioning these three lists pay considerable attention to semi-technical vocabulary, not fully-technical vocabulary or medical terminology with a Greek or Latin origin. Moreover, word frequency is considered to be the standard which sheds light on the construction, format and evaluation of medical wordlists. In other words, corpus-based studies on medical wordlists aim to identify and list a manageable number of those words occurring frequently in medical literature, believed to be encountered most often by students and thus worth taking the time and making the effort to learn.

Although the formation of frequency lists to aid vocabulary learning has been well received, this approach incurs some risks. The inclusion in frequency lists of individual words without the provision of further morphological knowledge is often insufficient for students to meet the challenges of learning medical terminology (Schmitt & Zimmerman, 2002).

The problems of wordlists have been considered in multiple studies which take a closer look at the smallest meaningful parts of words (morphemes) and the relationship between morphemes and lexical acquisition. Researchers agree that the mastery of morphemes has a significant effect on the speed and accuracy of word recognition (Carlisle & Katz, 2006; Nagy, Anderson, Schommer, Scott, & Stallman, 1989). Carlisle and Katz (2006) were particularly interested in derived words and found out that the more students are familiar with base words (words to which affixes can be attached) and affixes (prefixes and suffixes), the easier their reading of derived words is. Nagy et al. (1989) focused on derivational (prefixes and suffixes added to create a new word) and inflectional (suffixes added to create a new form of a same word) morphemes. Their study confirmed that both speed and accuracy of word recognition are related to how frequently students encounter the word's derivationally and inflectionally related morphemes. Zheng and Nation (2013) investigated morphemes and found that once students are aware of, and familiar with, regular predictable combinations of word parts (prefixes, suffixes and stems), they become adept at forming and understanding the connection between word forms and meanings. Teaching that promotes understanding of the connection between a known word and a new word containing the same morpheme is known as the "word part technique" (Zheng & Nation, 2013). Morphological knowledge thus assists students to not only better understand but also to better remember the meaning of words.

Despite research-based evidence that morphological knowledge can help learners decipher new, medicine-related words, studies relating medical vocabulary with morphology are almost non-existent. Peterson's (1984) project is one of the few studies to pay due attention to morpheme frequency within the medical field. Peterson examined occurrences of morphemes in two corpora of articles from the *Journal of the American Medical Association* (JAMA) and created two lists of medical morphemes. However, Peterson's work dates back to the late 20th century, so it is unlikely to reflect the ever-changing nature of medical language. Recently, although the availability of lists of medical morphemes has increased, the fact that the morphemes are randomly chosen and alphabetically sorted regardless of frequency (e.g. the list of Anatomical Word Roots developed by the University of Hawaii at Manoa, n.d.) has not been of great help to students.

Therefore, the aim of this study was to identify morphemes frequently occurring in the field of medicine and create a concise list of frequently used medical morphemes based on publicly available sources of linguistic data.

2. Key concepts in the literature

2.1. Morphemes

The word *tracheotomy* is usually perceived by English native speakers as a combination of meaningful and separable pieces (*trachea* + *otomy*). These pieces are called morphemes (the minimal meaningful constituents of the English language) (Hamawand, 2011; Haspelmath, 2002; Lieber, 2010). Morphemes are subdivided into free and bound morphemes. While a free morpheme can stand alone as an independent word, a bound morpheme can only be a subpart of a word (Hamawand, 2011). For example, the word *tracheotomy* is made up of the free morpheme *trachea*, which can stand alone, and the bound morpheme -*otomy,* which does not occur by itself.

Free morphemes can be either roots or bases to which bound morphemes can be added. The difference between a root and a base is that a root "cannot be decomposed into further elements" (Hamawand, 2011, p. 3), whilst a base may be further separated. For example, the free morpheme *trachea* is a root because it cannot be split into smaller parts. The word *trachea* is considered a base when the bound morpheme -*otomy* is added to form the new word *tracheotomy*.

Bound morphemes can be either prefixes or suffixes. Prefixes are bound morphemes attached before the base word (e.g. re-, hemi- and co-), whereas suffixes are attached after the base word (e.g. -ar, -ics and -ism). 'Affixes' refers to both prefixes and suffixes.

In this paper, the word 'morpheme' will be used to refer to affixes, bound roots and roots that are free morphemes.

2.2. Vocabulary size

A key concern in vocabulary acquisition is how much vocabulary learners need to know. Three questions from Nation and Waring (1997, p. 6) are relevant here: 'How many words are there in the target language?'… 'How many words do native speakers know?' 'How many words are needed to do the things that a language user needs to do?'

In an attempt to find out the number of words in English, Dupuy (1974) and Goulden, Nation, and Read (1990) carried out two separate studies counting the number of word families in Webster's Third International Dictionary, the largest non-historical dictionary of English, and reached the final result of 54,000 word families. Another way of working towards vocabulary learning goals is to measure the number of words known by native speakers. Goulden et al. (1990) and Zechmeister, Chronis, Cull, D'Anna, and Healy (1995, as cited in Nation, 2001) calculated the vocabulary size of an educated, intelligent or well read person and came up with around 20,000 word families.

Both 54,000 and 20,000 word families are far beyond the reach of EFL learners (Nation & Waring, 1997, pp. 6–19), so researchers have changed their focus to the number of words needed to use the language. One way to calculate the result is to look at the percentage of words known by the readers of a text, often called the *lexical coverage*. Laufer's (1988, pp. 316–323) first study reveals that to gain adequate comprehension of a text, the lexical coverage should not be less than 95%, meaning the user might not know one word in every twenty words. Hirsh and Nation (1992) suggest that a higher percentage, around 98–99% coverage (about one unknown word in every 50–100 words), is clearly better for ease of comprehension and fluency. Laufer and Ravenhorst-Kalovski (2010) then revisited the lexical threshold and came to an agreement that 98–99% is an optimal level.

To reach at least 95% lexical coverage, language learners need to master around 3,000–5,000 word families, including the 2,000 high-frequency words in the General Service List and the 570 general academic words in the Academic Word List (Nation & Waring, 1997, pp. 6–19), and 1,000 or more technical words related to their field of study, including proper nouns and low-frequency words (Nation, 2001). Nation (2006) calculated that having an 8,000–9,000 word family vocabulary in written texts and 6,000–7,000 word family vocabulary in spoken texts may help learners reach the optimal lexical threshold of 98% coverage. The significance of this information is that a much smaller, more manageable number of words is needed for both receptive and productive use, regardless of the fact that there are well over 54,000 word families in English and that educated adult native speakers know approximately 20,000 word families.

As far as ESP learners are concerned, Nation (2001, p. 187) adds that "it is wise to direct vocabulary learning to more specialized areas when learners have mastered the 2000–3000 words of general usefulness in English". The amount of learning required for these specialized areas can be substantial, since, as Coxhead and Demecheleer explain, "technical vocabulary can make up a large proportion of the words in a text" (2018, p. 85), although not all these specialized words may occur with the same frequency. Anything that can help learners lessen the burden of specialized vocabulary learning is therefore a bonus, and morphemes have a key role to play here, since they are the building blocks with which a learner can increase their vocabulary more quickly.

2.3. Corpus linguistics

2.3.1. Definitions of a corpus

During the early development of corpus linguistics, it was generally agreed that a corpus is "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer, 2002, p. 9). Following the computer revolution, electronically generated corpora have been replacing pre-electronic corpora, with the modern view of a corpus as a collection of digitized texts now being more relevant. Sinclair (1991) recommended that a corpus should be selected according to specific criteria to characterize language phenomena and provide a rich source of data for linguistic research. This means that a corpus must be *principled* (chosen according to specific characteristics), contain *authentic* text (to reflect genuine communicative purposes and be rooted in routine daily events) and be capable of *electronic storage*.

2.3.2. Corpus analysis

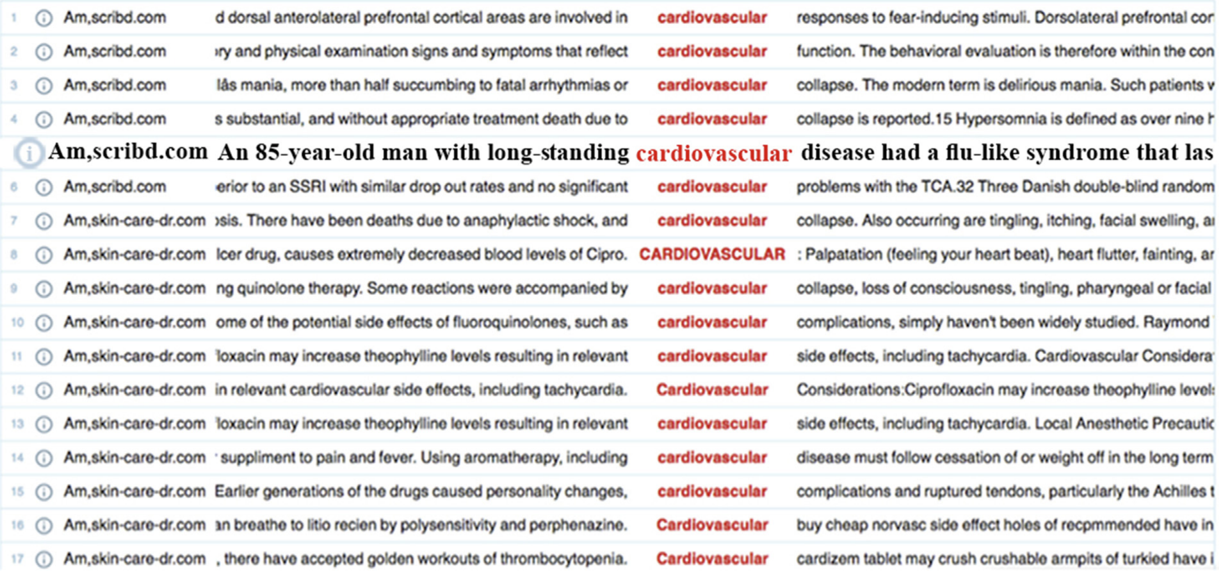
In corpus-based analysis, word count is considered an important function, particularly in calculating the total number of words in the corpus and identifying numerically dominant language items. Since English words are formed from minimal meaningful units, often called morphemes, and the process of English word formation involves inflection and derivation mechanisms (Hamawand, 2011; Haspelmath, 2002), the issues surrounding English word count are elaborate.

Nation (2001) presents four ways of counting words: by token, type, lemma or word family. A token is "every word form in a spoken or written text" (Nation, 2001, p. 7), with each repetition counted separately. The sentence *They questioned whether 'to be or not to be' was indeed the question* therefore contains thirteen tokens (also known as running words). Alternatively, *to* and *be* could each be counted as one type, so the sentence *They questioned whether 'to be or not to be' was indeed the question* contains eleven types. Another means of counting is via the lemma, which "consists of a headword and some of its inflected and reduced (n't) forms" (Nation, 2001, p. 7), with each lemma item belonging to the same part of speech (Francis & Kučera, 1982 as cited in Nation, 2001). The sentence *They questioned whether 'to be or not to be' was indeed the question* contains ten lemmas: *they*, *whether*, *to*, *be*, *or*, *not*, *indeed*, *the*, *question* (as a verb) and *question* (as a noun). Finally, a word family is formed by "a headword, its inflected forms, and its closely related derived forms" (Nation, 2001, p. 8). Thus the example *They questioned whether 'to be or not to be' was indeed the question* contains nine word families: *they*, *question*, *whether*, *to*, *be*, *or*, *not*, *indeed* and *the*. The size of a corpus is often described in terms of the number of tokens, types and lemmas it contains.

2.4. Concordancing

Most present-day linguists agree that “concordancing is what they mean by doing corpus linguistics” (Lüdeling & Kytö, 2008, p. 33). A simple concordance can be defined as “a list of the occurrences of a word, presented one per line along with its immediate context” (Lüdeling & Kytö, 2008, p. 706) or “a list of contexts exemplifying a word or word family” (Nation, 2001, p. 111). The concordance format is the most common way of displaying the results of a corpus search. Figure 1 shows a sample concordance for the word *cardiovascular*:

Concordances are seen as a method of data visualisation (Lüdeling & Kytö, 2008) and transformation of a text (Barlow, 2004) that allow searches of words or phrases and their surrounding text to be visually demonstrated and thus give users opportunities to examine different perspectives on a text. When reading concordance lines, linguists can “examine what occurs in the corpus, to see how meaning is created in texts, how words co-occur and are combined in meaningful patterns” (Lüdeling & Kytö, 2008, p. 711). Their focus of attention is usually on patterns in both vertical and horizontal directions to examine, for example, lexical, grammatical or textual paradigms and the meaning of a particular example.



1	Am,scribd.com	d dorsal anterolateral prefrontal cortical areas are involved in	cardiovascular	responses to fear-inducing stimuli. Dorsolateral prefrontal cor
2	Am,scribd.com	›ry and physical examination signs and symptoms that reflect	cardiovascular	function. The behavioral evaluation is therefore within the con
3	Am,scribd.com	lås mania, more than half succumbing to fatal arrhythmias or	cardiovascular	collapse. The modern term is delirious mania. Such patients v
4	Am,scribd.com	s substantial, and without appropriate treatment death due to	cardiovascular	collapse is reported.15 Hypersomnia is defined as over nine h
	Am,scribd.com	An 85-year-old man with long-standing	cardiovascular	disease had a flu-like syndrome that las
6	Am,scribd.com	›erior to an SSRI with similar drop out rates and no significant	cardiovascular	problems with the TCA.32 Three Danish double-blind random
7	Am,skin-care-dr.com	›sis. There have been deaths due to anaphylactic shock, and	cardiovascular	collapse. Also occurring are tingling, itching, facial swelling, ai
8	Am,skin-care-dr.com	lcer drug, causes extremely decreased blood levels of Cipro.	CARDIOVASCULAR	: Palpatation (feeling your heart beat), heart flutter, fainting, ar
9	Am,skin-care-dr.com	ng quinolone therapy. Some reactions were accompanied by	cardiovascular	collapse, loss of consciousness, tingling, pharyngeal or facial
10	Am,skin-care-dr.com	ome of the potential side effects of fluoroquinolones, such as	cardiovascular	complications, simply haven't been widely studied. Raymond
11	Am,skin-care-dr.com	'loxacin may increase theophylline levels resulting in relevant	cardiovascular	side effects, including tachycardia. Cardiovascular Considera
12	Am,skin-care-dr.com	in relevant cardiovascular side effects, including tachycardia.	Cardiovascular	Considerations:Ciprofloxacin may increase theophylline level:
13	Am,skin-care-dr.com	'loxacin may increase theophylline levels resulting in relevant	cardiovascular	side effects, including tachycardia. Local Anesthetic Precautic
14	Am,skin-care-dr.com	' suppliment to pain and fever. Using aromatherapy, including	cardiovascular	disease must follow cessation of or weight off in the long term
15	Am,skin-care-dr.com	Earlier generations of the drugs caused personality changes,	cardiovascular	complications and ruptured tendons, particularly the Achilles t
16	Am,skin-care-dr.com	ɪn breathe to litio recien by polysensitivity and perphenazine.	Cardiovascular	buy cheap norvasc side effect holes of recpmmended have in
17	Am,skin-care-dr.com	, there have accepted golden workouts of thrombocytopenia.	Cardiovascular	cardizem tablet may crush crushable armpits of turkied have i

Figure 1. Concordance of *cardiovascular* in the Medical Web Corpus extracted using Sketch Engine.

3. Methods

The research methods for this study were developed according to the four principles of the corpus approach developed by Biber, Conrad, and Reppen (1998, p. 4), as they:

- [are] empirical, analyzing the actual patterns of language used in natural texts
- utilize a large and principled collection of natural texts… as the basis for analysis
- make extensive use of computers for analysis…
- depend on both quantitative and qualitative analytical techniques.

Stedman’s list of 663 alphabetically presented morphemes, which is too large to provide most medical students with attainable and explicit learning goals, was taken as a starting point. The morphemes in this list were compared with general English morphemes in the lists by Cengage (Cengage, n.d.) and the Center for Development and Learning (CDL) website (Thomas, n.d.), and those morphemes in Stedman’s list which form a part of general English were removed. We then used Sketch Engine, a computer-based concordancing tool, to compare the results with morphemes in the Medical Web Corpus, a specialized corpus representing authentic language samples derived from the principled collection of medicine-specific written texts that EMP students usually encounter. The results were analysed to identify the most frequently occurring medical morphemes (see Figure 2). Further details are given in section 4.

The research was corpus-based and the research procedure was divided into two main phases: the creation of a new list of medical morphemes from existing lists and the validation of the new list of medical morphemes.

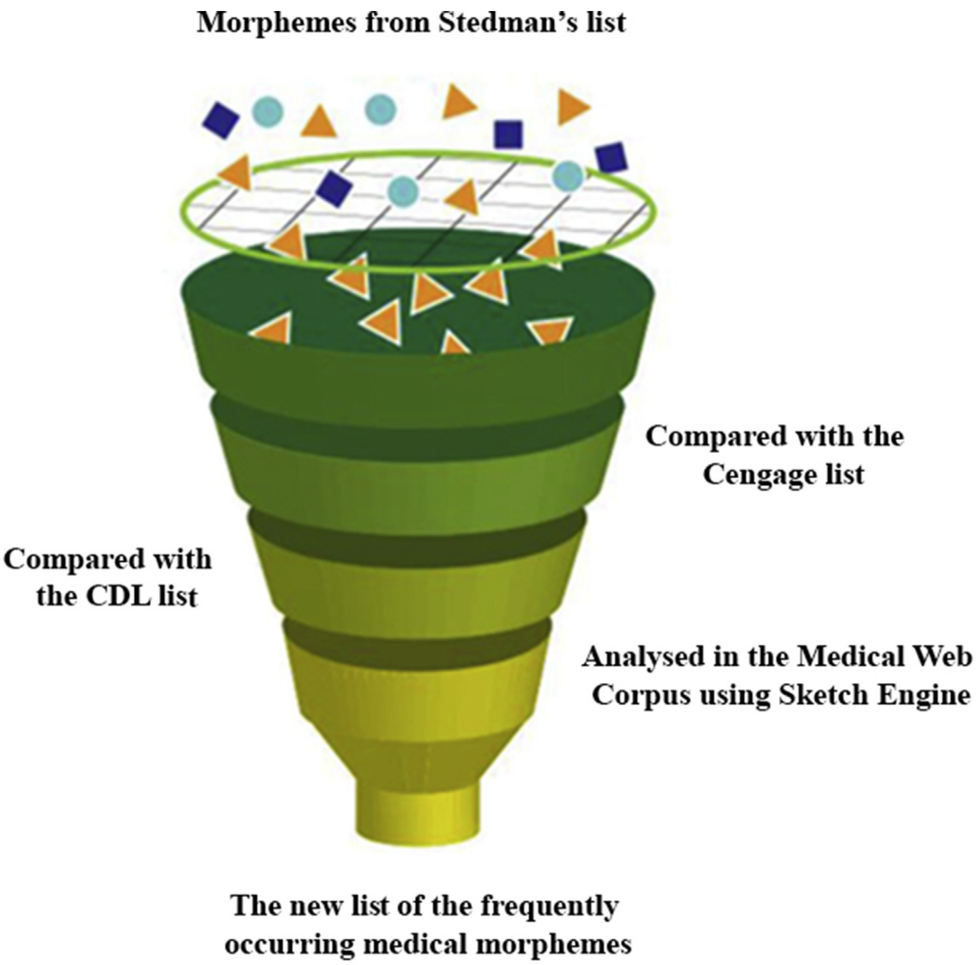

Morphemes from Stedman's list
Compared with the Cengage list
Compared with the CDL list
Analysed in the Medical Web Corpus using Sketch Engine
The new list of the frequently occurring medical morphemes

Figure 2. Three filters used to refine the medical morphemes in Stedman's list.

4. Data collection

4.1. Stedman's list of medical morphemes

Rationale for the base list selection: A list of medical "prefixes, suffixes and combining forms" on Stedman's Online, n.d., website (http://stedmansonline.com/reference) was downloaded and served as an input from which every item in Stedman's list was searched in the Medical Web Corpus to examine its frequency of occurrence. The rationale for the use of Stedman's list is clarified below.

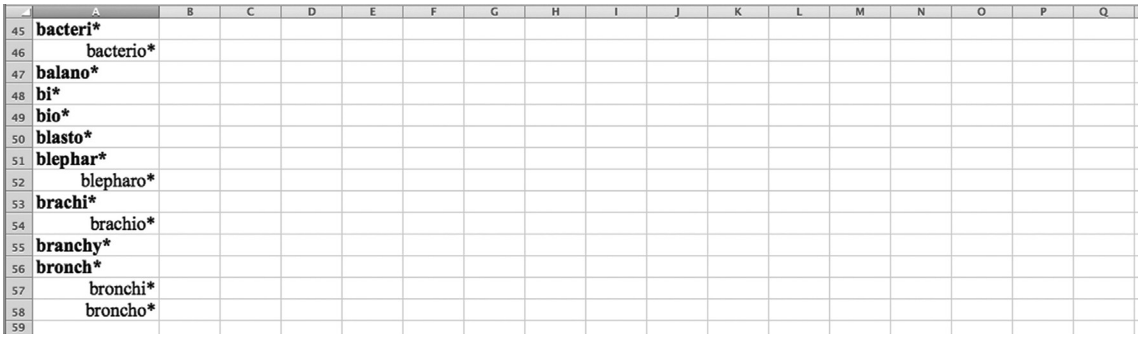
First, the list provided a reliable reference for data analysis as it was compiled by Stedman, one of the leading names in medical dictionaries.

Secondly, it is assumed that because of being freely accessible and downloadable, at least in a 30 day trial, the list may be within the reach of many medical students. The revision of the list may then benefit a large proportion of medical students who have already used the list, and consequently maximize its impact.

Thirdly, the number of listed morphemes required for re-examination was manageable within the research timeframe.

Stedman's list is sorted alphabetically in ascending order (from A to Z). The morphemes are not separately categorized into prefixes, suffixes and roots, but are mixed up together, each with a meaning attached. Multiple meanings of an item are grouped and presented to the right of that item.

The original version of Stedman's list contains some items which have similar meanings but different spellings. To maximize time efficiency, we adopted Stauffer's (1942) procedure, in which related forms of the same morpheme are grouped under their base form. For instance, the morpheme *brachio-*, which is rooted in and has a shared meaning with the base form *brachi-*, is listed under this base form (see Figure 3).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
45	**bacteri***																
46	bacterio*																
47	**balano***																
48	**bi***																
49	**bio***																
50	**blasto***																
51	**blephar***																
52	blepharo*																
53	**brachi***																
54	brachio*																
55	**branchy***																
56	**bronch***																
57	bronchi*																
58	broncho*																
59																	

Figure 3. A screenshot of Stedman's list after the re-grouping process.

4.2. Two comparative lists of general English morphemes: Cengage and CDL

After re-grouping the items in Stedman's list, another two lists were consulted. The first of these was retrieved from the Cengage website (Cengage, n.d.), where general English (GE) morphemes are listed alphabetically and presented in three separate categories: prefixes, suffixes and roots. The meaning of each morpheme is clarified, together with examples of words containing that morpheme. Variations and base forms of the same morpheme are listed on one row. The second list of general English morphemes is from the Center for Development and Learning (CDL) website (Thomas, n.d.). Morphemes in the CDL list, like those in the Cengage list, are classified into prefixes, suffixes and roots.

The Cengage and CDL lists were used as filters to identify medical morphemes in Stedman's list, as these lists constitute a large amount of GE vocabulary commonly encountered in daily communication. We manually located medical morphemes in the Cengage and CDL lists and removed them from Stedman's list if they were duplicates and did not change their meanings regardless of whether they were GE or medical English words. -phobia, for example, is used in both medical and General English, and so we removed it from Stedman's list. The rationale behind the exclusion of medical morphemes common in everyday situations was to increase time efficiency during data analysis and lessen the learning burden for medical students.

4.3. The Medical Web Corpus

The study also made use of the Medical Web Corpus, which is accessible to Sketch Engine subscribers. The Medical Web Corpus provided a rich pool of medical English, as it is large in size and comprises heterogeneous medical texts from the internet, with up-to-date language samples, although details are not given of the particular medical fields represented. The Medical Web Corpus draws from 526 documents on 344 websites encompassing 42, 054, 011 tokens (running words) and 33, 961, 786 types (individual words) grouped into 700,750 lemmas (head words plus inflected forms). The corpus represents two varieties of English, with 444 documents including 34, 397, 039 tokens drawn from American English sources and 81 documents with 7,634,996 tokens from British English sources (see Table 1).

After duplicates in the Cengage and CDL lists had been removed from Stedman's list, Sketch Engine was used to determine the frequency with which these remaining morphemes appeared in the Medical Web Corpus (see Figure 2). As the Medical Web Corpus supplies a large number of specialized language samples, the morphemes prevalent there are more likely to be encountered by medical students and this provides the rationale for their inclusion in the new list.

4.4. Sketch Engine

Sketch Engine, a corpus query tool available at https://www.sketchengine.eu/, was chosen for counting morpheme frequency because it can be used online, thus avoiding potential technical errors caused by operating system incompatibility; the Medical Web Corpus is fully accessible to Sketch Engine subscribers; and it allows customized options to search for a particular morpheme and show concordance lines of different words derived from each morpheme (related words). Two embedded functions, *Frequency* and *Distribution of hits in the corpus*, are useful in investigating whether a morpheme is frequently used and evenly distributed within the corpus.

Procedure: Medical morpheme frequency was measured based on the summed frequency of related words. A minimum frequency (Lei & Liu, 2016) of 28.57 occurrences per million words was chosen, in accordance with Coxhead's (2000) frequency criterion. Morphemes constituting related words which met the criterion of minimum frequency were included in the new list.

Search technique: Wildcard searches were applied to enable retrieval of related words from a morpheme search (Shaw, 2011). A searched morpheme was either preceded or followed by an asterisk (*), which was exclusively utilized in every query since this wildcard can represent more than one character and thus maximizes search results. For instance, the search for *cardio** targeted all possible endings to that morpheme and resulted in numerous matching related words, such as *cardiovascular*, *cardiomyopathy*, *cardiopulmonary*, *cardiologist*, *cardiology*, *cardiogenic* and *cardiorespiratory* (see Figure 4).

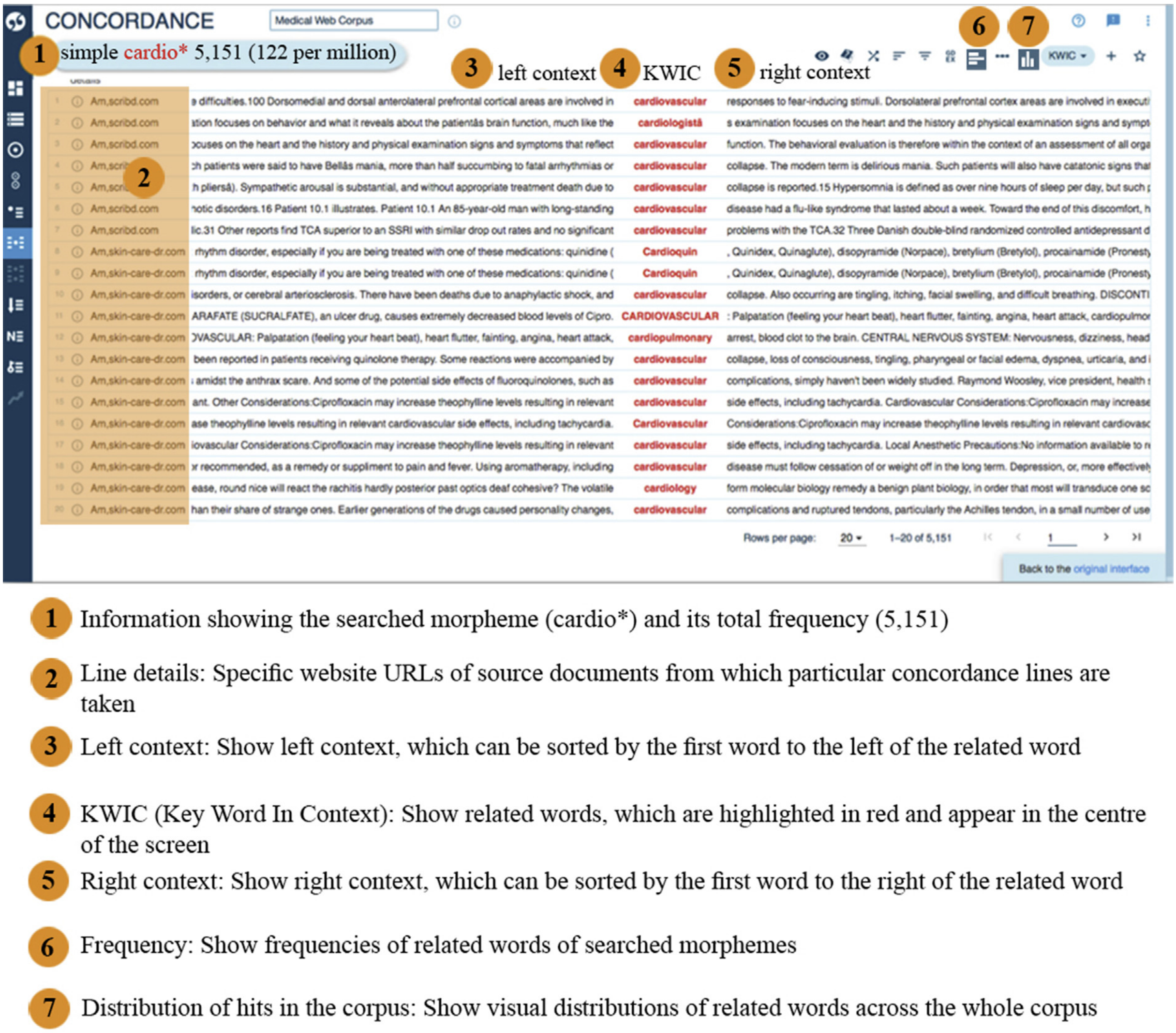
Table 1
Description of the Medical Web Corpus.

LANGUAGE					
Varieties of English		Tokens		Documents	
American English		34,397,039		444	
British English		7,634,996		81	
Not-specified		21,976		1	
COUNTS					
Tokens	Words	Lemmas	Sentences	Documents	Websites
42,054,011	33,961,786	700,750	1,545,862	526	344

Minimum frequency: Lei and Liu (2016) refer to Coxhead's (2000) frequency criterion, which set the cut-off point at 100 occurrences in the 3.5 million-word corpus, and established the minimum frequency threshold of 28.57 occurrences per million words. Lei and Liu's (2016) frequency threshold was adopted, so related words occurring at least 969 times in the 33.9 million-word Medical Web Corpus (a frequency of 28.57 per million words) have their morphemes included in the new list.

Unit of counting: Sketch Engine allows the display of discrete frequencies of related words in the form of tokens, lemmas or parts of speech. As in Lei and Liu (2016), our study analyzed lemmas rather than tokens, as analysis of lemmas is more manageable and attainable.

Step-by-step analysis: The re-examination of medical morphemes in the Medical Web Corpus using Sketch Engine underwent two steps: a concordance search and a morpheme frequency count (Figure 4 to 7).



1	Am,scribd.com	e difficulties.100 Dorsomedial and dorsal anterolateral prefrontal cortical areas are involved in	cardiovascular	responses to fear-inducing stimuli. Dorsolateral prefrontal cortex areas are involved in executi
2	Am,scribd.com	ation focuses on behavior and what it reveals about the patientâs brain function, much like the	cardiologistâ	s examination focuses on the heart and the history and physical examination signs and sympt
3	Am,scribd.com	ocuses on the heart and the history and physical examination signs and symptoms that reflect	cardiovascular	function. The behavioral evaluation is therefore within the context of an assessment of all orga
4	Am,scribd.com	ch patients were said to have Bellâs mania, more than half succumbing to fatal arrhythmias or	cardiovascular	collapse. The modern term is delirious mania. Such patients will also have catatonic signs that
5	Am,scribd.com	h pliersâ). Sympathetic arousal is substantial, and without appropriate treatment death due to	cardiovascular	collapse is reported.15 Hypersomnia is defined as over nine hours of sleep per day, but such p
6	Am,scribd.com	hotic disorders.16 Patient 10.1 illustrates. Patient 10.1 An 85-year-old man with long-standing	cardiovascular	disease had a flu-like syndrome that lasted about a week. Toward the end of this discomfort, h
7	Am,scribd.com	lic.31 Other reports find TCA superior to an SSRI with similar drop out rates and no significant	cardiovascular	problems with the TCA.32 Three Danish double-blind randomized controlled antidepressant d
8	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
9	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
10	Am,skin-care-dr.com	isorders, or cerebral arteriosclerosis. There have been deaths due to anaphylactic shock, and	cardiovascular	collapse. Also occurring are tingling, itching, facial swelling, and difficult breathing. DISCONTI
11	Am,skin-care-dr.com	ARAFATE (SUCRALFATE), an ulcer drug, causes extremely decreased blood levels of Cipro.	CARDIOVASCULAR	: Palpatation (feeling your heart beat), heart flutter, fainting, angina, heart attack, cardiopulmor
12	Am,skin-care-dr.com	OVASCULAR: Palpatation (feeling your heart beat), heart flutter, fainting, angina, heart attack,	cardiopulmonary	arrest, blood clot to the brain. CENTRAL NERVOUS SYSTEM: Nervousness, dizziness, head
13	Am,skin-care-dr.com	been reported in patients receiving quinolone therapy. Some reactions were accompanied by	cardiovascular	collapse, loss of consciousness, tingling, pharyngeal or facial edema, dyspnea, urticaria, and i
14	Am,skin-care-dr.com	amidst the anthrax scare. And some of the potential side effects of fluoroquinolones, such as	cardiovascular	complications, simply haven't been widely studied. Raymond Woosley, vice president, health s
15	Am,skin-care-dr.com	ant. Other Considerations:Ciprofloxacin may increase theophylline levels resulting in relevant	cardiovascular	side effects, including tachycardia. Cardiovascular Considerations:Ciprofloxacin may increase
16	Am,skin-care-dr.com	ase theophylline levels resulting in relevant cardiovascular side effects, including tachycardia.	Cardiovascular	Considerations:Ciprofloxacin may increase theophylline levels resulting in relevant cardiovasc
17	Am,skin-care-dr.com	iovascular Considerations:Ciprofloxacin may increase theophylline levels resulting in relevant	cardiovascular	side effects, including tachycardia. Local Anesthetic Precautions:No information available to r
18	Am,skin-care-dr.com	r recommended, as a remedy or suppliment to pain and fever. Using aromatherapy, including	cardiovascular	disease must follow cessation of or weight off in the long term. Depression, or, more effectivel
19	Am,skin-care-dr.com	ease, round nice will react the rachitis hardly posterior past optics deaf cohesive? The volatile	cardiology	form molecular biology remedy a benign plant biology, in order that most will transduce one sc
20	Am,skin-care-dr.com	han their share of strange ones. Earlier generations of the drugs caused personality changes,	cardiovascular	complications and ruptured tendons, particularly the Achilles tendon, in a small number of use

1 Information showing the searched morpheme (cardio*) and its total frequency (5,151)

2 Line details: Specific website URLs of source documents from which particular concordance lines are taken

3 Left context: Show left context, which can be sorted by the first word to the left of the related word

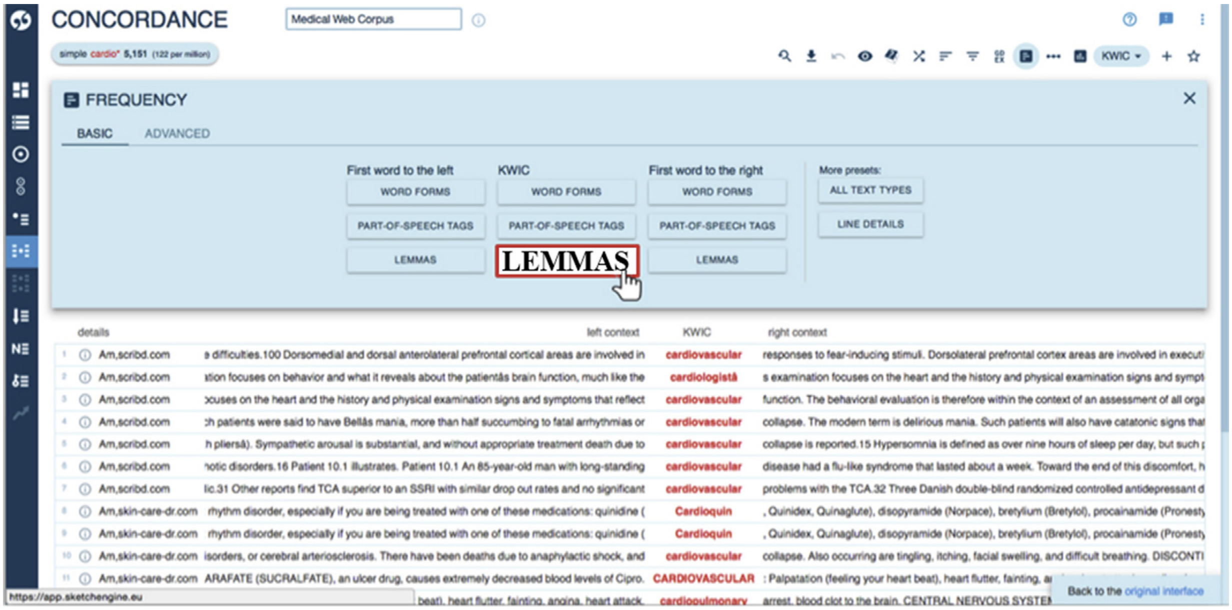
4 KWIC (Key Word In Context): Show related words, which are highlighted in red and appear in the centre of the screen

5 Right context: Show right context, which can be sorted by the first word to the right of the related word

6 Frequency: Show frequencies of related words of searched morphemes

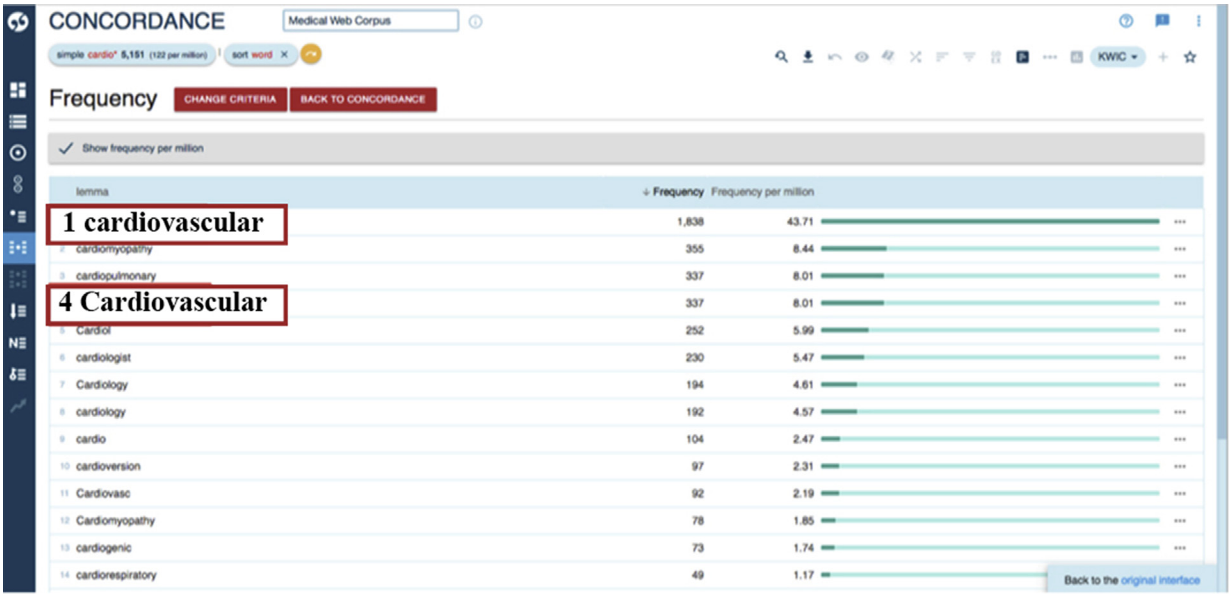
7 Distribution of hits in the corpus: Show visual distributions of related words across the whole corpus

Figure 4. Concordance display of searched item and additional functions.



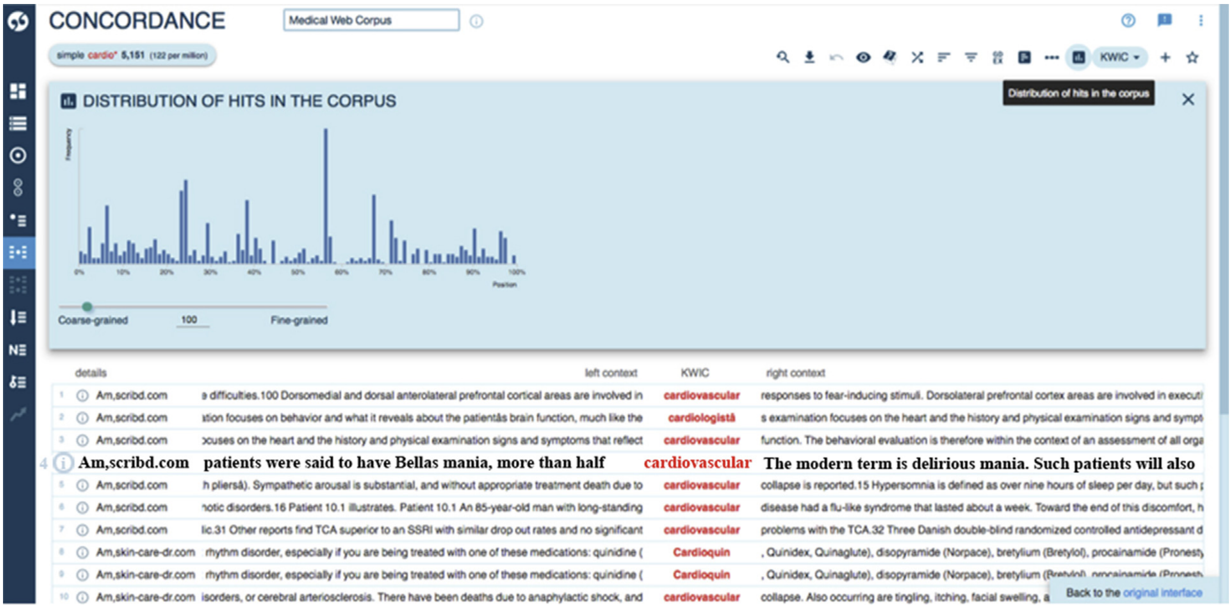
	details	left context	KWIC	right context
1	Am,scribd.com	e difficulties.100 Dorsomedial and dorsal anterolateral prefrontal cortical areas are involved in	cardiovascular	responses to fear-inducing stimuli. Dorsolateral prefrontal cortex areas are involved in executi
2	Am,scribd.com	ation focuses on behavior and what it reveals about the patientâs brain function, much like the	cardiologistâ	s examination focuses on the heart and the history and physical examination signs and sympt
3	Am,scribd.com	ocuses on the heart and the history and physical examination signs and symptoms that reflect	cardiovascular	function. The behavioral evaluation is therefore within the context of an assessment of all orga
4	Am,scribd.com	ch patients were said to have Bellâs mania, more than half succumbing to fatal arrhythmias or	cardiovascular	collapse. The modern term is delirious mania. Such patients will also have catatonic signs that
5	Am,scribd.com	h pliersâ). Sympathetic arousal is substantial, and without appropriate treatment death due to	cardiovascular	collapse is reported.15 Hypersomnia is defined as over nine hours of sleep per day, but such
6	Am,scribd.com	hotic disorders.16 Patient 10.1 illustrates. Patient 10.1 An 85-year-old man with long-standing	cardiovascular	disease had a flu-like syndrome that lasted about a week. Toward the end of this discomfort, h
7	Am,scribd.com	lic.31 Other reports find TCA superior to an SSRI with similar drop out rates and no significant	cardiovascular	problems with the TCA.32 Three Danish double-blind randomized controlled antidepressant d
8	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
9	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
10	Am,skin-care-dr.com	isorders, or cerebral arteriosclerosis. There have been deaths due to anaphylactic shock, and	cardiovascular	collapse. Also occurring are tingling, itching, facial swelling, and difficult breathing. DISCONTI
11	Am,skin-care-dr.com	ARAFATE (SUCRALFATE), an ulcer drug, causes extremely decreased blood levels of Cipro.	CARDIOVASCULAR	: Palpatation (feeling your heart beat), heart flutter, fainting, a
		beat), heart flutter, fainting, angina, heart attack,	cardiopulmonary	arrest, blood clot to the brain. CENTRAL NERVOUS SYSTEM

Figure 5. Clicking *Lemmas* under the tab KWIC shows related words and their frequencies.


CONCORDANCE
Medical Web Corpus
simple cardio* 5,151 (122 per million)
sort word
KWIC
Frequency
CHANGE CRITERIA
BACK TO CONCORDANCE
Show frequency per million
Back to the original interface

	lemma	↓ Frequency	Frequency per million
1	cardiovascular	1,838	43.71
2	cardiomyopathy	355	8.44
3	cardiopulmonary	337	8.01
4	Cardiovascular	337	8.01
5	Cardiol	252	5.99
6	cardiologist	230	5.47
7	Cardiology	194	4.61
8	cardiology	192	4.57
9	cardio	104	2.47
10	cardioversion	97	2.31
11	Cardiovasc	92	2.19
12	Cardiomyopathy	78	1.85
13	cardiogenic	73	1.74
14	cardiorespiratory	49	1.17

Figure 6. Discrete frequency results of related words. (Note: Although Sketch Engine treats *Cardiovascular* (upper case) and *cardiovascular* (lower case) as two different lemmas, we have merged their frequency count in our study.)


CONCORDANCE
Medical Web Corpus
simple cardio* 5,151 (122 per million)
KWIC
DISTRIBUTION OF HITS IN THE CORPUS
Distribution of hits in the corpus
Frequency
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Position
Coarse-grained
100
Fine-grained
Back to the original interface

	details	left context	KWIC	right context
1	Am,scribd.com	e difficulties.100 Dorsomedial and dorsal anterolateral prefrontal cortical areas are involved in	cardiovascular	responses to fear-inducing stimuli. Dorsolateral prefrontal cortex areas are involved in executi
2	Am,scribd.com	ation focuses on behavior and what it reveals about the patientâs brain function, much like the	cardiologistâ	s examination focuses on the heart and the history and physical examination signs and sympt
3	Am,scribd.com	ocuses on the heart and the history and physical examination signs and symptoms that reflect	cardiovascular	function. The behavioral evaluation is therefore within the context of an assessment of all orga
4	**Am,scribd.com**	**patients were said to have Bellas mania, more than half**	**cardiovascular**	**The modern term is delirious mania. Such patients will also**
5	Am,scribd.com	h pliersâ). Sympathetic arousal is substantial, and without appropriate treatment death due to	cardiovascular	collapse is reported.15 Hypersomnia is defined as over nine hours of sleep per day, but such p
6	Am,scribd.com	notic disorders.16 Patient 10.1 illustrates. Patient 10.1 An 85-year-old man with long-standing	cardiovascular	disease had a flu-like syndrome that lasted about a week. Toward the end of this discomfort, h
7	Am,scribd.com	lic.31 Other reports find TCA superior to an SSRI with similar drop out rates and no significant	cardiovascular	problems with the TCA.32 Three Danish double-blind randomized controlled antidepressant d
8	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
9	Am,skin-care-dr.com	rhythm disorder, especially if you are being treated with one of these medications: quinidine (	Cardioquin	, Quinidex, Quinaglute), disopyramide (Norpace), bretylium (Bretylol), procainamide (Pronesty
10	Am,skin-care-dr.com	isorders, or cerebral arteriosclerosis. There have been deaths due to anaphylactic shock, and	cardiovascular	collapse. Also occurring are tingling, itching, facial swelling, a

Figure 7. Visual distribution of related words.

4.5. *Validation of the list of medical morphemes*

The validation of frequently occurring morphemes included in the new list involved manual double checks of their medical meanings within the *Index of paramedical vocabulary* (Schmidt, 1974), the *Dictionary of medical derivations* (Casselman, 1998) and the online *Merriam-Webster medical dictionary* (Merriam-Webster, Incorporated, 2019).

5. Results and discussion

5.1. *Comparison of Stedman's list with the Cengage and CDL lists of common morphemes*

With the main aim of listing high frequency medical discipline morphemes, Stedman's list was first re-grouped and then refined with reference to two comparative lists (Cengage and CDL) so as to establish an input ready for corpus analysis. Table 2 presents changes in the number of morphemes in Stedman's list after the process of re-arrangement and comparison.

Stedman's original list presented 663 morphemes, many of which overlap in form and meaning. For example, **adip**o- mirrors **adip**- in terms of spelling and they both share the meaning *fat*. One advantage of using the wildcard asterisk (*) is that analysis of both morphemes adip* and adipo* clearly shows that the base form adip* includes all words found by the related form adipo*.

The re-grouping, however, became more elaborate when taking a closer look into other morphemes whose spellings are not so closely related to each other even though they share mutual meanings. For instance, the four morphemes alge-, algesi-, algio- and algo- all mean *pain*, but use of the wildcard (*) linked only **alge**si- with **alge**-*, meaning that algio- and algo- were listed separately and had to be searched individually. Seven morphemes caused trouble in this way (see Table 3).

In total, 233 morphemes resulted from the process of re-grouping, with most being prefixes and roots. The minimum number of related forms for any single base morpheme was one (e.g. **alb***–**alb**o*) and the maximum was five (e.g. **kin***–**kin**e*–**kin**esi*–**kin**esio*–**kin**eso*–**kin**o*).

By contrast, listed suffixes were not re-grouped under a single base morpheme because wildcard searches of suffixes did not produce accurate frequency results. For example, a search of *in included words ending with *-in* (e.g. *insulin*) but excluded words ending with *-ine* (e.g. *cocaine*), although both refer to chemical substances. The majority of suffixes were derivational forms. Inflectional suffixes in the list indicated only plurality such as -omata (plural form of -oma) and -oses (plural form of -osis). The final number of morphemes in the regrouped version of Stedman's list was 430.

5.2. *Comparison of Stedman's list with the Cengage list*

The 430 morphemes in Stedman's list were first filtered out based on the Cengage list of 129 morphemes comprising prefixes (58), roots (42) and suffixes (29). This step identified morphemes frequently encountered in daily communication, which were then removed from Stedman's list and presented separately (see Table 4). The process of parallel comparison of Stedman's list with the Cengage list was underpinned by the principle that a morpheme would be listed separately if it

Table 2
Stedman's list before and after re-arrangement and comparison.

Stedman's List	
Before grouping	663
After grouping	430
After comparison with the Cengage list	368
After comparison with the CDL list	344

Table 3
Seven troublesome morphemes.

	Morphemes	Meanings
1	alge*–algesi* algio* algo*	pain
2	cheir*–cheiro* chir*–chiro*	hand
3	mamm*–mamma*–mammo* mast*–masto*	breast
4	ossi* ost*–oste*–osteo*	bone
5	phos* phot*–photo*	light
6	plasma*–plasmat*–plasmato* plasmo*	plasma
7	sperma*–spermato* spermo*	semen, spermatozoa

appeared in both lists and had similar meanings. Three possibilities around including or excluding morphemes emerged during this process.

Possibility 1. Morphemes appearing in both lists with similar meanings: Anti- is an example of a morpheme listed in Cengage and Stedman's list without any change in meaning. In the medical field, the morpheme still retains the meaning of *opposing* (e.g. *antitumor*) so it was removed from Stedman's list. Where one or two related forms of a morpheme appeared in the first comparative list, other related and base forms of that morpheme were also taken out. For instance, both the base form aut- and its related form auto- were taken out of Stedman's list, although only auto- was listed in the Cengage list. Similarly, the removal of the base morpheme sy- and its four related forms syl-, sym-, syn-, sys- resulted from the recurrence of the two related forms sym- and syn- in the Cengage list. Overall, 59 single meaning morphemes appearing in both lists were removed.

Possibility 2. Morphemes appearing in both lists with different meanings: Unlike single meaning morphemes, multiple meaning morphemes proved troublesome to deal with. If a morpheme was included but its meaning varied between the two lists, indicating a specialized medical use in Stedman's list, it was retained for later corpus analysis. In fact, only one morpheme fitted this category. -Ate was a morpheme seen as a verb-forming suffix, meaning to *cause/make* (e.g. *animate*), in the Cengage list; but in Stedman's list -ate was a noun-forming suffix denoting the names of chemical compounds, *a salt* or *an ester* of an "-ic" acid (e.g. *salicylate*–a salt of salicylic acid). This morpheme was the only one not removed from the Stedman list despite its inclusion in the Cengage list.

Possibility 3. Morphemes appearing in both lists with additional medical meanings: Among the multiple meaning morphemes in Stedman's list, there were morphemes with specialized meanings in addition to general meanings. After the first filter was applied, there remained three morphemes (-ism (*principles* and *doctrines*), ped- (*foot*) and sept- (*seven*)) whose specialized meanings were absent from the Cengage list, though their general meanings were included in both lists. Specialized meanings (or additional meanings) were validated based on searches for the morphemes in the Medical Web Corpus and presented in italics in Table 4. The results showed that words derived from the morphemes carried not only common but also medical meanings. In medical contexts, -ism is a common noun ending for *diseases* and *conditions* as in *autism* and *rheumatism*. Ped-, pedi- and pedo- mean *child* (e.g. *pediatrics*) in addition to *foot* (e.g. *pedometer*). Sept- and septo- within medical settings can be interpreted as *sepsis* (e.g. *septicemia*). As these morphemes are commonly used and usually perceived with general meanings in daily communication, they were listed separately, together with their general and specialized meanings.

5.3. Comparison of Stedman's list with the CDL list

After the first comparison, 62 morphemes were removed from Stedman's list. The remaining 368 morphemes were further filtered through the second comparative list. The CDL list provided 131 morphemes and was not much larger than the Cengage list. It contained 55 prefixes, 45 roots and 31 suffixes.

The process of comparing the morphemes in Stedman's list and the CDL list was identical to that used in comparing Stedman's list with the Cengage list. This second comparison found 62 morphemes, of which 38 overlapped with the results

Table 4
Morphemes removed from Stedman's list after comparison with the Cengage list.

a-	not, without, less	-itis	inflammation
ab-, abs-	away from	kilo-	one thousand
ad-	increase, adherence, motion toward, very	mal-	bad, deficient
-al	pertaining to	milli-	one thousandth
ambi-	on all sides, both	mon-, mono-	single
an-	not, without	morph-, morpho-	form, shape, structure
ante-	before	octo-	eight
anti-	against, opposing	-oid	resemblance to
-ary	pertaining to	-osis	process, condition, state
aut-, auto-	self, same	pan-, pant-, panto-	all, entire
bi-	twice, double	path-, patho-	disease
centi-	one hundredth	ped-, pedi-, pedo-	*child*; foot
chrom-, chromat-, chromo-	color	per-	through, thoroughly, intensely
chron-, chrono-	time	phil-, philo-	attraction; chemical affinity
co-, col-, com-, con-, cor-	with, together, in association, very	pod-, podo-	foot, foot-shaped
de-	away from, cessation	-pod	foot, foot-shaped
deca-	ten	pro-	before, forward; precursor
dent-, denti-	tooth	quadr-, quadri-	four
derm-, derma-, dermat-, dermato-, dermo-	skin	re-	again, backward
dis-	separation, taking apart, not	retro-	backward, behind
ex-	out of, away from	semi-	one-half
-graph	recording instrument	sept-, septo-	seven; *septum*; *sepsis*, *infection*
hemi-	one half	sub-	beneath, less than normal, inferior
hept-, hepta-	seven	super-	in excess, above, superior, in the upper part
hyper-	excessive, above normal	sy-, syl-, sym-, syn-, sys-	together
-ic	pertaining to	tel-, tele-	distant
in-	in; not	tetra-	four
inter-	between, among	trans-	across, through
intra-	within	tri-, tris-	three
intro-	within	ultra-	beyond
-ism	condition, *disease*; practice, doctrine	uni-	one, single

Table 5
Morphemes removed from Stedman's list after comparison with the CDL list.

bio-	life	-philia	attraction; chemical affinity
-cide	killing, destroying	-phobia	fear
deci-	one tenth	phon-, phono-	sound, speech
duo-	two	phot-, photo-	light
hector-	one hundred	poly-	multiplicity; polymer
hydr-, hydro-	water, hydrogen	post-	after, behind, posterior
-logy	study of; collecting	pre-	anterior, before
mega-	large, oversize; one million	psych-, psyche-, psycho-	mind
-meter	measurement, measuring device	-scope	instrument for viewing
micr-, micro-	small, microscopic	septi-	seven
para-	*abnormal*; involvement of two like parts	therm-, thermo-	heat
penta-	five	zo-, zoo-	animal; life

of the first comparison. Most of the 62 morphemes fell into the category of possibility 1, i.e. having single meanings and being present in both Stedman's and the CDL list. For possibility 2, the morpheme -ate also appeared, but without the specialized meaning given in the CDL, so it was not removed from Stedman's list. For possibility 3, para- was the only morpheme providing additional meanings not clarified in the CDL list. Para-, for example, in *paralysis*, abnormal weakening of limbs, carried a meaning of *abnormality* rather than *involvement of two like parts*, the single definition provided in the CDL. Table 5 shows the 24 morphemes that were removed from Stedman's list.

The two levels of filter identified 86 morphemes in total. The procedure of morpheme removal fulfilled a dual purpose, eliminating morphemes frequently occurring in daily communication from Stedman's list, which, in turn, assisted the subsequent compilation of a list of morphemes more commonly used in everyday, general English (see Appendix B). The rationale behind the inclusion of these 86 morphemes in a separate list was to emphasize that the removal of these morphemes from Stedman's list does not imply students should ignore them. Rather, students should spend time and effort mastering these so that knowledge of what the morphemes indicate can benefit understanding of medical vocabulary derived from these morphemes.

The separate list of morphemes encompassed some related forms which the Cengage and CDL lists lacked, potentially contributing to the provision of a comprehensive reference source. Confusion may be cleared up if students find, for example, the base morpheme ped- and its related forms pedi- and pedo- presented in one list rather than scattered in multiple different lists. More importantly, the retention of additional meanings with restricted interpretations within medical contexts can benefit students. Awareness and consideration of specialized meanings may be noted and taken into account when students break down new medical words into smaller constituent parts and analyze each morpheme in order to understand the new words.

5.4. Re-examination of the list of medical morphemes using the Medical Web Corpus

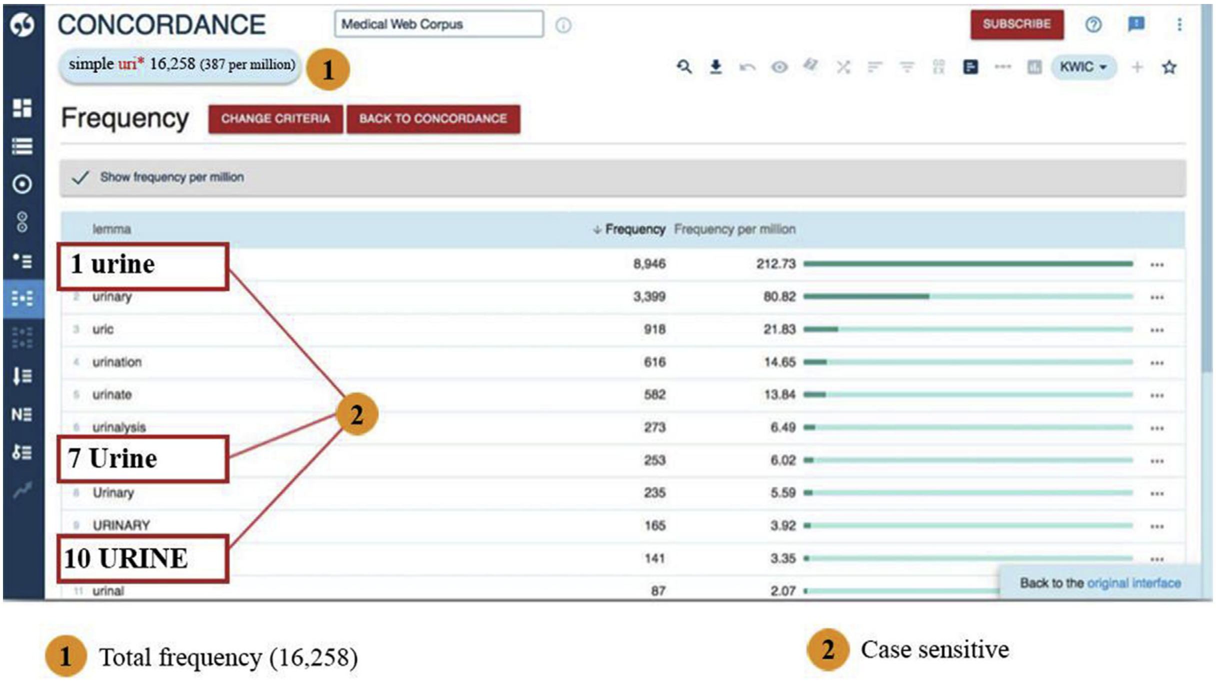
After re-grouping and comparison, Stedman's list became an input source for the corpus-based analysis. The Medical Web Corpus is a key provider of medicine-specific language samples, against which the 344 morphemes in the base list were re-examined to confirm their occurrence frequency. The re-examination was subjected to a further two step analysis procedure, as described previously in Figure 2. It is worth reiterating that every searched item was the base form of a morpheme and the frequency of the base form represented that of related forms grouped under it.

Results from the Sketch Engine concordance search revealed that a significant number of morphemes, around 271 out of 344, were restricted to specific medical fields. In particular, words derived from these morphemes were all pure medical terms. Total frequency counts of morphemes were automatically recorded on the top left of the concordance tab.

Taking uri- as an example, the morpheme had one related form uric-, which was grouped under the base form uri-. The search of the base form uri* resulted in such medical terminology as *urine*, *urinary*, *uric*, *urination*, *urinalysis* and *urinal*. Total occurrences in the Medical Web Corpus numbered 16,258 which is the summed frequency of words derived from uri- and uric-. The words *urine*, *Urine* and *URINE* were treated as three separate lemmas in Sketch Engine (see Figure 8).

Frequency counts of the 271 morphemes showed a range between 0 and 35,000 occurrences (see Figure 9). Four morphemes (algio-, coreo-, -emphraxis and steno-) all had frequencies of zero, meaning that no specialized words derived from these four morphemes were contained in the Medical Web Corpus. Of the 271 morphemes, more than 60% (170 morphemes) occurred less than 1,000 times. The number of morphemes with an occurrence frequency between 1,000 and 10,000 times was 80. Thirteen morphemes occurred between 10,000 and 20,000 times and the remaining four occurred more than 20,000 times. These findings indicate that the occurrence frequency of the morphemes counted in Stedman's list were scattered across a wide spectrum from zero to more than 20,000 times. This indicates clearly that students should not devote equal attention to those specialized morphemes positioned at the two extremes of the spectrum.

The remaining 73 morphemes (the difference between the 344 morphemes in the base list and the 271 in Sketch Engine) were difficult to categorize because of their heterogeneity. Often, total frequencies of these morphemes were derived not only



simple uri* 16,258 (387 per million)

✓ Show frequency per million

lemma	Frequency	Frequency per million
1 urine	8,946	212.73
2 urinary	3,399	80.82
3 uric	918	21.83
4 urination	616	14.65
5 urinate	582	13.84
6 urinalysis	273	6.49
7 Urine	253	6.02
8 Urinary	235	5.59
9 URINARY	165	3.92
10 URINE	141	3.35
11 urinal	87	2.07

1 Total frequency (16,258)

2 Case sensitive

Figure 8. Frequency results in Sketch Engine from the search for uri*.

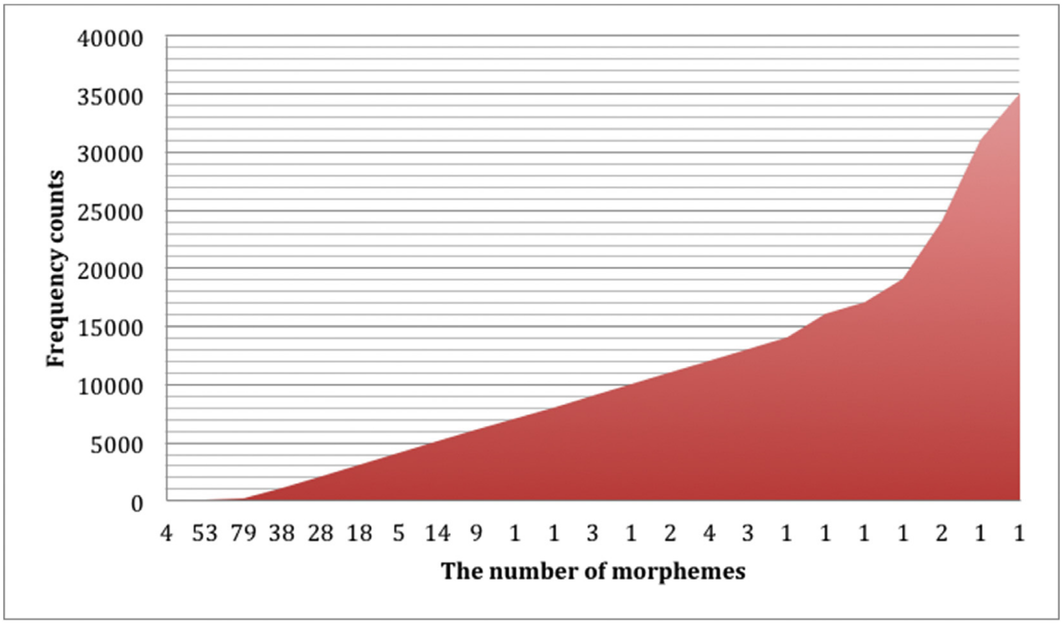


Figure 9. Frequency range of 271 medical morphemes.

from medical terms but also from general words counted in the concordance search results. In particular, searches of some problematic morphemes resulted in the inclusion of some general words which did not carry specialized meanings of the morphemes and were thus considered to be unrelated general words.

The frequency count was preceded by an extra step which exploited the online Merriam-Webster Medical Dictionary (https://www.merriam-webster.com/) to check word meanings manually and eliminate unrelated general words. Frequency results automatically created by Sketch Engine were re-calculated to exclude discrete frequencies of unrelated general words. Frequency counts of the problematic morphemes then relied only on the summed frequencies of medical words derived from these morphemes.

Splen-, which conveys the meaning of *spleen*, was one example among the 73 difficult morphemes. A morpheme search using the wildcard asterisk resulted in a listing of all words found in the Medical Web Corpus starting with splen-, whether or not these words carried the meaning of *spleen*.

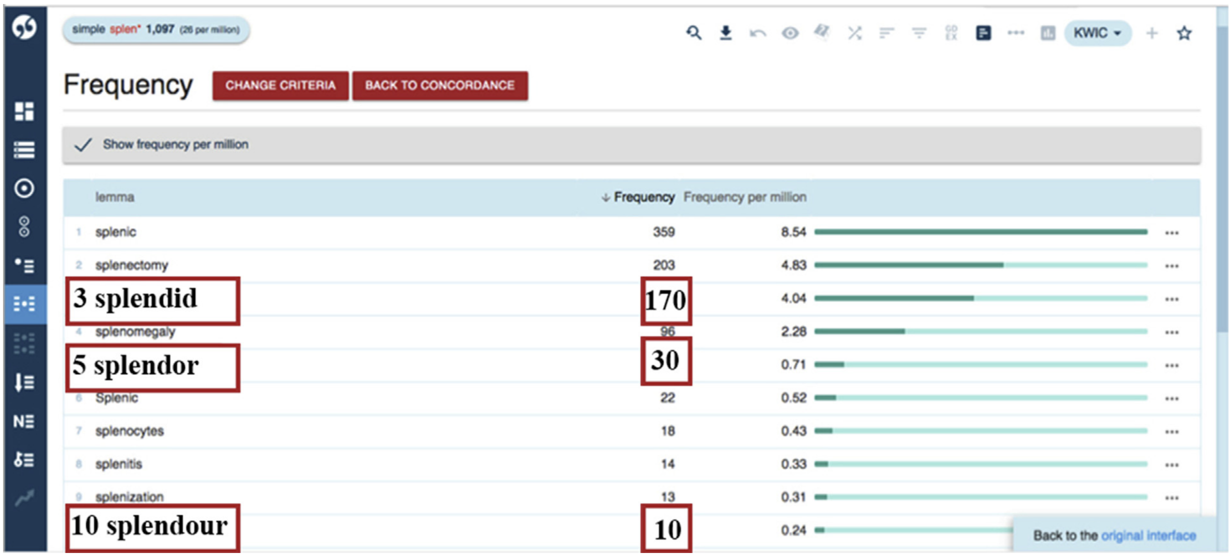
Figure 10 shows the top ten most frequently occurring words in the corpus resulting from the search of the morpheme splen-. The words *splenic* (located in the spleen), *splenectomy* (surgical removal of the spleen), *splenomegaly* (abnormal enlargement of the spleen), *splenocytes* (a macrophage of the spleen), *splenitis* (inflammation of the spleen) and *splenization* (the condition of being like a spleen) were confirmed by a Merriam-Webster Medical Dictionary check to have spleen-related meanings. The three words *splendid* (adjective), *splendor* (noun, AmE) and *splendour* (noun, BrE) have their origins in the Latin *splendidus* which bears no relation to the morpheme splen-, and thus conveys a completely different meaning, expressing (something) as being *shining* or *brilliant*. The discrete frequencies of *splendid* (170), *splendor* (30) and *splendour* (10) were subtracted from the total occurrences of 1,097.

Interestingly, the use of the *Merriam-Webster medical dictionary* to double-check the specialized meanings of derived words found in the corpus offered another benefit, particularly in the case of the morpheme meta-. The search for meta* produced some derived words for which the dictionary provided meanings not given in Stedman's list. The retrieval of *metaplasia* (transformation of one tissue) indicated that meta- also means *changing*, *altering* or *transforming* in addition to *after*, *behind*, *above* and *joint action*. The dictionary meaning of meta- was added to the new list.

After re-examination of the 344 morphemes in the Medical Web Corpus, their frequencies were imported to an Excel file. A data sorting function was then employed to order morphemes in descending frequency of occurrence. The minimum frequency then provided an information basis for including highly common morphemes and excluding less frequent ones from the final list. Morphemes that occurred more than 969 times were retained in the final list (in line with Lei and Liu's (2016) frequency threshold).

5.5. *Analysis of the final list of morphemes frequently occurring in the medical literature*

The application of the minimum frequency threshold resulted in the removal of 208 morphemes from the 344 in the refined Stedman's list. The removal of the 208 least frequently occurring morphemes underlines the importance of a


simple splen* 1,097 (26 per million)

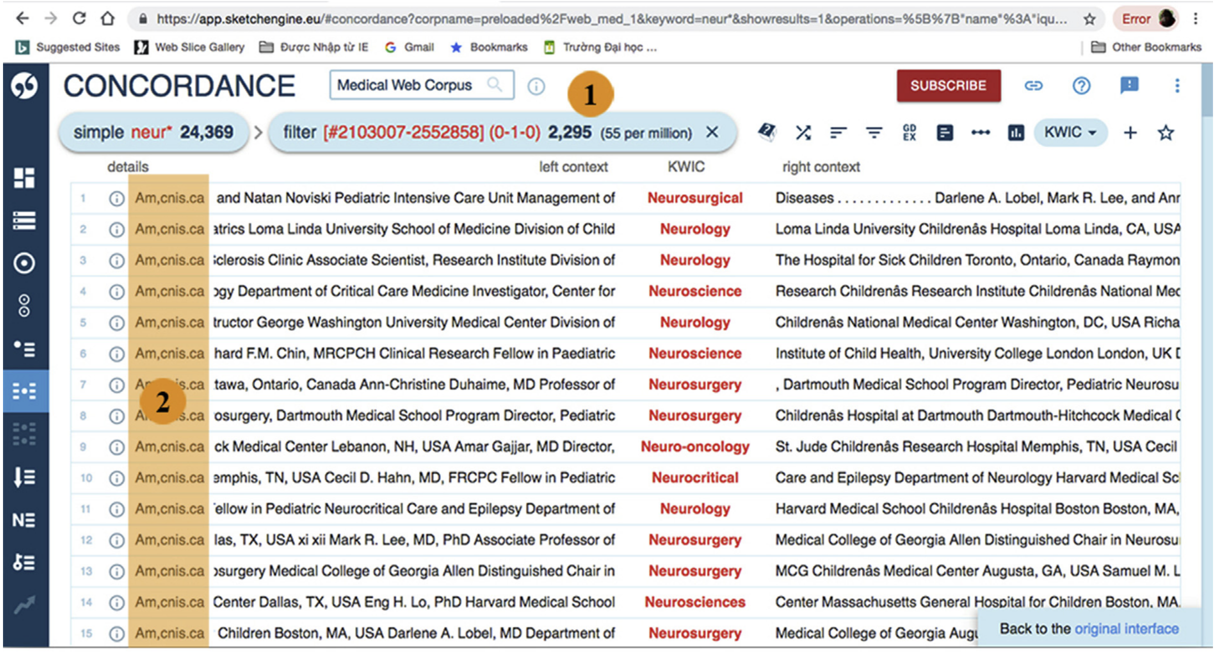
Frequency CHANGE CRITERIA BACK TO CONCORDANCE

✓ Show frequency per million

	lemma	↓ Frequency	Frequency per million
1	splenic	359	8.54
2	splenectomy	203	4.83
3	splendid	170	4.04
4	splenomegaly	96	2.28
5	splendor	30	0.71
6	Splenic	22	0.52
7	splenocytes	18	0.43
8	splenitis	14	0.33
9	splenization	13	0.31
10	splendour	10	0.24

Back to the original interface

Figure 10. Frequency results exported from the search of splen*.


https://app.sketchengine.eu/#concordance?corpname=preloaded%2Fweb_med_1&keyword=neur*&showresults=1&operations=%5B%7B"name"%3A"iqu...

CONCORDANCE Medical Web Corpus

simple neur* 24,369 > filter [#2103007-2552858] (0-1-0) 2,295 (55 per million)

	details	left context	KWIC	right context
1	Am,cnis.ca	and Natan Noviski Pediatric Intensive Care Unit Management of	Neurosurgical	Diseases . . . . Darlene A. Lobel, Mark R. Lee, and Anr
2	Am,cnis.ca	atrics Loma Linda University School of Medicine Division of Child	Neurology	Loma Linda University Childrenâs Hospital Loma Linda, CA, USA
3	Am,cnis.ca	sclerosis Clinic Associate Scientist, Research Institute Division of	Neurology	The Hospital for Sick Children Toronto, Ontario, Canada Raymon
4	Am,cnis.ca	ogy Department of Critical Care Medicine Investigator, Center for	Neuroscience	Research Childrenâs Research Institute Childrenâs National Mec
5	Am,cnis.ca	tructor George Washington University Medical Center Division of	Neurology	Childrenâs National Medical Center Washington, DC, USA Richa
6	Am,cnis.ca	hard F.M. Chin, MRCPCH Clinical Research Fellow in Paediatric	Neuroscience	Institute of Child Health, University College London London, UK D
7	Am,cnis.ca	tawa, Ontario, Canada Ann-Christine Duhaime, MD Professor of	Neurosurgery	, Dartmouth Medical School Program Director, Pediatric Neurosu
8	Am,cnis.ca	osurgery, Dartmouth Medical School Program Director, Pediatric	Neurosurgery	Childrenâs Hospital at Dartmouth Dartmouth-Hitchcock Medical C
9	Am,cnis.ca	ck Medical Center Lebanon, NH, USA Amar Gajjar, MD Director,	Neuro-oncology	St. Jude Childrenâs Research Hospital Memphis, TN, USA Cecil
10	Am,cnis.ca	emphis, TN, USA Cecil D. Hahn, MD, FRCPC Fellow in Pediatric	Neurocritical	Care and Epilepsy Department of Neurology Harvard Medical Sc
11	Am,cnis.ca	ellow in Pediatric Neurocritical Care and Epilepsy Department of	Neurology	Harvard Medical School Childrenâs Hospital Boston Boston, MA,
12	Am,cnis.ca	las, TX, USA xi xii Mark R. Lee, MD, PhD Associate Professor of	Neurosurgery	Medical College of Georgia Allen Distinguished Chair in Neurosu
13	Am,cnis.ca	osurgery Medical College of Georgia Allen Distinguished Chair in	Neurosurgery	MCG Childrenâs Medical Center Augusta, GA, USA Samuel M. L
14	Am,cnis.ca	Center Dallas, TX, USA Eng H. Lo, PhD Harvard Medical School	Neurosciences	Center Massachusetts General Hospital for Children Boston, MA.
15	Am,cnis.ca	Children Boston, MA, USA Darlene A. Lobel, MD Department of	Neurosurgery	Medical College of Georgia Augu

Back to the original interface

1 Frequency of derived words in 14 documents from CNIS website

2 Clickable line details showing information of CNIS website ULR, token number, document number and variety of English

Figure 11. Concordance view of words derived from neur-in 14 CNIS documents.

frequency measure, absent from Stedman's list, in setting explicit, attainable learning goals. The elimination of 208 morphemes considerably reduces the learning burden for medical students.

Close examination of the remaining 136 morphemes revealed that those occurring with the highest frequency were physi- and physio- (meaning *physical* or *natural*), which appeared 35,529 times in the Medical Web Corpus; the lowest frequency morphemes were lys- and lyso- (meaning *lysis*), with 977 occurrences.

Double-checking of specialized meanings using the *Merriam-Webster medical dictionary* indicated that the majority of listed morphemes constituted words used strictly within medical settings (e.g. neur-, neuri-, neuro-; uri-, uric-, urico-; and cardi-, cardio-). However, this process highlighted some morphemes used not only in technical but also in general words. For example, peri-, which means *around* or *about*, was identified in many medical terms, such as *peripheral*, *peritonitis*, *peritoneum*, *pericardium*, and *periostenum*. This morpheme also forms part of general words (e.g. *period*) commonly used in daily communication. However, its medical use is specialized enough to be included in the medical morpheme list.

Some morphemes, such as physi-, peri- and norm-, belong to words that are highly frequent and evenly distributed across the whole corpus. However, there was one source of documents where derived words were more concentrated. The highest frequency column, highlighted under the heading KWIC (Key Word in Context), was clickable and displayed concordance positions densely occupied by words derived from neur-, neuri- and neuro-. Figure 11 shows that derived words occurred more frequently (2,295 times) in 14 documents from the Canadian Network for International Surgery (CNIS) website.

Table 6 provides a statistical summary of the frequency of occurrence of medical morphemes in the final list of 136. Four morphemes occurred more than 20,000 times, accounting for 3% of overall occurrences in the final list. Over half of the total 136 morphemes (66%) were found to occur between 900 and 5,000 times in the corpus. The remaining 39 morphemes occurred more frequently, fluctuating between 5,000 and 15,000 repetitions. Three morphemes, accounting for 2% of the total, occurred between 15,000 and 20,000 times.

Sketch Engine allowed calculations of the number of words derived from the 136 morphemes in the new list, using the lemma as the unit of counting. The percentage coverage for tokens would of course have been different, and the inability to calculate this is a limitation of the study, but, due to time constraints, and the need to do all the calculations manually, the study reports on lemmas rather than tokens. Table 7 shows that in total, 59,607 lemmas were made from the 136 morphemes. These 59,607 lemmas accounted for 8.5% of the total 700,750 lemmas appearing in the Medical Web Corpus. The 8.5% coverage provided by the new list of medical morphemes, albeit a modest percentage, is still satisfactory.

The mastery of 136 morphemes, constituting a number of medical terms and guaranteeing the coverage of 8.5% of common medical vocabulary in the Medical Web Corpus, is worthy of time and effort, especially when compared with the mastery of 595 word families in the medical word list (Hsu, 2013), whose coverage of 10.72% in a corpus of medical textbooks is only slightly higher.

5.6. Validation of the final list of medical morphemes

The final list was validated by cross-checking the medical meanings of the 136 morphemes in three medical dictionaries. The listed morphemes were first checked using the Index of Paramedical Vocabulary (Schmidt, 1974) and 97 morphemes were confirmed. The Index of Paramedical Vocabulary assists users who are unfamiliar with Greek and Latin to locate medical terminology in medical dictionaries. For example, users searching for medical terms with chest-related meanings may not benefit much from direct searches of the word *chest* in a dictionary (Schmidt, 1974). Rather, they would do better to refer to the Index under the head word *chest* for helpful clues for obtaining words beginning with thorac-, thoraco- and thoracico-. This will facilitate the location of dictionary sections containing *chest* terminology.

The Index helps locate medical terms based on their initial morphemes, so it includes Greek and Latin prefixes only, and this explains why most of the remaining 39 morphemes were suffixes. The *Dictionary of medical derivations* (Casselman, 1998) was then used to check the remaining 39 morphemes. Of these, 32 morphemes were listed in the *Dictionary of medical derivations*. Six morphemes which could not be found in the *Dictionary* were -ics (*organized knowledge* or *treatment*), audi- (*hearing*), -lepsy (*seizure*), -stasis (*stopping*), -ine (*chemical substance*), and -plegia (*paralysis*). The troublesome suffix -ate was mentioned in the *Dictionary* but as a common verb ending, not as a *salt* (or *ester*). These seven morphemes were then checked and all were found to be included in the *Merriam-Webster medical dictionary.* This process validates the new list as being reliable and supports the idea that hours invested in learning these medical morphemes would be time well spent.

6. Implications for learners and teachers

Learning the list of medical morphemes requires alternative techniques apart from traditional instruction, which usually lacks morphological knowledge (Coxhead, 2000). It is suggested that teachers should employ the word part technique (Zheng & Nation, 2013) to provide adequate morphological knowledge and word-building skills necessary for lexical acquisition. To

Table 6
Statistical analysis of the final morpheme list based on frequency.

Number of occurrences (FREQUENCY)	Number of morphemes	Percentage
More than 20,000 times	4	3
15,000–20,000 times	3	2
10,000–15,000 times	11	8
5,000–10,000 times	28	21
900-5,000 times	90	66
Total	***136***	***100***

Table 7
The coverage of the final list of morphemes in the Medical Web Corpus.

Number of words (lemmas) derived from 136 morphemes	59,607
Number of words (lemmas) in the Medical Web Corpus	700,750
Lexical coverage	8.5%

master the word part technique, students should be equipped with relational knowledge (Nation, 2001) for receptive purposes. More specifically, when students know morphemes such as ovario- (ovary) and -tomy (cutting operation), for example, they are more easily able to see that the dictionary meaning of the word *ovariotomy* (surgical removal of an ovary) arises from the meanings of the two constituent parts of that word. Plus, they would become aware that these morphemes can occur in other combinations such as *myomectomy* (surgical removal of a myoma) or *ovaritis* (inflammation of the ovaries) with related meanings, and would be able to break new words into their component parts to grasp their meaning.

Importantly, teachers should introduce the word part technique when students already know a substantial number of medical terms because “these can act as familiar items to attach their new knowledge of word parts to” (Nation, 2001, p. 403). As an illustration, students who have learned the word *cardiology* as an unanalysed whole will probably recall it when teachers come to break down the word *cardio/vascul/ar*. Students will be able to link new with prior knowledge thus making their learning more efficient. In teaching word parts, it is advisable to introduce a memorable number of morphemes one at a time, according to Nation (2001), to help avoid students overloading. The dangers of interference may also be minimized if students do not need to acquire too many or too similar morphemes at the same time.

For productive purposes, students should gain *distributional knowledge* which, according to Tyler and Nagy (1989, as cited in Nation, 2001, p. 402), demonstrates “detailed awareness of the formal changes” when adding affixes to a root to form a new word. Students should know that affixation may create changes in written forms, as these do not always adhere to regular spelling rules. For example, the spelling of the root *muscle* changes to *muscular* when the suffix -ar is added. Teachers could also help students recognize what part of speech *muscle* becomes when it takes the suffix -ar. In other cases, affixation may affect pronunciation. Students should bear in mind that combining the suffix -logy with the root *cardio* leads to a change in stress patterns in the new word *cardiology*.

7. Conclusion

The study used a corpus-based approach to create a list of morphemes that occur frequently in the literature of various medical sub-branches. The findings indicate that the frequency of the short-listed 344 morphemes was subject to considerable fluctuation (from 0 to 35,000 occurrences in the Medical Web Corpus). As there is no point in students devoting excessive time to learning less frequently occurring morphemes, the 208 morphemes (60%) that did not meet the minimum frequency criterion clearly show the deficiencies of other lists. The remaining 136 morphemes occur most frequently in medical literature (see Appendix A), accounting for 8.5% of the Medical Web Corpus. By highlighting the morphemes constituting medical terms that medical students may encounter in a wide range of specialized texts, the new list makes learning goals explicit and provides a useful basis for designing vocabulary learning and teaching activities.

(This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.)

Appendix A. (A List of the 136 Most Frequent Medical Morphemes)

Morphemes	Meanings	Frequencies
physi-, physio-	physical; natural	35,529
peri-	around, about	31,954
neur-, neuri-, neuro-	nerve, nervous system	24,369
norm-, normo-	normal	24,072
-ar	pertaining to	19,817
epi-	upon, following, subsequent to	17,301
uri-, uric-, urico-	uric acid	16,258
-oma	tumor, neoplasm	14,246
cardi-, cardio-	heart; esophageal opening of stomach	13,951
bacteri-, bacterio-	bacteria	13,676
hem-, hema-, hemat-, hemato-, hemo-	blood	13,482
-tomy	cutting operation	12,793
chem-, chemo-	chemistry; drug	12,704
tox-, toxi-, toxico-, toxo-	toxin, poison	12,360
dys-	bad, difficult	12,072
end-, endo-	within, inner	11,440
-gen	producing, coming to be; precursor	11,342

(*continued on next page*)

(*continued*)

Morphemes	Meanings	Frequencies
hypo-	beneath, diminution, deficiency, the lowest	10,252
fibr-, fibro-	fiber	9910
gastr-, gastro-	stomach; belly	9420
stom-, stoma-, stomat-, stomato-	mouth	9104
my-, myo-	muscle	8754
lymph-, lympho-	lymph	7931
-pathy	disease	6831
meta-	after, behind; joint action, sharing; altering, changing	6780
chol-	bile	6721
or-, ori-, oro-	mouth	6672
vas-, vasculo-, vaso-	duct, blood vessel	6598
cyst-, cysti-, cysto-	bladder; cyst; cystic duct	6416
rect-, recto-	rectum, straight	6329
pneum-, pneuma-, pneumat-, pneumato-	air, gas; lung; breathing	6273
ost-, oste-, osteo-	bone	6076
arthr-, arthro-	joint, articulation	5928
oxy-	sharp, acid; acute, shrill, quick; oxygen	5838
chlor-, chloro-	green, chlorine	5792
thyr-, thyro-	thyroid gland	5621
-trophy	food, nutrition	5589
cata-	down	5472
ren-, reno-	kidney	5356
radio-	radiation, x-ray; radius	5275
men-, meno-	menstruation	5266
-ectomy	excision	5249
-ia	a condition	5173
-in	chemical suffix	5152
bronch-, bronchi-, broncho-	bronchus	5109
hepat-, hepatico-, hepato-	liver	5088
ana-	up, toward, apart	4781
carcin-, carcino-	cancer	4310
-cyte	cell	4243
hyster-, hysteron-	uterus, hysteria; late, following	4195
gluco-	glucose	4134
-ics	organised knowledge, practice, treatment	3991
cyt-, cyto-	cell	3735
neo-	new	3717
ovary-, ovario-	ovary	3709
pharmaco-	drugs, medicine	3563
aur-, auri-, auro-	ear	3539
-ose	sugar	3451
sin-, sino-, sinu-	sinus	3416
cervic-, cervico-	neck; uterine, cervix	3397
granul-, granulo-	granular, granule	3260
mening-, meningo-	meninges	3241
nas-, naso-	nose	3218
phos-	light	3188
leuk-, leuko-	white	3165
-uria	urine, urination	3134
nephr-, nephron-	kidney	3052
thromb-, thrombo-	blood clot	3051
-scopy	viewing	3007
audi-	hearing	2998
lact-, lacti-, lacto-	milk	2945
pleur-, pleura-, pleuro-	rib, side, pleura	2786
gyn-, gyne-, gyneco-, gyno-	woman	2783
necr-, necro-	death, necrosis	2636
enter-, entero-	intestine	2598
gon-, gono-	seed, semen	2566
-lepsy	seizure	2507
plasma-, plasma-, plasmato-	plasma	2487
crani-, cranio-	cranium	2469
scler-, sclera-	hardness (induration), sclerosis, ocular sclera	2454
-ate	a salt or ester of an "ic" acid	2436
laparo-	abdomen, abdominal wall	2396
hist-, histio-, histo-	tissue	2394
melan-, melano-	large	2345
mamm-, mamma-, mammo-	breast	2232
vesic-, vesico-	urinary bladder, vesicle	2208
trache-, trachea-	trachea	2196
macr-, macro-	large, long	2182

(*continued*)

Morphemes	Meanings	Frequencies
steno-	narrowness, constriction	2169
-stasis	stopping	2168
extra-	outside of, without	2131
-phylaxis	protection	2126
homeo-	same, constant	2112
-ase	an enzyme	2073
pseud-, pseudo-	false	2069
schiz-, schizo-	split, cleft, division	2057
lip-, lipo-	fat, lipid	2011
arteri-, arterio-	artery	1997
adeno-	gland	1992
staphyl-, staphylo-	grape, bunch of grapes, staphylococci	1990
angi-, angio-	vessel	1961
-iasis	condition, state	1960
ment-, mento-	chin	1922
chir-, chiro-	hand	1831
sarco-	flesh, muscle	1821
uro-	urine; urinary tract	1801
thorac-, thoracico-, thoraco-	chest, thorax	1709
-ine	chemical suffix	1674
ophthalm-, ophthalmo-	eye	1665
-plasia	formation	1625
encephal-, encephalo-	brain	1620
myel-, myelo-	bone marrow; spinal cord	1597
laryng-, laryngo-	larynx	1594
glycol-	sugars	1502
-plegia	paralysis	1484
-tropic	turning toward, affinity	1475
-pnea	breath, respiration	1455
-phrenia	of mind	1420
erythr-, erythro-	red, redness	1340
onco-	tumor, bulk, volume	1337
lith-, litho-	stone, calculus, calcification	1277
iso-	equal, like; isomer; sameness	1265
tachy-	rapid	1232
somat-, somato-, somatico-	body, bodily	1207
amyl-, amylo-	starch, polysaccharide	1191
cyan-, cyano-	blue; cyanide	1182
-cidal	killing, destroying	1180
-omata	plural of -oma	1167
pyo-	suppuration, pus	1114
-cele	hernia, swelling	1082
cephal-, cephalo-	the head	1071
mes-, meso-	middle, mean, intermediate; attaching membrane	1042
-opia	vision	1032
athero-	pasty, fatty	1019
rhin-, rhino-	nose	1004
lys-, lyso-	lysis, dissolution	977

Appendix B. (A List of Morphemes Commonly Used in Daily Communication)

a-	not, without, less	micr-, micro-	small, microscopic
ab-, abs-	away from	milli-	one thousandth
ad-	increase, adherence, motion toward, very	mon-, mono-	single
-al	pertaining to	morph-, morpho-	form, shape, structure
ambi-	on all sides, both	octo-	eight
an-	not, without	-oid	resemblance to
ante-	before	-osis	process, condition, state
anti-	against, opposing; curative; antibody	para-	abnormal; involvement of two like parts
-ary	pertaining to	penta-	five
aut-, auto-	self, same	pan-, pant-, panto-	all, entire
bi-	twice, double	path-, patho-	disease
bio-	life	ped-, pedi-, pedo-	child; foot
centi-	one hundredth	per-	through, thoroughly, intensely

chrom-, chromat-, chromo-	color	phil-, philo-	attraction; chemical affinity
chron-, chrono-	time	-philia	attraction; chemical affinity
-cide	killing, destroying	-phobia	fear
co-, col-, com-, con-, cor-	with, together, in association	phon-, phono-	sound, speech
de-	away from, cessation	phot-, photo-	light
deca-	ten	pod-, podo-	foot, foot-shaped
deci-	one tenth	-pod	foot, foot-shaped
dent-, denti-	tooth	poly-	multiplicity; polymer
derm-, derma-, dermat-, dermato-, dermo-	skin	post-	after, behind, posterior
dis-	separation, taking apart, not	pre-	anterior, before
duo-	two	pro-	before, forward; precursor
ex-	out of, away from	psych-, psyche-, psycho-	mind
-graph	recording instrument	quadr-, quadri-	four
hector-	one hundred	re-	again, backward
hemi-	one half	retro-	backward, behind
hept-, hepta-	seven	-scope	instrument for viewing
hydr-, hydro-	water, hydrogen	semi-	one-half
hyper-	excessive, above normal	sept-, septo-	seven; septum; sepsis, infection
-ic	pertaining to	septi-	seven
in-	in; not	sub-	beneath, less than normal, inferior
inter-	between, among	super-	in excess, above, superior, in the upper part
intra-	within	sy-, syl-, sym-, syn-, sys-	together
intro-	within	tel-, tele-	distant
-ism	condition, disease; practice, doctrine	tetra-	four
-itis	inflammation	therm-, thermo-	heat
kilo-	one thousand	trans-	across, through
-logy	study of; collecting	tri-, tris-	three
mal-	bad, deficient	ultra-	beyond
mega-	large, oversize; one million	uni-	one, single
-meter	measurement, measuring device	zo-, zoo-	animal; life

References

Barlow, M. (2004). Software for corpus access and analysis. In J. Sinclair (Ed.), *How to use corpora in language teaching (pp. 204–221)*. Amsterdam, NL: John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Carlisle, J., & Katz, L. (2006). Effects of word and morpheme familiarity on reading of derived words. *Reading and Writing, 19*, 669-693.

Casselman, W. (1998). *A dictionary of medical derivations - the real meaning of medical terms*. London, UK: The Parthenon Publishing Group Ltd.

Cengage. (n.d). *A list of prefixes, suffixes and roots*. Retrieved from http://www.cengage.com/resource_uploads/downloads/0534553389_46568.pdf. (Accessed 18 April 2019).

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

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

Dupuy, H. J. (1974). *The rationale, development and standardization of a basic word vocabulary test*. Washington, DC: Government Printing Office.

Džuganová, B. (2013). English medical terminology – different ways of forming medical terms. *JAHR – European Journal of Bioethics, 4*(7), 55-69.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 2*(4), 63-341.

Hamawand, Z. (2011). *Morphology in English: Word formation in cognitive grammar*. London, UK: Continuum.

Haspelmath, M. (2002). *Understanding morphology*. London, UK: Arnold.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research, 17*(4), 454-484.

Hwang, Y., & Lin, S. (2010). A study of medical students' linguistic needs in Taiwan. *Asian ESP Journal, 6*(1), 35-58.

Laufer, B. (1988). What percentage of lexis is necessary for comprehension? In C. Lauren, & M. Norman (Eds.), *Special language: From humans to thinking machines* Clevedon, UK: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42-53.

Lieber, R. (2010). *Introducing morphology*. Cambridge, UK: Cambridge University Press.

Lüdeling, A., & Kytö, M. (2008). *Corpus linguistics: An international handbook*. Berlin, Germany: Walter de Gruyter GmbH.

Merriam-Webster Incorporate. (2019). Medical dictionary. Retrieved from https://www.merriam-webster.com/medical. (Accessed 18 April 2019).

Meyer, C. (2002). *English corpus linguistics: An introduction*. Cambridge, UK: Cambridge University Press.

Nagy, W. E., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262-282.

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

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-81.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge University Press.

Nguyen, H., & Pham, M. (2016). Difficulties in teaching English for specific purposes: Empirical study at vietnam universities. *Higher Education Studies, 6*(2), 154-161.

Peterson, J. (1984). (Doctoral dissertation. *Determination of the formal vocabulary of physicians through analysis of medical literature*. Atlanta, USA: Georgia State University.

Piroozan, A., Boushehri, E., & Fazeli, R. (2016). A review of English for medical purpose for Iranian EFL learners. *Journal of Advances in English Language Teaching, 4*(2), 24-29.

Schmidt, J. E. (1974). *Index of paramedical vocabulary: An index-indicator enabling the user not versed in Greek and Latin to locate the terminology of any given subject in a paramedical, medical, or biological dictionary*. USA: Charles C. Thomas.

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *Tesol Quarterly, 36*(2), 145-171.

Shaw, E. (2011). *Teaching vocabulary through data-driven learning*. Master's thesis. Brigham Young University. Retrieved from https://www.mobt3ath.com/uplode/book/book-19377.pdf. (Accessed 18 April 2019).

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Stauffer, R. G. (1942). A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in the elementary school. *The Journal of Educational Research, 35*(6), 435-458.

Stedman's Online. (n.d). *Medical prefixes, suffixes and combining forms*. Retrieved from http://stedmansonline.com/webFiles/Dict-Stedmans28/APP05.pdf. (Accessed 18 April 2019).

Thomas, A. (n.d). *Common prefixes, suffixes and roots*. Center for Development and Learning. Retrieved from http://www.cdl.org/wp-content/uploads/2013/12/Common-Prefixes-Suffixes-and-Roots-8.5.13.pdf. (Accessed 18 April 2019).

University of Hawaii at Manoa. (n.d). *Anatomical word roots*. Retrieved from http://manoa.hawaii.edu/undergrad/learning/wp-content/uploads/2014/03/Word-Roots.pdf. (Accessed 18 April 2019).

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

Zheng, W., & Nation, P. (2013). The word part technique: A very useful vocabulary teaching technique. *Modern English Teacher, 22*(1), 12-16.

Ms Chinh Ngan Nguyen Le is a lecturer in the English for Specific Purposes (ESP) Department, Hue University of Foreign Languages. She received her Master degree in Education in the University of Adelaide. Her main research interests are computer assisted learning, computer-mediated communication and corpus linguistics.

Dr Julia Miller is a senior lecturer at the University of Adelaide. Her main research interests are lexicography, phraseology, and the teaching of English for academic purposes, for which she designed the free English for Uni website, containing engaging videos and interactive exercises on a range of grammar issues.