

ORIGINAL ARTICLE

A Corpus-Based Evaluation of English Programs for Medical Education in Iran

Ebrahim Samani^{1*}, Nooreen Noordin[†], Razieh Bagheripour[‡]

^{1*} **Corresponding author**, English Language Teaching Dept., Farhangian University, Tehran, Iran,

e.samani@cfu.ac.ir

[†] Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

[‡] English Language Teaching Dept., Payame-Noor University, Tehran, Iran

Abstract

English for Specific Purposes (ESP) has grown in importance in these years. English for Medical Purposes (EMP) as one of its types plays a vital role for medical students. This study was conducted to evaluate the textbooks taught in teaching English in different programs of medical universities. The study focused on six textbooks widely used in medical universities. The method for the analysis was corpus linguistics. The researchers of the study compiled the English for Medical Programs corpus (EMP) as well as a corpus from the videos of YouTube with medical content. They examined the corpus of the textbook against YouTube corpus as well as an already available corpus known as Medical Web Corpus. Sketch Engine was used for the analysis. According to the findings, three corpora were compatible. The further qualitative information indicated that the words that were typical of medical web and YouTube corpora and didn't exist in EMP corpus are already known by medical students. Findings also indicated that the focus needs to be on general English, as students need this vocabulary for communicating with the international community.

Received

10 Jun 2026

Revised

10 July 2026

Accepted

20 July 2026

Published

1 August 2026

Keywords:

EMP, YouTube, Corpus Linguistics, Sketch Engine, Medical Web Corpus

Introduction

English plays an important role in higher education around the world, and the demand for English language courses has been increasing for years. In non-English countries, this language, which is mentioned as *lingua franca* for all sciences, has emerged as a special course named English for specific purposes (ESP) (Boniadi, A., M., & Ghojzadeh, et al., ۲۰۱۳). ESP includes several types, one of which is English for medical purposes (EMP). It is vital for medical students to know the technical terms of medicine because much of the medical information in the world, including textbooks, articles, and new findings, is expressed in English. One of the most important aspects of English courses for medical purposes is learning specialized medical vocabulary. To teach these specialized words to medical students, many researchers have tried to make a list of medical words based on medical texts. In other words, the studies of most researchers have been based on the corpus to make a list of the most important high-frequency words in the medical literature, the learning of which is worth the time and effort for students (Le and Miller, ۲۰۲۰). Unfortunately, it is seen that many medical students in Iran, even our doctors, do not demonstrate satisfactory performance, especially in spoken and listening skills. According to the literature on ESP and corpus linguistics, various factors such as not examining students' needs, lack of appropriate content, lack of using appropriate technology in classrooms, inefficient English language teaching system in schools, and so on have been highlighted as the reason for this deficiency. One of the ways to address the needs of students is the evaluation of medical textbooks through corpus linguistics. Farhadi (۲۰۰۷) pointed that the most common words that students need in their field should be identified by corpus analysis and then these corpora can be taught in ESP courses.

Literature Review / Background

English for specific and medical Purposes (ESP&EMP)

English for Specific Purposes (ESP) has a long history and has been an important field in English language studies since the ۱۹۶۰s. In fact, the basis of the emergence of ESP is the answer to the question of why a particular learner needs to learn English (Dudley-Evans et al., ۱۹۹۸). Paltridge and Starfield (۲۰۱۴) described ESP as teaching and learning English where the goal is to use English in a particular field. The content and purpose of the ESP courses are based on the specific needs of the learners. But the main point in teaching English for special purposes is to identify the needs of learners and use materials tailored to their needs.

According to Belcher (۲۰۰۹), there are several types of ESP because there are different learners' needs. For instance, English for Academic Purposes (EAP) is adapted to the needs of educational learners; English for

Occupational Purposes (EOP) is based on professional needs of learners, which itself is divided into other subcategories including English for Business Purposes (EBP); English for Legal Purposes (ELP) and English for Medical Purposes (EMP).

Knowledge of English plays a role in all professions, but it seems to play a distinct role in medicine. Medical students need to obtain a lot of information and data from international sources and use them in diagnosing and treating diseases. But unfortunately, many doctors and medical students in Iran are not satisfied with teaching English in universities. Therefore, there are myriad studies in the field of teaching English to medical students to find problems and try to solve them (Khalili & Tahririan, ۲۰۲۰). For example, Hekmati and Navidinia (۲۰۱۶) stated that teaching English in Iranian universities is mostly based on the 'grammar translation' method and the textbooks are consistent with this approach. To solve this problem, they studied a new method for teaching English to medical students, in which they used various media such as videos, audio files, and real medical texts as the course teaching material. According to their findings, the utilization of this method improved all four language skills of students. The students were much more satisfied with this method than traditional ones.

It seems that students around the world would rather have an efficient textbook as a source of learning. And here again, a need for an effective book in teaching language to medical students is felt. An example demonstrating students' tendency to have an efficient book in a national context is the study conducted by Khalili and Tahririan (۲۰۲۰) who studied the challenges of teaching English to medical students according to the needs, lacks and expectations of this group of students. They found that English language courses for medical purposes (EMP) did not meet the needs of the students, and they were dissatisfied with the textbooks used in these courses, especially old books, and expected that more of the authentic material should be used in these classes.

Another study highlighting the critical role of efficient books is Nezakatgoo and Behzadpoor (۲۰۱۷) who studied the important challenges in teaching ESP at medical universities of Iran. One of the challenges they mentioned was finding the appropriate content to teach in these courses. According to these researchers, the existing books either do not correspond to the students' culture or are not based on their needs, and this issue reduces the motivation of students in ESP courses. As the next example, Vosoughi et al. (۲۰۱۳) considered one of the EAP textbooks for medical students in Iran "English for University Medicine Students I". They selected three criteria to evaluate this book: objective evaluation, review of the content and evaluation of lesson formats. They found that this book doesn't meet the essential needs of medical students. They concluded that fundamental changes should be made to medical students' course books for learning English.

There is also dissatisfaction with English language courses among medical students of other countries. In a study in the Romanian context, Pavel (۲۰۱۴) stated that teaching English in medicine should be tailored to the students' needs. According to the researcher, there are very few English medical textbooks in Romania that meet only some of the medical students' needs. He developed a new curriculum based on various tasks in the field. Ibrahim (۲۰۲۰) also indicated that Sudanese medical students and instructors are not satisfied with English language courses. And an important problem for these students is the lack of improvement in their English communication skills in these courses. He suggested that to improve these courses, fundamental changes should be made to the content based on the students' needs.

As mentioned above, several studies conducted in the field of ESP have indicated that medical students are dissatisfied with the content taught and its incompatibility with their needs. It is agreed that corpus linguistics is a common means of identifying the words that students need to focus on ESP courses (Farhadi, ۲۰۰۷).

Corpus linguistics as the theoretical framework of the study

Corpus is actually a Latin word meaning body. Corpus in linguistics is a collection of texts stored in an electronic database and usually contains thousands of words (Baker, ۲۰۰۶). Corpus linguistics was introduced by structuralists before Chomsky (McEnery et al., ۲۰۰۶). The definition of corpus linguistics is in fact the study of language based on examples of real-life language use. Kacetl and Frydrychova-Klimova (۲۰۱۵) defined corpus as an electronic database of authentic materials such as transcripts of spoken language. Flowerdew (۲۰۱۱) has stated some criteria for defining corpus in his book, including that a corpus must contain natural and real data, represent a particular language or genre, and be designed for a specific linguistic purpose.

There are different types of corpora. Each corpus is created for a specific purpose and the types of corpora depending on their purpose include specialized, general, comparable, parallel, historical, learner, pedagogic and monitor. A general corpus may be created from a variety of texts. It may be used to produce general language teaching materials, and is sometimes referred to as a reference corpus. British National Corpus (BNC), for example, is a well-known general corpus that contains about one hundred million words from a variety of written and spoken texts. On the other hand, a special corpus is designed for a specific purpose and its references are from special texts. Michigan Corpus of Academic Spoken English (MICASE) is one of these corpuses (Hunston, ۲۰۰۲). Since its inception, Corpus has been used for a variety of applications, including dictionary creation, production of language teaching materials, a suitable tool for language learning and translation studies, etc. (Baker, ۲۰۰۶; Reppen, ۲۰۱۰; Flowerdew, ۲۰۱۲). Teaching and learning English are greatly enhanced by using authentic material extracted from the language corpus

(Phoocharoensil, ۲۰۱۲). In particular, a corpus is widely used in ESP, especially for extracting high-frequency words in a specific field (Boulton, ۲۰۱۶).

Corpus linguistics has been used in many studies in recent years (Baker, ۱۹۹۵; Klimova, ۲۰۱۴; Zhu & Wang, etc.). In fact, studying the real language in the right context can provide useful information to the researchers (Del Mar Sánchez Ramos, ۲۰۲۰). As a case in point, in a study conducted by Phoocharoensil (۲۰۱۲), ۱۷ Thai students took part in a course to learn grammar through corpus data. The results showed that the students were very satisfied with this method of teaching grammar and stated that corpus is a very useful tool for teaching grammar and, despite traditional methods, motivated them to learn grammar. Del Mar Sánchez Ramos (۲۰۲۰) conducted research to determine whether corpus-based teaching is useful in translating medical texts and whether the use of corpus management tools is appropriate to analyze the corpora. For this purpose, ۴۰ students participated in this study and collected their own corpora and analyzed them with the LancsBox tool. The results indicated that a corpus is a useful tool in translating medical texts and students had a positive attitude towards using corpus management tools as they were useful tools “to solve linguistic difficulties”.

One of the most important challenges in learning a language is learning vocabulary. And in this regard, one of the most important concerns of educational planners and language teachers, especially in ESP, is which words to focus on in their courses (Hsu, ۲۰۱۸). To this end, some researchers and teachers have prepared vocabulary lists in special fields of study based on corpora. For instance, Otto (۲۰۲۱) used a corpus-oriented system to select important specific words for civil engineering. He produced two corpora, one of which was compiled from professional documents in the field and the other was collected from the civil engineering students. Finally, he produced a list of ۱۸ words that contained the most common words in writing for the field of civil engineering.

Hsu also (۲۰۱۸) collected a list of the most widely used words in Traditional Chinese Medicine textbooks. To this end, he compiled English-language textbooks on traditional Chinese medicine under ۲۲ subjects and produced a corpus of ۱۳ million words. He then prepared a list of English specialized words in traditional Chinese medicine, using the word frequency scale of the Contemporary American English Corpus (COCA) and the British National Corpus (BNC). This English word list can be a suitable resource for preparing the material needed to teach the technical vocabulary of traditional Chinese medicine. According to the results of his research, the first ۱۰۰۰۰ words of the Contemporary American English Corpus (COCA) and the British National Corpus (BNC) matched about ۹۸% of the corpus words he had prepared based on traditional Chinese medicine textbooks.

In another study on English medical morphemes, Le and Miller (۲۰۲۰) compiled a list of common medical morphemes based on some sources available to the public. They compared Stedman’s list, which contains

۶۶۳ medical morphemes, to the Cengage list of general English morphemes and also to the Center for Development in Learning (CDL) list of general English morphemes. Then they removed the morphemes from Stedman's list that were on Cengage and CDL lists. After that they compared the remaining morphemes in Stedman's list using sketch engine with the Medical Web Corps to determine the frequency of the remaining morphemes. Of the remaining ۳۴۴ morphemes, ۲۰۸ were low-frequency words in the Medical Web Corps, which were removed from the list, leaving ۱۳۶ words, which cover about ۸,۵ percent of the Web Corps medical words and are frequently repeated in medical texts. And knowing them is essential for medical students and is worth the time and effort.

Some researchers have developed specialized vocabulary lists for teaching other languages. For example, Rezvani et al. (۲۰۱۶) have provided a list of Persian academic words (PAWL) to assist non-native learners in learning Persian vocabulary and also to be useful in producing the materials needed to learn this language. They made the PAWL list, which consists of ۵۳۹ academic words in ۷ subcategories, based on a corpus compiled from several books and articles, and the criteria for selecting words in PAWL was the frequency and range of words in the corpus.

Some researchers use written texts such as books and articles to create corpora, but others use spoken language such as videos. Mattiello (۲۰۱۹), for example, created a corpus that consisted of Ted Talk videos on cancer. Her aim was to examine the use of language strategies by Ted's speakers to transfer medical knowledge to professionals and non-professionals. In the end, the authors acknowledged that Ted Talks could play an important role in conveying scientific-medical content to the public, and that the use of some language strategies, such as metaphor, helps to simplify scientific content. In addition, they concluded that Ted Talks could be used as a suitable material for teaching English in ESP courses, especially English for medicine.

In a similar case, Kaceti and Frydrychova-Klimova (۲۰۱۵) conducted research on English words in video clips focusing on travel and Tourism. They prepared their corpus from a collection of videos related to travel and tourists. Sketch Engine was the tool in facilitating the extraction of words with a high frequency related to travel and tourism. Finally, they compared specialized vocabulary in video clips to the British National Corpus (BNC) and provided a list of the most used words in video clips about tourism. This list can be used as material in English for travel and tourism courses. Coats (۲۰۱۹) also tried to create a corpus of American language diversity in different regions based on videos posted on YouTube focusing on Native Americans. Then, the diversity of corpus structure and words was analyzed using different techniques. The results demonstrated that the frequency of used words varied in different areas and therefore the regional diversity of spoken language is obvious. According to what was mentioned earlier, it seems that there is a gap in the literature on textbooks evaluation through corpus linguistics in Iran's context and especially focusing on textbooks taught in EMP courses. Moreover, there isn't any medical corpus in the

world collected from oral sources. To address these gaps, the current study examines the following questions:

١. What are the most frequent medical words in the medical videos of the most subscribed medical channels on YouTube?

٢. What are the most frequent medical words in the selected medical books which are taught in English for medical purposes (EMP) courses in Iran?

٣. How consistent is the corpus represented in this study with the Medical Web Corpus (an international corpus)?

٤. Is there an association between the frequency of medical words in the corpus of the selected books and the corpus from YouTube videos?

٥. What are the words that are typical of the YouTube corpus that is not present in the EMP? And how much important are they to be included in EMP?

Method

Material and Instruments

Medical Web Corpus (MWC)

One of the tools used in this study was Medical Web Corpus (a well-established and international corpus), which is accessible to Sketch Engine users. The Medical Web Corpus is rich in medical English because it is large in size and consists of various types of medical texts, including up-to-date examples, although it does not provide details of the special medical fields. This corpus is taken “from ٥٢٦ documents on ٣٤٤ websites encompassing ٤٢, ٠٥٤, ٠١١ tokens (running words) and ٣٣, ٩٦١, ٧٨٦ types (individual words) grouped into ٧٠٠, ٧٥٠ lemmas (headwords plus inflected forms)”. The corpus contains two types of English, “with ٤٤٤ documents including ٣٤, ٣٩٧, ٠٣٩ tokens drawn from American English sources and ٨١ documents with ٧, ٦٣٤, ٩٩٦ tokens from British English sources” (Le and Miller, ٢٠٢٠, p١٠٧).

YouTube Medical Corpus (YouTube Corpus)

In order to collect medical corpus in the context of this study, the authors selected the most subscribed medical YouTube channels and downloaded the entire content of these channels. Table ١ indicates several examples of these channels:

Table ١. Some of the most visited YouTube channels used for compiling medical YouTube corpus in this study

#	channels				
۱	Med school insider	۷	Dr najeeb lectures	۱۳	Doctor mike
۲	Registered nurse rn	۸	Chubbyemu	۱۴	Khan Academy health
۳	Kati morton	۹	Armando hasudungan	۱۵	Osmosis
۴	Nucleus medical media	۱۰	FreeMedEd	۱۶	One-minute medical school
۵	Harvard medical school	۱۱	Interactive biology		
۶	Medcram videos	۱۲	Ted Talks		

As it is clear from Table ۱, these channels provide videos about the variety of topics related to medical fields including but not limited to medical school, microbiology, pharmacology, cardiology, general healthcare. The videos that were downloaded from these websites were uploaded to the otter.ai website which provides the users with the transcripts of the videos. The transcription of videos then was compiled in one notepad document (about ۳ million words).

English in Medical Program (EMP) corpus

In order to collect the medical corpus, six widely used textbooks to teach English in different programs of medical universities in Iran were selected. The list of these textbooks was given in Table ۲:

Table ۲. Books that were used in compiling EMP corpus

#	Title	Author	Publication
۱	English for the students of Medicine	Fatemeh Fazelli Laleh Khojasteh Ali Mahboudi Nasrin Shokrpour	Publications of Shiraz University of medical sciences
۲	English for the students of Medicine	Mahmood Reza Atai Leila Shoja Masoumeh Kafshgarsouteh Mandana Zolghadri	the organization for researching and composing university textbooks in the humanities (SAMT)
۳	English for the students of Medicine (I)	Reza Deedari	the organization for researching and

		Seyyed Mohammad Zia Hossaini	composing university textbooks in the humanities (SAMT)
۴	English for the students of Medicine (II)	Mohammad Tahririan Fereshteh Mehrabi	Hassan the organization for researching and composing university textbooks in the humanities (SAMT)
۵	English for the students of Medicine (II)	Mohammad Tahririan	Hassan the organization for researching and composing university textbooks in the humanities (SAMT)
۶	English for the students of Medicine	Mohammad Tahririan A.Ameri Golestan M.A. Tahririan	Hassan the organization for researching and composing university textbooks in the humanities (SAMT)

The books that were mentioned in Table ۶ were scanned and then the scanned documents were uploaded to ocr.space website that converts images and PDF documents to text free of charge. The free plan of this website allows the maximum upload of ۱۰ pages in every attempt. As a result, the scanned pages which reached ۱۴۷۴ pages were fed into the website in size of ten pages each time. The outputs were saved as a word document and then all the outputs were compiled in a word document for analysis. To make sure that the output as the result of OCR is an exact duplicate of the original document, the authors reviewed all the six books and their equivalent OCR output to make sure of exact match and corrected the cases that were inconsistent.

Otter.ai

It is a company that offers applications on its website which use artificial intelligence. It turns the speeches into written transcriptions. Through the software of this company which is called otter, the users are able to generate written transcriptions of the speeches. The authors subscribed to the pro account of this software that provides ۶۰۰۰ minutes of audio and video transcription each month for one year. Figure ۱. Provides a snapshot of the home page of the website:

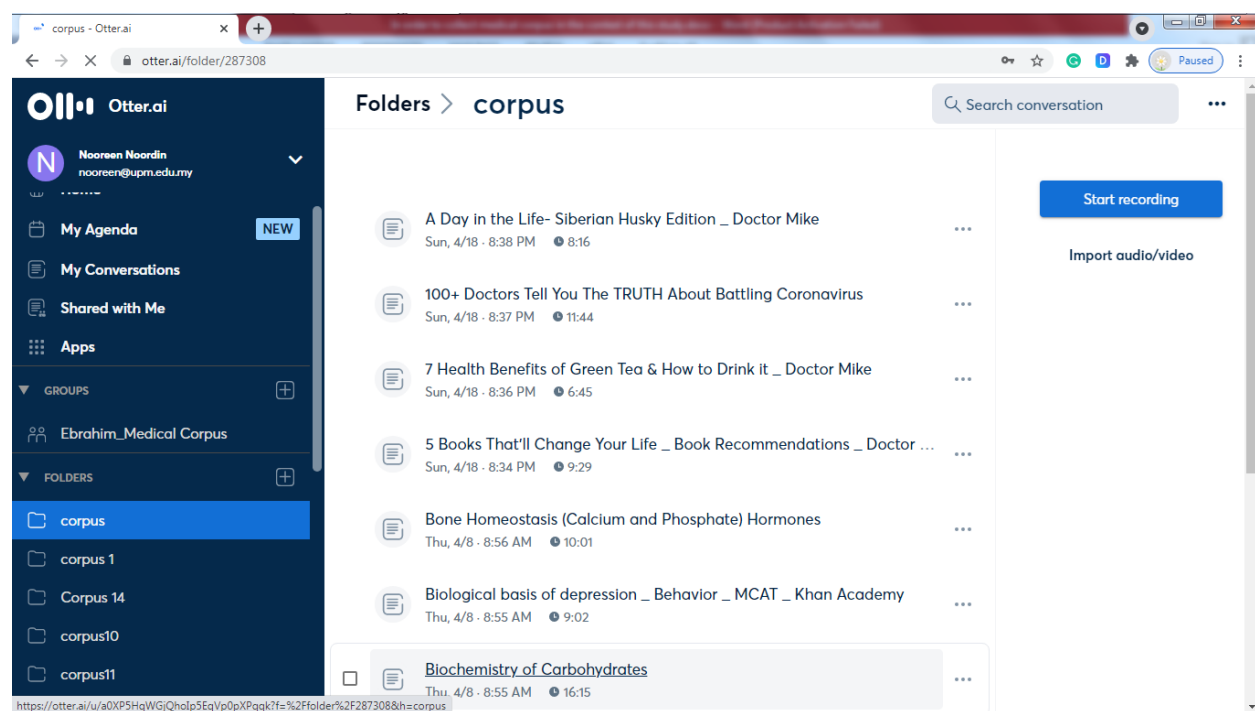


Figure 1. A snapshot of the dashboard page of the Otter.ai website

As Figure 1 shows, the website is completely user-friendly and the users just need to import audio/video documents into the website and then the software provides the transcripts that can be edited or exported in different formats such as txt, docs., pdf., and srt. The accuracy of the website in transcribing the audio and video documents is quite high and as far as the authors remember, the errors of mismatch were negligible if there was any.

Text Fairy (OCR Text Scanner)

Text fairy is an application that converts images into texts. The authors used this application to convert scanned images of the selected books to text format because the electronic versions were not available.

Sketch Engine

Sketch Engine is an online tool that is available at <https://www.sketchengine.eu/>. It gives us different information about the properties of language such as words counts and frequencies. Those who subscribe to the Sketch Engine website can also have access to Medical Web Corpus; Sketch Engine is actually a great tool to study how language works. Its algorithms are designed to analyze real texts and instantly

determine which words are common, which are rare, which are unusual, and so on (Le and Miller, ۲۰۲۰). The largest corpora that are available on this company's website consists of ۵۰,۰۰۰,۰۰۰,۰۰۰ words. The website helps researchers and users in uploading their own texts and preparing them for the analysis. Figure ۲ shows a snapshot of the dashboard page of the website:

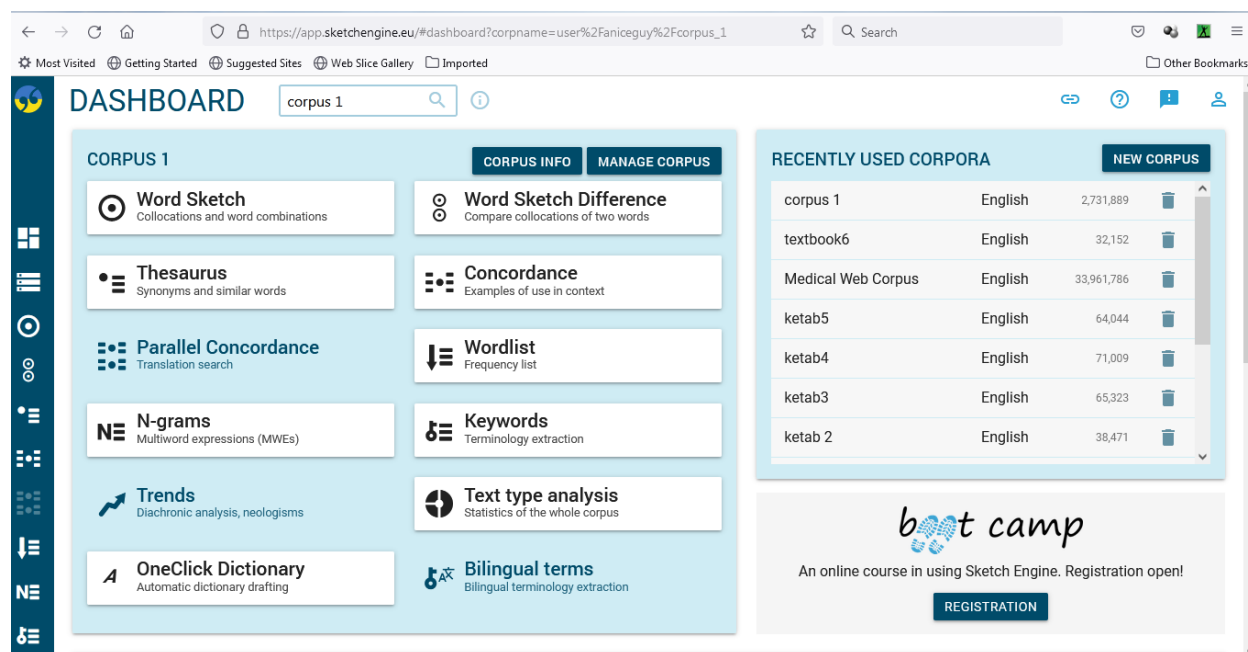


Figure ۲. A snapshot of the dashboard page of the Sketch Engine website

As Figure ۲ indicates, the website provides solutions in the area of full-text search, terminology extraction, document classification and categorization, data mining and information retrieval. As it is clear from Figure ۲, the authors used only the wordlist and keywords options of the website for their analysis.

Procedure

This study aimed at examining the textbooks that are common for teaching English in different programs of the medical fields in the medical universities of Iran. At the beginning of the study, the authors purchased the common books and then the books were scanned. In order to convert the scanned pages into word documents, the authors selected ocr.space as the website that is correct, reliable, and, user friendly. The extra information was also deleted from the corpus like the front and last pages of the books that are not related to the main content and also any footprints related to the website. The text of all the books was saved in one file that is called EMP corpus in the context of this manuscript.

Another task in this study was preparing the corpus of video and audio materials that were available in the most subscribed channels on YouTube. Since the researchers used the pro account of Otter software, there was a monthly limit of ٦٠٠٠ minutes on the transcription; As a result, the authors had to wait for several months for the transcription of all the collected materials. The website transcribes the audio and video materials separately and provides an output that can be exported in different formats. The authors also had access to web medical corpus as the result of their subscription to the sketch engine website. This website provided a forum for not only organizing the data in one place, but also facilitating the analysis of the mentioned corpora, some of the tasks that the users can perform through the website on data are:

Word sketch: The word sketch processes the words collocates and other words in their surroundings. It can be used as a one-page summary of the word's grammatical and collocational behavior. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc.

Concordance: The Concordance is the most powerful tool with a variety of search options. It can find words, phrases, tags, documents, text types or corpus structures and displays the results in context in the form of a concordance. The concordance can be sorted, filtered, counted and processed further to obtain the desired result. Despite being the most powerful tool, the concordance used with large corpora may find so many results that it can be tedious to analyze and interpret them.

Keywords and term extraction — identifying typical words

Sketch Engine supports monolingual and bilingual term extraction. Keyword and term extraction are used to extract terminology for use in translation and interpreting extract single word and multiword units which are typical of a corpus/document/text or which define its content or topic, and compare two corpora/documents/texts by identifying what is unique in the first corpus compared to the second one.

Results

In order to answer the first and the second research question, the researchers used the wordlists function of the sketch engine website, it was configured in a way that results include only lemmas (headwords plus inflected words) excluding prepositions, adverbs, article, etc. The results of the wordlists analysis of three corpora in the study are given in Table ٣.

Table ۳. The most frequent words in YouTube, Medical Web, and EMP corpora

YouTube corpus			Medical Web corpus			EMP corpus		
Words	AF	FinM	Words	AF	FinM	Words	AF	FinM
cell	۱۱۳۱۸	۳۷۱۳	patient	۹۶۴۱۹	۲۲۹۲	patient	۱۷۲۸	۴۶۲۴
blood	۸۹۹۰	۲۹۰۰	case	۷۱۴۵۹	۱۶۹۹	disease	۱۱۴۰	۳۰۰۰
patient	۶۷۸۸	۲۲۲۷	disease	۶۵۸۲۸	۱۵۶۵	cell	۱۱۳۲	۳۰۲۹
right	۶۳۱۹	۲۰۷۳	time	۵۶۰۰۶	۱۳۳۱	blood	۱۱۲۴	۳۰۰۸
system	۴۸۷۶	۱۶۰۰	treatmen t	۵۳۷۴۲	۱۲۷۷	body	۸۵۰	۲۲۷۴
body	۴۶۸۲	۱۵۳۶	year	۴۲۷۲۸	۱۰۱۶	cancer	۷۶۶	۲۰۰۰
time	۴۶۷۸	۱۵۳۵	day	۴۰۵۶۱	۹۶۴	word	۶۹۵	۱۸۶۰
thing	۴۵۷۴	۱۵۰۰	cell	۳۷۲۴۰	۸۸۵	infection	۵۷۰	۱۵۲۵
heart	۴۴۰۸	۱۴۴۶	child	۳۶۵۹۵	۷۸۰	treatmen t	۵۴۲	۱۴۵۰
people	۳۵۴۴	۱۱۶۲	blood	۳۶۳۸۱	۸۶۵	part	۵۳۵	۱۴۳۱
part	۳۳۲۲	۱۰۹۰	pain	۳۵۸۳۰	۸۵۱	drug	۴۹۳	۱۳۱۹
level	۳۰۰۱	۹۸۴	body	۳۴۷۵۷	۸۲۶	paragrap h	۴۵۷	۱۲۲۳
muscle	۲۹۷۴	۹۷۵	symptom	۳۳۲۹۸	۷۹۱	people	۴۴۹	۱۲۰۱
way	۲۸۲۷	۹۲۷	effect	۳۲۴۶۶	۷۷۲	system	۴۴۰	۱۱۷۷
lot	۲۸۱۴	۹۲۳	study	۳۱۱۰۹	۷۳۹	heart	۴۲۳	۱۱۳۲

The results in Table 3, are reported in absolute frequency (AF) and frequency in millions (FinM) to facilitate comparison and further analysis. As the table shows, only 10 frequent words have been reported for all three corpora due to space limitations in the manuscript. As it is clear from the table, all three corpora of YouTube, medical, and EMP are similar to some extent in terms of the most frequent words. In order to examine consistency, the top 100 frequently used words in YouTube and medical web corpora were selected and pasted into two columns of an excel worksheet. The following direction in excel was followed to find similar cells in both columns

Conditional formatting-highlight cell rules-duplicate values

According to the results, 48 words out of 100 top frequently mentioned words in two corpora of medical web and EMP are similar which yields a consistency of 61 percent which shows that the YouTube corpus collected in this study is consistent with the medical web corpus.

There was also consistency between MWC and EMP corpora, there was also correspondence between the YouTube, web corpus, and EMP corpus because the conditional formatting function of the Excel didn't find any inconsistencies between EMP corpus and the two other corpora.

As the next task and in order to answer the fourth research question in this study, keywords and term extraction option that is available in the sketch engine was used. As it was mentioned, this option identifies the cases that are typical of one corpus, however, they don't exist or they are of low frequency in another one. After comparing YouTube and EMP corpora, the words that were typical of the YouTube corpus were extracted. Table 4 represents a fraction of these words:

Table 4. A fraction of words that were typical of the YouTube corpus, but they were nonexistent in EMP corpus

#	words	#	words	#	words	#	words
1	fibre	11	cerebellum	21	laterally	31	magnesium
2	posterior	12	cranial	22	coronavirus	32	follicle
3	depolarization	13	amino	23	tyrosine	33	thumb
4	dorsal	14	recognize	24	glucagon	34	ventral
5	covid	15	bond	25	shift	35	interleukin
6	medial	16	depolarize	26	hypo	36	reticular
7	inferior	17	brainstem	27	fossa	37	sulcus
8	track	18	neutrophil	28	repolarization	38	bipolar
9	ganglion	19	med	29	midbrain	39	humerus
10	thalamus	20	subscribe	30	specialize	40	Intrinsic

In order to find the importance of these words and whether the students in medical universities already know them, a sample of ۱۱ medical students was selected based on convenience sampling from Bam and Qom University of medical sciences. They were in their fifth and seventh years at university. They reviewed a sample of ۱۳۰ words and responded by saying whether or not they know the word, and how important they think the word is for their English-specific knowledge. Table ۵ represents the response to the questionnaire.

Table ۵. Percentage of students who didnt know the word and considered the words as important

Students	Wotds that they donot know	Words that are important
۱۱	۱۱	۴۳

As Table ۵ indicates just ۱۱ percent of those words that were typical of the YouTube corpus were unfamiliar to our sample of students and ۴۳ percent of the words were considered as unimportant by the students.

Discussion

This study was conducted to evaluate the books that are taught in English courses for different programs of medical universities using corpora. The outcomes of this study were a corpus that was compiled based on medical video clips (YouTube Medical Corpus) and a corpus based on the books taught in English courses in medical universities (EMP corpus). Kaceti and Frydrychova-Klimova (۲۰۱۵) conducted similar research in that the outcome of their study was a corpus of the most frequent words which were appeared in the video clips about travel and tourism. In a similar way, the result of Rezvani et al. (۲۰۱۶) was the first Persian Academic Word List (PAWL) which consists of the most frequent academic words in Persian. Their list was created based on several books and research articles in different disciplines.

In this study, the YouTube corpus was compared to the MWC and then compared to the EMP corpus. The researchers focused on ۸۰۰ words with top frequency in each corpus. As a result of this comparison the finding was that the most frequently used words in all three corpora of Medical Web, EMP, and YouTube are consistent to some extent and they were a consistency of ۶۱ percent between the YouTube corpus the MWC. These results were similar to that of Hsu (۲۰۱۸) who stated that knowing the first ۱۰۰۰ words of the Contemporary American English Corpus (COCA) and the British National Corpus (BNC) equals about ۹۸% of the traditional Chinese medicine texts. Our findings are also in agreement with Le and Miller (۲۰۲۰) who found that about ۸۰ percent of the Medical Web Corpus words are frequently repeated in the medical literature.

The consistency also indicates that high-frequency words in medical videos, which are considered authentic material, are taught to the medical students in Iran via EMP books. Therefore, the results of our research are in contrast to the findings of Hekmati and Navidinia (۲۰۱۶) and Khalili, and Tahririan (۲۰۲۰) who describe English language textbooks in medicine as non- authentic. It seems that the problem of these books is not the authenticity of the words, but it can be the way of designing words and structures for learning or some other things. For example, Hekmati and Navidinia (۲۰۱۶) stated that teaching English in Iranian universities is mostly based on the ‘grammar translation’ method and the textbooks for these courses have been written accordingly. As another instance, the students under study in Khalili & Tahririan (۲۰۲۰) stated that one of their problems with terminology was the lack of teachers’ knowledge about specialized words. And another problem with words was that some of them were listed and introduced out of context in their textbooks. They also stated that they were not taught the pronunciation of words well, and this leads to less vocabulary learning.

Another finding was that some of the most frequent words in the YouTube corpus were not found in the EMP corpus. And after administering a questionnaire to medical students, it was found that most students already know about ۸۹ percent of these words. So, just ۱۱ percent of those words that were typical of YouTube corpus were unfamiliar to the students and ۴۳ percent of these words were considered as unimportant by the students. Since many students already knew these words, without being mentioned in English medical textbooks, it can be concluded that students had studied these words in other medical textbooks and used these words in other specialized medical courses. Considering the earlier finding of the authenticity of textbooks in terms of words and also the statements of the students regarding the words that were typical of YouTube corpus, but were absent in EMP corpus, the authors of the study believe that the lack of proficiency of medical students in communicating efficiently is not related to field-specific words but a deficiency in general and semi-specific words. However, since the study was limited in a number of ways as will be mentioned in the conclusion, the results from this analysis should therefore be treated with considerable caution.

Conclusion

This study aimed at examining the textbooks that are common for teaching English in the medical universities of Iran. For this purpose, two corpora were created; one corpus was prepared from ٦ books that are taught in English for medical courses in Iran universities, and in this research, it was called the English in Medical Program (EMP) corpus, and another corpus was prepared from medical video clips from YouTube, and it was called YouTube Medical Corpus. The findings of this study showed that the corpora prepared in this study have a high consistency with the Medical Web corpus (MWC) and this indicates that these corpora can be used as valid material in teaching English to medical students. Thus, English teachers at medical schools can use these corpora. Also, they can be used as sources for textbook writers and researchers.

As another finding of the study, the three corpora prepared from textbooks, video clips, and Medical Web are largely consistent. This finding supports the idea that the words used in the English for medicine textbooks have good authenticity. However, students' dissatisfaction with the English medical textbooks, mentioned in the works of previous researchers, indicates that there are still problems in these books that need to be further investigated.

The last finding was that some of the most frequent words in the YouTube corpus were not found in the EMP corpus. And this result led the researchers of this study to find out if students knew the words. After investigating, it was found that medical students already knew most of these words. Therefore, it can be concluded that the authors of English medical textbooks probably did not feel the need to include these words in the textbooks. The students were not familiar with Only ١١ percent of these words which can be added to the text of the English for medicine books to achieve a more complete collection.

Further qualitative information in this study indicates that there are still problems with Medical English books that need to be addressed like they are not attractive enough, there are no pictures in the books, their texts are too hard, sentences structures are difficult for students, grammar in these books is not well presented, there are not enough practical exercises for words and grammar in these books, and video and audio files are not used in language classes. It seems that inclusion of sample conversations in these books between the patient and the doctor with audio or video files as well as using attractive pictures that are related to the words of each lesson, teaching the words in the form of authentic texts not separate lists, teaching grammar in a practical way, and including more practical exercises is a positive step to address students' needs.

It also seems that medical students have problems with general English because the English language they learned in high school was not practical. And they need more general language to better understand English texts and communicate with English-speaking patients. Therefore, it is suggested that more general language courses be established for this group of students and then they study specialized English.

This study was limited in that the videos were taken only from YouTube. Further related videos from other sources can be explored in future research. This can add to the richness of the YouTube corpus and introduce it as an internationally recognized corpus and can contribute to scholars as an oral corpus alongside other corpora from written resources. The second limitation is the number of books reviewed. In the next studies, it could be possible to examine all medical English books in all medical universities of the country. And finally, the qualitative information was collected from a sample of medical students from two universities. The inclusion of students of other universities and even the students of other fields of study in different universities of medical sciences may yield different and more reliable results.

References

- Baker, P. (۲۰۰۶). *Glossary of corpus linguistics*. Edinburgh University Press.
- Belcher, D. (۲۰۰۹) What ESP is and can be: An introduction. In D. Belcher (ed.), *English for Specific Purposes in Theory and Practice ۱ – ۲۰*. Ann Arbor, MI: University of Michigan Press.
- Boniadi, Alireza; Ghojzadeh, Morteza; Rahmatvand, Neda. (۲۰۱۳). Problems of English for Specific Purpose course for medical students in Iran. *Khazar Journal of Humanities & Social Sciences*. Vol. ۱۶ (۱), ۴۸-۵۵.
- Boulton, A. (۲۰۱۶). Integrating corpus tools and techniques in ESP courses. *ASp. la revue du GERAS* (۶۹), ۱۱۳-۱۳۷.
- Coats, S. (۲۰۱۹). A corpus of regional American language from YouTube. In Navarretta, C. et al. (Eds.), *Proceedings of the ۴th Digital Humanities in the Nordic Countries Conference, Copenhagen, Denmark, March ۶-۸, ۲۰۱۹* (pp. ۷۹-۹۱). Aachen, Germany: CEUR.
- Del Mar Sánchez Ramos, M. (۲۰۲۰). Teaching English for Medical Translation: A Corpus-Based Approach. *Iranian Journal of Language Teaching Research*, ۸(۲), ۲۵-۴۰.
- Dudley-Evans, T., St John, M. J., & Saint John, M. J. (۱۹۹۸). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge university press.
- Flowerdew, L. (۲۰۱۱). *Corpora and language education*. Springer.
- Hekmati, N., & Navidinia, H. (۲۰۱۶). English for specific purposes: Proposing an innovative approach to teaching English to medical students. *Future of Medical Education Journal*, ۶(۱), ۱۹-۲۴.
- Hsu, W. (۲۰۱۸). The most frequent BNC/COCA mid-and low-frequency word families in English-medium traditional Chinese medicine (TCM) textbooks. *English for Specific Purposes*, ۵۱, ۹۸-۱۱۰.
- Hunston, S. (۲۰۰۲). *Corpora in applied linguistics*. Ernst Klett Sprachen.
- Ibrahim, H. H. (۲۰۲۰). Needs Analysis as a Prerequisite for Designing an ESP course for Medical Students. *Open Journal of Modern Linguistics*, ۱۰(۲), ۸۳-۱۰۳.
- Kacatl, J., & Frydrychova-Klimova, B. (۲۰۱۵). English vocabulary in video clips on travel and tourism. *Procedia-Social and Behavioral Sciences*, ۱۸۲, ۳۶۴-۳۶۸.
- Khalili, S., & Tahririan, M. H. (۲۰۲۰). Deciphering Challenges of Teaching English for Specific Purposes to Medical Students: Needs, Lacks, Students' Preferences, and Efficacy of the Courses. *Teaching English Language*, ۱۴(۱), ۳۶۵-۳۹۴.
- Le, C. N. N., & Miller, J. (۲۰۲۰). A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. *English for Specific Purposes*, ۵۸, ۱۰۲-۱۲۱.
- Mattiello, E. (۲۰۱۹). A corpus-based analysis of scientific TED Talks: Explaining cancer-related topics to non-experts. *Discourse, Context & Media*, ۲۸, ۶۰-۶۸.

- McEnery, T., Xiao, R., & Tono, Y. (۲۰۰۶). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
- Nezakatgoo, B., & Behzadpoor, F. (۲۰۱۷). Challenges in teaching ESP at medical universities of Iran from ESP stakeholders' perspectives. *Iranian Journal of Applied Language Studies*, ۹(۲), ۸۹-۸۲.
- Otto, P. (۲۰۲۱). Choosing specialized vocabulary to teach with data-driven learning: An example from civil engineering. *English for Specific Purposes*, ۶۱, ۳۲-۴۶.
- Paltridge, B., & Starfield, S. (۲۰۱۴). *The handbook of English for specific purposes*. John Wiley & Sons.
- Pavel, E. (۲۰۱۴). Teaching English for Medical Purposes. *Bulletin of the Transilvania University of Braşov Series VII: Social Sciences, Law*, volume ۷, p. ۳۹-۴۶.
- Phoocharoensil, S. (۲۰۱۲). Language corpora for EFL teachers: An exploration of English grammar through concordance lines. *Procedia-Social and Behavioral Sciences*, ۶۴, ۵۰۷-۵۱۴.
- Reppen, R. (۲۰۱۰). *Using corpora in the language classroom*. Cambridge University Press.
- Rezvani, R., Gholtash, A., & Zamani, G. (۲۰۱۶). The First Corpus-Based Persian Academic Word List: Development and Pedagogical Implications. *Journal of Teaching Persian to Speakers of Other Languages*, ۵, ۴۳-۶۴.
- Vosoughi, M., Sharifabad, E. D., & Raftari, S. (۲۰۱۳). Investigating current status of English for academic purposes (EAP) in Iran; Revisiting ESP specificity continuum. *International Journal of Applied Linguistics and English Literature*, ۲(۳), ۵۴-۶۱.
- Zhu, C., & Wang, H. (۲۰۱۱). A corpus-based, machine-aided mode of translator training: ClinkNotes and beyond. *The Interpreter and Translator Trainer*, ۵(۲), ۲۶۹-۲۹۱.