INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 11 No. 3, January 2022, pp. 577-588

Available online at:
https://ejournal.upi.edu/index.php/IJAL/article/view/37103

https://doi.org/10.17509/ijal.v11i3.37103

Establishing a COVID-19 lemmatized word list for journalists and ESP learners

Hadeel Saed[1], Riyad F Hussein[1], Ahmad S Haider[1*], Saleh Al-Salman[1], and Iyad M. Odeh[2]

[1]English Language and Translation Program, Department of English Language and Translation, Faculty of Arts and Science, Applied Science Private University, Al Arab St 21, Amman, Jordan
[2]Quality and Training Department, Islamic Hospital, Al-Shariaah College St. 2, Amman, Jordan

ABSTRACT

The aim of this research is two-fold; first, to explore the most frequent COVID-19 inspired words in medical news reporting contexts, and second, to classify them into different categories. This paper adopts a corpus-based approach to build a lemmatized academic word list (AWL) inspired by the COVID-19 pandemic. Factiva was used to retrieve the pandemic-related articles published in News Rx from January 1 - October 31, 2020. A total number of 18,249,093-word corpus was compiled. The corpus linguistic software program Wordsmith (WS-6) (Scott, 2012) was used to generate a word list based on the complied corpus. Subsequent to compiling, lemmatizing, and analyzing the AWL, six categories were identified, namely, acronyms and abbreviation, diseases, COVID-19, biology, medicine, and scientific disciplines, all of which are of essential use for media workers, ESP learners of journalism, medicine, nursing, pharmacy, and allied health sciences. Building such a discipline-specific glossary will be of special pedagogical value for health journalists, textbook writers and curriculum designers, instructors, and ESP learners in the health sciences field. One of the major contributions of this research is establishing lemmas of a large set of AWL. This set can be utilized by news media workers, health communication specialists, and ESP learners. Lemmatization will ensure rapid dissemination of the word list and its integration in the linguistic system through derivation and other word-formation processes.

Keywords: COVID-19; ESP; journalism; lemmatization; pedagogy

First Received:	**Revised:**	**Accepted:**
27 July 2021	30 December 2021	10 January 2022
Final Proof Received:		**Published:**
18 January 2022		30 January 2022

How to cite (in APA style):
Saed, H., Hussein, R. F., Haider, A. S., Al-Salman, S., & Oded, I. M. (2022). Teachers' attitude towards translanguaging practice and its implication in Indonesian EFL classroom. *Indonesian Journal of Applied Linguistics, 11*(3), 577-588. https://doi.org/10.17509/ijal.v11i3.37103

INTRODUCTION

Interest in building specialized datasets and accompanying research in English for Specific Purposes dates back to the 1960s when ESP studies emerged. Ever since ESP "has established itself as a viable and vigorous movement within the field of TEFL/TESL. It has become international in scope and specific in focus" (Johns & Dudley-Evans, 1991, p. 297). The discipline has been developing ever since to establish itself in teaching-learning pedagogy. ESP has been booming in the last two decades as an approach and a discipline (Ramírez, 2015). The empirical nature and interdisciplinarity of the ESP research have become evident since its early beginnings (see Hsu, 2011; Loong & Chan, 2012). With the English language being the global lingua franca, English for specific purposes continued to thrive over the years to assume a leading role in language studies (Williams, 2014). It is noteworthy that the ESP literature has always been abreast of the academic word lists in different disciplines. These include civil engineering (Gilmore & Millar, 2018),

* Corresponding Author
Email: Ah_haider86@yahoo.com

nursing (Yang, 2015), environment (Liu & Han, 2015), research writing (Hussein et al., 2021), finance (Tongpoon-Patanasorn, 2018), and art (Wang, 2017). This study discusses the need for creating field-specific word lists associated with medical news reporting in the post-COVID-19 era.

With the massive effects of the novel coronavirus pandemic, the international community has been harshly stricken by its devastating effects indiscriminately. Such a global disaster inspired researchers in all disciplines to respond, allowing ESP specialists to probe into the short-term and long-term impact of COVID-19 on the academic community. Research endeavors have been accelerated ever since to cover aspects of health communication, medical news reporting, economy, psychology, human well-being, among others.

Given the value of EAP/ESP in present-day academic research and emanating from the global impact of COVID-19 on our daily life, the current study focuses on the importance of compiling a coronavirus academic word list in the post-COVID-19 era. In more specific terms, the coronavirus crisis has impacted all aspects of life (Al-Salman & Haider, 2021b; Almahasees et al., 2021; Haider & Al-Salman, 2020), requiring close follow-up and coverage by international and local media outlets (Al-Salman & Haider, 2021c).

The multi-faceted and interdisciplinary domain of journalism tackles the sweeping spread of the COVID-19 pandemic. According to The United Nations (2020), journalists worldwide have thrived on the COVID-19 unprecedented spread providing a news story for them. In the same vein, The World Health Organization (2020) stressed the vital role of journalists in disseminating information about the vaccine and its effectiveness in this exceptional era of humankind history.

On the roles of journalism's 24/7 wide coverage of the COVID-19 pandemic and its devastating effects worldwide, different international bodies valued the relentless efforts made by journalists and news reporters in international news outlets.

For this to be conducted efficiently, journalists who do not have a background in the medical field will most likely find the word list compiled in this study of paramount importance as it provides them with the specialized terminology used in this particular field.

The significance of the study derives from its in-depth search to create a specialized word list extracted from a large corpus of 18,249,093 medical terms as tokens/running words, which have been short-listed in 550 post-COVID-19 most recognized medical news terms. Therefore, the study will be of special value for news media reporters and journalists, health communication specialists, students of medicine, textbook writers and curriculum designers, instructors, and ESP specialists in the medical field. Additionally, it is one of the first studies to establish a COVID-19 repository of specialized vocabulary that would help medical news reporters and ESP learners use this varied lexical stock effectively in a sub-area of the medical and communication fields.

In sum, the study sets to address the following two research questions:

1. What are the most frequent COVID-19 inspired academic words in the compiled medical news reporting corpus?
2. How can these terms be classified into different categories according to domains?

REVIEW OF LITERATURE

The act of determining the most relevant academic lists in English for academic purposes is both demanding and challenging (Coxhead, 2000). Compiling a short list of the most useful technical terms in a certain field is vital, but it is an intricate process that requires in-depth knowledge, experience, and academic expertise. This approach has become increasingly significant due to the growing interest in ESP and EAP as a pedagogical resource in the last few decades. Citing West (1953), Coxhead (2000, p. 213) reports that "the General Service List (GSL) developed from a corpus of 5 million words with the needs of ESL/EFL in mind, contains the most widely used 2,000-word families in English". Typically, tens of millions of tokens/running words are screened to identify a smaller list of technical vocabulary (i.e., running words), which will subsequently be reduced to a 1000-word list. Csomay and Prades (2018) argue that the knowledge of core academic vocabulary is a key element to academic success as they provide learners with a glossary of words that can enhance their knowledge base. To improve the academic vocabulary knowledge among students, Green (2019) compiled an academic word list for grammar, showing the interaction between lexis and grammatical patterns known as lexicogrammar. On the other hand, Wang et al. (2008) reported that isolating the most frequently used medical academic word list (MAWL), which is based on a corpus of 1093011 running words of medical research articles, may be utilized in curriculum design for course materials and medical terminology.

According to Green and Lambert (2018), discipline-specific wordlists, collocations, and word families are bound to boost students' understanding at the pre-tertiary level of the variation of English across different disciplines. The use of collocations in corpus-based language learning research was addressed by Gablasova et al. (2017), who investigated how two words in a corpus come as collocates. On a different health-related level, Yang (2015) stressed the importance of creating an academic word list in the field of nursing to improve nursing students' reading and writing skills, together

with helping instructors and researchers design teaching material.

Technical words, which are used in specialized disciplines, vary from one field to another. While they may be very common in one subject area, they may not be as common in another discipline. Low-frequency words cover about 5% of the running words in an academic text (Yang, 2015). As reported by Durrant (2016), low-frequency word lists pose a challenge for EAP students, which makes it necessary for researchers to maximize their corpus-based research endeavors to produce various types of academic word lists which will enhance students' academic vocabulary learning. Some scientific research articles' word lists are created from the frequently used words in titles and abstracts of research articles (Masaya, 2020). In a pharmacy academic word list (PAWL) compiled from a corpus of 3,458,445 tokens, Heidari et al. (2020) pointed out that specialized word lists in a given discipline can be utilized as a reference for designing EAP course syllabi for foreign students.

According to Sarré and Whyte (2017), instructors having modern foreign language training discuss the language needs of engineers, lawyers, and doctors to develop students' vocabulary lists and glossary in the EFL classrooms. The new developments in ESP teaching and learning research have contributed to building multi-disciplinary academic word lists to serve the needs of language learners globally.

This conflation of research and pedagogical practice in ESP has significantly enriched the lexical storage available to students and researchers in different disciplines (Abu Melhim, 2013; Johns, 2013). Further developments in ESP were highlighted by Negro Alousque (2016), who reported on the evolution of research in ESP to be expanded to include genre analysis in theoretical and applied sciences. The interaction between theoretical and applied linguistics is conducive to enriching ESP research through building academic word lists, which will enhance the students' academic performance in various areas of specialization. Identifying and teaching target vocabulary in ESP courses was investigated by Nekrasova-Beker et al. (2019), who stressed the need for establishing field-specific specialized vocabulary to help students understand the content and interact more effectively in academic discussions pertaining to a specific discipline. Establishing specialized corpora for teaching vocabulary is necessary for pedagogical and research purposes by using corpus-based techniques, a trend which has been adopted in different disciplines including engineering, business, language and linguistics, information technology, nursing, pharmacy, and medicine, among others. Adequate knowledge of specialized materials requires creating academic word lists, reduced to a smaller list of unique words/types, and ending up with a short list of specialized terms. Hsu (2014) maintained that achieving reasonable comprehension of engineering textbooks, L2 students should be familiar with approximately 5000 word families in addition to proper nouns and word abbreviations.

New word lists inspired by the COVID-19 pandemic were found in the work of Akut (2020), who analyzed the morphological structure of COVID-19 generated neologisms. In the same vein, Roig–Marín (2020) reported that a word list of more than COVID-19- related neologisms have gained currency as they have widely spread during the pandemic. The new lists have undergone some morphological changes of word-formation processes, including compounding, blending, clipping, affixation, and acronyms. Likewise, many of the emerging neologisms have been borrowed by other languages as loanwords or calques (i.e., loan translations). Incidentally, many common COVID-19 terms such as vaccine, virus, ventilators, quarantine, pandemic, social/physical distancing, closures, lockdowns, among others, are pre-existing words that have been in use for a long time; however, they have been more frequently used by non-specialists during the coronavirus outbreak due to their immediate association with the disease. More importantly, these expressions have taken a new turn by combining with other words to form new expressions such as corona babies, covidiots (Roig–Marín, 2020).

METHOD

For the purpose of this study, we compiled a corpus of 18,249,093 words from 36,365 medical news articles. The corpus of this study was based on different criteria, including the type of articles text source (Medical news articles), text source (NewsRx), text topic (COVID-19), text language (English), and time span (January 2020- October 2020). Our choice of medical news articles to establish academic word lists may be attributed to two reasons. First, the compiled medical words address a mixture of specialized and non-specialized audience, and second, medical news may have a higher circulation rate when compared to academic journal articles. NewsRx is selected because the main interest of this paper is not only to generate an academic world list to be used by ESP students and academics but also by news media workers, health communication specialists, and journalists who write widely on the topic but do not have enough background in the medical jargon in general and the COVID-19 inspired terminology in particular.

Why NewsRx?

NewsRx has been chosen as a news source since it is one of the most pronounced media outlets that publish health news. Our choice is based on its outstanding record and leading role as a media outlet

and technology-based company focusing on digital media, printed media, and health news. Being known for providing updates on ongoing research and discoveries through selecting relevant and trustworthy information, News RX remains a reliable database for daily and weekly medical news and pharmaceutical research. With over 7000 articles published a day, News Rx is a world-renounced database and a reliable source adhering to strict standards of accuracy and objective reporting style. In light of the above, we settled on the News Rx data, marked with the seal of quality, as the only source of information for the current piece of research.

Another reason for our choice of News Rx is the nature of our paper which reflects a wide range of diversity at two levels: (1) the varied compiled corpus and its categorization, and (2) the wide spectrum of readership with a genuine interest in the content of the paper. Thus, the comprehensive orientation of the paper seeking to establish a COVID-19 lemmatized word list in different domains is of special interest to the ESP readership in all COVID-19 related categories. This diversity in the paper focus and readership can be easily accommodated in the News Rx database with its interdisciplinary coverage of media, technology, health news, business, knowledge, and discoveries globally.

Data Collection and Corpus Compilation

The corpus of data for this study was compiled from scratch using the Factiva news database. Factiva, produced by Dow Jones, provides users with access to browse through different global content sources in different languages and enables them to conduct structured searching (Haider, 2019). The reason for selecting Factiva as the data gathering source stems from the fact that it provides users with a wide range of information from "more than 32,000 sources from nearly every country worldwide in 28 languages, including over 450 continuously updated newswires" (Al-Abbas & Haider, 2020, p. 315).

Since this study aims to establish a COVID-19 related AWL, two core query terms were used, namely corona and COVID. The choice of these two terms is important given the clearly defined purpose of building the corpus.

Looking for the following query terms (corona OR COVID) in NewsRx using Factiva between (1/1/2020) and (31/10/2020) resulted in 36,365 articles with a total word count of 18,249,093 (Table 1).

The corpus of the current study has been processed and dealt with as an integrated whole to cover the period of 10 months. In other words, the authors adhered to an overall diachronic approach rather than a synchronic one to establish a homogenous number of monthly articles. Accordingly, the researchers did not pay much attention to balance, which is defined as having equal proportions of the different sub-corpora (Hunston, 2008). The researchers collected all articles that met the corpus compilation criteria irrespective of their number each month, which vary according to development in the spread of the virus and health measures taken, and so on.

Table 1
The Size of the Compiled Corpus

Month	Number of Articles	Running words
January	41	23,213
February	176	111,978
March	1097	570,137
April	4792	2,253,980
May	7497	3,630,422
June	6064	3,106,146
July	5797	2,988,999
August	6027	3,108,208
September	1481	722,860
October	3393	1,733,150
Total	36365	**18,249,093**

Therefore, the variation in the number of articles came as a reflection of newsworthiness and facts on the ground about the virus. In other words, it is quite normal to have a small number of articles discussing COVID-19 at the very beginning of the crisis in January when compared to soaring numbers in May with a spike in cases. For example, in May, the number of articles was 7497 (3,630,422 words) compared to January with 41 articles (23,213 words). Categorizing the data according to the month when the article has been published gives a representation of the data as it happened and how it impacted the international community during that specific one-month span. However, our findings represent the total and comprehensive word list covering the 10-month period holistically.

Analytical Software

In this paper, we use Wordsmith 6.0 (WS6) (Scott, 2012), which is a corpus linguistic software developed by Mike Scott. Using the software, researchers can carry out frequency analysis, concordance (KWIC) analysis, collocation analysis, cluster analysis, keyword analysis, among others. In this study, we mainly use the frequency tool to generate a list of all words that occurred in the corpus under investigation.

FINDINGS & DISCUSSION

Word Frequency

To answer the first research question related to the most frequent COVID-19 inspired academic words in the 18,249,093-word corpus, we used Wordsmith 6 (WS6) to generate a word list of all words in the

corpus along with their frequencies. It is worth noting that we used a stop list to filter out function/grammatical words such as articles, prepositions, conjunctions, pronouns, among others. We have also removed the general English words manually. A lemma list was also used to combine related words with their roots (see Figure 1).

Figure 1

Lemmatization of the Words' Virus' and 'Medicine'

Lemma Forms

variants	frequency
VIRUS	21,993
VIRUSES	41,583
VIRAL	40,726
VIROLOGIC	42
VIROLOGICAL	156
VIROLOGICALLY	9
VIROLOGIST	51
VIROLOGISTS	44
VIROLOGY	35,449
VIRUSTATIC	2
NIDOVIRALES	5,514
ANTIVIRAL	3,332
ANTIVIRALS	734

Lemma Forms

variants	frequency
MEDICINE	38,605
MEDICINES	721
MEDIC	5
MEDICS	17
MEDICAL	95,771
MEDICALS	2
MEDICAMENT	26
MEDICAMENTS	2
MEDICATION	703
MEDICATIONS	954
BIOMEDICINE	447
BIOMEDICAL	1,115
BIOMEDICALE	2
BIOMEDICALLY	3
BIOMEDICALS	1
BIOMEDICAS	1
BIOMEDICHE	2
BIOMEDICINES	1
BIOMEDOMICS	3
TELEMEDICINE	3,090
MEDICARE	594

Figure 1 shows the most 20 frequent words and their lemmas. The most frequent words in the corpus are virus 149,635, medicine 142,065, corona 119,776, health 91,258, FDA 69,456, and CDC 68,940. The most salient lemmas of virus are viruses 41583, viral 40726, and virologic 42, and the most noticeable lemmas of medicine are medical 95771, medication/s 1657, Medicare 594, and medic/s 22. The most frequent lemmas of health are healthcare, 12448, healthy 2168, and telehealth 1411

We considered the 550 most frequent words in the corpus after excluding function and general English words (see Excel sheet 1 https://data.mendeley.com/datasets/j8256h483m/1). Figure 2 below shows the 20 most frequent words in the corpus.

Figure 2

A Wordlist of the 20 Most Frequent Words with Their Lemmas

N	Word	Freq.	Lemmas
1	virus	149,635	virus[21993] viruses[41583] viral[40726] virologic[42] vi
2	medicine	142,065	medicine[38605] medicines[721] medic[5] medics[17] me
3	corona	119,776	corona[1307] coronas[4] coronavirus[110038] coronaviru
4	health	91,258	health[74955] healthcare[12448] healthy[2168] healthier
5	fda	69,456	
6	cdc	68,940	
7	covid	67,347	covid[62714] cov[4633]
8	patient	67,156	patient[9470] patients[56792] patience[17] outpatient[87
9	disease	52,631	disease[34544] diseases[18087]
10	infect	51,778	infect[588] infected[6731] infecting[252] infects[162] infe
11	respire	47,791	respire[0] respiring[1] respir[4] respirable[3] respirating[
12	severe	46,423	severe[42395] severer[6] severest[2] severance[25] severi
13	pandemic	45,559	pandemic[44056] pandemics[1503]
14	risk	43,335	risk[40212] risked[3] risking[56] risks[2140] high-risk[92
15	deliver	42,040	deliver[36764] delivered[810] delivering[827] delivers[38
16	rna	41,957	
17	acute	37,156	
18	hospitalize	35,226	hospitalize[6] hospitalized[2729] hospitalizes[2] hospital
19	syndrome	35,116	syndrome[34930] syndromes[186]
20	clinic	34,856	clinic[0] clinical[29760] clinic[1783] clinics[882] clinicall

Figure 2 includes acronyms such as FDA, CDC, and RNA, which can be easily recognized not only by people in English-speaking countries but also by medics and ESP students in the medical profession all over the world. These acronyms were recurrent on different media outlets millions of times and have thus become internationalized. Other words in Figure 2 which should not go unnoticed are patient, infect, respire, and pandemic because of their direct relevance to COVID-19. Indeed, the high

frequency of these words did not come as a surprise due to their paramount importance in the context of COVID-19. This necessitates their acquisition by allied health personnel who are interested in understanding COVID-19-related issues or writing about them.

Corpus Categorization

Subsequent to generating a list of 550 most frequent words in the corpus, each of the four authors of the current research has individually been tasked with classifying the list into categories according to themes. Five thematic categories were shared by all, whereas one additional category, namely scientific disciplines, was suggested by three of the four authors. Consequently, we ended up with six categories (https://data.mendeley.com/datasets/j8256h483m/1). These include: acronyms and abbreviations (see Excel sheet 2), diseases (see Excel sheet 3), COVID-19 (see Excel sheet 4), biology (see Excel sheet 5), medicine (see Excel sheet 6) and scientific disciplines (see Excel sheet 7). Of the six categories, it is worth noting that the category titled 'scientific disciplines' was derived directly from the wordlist of the original corpus (outside the lemmatized 550 most frequent words) by considering the words that end with the suffix'_logy.'

In the subsections below, we discuss the 20 most frequent words in the six categories except the category of diseases which comprised only 15 words.

Acronyms and Abbreviations

This category comprises COVID-19 related acronyms and abbreviations (see Excel sheet 2 for the full list https://data.mendeley.com/datasets/j8256h483m/1). Acronyms and abbreviations can be defined as shortened forms of longer expressions. Table 2 shows the 20 most frequent shortened expressions in the corpus.

Table 2

The 20 Most Frequent Acronyms and Abbreviations Related to COVID-19 in the Corpus

No.	Acronym and Abbreviation	Freq.	Full Expressions
1	FDA	69,456	Food and Drug Administration
2	CDC	68,940	Centers for Disease Control and Prevention
3	RNA	41,957	Ribonucleic acid
4	MLCF	22,126	Medical Leadership Competency Framework
5	TB	4,693	Tuberculosis
6	CT	3,391	computed tomography
7	ICU	2,727	intensive care unit
8	IGG	2,131	Immunoglobulin G
9	RT-PCR	2,124	Reverse transcription polymerase chain reaction
10	PPE	1,950	Personal protective equipment
11	PCR	1,815	polymerase chain reaction
12	HIV	1,501	human immunodeficiency virus
13	ARDS	1,479	Acute respiratory distress syndrome
14	AG	1,469	Antigen
15	DNA	1,359	Deoxyribonucleic acid
16	IGM	1,166	Immunoglobulin M
17	HCQ	936	Hydroxychloroquine
18	SOC	905	Standards of Care
19	RBD	896	Receptor Binding Domain
20	BV	862	Bacterial vaginosis

Table 2 shows that the most frequent acronyms and abbreviations in this category are FDA 69456, CDC 68940, RNA 41957, TB 4693, ICU 2727, RT-PCR 2124, HIV 1501, ARDS 1479, and HCQ 936. For instance, FDA's function is to protect public health by ensuring the safety of human drugs. CDC, on the other hand, is concerned with saving lives and protecting people from health hazards. The same applies to other acronyms in the corpus, such as ICU, RT-PCR, HIV, and HCQ. Intensive Care Units ICU provides health care for critically ill patients, as is the case with COVID-19 urgent cases. It can be argued that this virus is strongly related to ICU as quite a good number of coronavirus patients were admitted to it. RT-PCR is a technique that detects the virus; needless to say that tens if not hundreds of millions of people have taken this test in the last thirteen months or so. Finally, the acronym HCQ is the drug that people used to take in the early stages of COVID-19 to treat the virus and should therefore be part of ESP students' and specialists' active lexicon.

Diseases

This category includes disease-related words. Figure 3 shows the words included under this category along with their frequencies (see Excel sheet 3 for the full list https://data.mendeley.com/datasets/j8256h483m/1).

Figure 3
A Wordlist of the Category of 'Diseases'

N	Word	Freq.
1	disease	52,631
2	cancer	9,467
3	sars	7,227
4	pneumonia	5,531
5	ill	4,723
6	diabetes	4,691
7	influenza	3,642
8	obese	3,066
9	mers	2,501
10	diarrhea	1,183
11	allergy	1,007
12	ebola	802
13	asthma	784
14	malaria	659
15	hepatitis	622

There is a direct link between COVID-19 and other diseases. People with chronic diseases such as cancer, diabetes, arthritis, and asthma are the most vulnerable to COVID-19. Other ill people may have symptoms of COVID 19, which include respiratory problems and other signs of sickness. The most frequent words in this category are cancer 9467, SARS 7227, and pneumonia 5531. Diseases with relatively less frequency are ebola 802, asthma 784, malaria 659, and hepatitis 622. These diseases should be taken into account by people in the ESP field or students in schools of medicine, pharmacy, or nursing.

COVID-19

In this category, COVID-19 related words are listed. Figure 4 below shows the 20 most frequent words in the corpus (see Excel sheet 4 for the full list https://data.mendeley.com/datasets/j8256h483m/1).

Figure 4
A wordlist of the 20 most frequent words in the COVID-19 category

N	Word	Freq.	N	Word	Freq.
1	virus	149,635	11	epidemy	29,889
2	corona	119,776	12	test	27,038
3	covid	67,347	13	immunize	20,723
4	infect	51,778	14	vaccine	18,808
5	respire	47,791	15	spread	11,844
6	severe	46,423	16	transmit	10,501
7	pandemic	45,559	17	try	10,185
8	acute	37,156	18	novel	9,170
9	syndrome	35,116	19	distance	8,239
10	contact	31,815	20	screen	7,680

Figure 4 shows that the words virus 149635, corona 119776, and COVID-19 67347 were the most frequent in this category. Words with relatively less frequency in this category are transmit 10501, novel 9170, distance 8239, and screen 7680. These words reflect activities related to COVID-19. For instance, COVID-19 can be transmitted from human to human, and the word distance and social distancing are commonly used these days worldwide to prevent the spread of coronavirus. Other equally important words in this category include respire, pandemic, epidemy, immunize, and vaccine. Some of the lemmas of respire are respiratory 46601, respirators 719, respirator 340, respiration 113, and respirations 2. Some of the lemmas of epidemy are epidemiology 17293, epidemics 8527, epidemiological 2297, epidemiologic 280, epidemiologist/s 198, epidemiologically 32. Listed vaccine lemmas are only vaccines and vaccinations, and immunize lemmas are immunology 4994, immunity 2633, immunogenicity 212, immunodeficiency 157, and immunobiology 34.

Biology

The sample list of 20 biological terms given in Figure 5 is only a small portion of a larger list of terms that fall under this category (see Excel sheet 5 for the full list https://data.mendeley.com/datasets/j8256h483m/1).

Figure 5
A Wordlist of the 20 Most Frequent Words in the 'Biology' Category

N	Word	Freq.	N	Word	Freq.
1	cell	16,441	11	enzyme	3,795
2	protein	12,518	12	biotech	3,792
3	body	11,869	13	peptide	2,798
4	gene	9,022	14	species	2,630
5	lung	7,548	15	tissue	2,316
6	biology	6,069	16	plasma	2,305
7	molecule	5,612	17	kidney	2,298
8	tract	4,968	18	liver	2,264
9	genome	4,398	19	chest	2,241
10	microbe	3,983	20	lymph	2,092

Words such as cell 16441, protein 12518, body 11869, and gene 9022 are indicative of a close relationship between biology and coronavirus. Relatively less frequent words in this category are lungs, and molecule with a frequency count of 7548 and 5612, respectively. Other biologically-related words with a lower frequency include liver 2264, chest 2241, and lymph 2092. Furthermore, a representation of the biologically-based terms through lemmas shows a noticeable exemplification of lemmatization where productive word-formation derivational processes have been identified. For example, the word gene in this category is relatively high, and some of its lemmas are genetics 3010, gene 2266, genetic 1955, genes 1617, and genetically 173. Other terms such as microbe and lymph reflected a high level of lemmatization processes.

Medicine

Medicine-related terms are by far the richest in all categories under investigation (see Excel sheet 6 for

the full list https://data.mendeley.com/datasets/j8256h483m/1). Figure 6 shows the 20 most frequent words in this category.

Figure 6
A Wordlist of the 20 Most Frequent Words in the 'Medicine' Category

N	Word	Freq.	N	Word	Freq.
1	medicine	142,065	11	therapy	21,087
2	health	91,258	12	drug	16,696
3	patient	67,156	13	symptom	16,107
4	risk	43,335	14	condition	14,509
5	hospitalize	35,226	15	diagnose	12,311
6	clinic	34,856	16	die	11,869
7	prevent	34,256	17	protect	9,797
8	care	25,440	18	pharm	9,686
9	case	24,718	19	mental	9,480
10	treat	22,321	20	emergent	7,603

Figure 6 shows that the word medicine occurred 142065 times. The lemmas related to this word include: medical 95771, medicine 38605, medications 954, medicines 721, medication 703, medicament 26, medics 17, medic 5, medicals 2, medicaments 2, among others. The word health shows a high frequency of 91,528, and its lemmas health 74955, healthcare 12448, healthy 2168, telehealth 1411, healthier 271, and healthiest 5. This massive bulk of medical terminology with a remarkably high count of frequencies is indicative of the inter-relatedness of medicine with COVID-19.

Scientific Disciplines

This particular domain is viewed as the knowledge base for students and specialists in medical sciences and allied health sciences (see Excel sheet 7 for the full list https://data.mendeley.com/datasets/j8256h483m/1). Figure 7 shows a sample list of the 20 most frequent words in the corpus representing this domain.

Figure 7
A Wordlist of the 20 Most Frequent Words in the 'Scientific Disciplines' Category

N	Word	Freq.	N	Word	Freq.
1	medicine	142,065	11	therapy	21,087
2	health	91,258	12	drug	16,696
3	patient	67,156	13	symptom	16,107
4	risk	43,335	14	condition	14,509
5	hospitalize	35,226	15	diagnose	12,311
6	clinic	34,856	16	die	11,869
7	prevent	34,256	17	protect	9,797
8	care	25,440	18	pharm	9,686
9	case	24,718	19	mental	9,480
10	treat	22,321	20	emergent	7,603

Figure 7 shows that virology is the most directly related science to coronavirus with a frequency of 35525, followed by epidemiology 17306 and immunology 5028. Interestingly all other sciences down the list are inherently linked to the COVID-19 clinical management. This is particularly important in coronavirus-inflicted patients with a record of pre-existing diseases which may lead to life-threatening complications, where sciences like biotechnology 1899, microbiology 1733, cardiology 1475, pathology 1003, pharmacology 936, and pulmonology 898 come into play. In sum, compiling such a varied academic word list is an added value for ESP medical students.

DISCUSSION

A distinctive feature of the AWL compiled in this research is its emphasis on the lemmas as derivatives of key words of direct relevance to the COVID-19 pandemic. The fact that different word-formation processes are used to generate nouns, adjectives, compounds, and blends will enrich the ESP lexical repository for medical students, lab technicians, nursing staff, and other allied health sciences personnel. These lemmas are also directed to advanced ESP experts, specialists, and graduate students who embark on reading or writing advanced papers, reviews, or reports on COVID-19. Their knowledge of these lemmas can help ESP learners overcome difficulties in comprehending and using academic words.

Most acronyms and abbreviations that appeared in the word list are used by millions of doctors and Coronavirus specialists across the world. Their relevance to ESP specialists and students' knowledge cannot be underestimated because, for people to understand the content of what they read, they should know these forms and their full meaning. Knowledge of acronyms or abbreviations enables COVID-19 specialists to read and write on the topic. They should therefore become part of the word stock of ESP specialists or advanced students in the field of medicine, pharmacy, or nursing.

It is well known that there is a relationship between COVID-19 and other diseases, which determines the severity or magnitude of the virus on various people proportionate to age and pre-existing diseases. Consequently, symptoms on affected people vary from mild headache to high fever and severe pulmonary complications. This makes this particular category of direct relevance to ESP students in the medical-allied health sciences.

The corpus includes words of direct bearing on COVID-19, which constitutes the core of this research. Doctors, as well as medical and pharmaceutical researchers, use some words interchangeably to refer to the same disease. So, ESP students must know these words and not hesitate to

use corona to mean virus and virus to mean COVID-19 or corona to mean COVID-19.

With its definition, biology is a natural science that studies living organisms, their physical structure, genetic relations, development, and evolution; this category has an immediate bearing on the COVID-19-generated terms and their lemmas. Biology-related terms which have direct relevance to the COVID-19 pandemic are markedly significant to ESP students in the medical field, especially those of medicine and nursing, and bio-medical laboratory specialists, among others.

With COVID-19 being a life-threatening global pandemic, its effects are diverse and often unpredictable. Consequently, a wide spectrum of medical terms associated with different medical-support institutions, research centers, hospitals, pharmaceutical industries, drugs, among others, have a big share in this discipline.

The corpus of medical terminology can enrich the ESP medical repository, which will consequently boost the medical students' academic word lists. Furthermore, this word list, along with others, can be shared by health specialists and workers globally.

The academic value of the scientific-discipline-related terms derives from their being a knowledge base for students and specialists in the medical sciences and related fields. As such, these words can be an integral part of the ESP medical students' terminology. It is well-known that sciences such as physiology, anatomy, pathology, microbiology, biochemistry, and genetics are closely related to coronavirus. Therefore, it is extremely important for ESP students not only to be familiar with these words but also to use them effectively. Consequently, a purely scientific discussion of COVID-19 should make reference to those sciences.

The present study, which provides a COVID-19 inspired (AWL), is expected to be of value for a wide range of ESP learners. Furthermore, it might have far-reaching pedagogical implications for textbook designers and researchers in medicine, pharmacy, nursing, and related disciplines. The corpus-based AWL emerging from the current study will not be limited to the medical lexical inventory but will include all other domains which the COVID-19 pandemic has hit. A very important discipline to contend with as a key player is journalism and the role of journalists and news reporters in the COVID-19 crisis. The present study will be extremely beneficial for stakeholders in this field, including media and communication students.

The multi-dimensional implications of the pandemic created new realities, which led to coining new words as neologisms (Al-Salman & Haider, 2021a). These coinages provide another representation of ESP in interacting with global affairs. The emergence of COVID-19-related words shows how the ESP paradigm is multi-disciplinary as it thrives on establishing world lists in all disciplines and domains which affect people's lives. They provide new glossaries and lexical inventory that fill gaps arising from the introduction of new concepts associated with a new state, development, innovation in a given field. This requires an immediate linguistic response by coining new terms or re-introducing existing ones but sometimes in a different sense or additional meaning. More often than not, a good number of these neologisms, especially those of a scientific/technical nature, are classified as academic word lists (AWLs) in the specific/special domain they belong to. The chances of long-term vs. short-term survival for these coinages are unpredictable as they might leave their footprints for decades or simply function as nonce words that will disappear soon after the event/occasion with which they are associated comes to a close (Crystal, 2008).

A strong point of this research is its use of a large corpus of more than 18 million words. Thus, it distinguishes itself in that the obtained word list is exhaustive and can fulfill the needs of the targeted population, such as medical news reporters, health communication specialists, researchers, instructors, and graduate students majoring in media, health communication, and medicine. The word list is useful for academic medical researchers, especially those conducting research on COVID-19. They can become more aware of terms pertinent to this area which consequently enables them to read and understand texts related to COVID-19. In addition, this word list can provide curriculum designers with some tips concerning the most frequent and useful words in this field. It can help journalists and graduate students identify and use the most relevant AWL in the context of coronavirus.

The researchers believed that timely and sustained exposure to this word list can contribute to the acquisition of these terms and consequently makes searching for them a smooth task. With further use and exposure to the compiled list, medical news reporters, learners, and instructors whose proficiency is noticeably lacking can largely benefit from the established list reported here. What particularly distinguishes this piece of research is establishing a lemmatized word list. This can be utilized by news reporters in the medical field on the one hand and by ESP students, but perhaps more so by the former group.

CONCLUSION

Based on the COVID-19-inspired word list, the present study, its findings, and implications testify to the major role of medical news ESP terminology in serving a large sector of the international community fighting the pandemic. It is this global dimension of the research topic in question and the interdisciplinary texture of its theme in response to the COVID-19 crisis that renders the current study relevant and timely. The research findings and results have been responsive and in harmony with

the thesis and research questions outlined earlier. The massive list of 18,249,093-word corpus is responsive to the first research question, namely 'what are the most frequent COVID-19 inspired academic words in the compiled medical news reporting corpus?'. On the other hand, the fact that compiled lists have been categorized and lemmatized is consistent with the second research question, which calls for classifying the data into different categories according to domains.

The fact that the large corpus has been short-listed in 550 post-COVID-19 most recognized medical news terms made the results more visible and accessible to the different stakeholders. Consequently, and as indicated earlier, the results study will be of special interest and value for news media reporters and journalists, health communication specialists, students of medicine, textbook writers and curriculum designers, instructors, and ESP specialists in the medical field. This partnership between researchers from different disciplines will contribute to enhancing applied linguistics studies and research from different perspectives. This interplay between specialists in the health sector, linguists and practitioners in second language theory and practice, journalists, news media workers, health communication specialists, and ESP learners will ensure rapid dissemination of this lemmatized word list. Such takeaways are particularly significant due to the fact that current research findings are relevant, timely, and responsive to the needs of a large sector of the international community fighting the pandemic, each in a different capacity but all dedicated to the one goal of beating the killer virus.

The six categories comprising acronyms and abbreviations, diseases, COVID-19, biology, medicine, and scientific disciplines will prove helpful in creating a common-core terminology where all work and ESP research teams worldwide will be on the same page by using the same terms concerning the virus. This word list will give all stakeholders dealing with COVID-19 across the globe an added value based on their contribution to this endeavor. The list of beneficiaries will include, but will not be limited to, experts and practitioners in health communication, researchers, students, ESP specialists in the medical field, curriculum designers, textbook writers, and journalists, among others. Furthermore, experts in psychology, mental health, and human behavior will be better informed about the psychological and social impact of COVID19 on people's behavior and well-being.

Despite the fact that this corpus comprised more than 18 million words, there were still a few limitations that need to be outlined here; the first stems from the fact that the data were derived from one online news outlet, namely NewsRx, where thousands of articles on COVID-19 are regularly posted. Second, the data were compiled over a relatively short time span, from January 1, 2020, through October 31, 2020. Third, the search queries were restricted to corona virus-related words. Finally, coronavirus terms prior to January 1 were not queried. Because of these limitations, the results cannot be generalized to the whole data, namely the coronavirus word list. Other researchers might be interested in establishing a pre-COVID-19 era word list. This is a recommendation for other researchers interested in this topic to pursue. As stated above, the collected data were derived from an electronic news outlet. One wonders if the same or similar AWL word list will be established if the data or corpus were based on articles published in academic or medical-related journals published by Wiley, Springer, Elsevier, Taylor & Francis, Sage, Oxford University Press, Cambridge University Press, and Nature, to mention a few.

To conclude, we call for more research projects to be conducted to cope with the continuing effects of COVI -19 so that more academic word lists should be generated to boost the existing ones through utilizing more online medical news outlets. After all, further research in this direction will emphasize the vital and vigorous role of the ESP movement, past, present, and future. With its multi-disciplinary composite, the current study will contribute significantly to second language theory and practice, applied language research in general, and ESP in particular.

REFERENCES

Abu Melhim, A. (2013). Exploring the historical development of ESP and its relation to English language teaching today. *European Journal of Social Sciences*, *40*(4), 615-627. https://doi.org/https://doi.org/10.5539/elt.v7n1p50

Akut, K. B. (2020). Morphological analysis of the Neologisms during the COVID-19 pandemic. *International Journal of English Language Studies*, *2*(3), 01-07. https://doi.org/https://doi.org/10.32996/ijels.2020.2.3.11

Al-Abbas, L. S., & Haider, A. S. (2020). The representation of homosexuals in Arabic-language news outlets. *Equality, Diversity Inclusion: An International Journal*, 1-29. https://doi.org/https://doi.org/10.1108/EDI-05-2020-0130

Al-Salman, S., & Haider, A. S. (2021a). COVID-19 trending neologisms and word formation processes in English. *Russian Journal of Linguistics*, *25*(1). https://doi.org/10.22363/2687-0088-2021-25-1-00-00

Al-Salman, S., & Haider, A. S. (2021b). Jordanian university students' views on emergency online learning during COVID-19. *Online Learning*, *25*(1), 286-302.

https://doi.org/https://doi.org/10.24059/olj.v25i1.2470

Al-Salman, S., & Haider, A. S. (2021c). The representation of Covid-19 and China in Reuters' and Xinhua's headlines. *Search (Malaysia)*, *13*(1), 93-110.

Almahasees, Z., Mohsen, K., & Omer, M. (2021). Faculty's and students' perceptions of online learning during COVID-19. *Frontiers in Education*. https://doi.org/https://doi.org/10.3389/feduc.2021.638470

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics* (6 ed.). Blackwell Publishing.

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, *33*, 100-118. https://doi.org/https://doi.org/10.1016/j.jeap.2018.02.003

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, *43*, 49-61. https://doi.org/http://dx.doi.org/10.1016/j.esp.2016.01.004

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language learning*, *67*(S1), 155-179. https://doi.org/https://doi.org/10.1111/lang.12225

Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes*, *51*, 1-17.

Green, C. (2019). Enriching the academic wordlist and Secondary Vocabulary Lists with lexicogrammar: Toward a pattern grammar of academic vocabulary. *System*, *87*, 1-10. https://doi.org/https://doi.org/10.1016/j.system.2019.102158

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes*, *35*, 105-115. https://doi.org/https://doi.org/10.1016/j.jeap.2018.07.004

Haider, A. S. (2019). Using corpus linguistic techniques in (Critical) discourse studies reduces but does not remove bias: Evidence from an Arabic corpus about refugees. *Poznan Studies in Contemporary Linguistics*, *55*(1), 89-133. https://doi.org/ https://doi.org/10.1515/psicl-2019-0004

Haider, A. S., & Al-Salman, S. (2020). Dataset of Jordanian university students' psychological health impacted by using e-learning tools during COVID-19. *Data in Brief*, *32*, 1-8. https://doi.org/https://doi.org/10.1016/j.dib.2020.106104

Heidari, F., Jalilifar, A., & Salimi, A. (2020). Developing a corpus-based word list in pharmacy research articles: A focus on academic culture. *International Journal of Society, Culture & Language*, *8*(1), 1-15.

Hsu, W. (2011). A business word list for prospective EFL business postgraduates. *The Asian ESP Journal*, *7*(4), 63-99. https://doi.org/https://doi.org/10.1016/j.esp.2011.04.005

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*, 54-65. https://doi.org/https://doi.org/10.1016/j.esp.2013.07.001

Hunston, S. (2008). Collection strategies and design decisions. In A. Ludeling & M. Kyto (Eds.), *Corpus Linguistics: an international handbook* (Vol. 1, pp. 154-167). Mouton de Gruyter.

Hussein, R. F., Haider, A. S., & Al-Sayyed, S. (2021). A corpus-driven study of terms used to refer to articles and methods in research abstracts in the fields of economics, education, english literature, nursing, and political science. *Journal of Educational Social Research*, *11*(3), 119-131. https://doi.org/https://doi.org/10.36941/jesr-2021-0056

Johns, A. M. (2013). The history of English for specific purposes research. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (Vol. 5, pp. 1-47). Wiley-Blackwell.

Johns, A. M., & Dudley-Evans, T. (1991). English for specific purposes: International in scope, specific in purpose. *TESOL quarterly*, *25*(2), 297-314. https://doi.org/https://doi.org/10.2307/3587465

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, *39*, 1-11. https://doi.org/http://dx.doi.org/10.1016/j.esp.2015.03.001

Loong, Y. C., & Chan, L. (2012). A study of vocabulary learning strategies adopted by dentistry students in Hong Kong in learning specialized dental vocabulary. *The Asian ESP Journal*, *8*, 28-49.

Masaya, K. (2020). Making a scientific research article word list. *言語教育研究*(30), 73-98.

Negro Alousque, I. (2016). Developments in ESP: from register analysis to a genre-based and CLIL-based approach. *LFE. Revista de Lenguas para Fines Específicos*, *22*(1), 190-

212. https://doi.org/https://doi.org/10.20420/rlfe.2016.310

Nekrasova-Beker, T., Becker, A., & Sharpe, A. (2019). Identifying and teaching target vocabulary in an ESP course. *TESOL Journal*, *10*(1), e00365. https://doi.org/https://doi.org/10.1002/tesj.365

Ramírez, C. G. (2015). English for specific purposes: Brief history and definitions. *Revista de Lenguas Modernas*, (23), 379-386 https://doi.org/https://doi.org/10.15517/rlm.v0i23.22359

Roig–Marín, A. (2020). English-based coroneologisms: A short survey of our Covid-19-related vocabulary. *English Today*, 1-3. https://doi.org/https://doi.org/10.1017/s0266078420000255

Sarré, C., & Whyte, S. (2017). *New developments in ESP teaching and learning research*. Research-publishing. net.

Scott, M. (2012). *WordSmith tools version 6*. In Lexical Analysis Software.

The United Nations. (2020). *COVID-19: an unprecedented news story for journalists*. https://www.un.org/en/coronavirus/covid-19-journalists-biggest-story-their-lifetime

The World Health Organization. (2020). Tips for professional reporting on COVID-19 vaccines. https://www.who.int/news-room/feature-stories/detail/tips-for-professional-reporting-on-covid-19-vaccines

Tongpoon-Patanasorn, A. (2018). Developing a frequent technical words list for finance: A hybrid approach. *English for Specific Purposes*, *51*, 45-54. https://doi.org/https://doi.org/10.1016/j.esp.2018.03.002

Wang, J., Liang, S.-l., & Ge, G.-c. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442-458. https://doi.org/https://doi.org/10.26686/wgtn.12560297

Wang, P. (2017). A Corpus-based Study of English Vocabulary in Art Research Articles. *Journal of Arts Humanities*, *6*(8), 47-53. https://doi.org/http://dx.doi.org/10.18533/journal.v6i8.1255

West, M. (1953). *A general service list of English words*. Longman.

Williams, C. (2014). The future of ESP studies: building on success, exploring new paths, avoiding pitfalls. *ASp. la revue du GERAS*, (66), 137-150. https://doi.org/https://doi.org/10.4000/asp.4616

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27-38.