BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Building the Medical Lexicon: A Corpus-Based Approach to Optimising Medical Terminology Acquisition for Pre-Health Science Students

Amira Abdullah Alshehri

Department of Languages and Translation, College of Education and Arts, University of Tabuk, Tabuk 47512, Saudi Arabia

ABSTRACT

Pre-health science students, especially those learning English as a Foreign Language (EFL), encounter substantial challenges in acquiring medical vocabulary, as they must master specialised terminology while developing general English proficiency. This study addresses this issue by proposing a frequency-based, corpus-driven approach to streamline medical vocabulary acquisition and reduce cognitive load. Focusing on the skeletal system chapter of a medical textbook, the research categorises terms into four groups—foundational, intermediate, advanced, and deferred—based on their frequency and relevance within medical discourse. Using Sketch Engine as the primary corpus tool for analysis, high-frequency terms are prioritised in early instruction to build a strong foundation, while more complex terms are introduced incrementally to support progressive knowledge development. Rare, highly technical terms are deferred to advanced stages, ensuring students engage with essential terminology at appropriate learning points. The study provides a practical, data-driven framework adaptable across other medical domains, offering a scalable model to enhance EFL students' vocabulary acquisition. By aligning instruction with frequency-based categories, educators can better manage students' learning burden, promote retention, and ensure mastery of critical concepts, while also guiding curriculum planning to foster gradual and structured learning progression.

Keywords: Lexicon; Frequency-Based Approach; Medical Terminology; Vocabulary Acquisition; Corpus Analysis

*CORRESPONDING AUTHOR:
Amira Abdullah Alshehri; Department of Languages and Translation, College of Education and Arts, University of Tabuk, Tabuk 47512, Saudi Arabia; Email: aa.alshehri@ut.edu.sa

ARTICLE INFO
Received: 2 October 2024 | Revised: 29 October 2024 | Accepted: 30 October 2024 | Published Online: 9 December 2024
DOI: https://doi.org/10.30564/fls.v6i6.7400

CITATION
Alshehri, A.A., 2024. Building the Medical Lexicon: A Corpus-Based Approach to Optimising Medical Terminology Acquisition for Pre-Health Science Students. Forum for Linguistics Studies. 6(6): 558–574. DOI: https://doi.org/10.30564/fls.v6i6.7400

1. Introduction

The role of English as an international language is prominent not only in facilitating communication between people of different linguistic backgrounds but also in other fields such as education, academia and publications. In the medical field, most textbooks, journals, and conferences are conducted in English, which has led to the emergence of English for Medical Purposes (EMP), particularly in countries where English is taught as a foreign language [1]. In Saudi Arabia, for instance, English serves as the medium of instruction in health-related programmes such as medicine, nursing, and pharmacology. Therefore, future health professionals need preparation to communicate effectively in clinical settings, as precise language use is critical to preventing misunderstandings that could affect patient well-being. Recognising this need, universities in Saudi Arabia have integrated a Medical Terminology course alongside intensive English language courses for first-year students.

However, EFL students face a real challenge: they must master a vast number of medical terms while they are developing proficiency in English [1]. There is a clear difference between general English vocabulary and medical terminology. The latter is a specialised language used by healthcare professionals to describe the body's anatomy, pathologies, procedures, and treatments [2]. The field contains around 60,000 distinct terms, with approximately 70% rooted in Greek and Latin origins [1]. Hence, the need for memorization is immense. As highlighted by [2], students at this stage are required to memorize medical terms and understand their meanings and applications in a clinical setting. This dual task of memorisation and practical application presents a substantial cognitive burden for EFL students, impacting both their academic performance and ability to function confidently in real-world healthcare environments.

For EFL teachers, the complexity and large volume of medical terms pose a pedagogical challenge regarding introducing and prioritising medical vocabulary to ensure students build a foundational understanding and successfully guide them toward learning more advanced terms in their future studies. Identifying which terms are essential for early learning becomes critical in this context, as teachers face the dilemma between balancing the need for students to acquire key medical concepts and easing their cognitive load to maximise their ability to comprehend and retain them [3,4]. Unfortunately, medical terminology courses often lack structured scaffolding that focuses on organising medical vocabulary effectively. Students are introduced to large volumes of specialised terms all at once, making it difficult to identify which terms are essential for early learning. Without clear guidance on how to prioritise vocabulary, EFL students often feel overwhelmed, which hinders both comprehension and retention. As Wang and Reynolds [5] explain, learners thrive when supported by environmental and social scaffolding, which enables them to develop gradually and build competencies over time. Sequencing learning materials to match students' capacities is essential for reducing cognitive overload and promoting effective vocabulary acquisition.

Therefore, the main concern of this research is that pre-health science EFL students[1] need help when confronted with a large amount of medical terminology presented in their medical foundational course, Medical Terminology. Drawing from the researcher's experience as an EFL instructor, it is clear that this difficulty is compounded by the course syllabus that consists of 15 chapters, each introducing 230 to 300 medical terms. Unfortunately, the course specification does not clarify which terms are essential, leaving students with the assumption that they must learn all the terms. This results in an overwhelming total of 3,000 to 4,500 terms in a single academic term. Such a large volume of specialised vocabulary is difficult to grasp and places an undue cognitive load on students still developing English proficiency [3]. This overload can hinder their ability to form a mental lexicon of essential medical terms, and in this situation, EFL teachers' critical challenge is how to identify and prioritise the medical terms that should be introduced at the beginner level. Therefore, the present research aims to develop a data-driven system for categorising and prioritising medical

1 First-year students in health-related programs at Saudi universities are often referred to as pre-health science students, as their grades in the first year determine which health-related major they can progress to.

terms that is based on their relative frequency in a specialised medical corpus. Through corpus analysis tools, this study also aims to create a systematic method for determining which medical terms should be introduced early and which can be deferred. This approach will streamline vocabulary instruction and reduce the cognitive load on pre-health science students, helping them focus on the most critical terms first.

The skeletal system was selected as a case study because it represents one of the core chapters in the medical terminology textbook, and the decision to focus on this chapter allows for the findings of this study to be generalised to other chapters of the textbook, thus providing a broader framework for the entire curriculum. Accordingly, the research question is:

How are medical terms in the skeletal system chapter prioritised and categorised using a systematic, corpus-based approach to facilitate the development of a medical lexicon for pre-health science students, ensuring they focus on the most essential and foundational terminology at an early stage of their learning?

A corpus analysis will be conducted to examine the relative frequency of terms in a medical corpus, which is an objective and systematic method of prioritising medical terms. By analysing how often terms occur in the corpus, the research can identify which words are most encountered in medical discourse, making them ideal candidates for early introduction to students.

This study distinguishes itself from previous research in English for Medical Purposes (EMP) by introducing a corpus-based approach to categorise and prioritise medical terms according to their frequency of use. Traditional methods, such as rote memorisation and long, undifferentiated term lists, often overwhelm students, offering little guidance on where to focus their efforts. In contrast, this research offers a frequency-based categorisation that supports the progressive building of a medical lexicon, ensuring that students at the early stages of learning medical terminology engage with high-frequency, essential terms first. This structured prioritisation not only helps reduce cognitive overload but also enables students to develop a strong foundation, preparing them to gradually progress toward more specialised vocabulary. The following section explores the theoretical framework underpinning this study, with a focus on incremental learning, corpus linguistics, and frequency-based approaches. These principles inform the categorisation and prioritisation of medical terms, providing a systematic way to support the progressive development of a medical lexicon for pre-health science students.

2. Literature Review

This research's theoretical foundation draws heavily from vocabulary acquisition principles and corpus linguistics, focusing on frequency-based approaches to vocabulary prioritisation. Key to this foundation is Webb and Nation's [4] argument that vocabulary should be introduced incrementally, with high-frequency words given precedence due to their broad application and essential role in language comprehension. This incremental approach ensures that learners acquire essential terms before gradually advancing to less-frequent vocabulary, a theory that directly informs the categorisation of medical terms in this study.

In literature, there is a consensus that high-frequency vocabulary and collocations tend to be learned more easily. Learners are exposed to such vocabulary repeatedly in different contexts, which enhances its acquisition and retention [6] [7]. Therefore, frequency is regarded as a key factor in determining which words and collocations should be emphasised in teaching. Several studies in second language vocabulary acquisition have argued that high frequency is a feature curriculum developers should consider when developing English for academic purposes courses.

In the context of specialised fields such as English for Medical Purposes (EMP), the principles of frequency still apply, but there is a key difference: medical terms are much less frequent compared with general vocabulary due to their specialised nature [8]. When a comparison is made, for example, between a medical and general English corpora, we find that medical terms have very low relative frequency. Dang & Webb [9] and Le & Ha [10] provide evidence that medical terminology places significant lexical demands on learners due to its specialised vocabulary and technical nature. However, despite their low frequency

in general language, these terms are crucial for students in medical fields. This reinforces the need for a frequency-based approach in specialised domains like medicine.

Nevertheless, even within medical vocabulary, there is a frequency hierarchy, ranging from very high to very low frequency. Blood, bone, skull, artery, vein, capillaries, lymph, and larynx are examples of medical terms students encounter very often, making them essential for understanding the core concepts of the field. These high-frequency medical terms play a role similar to high-frequency general English words in that they are necessary for communication within the domain [7]. However, as health science learners progress in their specialised domains, they will encounter more complex and lower-frequency terms (e.g. osteochondroma or autologous bone marrow transplant) which are critical for advanced understanding. To deal with the complexity with low-frequency terms, educators need to implement a strategic approach to vocabulary acquisition. For example, the gradual introduction of technical vocabulary according to frequency can help learners build their knowledge systematically. According to Webb & Nation [4], by focusing on high-frequency medical terms first, learners can develop a solid foundation before moving on to more specialised and infrequent words.

Hsu [11] also emphasises the importance of prioritising high-frequency terms in teaching medical terminology, suggesting that students will maximise their study efforts by concentrating on the words that appear most often in textbooks, lectures, and clinical settings, without being overwhelmed by less frequent or highly specialised vocabulary. The frequency approach along with another called incremental learning can optimise students' ability to build their technical lexicon. Incremental learning refers to the gradual process through which learners build their knowledge of words through repeated exposure and engagement with words in different contexts [12–14]. Studies have shown that incremental learning improves technical vocabulary retention and enhances students' ability to use these terms in real-life medical contexts, such as case studies and patient communication [15].

Teachers adopting incremental learning and frequency approaches introduce words to learners gradually and in stages, beginning with high-frequency, essential terms before advancing to more specialised and lower-frequency terms. This gradual process allows for consolidating knowledge and makes it easier for students to integrate new vocabulary into their existing lexicon and lessen the overwhelming feeling that comes with many unfamiliar words [3,4]. When applied to medical terminology, this approach becomes critical for EMP students, who often encounter numerous highly technical and unfamiliar terms derived from Latin and Greek [16]. Webb & Nation [4] suggest that educators should introduce foundational medical terms gradually so that students can build a solid knowledge base before progressing to more advanced terms, such as those related to specific procedures or complex conditions.

The role of the corpus in analysing vocabulary provides valuable information in terms of technicality and frequency [17]. When comparing a specialised corpus, such as the Medical Web Corpus, with a general one, such as the British National Corpus (BNC) or The English Web Corpus (enTenTen), the outcomes can help researchers in several ways. For instance, a researcher can determine whether a term is frequently used in the medical field but uncommon in general discourse by calculating the frequency of a word in both corpora. This analysis supports the idea of targeting high-frequency terms in medical education. If a word occurs in the medical corpus more significantly than the general English corpus, then the word can be classified as a technical term. Also, the frequency results within the specialised corpus can reveal whether a technical term is more specialised than another within a sub-domain such as skeletal, cardiac, or lymphatic domains. In other words, the key advantage of this method is its objectivity; instead of relying on subjective judgments about whether a word "sounds" technical, corpus-based methods provide quantitative data that can be used to support or refute such claims.

Hsu [11] also conducted a large-scale study to create a Medical Word List (MWL) for undergraduate medical students. He developed MWL by examining a corpus of 155 medical textbooks with a combined word count of over 15 million, covering 31 different medical topic areas. He contends that medical students need this list to successfully understand and engage in medical books. This highlights

the importance of corpus-based methods in identifying critical vocabulary, as both the MWL and the current research target high-frequency medical terms to optimise learning outcomes for students. According to Hsu [11], a list of high-frequency medical terms will reduce the students' learning burden, described in Barclay & Pellicer-Sánchez [18] and improve their ability to understand and retain essential medical vocabulary.

In the same way that the MWL was designed to bridge the gap between general and specialised vocabulary, the present research aims to categorise medical terms into foundational, intermediate, advanced, and deferred categories to guide the learning progression of pre-health science students. This method ensures that students are first introduced to the most relevant and frequently encountered terms in order to build a solid vocabulary base before moving on to more specialised and less common terms.

3. Methodology

This study employs a quantitative, corpus-based research design, focusing on the analysis of medical terms using frequency data. The design aims to systematically identify, categorise, and prioritise essential medical terms relevant to pre-health science students. By applying a data-driven methodology, the study minimises subjectivity in term selection, ensuring that high-frequency terms, which are more likely to be encountered in medical discourse, are introduced early in the students' learning journey. The frequency-based approach offers a structured framework to optimise the acquisition of critical terminology, reducing cognitive load and enhancing retention. This design aligns with the research objective of developing a progressive method for building a foundational medical lexicon, which can be applied to other medical domains beyond the skeletal system.

3.1. Data Collection

The medical terms for this study were collected from Ehrlich et al.'s textbook [19] chapter on the skeletal system, which serves as the main course textbook for pre-health science students at the researcher's institution due to its comprehensive coverage of medical terms related to human body systems. A total of 226 terms were extracted, divided into two lists: 109 single-word terms and 117 multi-word terms. These bolded terms were selected because they represent key medical concepts emphasised by the textbook authors. According to the textbook, boldface is used to identify primary terms that are critical for student learning. However, secondary terms, which appear in orange italics to clarify the meanings of primary terms, were also reviewed to ensure that no essential terminology was overlooked. This approach ensures that, while the focus remains on bolded terms, the broader context of medical terminology is considered. The selection criterion, based on bolded terms, aligns with the study's objective to investigate which terms should be prioritised for pre-health science students. Examples of the extracted terms are provided in **Table 1**, with the full lists available in Appendies A and B.

Table 1. Single- and multiple-word medical terms.

Single-Word Terms	Compound Word Terms
bursa	Ankylosing spondylitis
chiropractor	Axial skeleton
femur	Bone grafting
ilium	Fibrous joints
Ligaments	osteoblast cells
mandible	pectoral girdle
Osteoarthritis	Rheumatoid arthritis

The purpose of creating two separate lists for single-word and multi-word terms was to ensure accurate frequency per million (also known as relative frequency) analysis. The corpus analysis tool used in this study treats each word in a corpus as an independent unit, making relative frequency analysis straightforward for single-word terms. However, multi-word terms need to be analysed as sequences of words. If they were not placed in separate files, each word in a multi-word term would have been analysed individually, leading to inaccurate relative frequency counts, as the analysis would treat each part of the term as a separate entity rather than considering the full expression. By separating the lists, the multi-word terms could be analysed accurately, ensuring that their relative frequency was calculated correctly. This method was crit-

ical to maintaining the precision and integrity of the data analysis.

3.2. Corpus Software

Sketch Engine, a powerful corpus analysis tool, was used to conduct the relative frequency analysis in this study due to its extensive database and adaptability for specific domain research, ensuring the reliability of frequency-based approaches in medical terminology [20]. The default configuration of Sketch Engine was used, as it provided a straightforward and effective method for identifying relevant medical terms. For single medical terms, the tool was set to identify words only, with a minimum frequency threshold of 5, and the results were displayed as a simple list. For multi-word medical terms, the N-gram length was limited to 2 to 3 words, with a minimum frequency of 4, as the terms in the study did not exceed three words. These default settings were appropriate for the study's objective of prioritising frequently encountered terms, ensuring that the analysis remained focused on essential vocabulary while minimising noise from low-frequency terms.

The Medical Web Corpus, provided by Sketch Engine and containing approximately 33 billion words, was selected as the sole corpus for this analysis [20]. This specialised corpus contains authentic language samples systematically collected from medicine-specific texts that EMP students commonly encounter [21]. The corpus ensures that the results reflect key vocabulary relevant to foundational courses. By relying on the relative frequency from this corpus, the study aimed to prioritise medical terms based on their commonality within medical contexts, eliminating the need for a comparison with general language corpora like The English Web Corpus (enTenTen). This approach allowed for a focused analysis on terms essential to the students' understanding of the skeletal system in medical settings.

3.3. Data Analysis and Term Categorisation

In this study, relative frequency analysis was used to categorise medical terms into four groups: foundational, intermediate, advanced, and deferred terms. Relative frequency, as defined by Leech, Rayson, and Wilson [22], is a normalised measure of how often a word or phrase appears in a corpus, expressed as occurrences per million words. This measure allows for accurate comparisons of term usage across a large corpus. Using the Medical Web Corpus [20], the study identified terms commonly used in medical discourse to prioritise those most relevant for pre-health science students.

To ensure systematic categorisation, a percentile-based approach was applied. This approach ranks items according to their relative standing within the dataset and is particularly effective when dealing with uneven data distributions, as is typical with corpus data [23,24]. Quartiles, which divide data into four equal parts, represent 25% of the ranked terms per group. Using quartiles ensures systematic grouping, as each threshold reflects the natural distribution of the data, providing objective and proportional classification based on data-driven criteria. This method also reflects a logical progression for introducing terms across stages of learning, aligning with the study's educational objectives. Quartile-based grouping is a well-established statistical method, commonly employed in educational research, ensuring reliable and meaningful categorisation [25]. This method facilitates objective analysis by dividing data proportionally, allowing for meaningful interpretation of complex datasets. By grounding the categorisation in these established practices, the use of percentiles ensures that the thresholds are not arbitrary but follow recognised methods for data classification. The categories based on percentiles are as follows:

1. Foundational Terms: Percentiles above 0.75 (top 25% of terms).
2. Intermediate Terms: Percentiles between 0.5 and 0.75 (next 25%).
3. Advanced Terms: Percentiles between 0.25 and 0.5 (next 25%).
4. Deferred Terms: Percentiles below 0.25 (bottom 25%).

Alternative methods, such as fixed frequency cut-offs or mean-based classification, were considered but deemed unsuitable for this study. Fixed cut-offs can be subjective, relying on the researcher's judgment, and mean-based

classification assumes a normal distribution, which is not typical for corpus data [26,27]. In contrast, the percentile-based approach ensures flexibility and data-driven thresholds that reflect the actual distribution of terms [28]. Baker [29] also highlights that this approach is particularly effective in educational contexts, as it helps to identify key terms that are central to understanding a specific content domain.

To further ensure the accuracy and pedagogical soundness of the categorisation, a faculty member with a PhD in medicine reviewed the terms. This expert validation confirmed that the sequencing aligned with students' learning needs and supported the gradual development of their medical lexicon. The expert also recommended deferring surgical procedures and treatments to advanced stages, ensuring that the progression from foundational to deferred terms reflects a logical learning sequence. This approach prevents students from being overwhelmed at early stages while building the necessary vocabulary for more specialised content.

4. Results and Discussion

The results of the corpus analysis for the 109 single-word terms revealed that 108 terms had significantly varying relative frequencies, ranging from standard terms like arthritis with a relative frequency of 105.5 to specialised terms like osteochondroma and atherosclerosis, both with a relative frequency of 0.04756. This variation is critical as it highlights the disparity in how often these terms appear in medical texts, helping identify the terms students are most likely to encounter regularly in foundational learning contexts. High-frequency terms, such as arthritis, are essential for building a solid medical lexicon early on, while lower-frequency terms, like osteochondroma, are more specialised and can be deferred to advanced stages of learning. This distinction aligns with the research objective of optimising students' learning process by prioritising accessible and essential vocabulary while reserving complex terms for future study, thereby reducing cognitive overload. The term costals was excluded from the analysis due to its frequency of 0, suggesting its rarity in the target corpus. Despite its anatomical importance (referring to the ribs), its absence in the Medical Web Corpus indicates that it is not widely referenced in medical texts for this domain. Instead, at this early stage of learning, terms such as rib (e.g., true ribs, false ribs, and floating ribs) are more commonly used, making them more suitable for students to grasp within the skeletal system's context.

For the 117 multi-word terms, relative frequency analysis yielded 67 terms, with values ranging from 10.93 for lower extremities to 0.09512 for spiral fracture, confirming it as the lowest observed value. The remaining 50 terms were excluded because their occurrences fell below the minimum threshold of 4, as determined by Sketch Engine's default settings.

After obtaining the relative frequencies for both single-word and multi-word terms, all terms were combined into a single Excel sheet, with ranks sorted in ascending order according to relative frequency. Percentiles were calculated using the following formula:

P = PERCENTRANK.INC (Array, X)

Where:

- Array represents the range of all relative frequency values.
- X is the specific value for which the percentile is calculated.

The percentile analysis revealed values ranging from 1.00 for arthritis to 0.011 for chondromalacia. Notably, terms like osteochondroma and atherosclerosis, which have the lowest relative frequency of 0.04756, received a percentile value of 0, reflecting their extreme level of specialisation within the skeletal system domain. These percentile values provide educators with insights into the appropriate sequence for teaching medical terms by distinguishing which terms should be introduced early and which are better suited for advanced stages of learning. The focus on high-frequency terms ensures that students first acquire core medical vocabulary, reducing cognitive overload while preparing them for more specialised vocabulary in future studies. As a result, terms such as osteochondroma and atherosclerosis were included in the deferred list, as they represent conditions students are likely to encounter only in advanced studies.

4.1. Term Categorisation and Sub-Domain Insights

Based on the term categorisation criteria outlined in Section 3.3, a total of 175 terms were classified into four categories:

1. 44 foundational terms, representing 25% of the total number of analysed terms.
2. 43 intermediate terms, representing 24.5% of the total number of analysed terms.
3. 41 advanced terms, representing 22.3% of the total number of analysed terms.
4. 49 deferred terms, representing 28% of the total number of analysed terms.

Upon reviewing the categorisation of terms, the following adjustment was made: the term axial skeleton was originally categorised as a deferred term due to its low relative frequency. However, given its close relationship to the term appendicular skeleton, which was classified as advanced, axial skeleton was re-categorised as an advanced term. This adjustment ensures that both terms, which describe major structural components of the skeletal system, are introduced together to provide students with a more coherent understanding of the skeletal framework. A systematic review of the categorisation process was conducted to ensure consistency across all terms, with particular attention to conceptual relationships between terms. While similar adjustments were considered, axial skeleton was the only term requiring re-categorisation based on its specific connection with appendicular skeleton.

As a result of this re-categorisation:

- The advanced list increased from 41 to 42 terms, representing 22.3% of the total terms analysed.
- The deferred list was reduced from 49 to 48 terms, representing 28% of the total terms analysed.

For the full lists, see Appendiex C, D, E, and F. **Table 2** summarizes the distribution of terms across the four categories, broken down by sub-domain:

Table 2. Summary of Term Categorisation by Sub-Domain and Percentages.

Category	Anatomy Terms	Pathology Terms	Procedure Terms	Treatment Terms	Specialism Terms	Total Terms	Percentage of Total (%)
Foundational	28	10	3	3	0	44	25%
Intermediate	21	14	5	2	2	43	24.5%
Advanced	23	11	3	1	4	42	22.3%
Deferred	17	17	12	0	0	48	27.1%

The categorisation of terms across sub-domains reflects a structured approach to introducing medical terminology that aligns with pedagogical expectations for EFL students. Anatomy-related terms dominate the foundational category (63.6%), confirming the importance of building students' understanding of the skeletal system's structure in the initial stages. The prioritisation is consistent with educational theory, which emphasises starting with tangible concepts that provide essential scaffolding for more complex material [4].

As students progress into intermediate learning, the focus shifts toward pathology (32.6%) and procedures (11.6%), indicating that students are expected to apply anatomical knowledge to clinical contexts. This gradual exposure to more specialised vocabulary mirrors best practices in incremental learning, where the sequence of content reflects growing depth of knowledge to promote retention and understanding.

Advanced terms represent a balanced mix of anatomy (54.8%) and pathology (26.2%), with a growing presence of specialism-related terms (9.5%). These terms introduce highly specialised vocabulary that builds on foundational and intermediate knowledge. This phase prepares students for advanced healthcare education, ensuring they can engage with specific medical contexts, such as diagnostic procedures or rare conditions. The minimal focus on treatment-related terms (2.4%) suggests that advanced learning emphasises understanding complex skeletal conditions before therapeutic interventions.

In the deferred category, pathology-related terms

(35.4%) and procedure-related terms (25%) dominate. These terms are reserved for advanced study due to their complexity and infrequent use in early medical discourse. Deferring these terms aligns with reducing cognitive overload, ensuring students encounter complex concepts only after mastering foundational material.

These findings align with previous research on medical vocabulary acquisition for EFL students, which stress the importance of introducing accessible, high-frequency terms first to reduce cognitive overload [9] [4]. Research indicates that students benefit from incrementally engaging with vocabulary, beginning with essential terms that form the foundation for understanding more specialised concepts. This approach aligns with incremental learning models that promote retention through staged acquisition [6]. Furthermore, studies in medical education highlight that building a solid anatomical foundation is critical for medical students. Students perceive anatomical knowledge as essential for understanding clinical practices and conditions encountered later, such as pathologies and treatments, reinforcing the importance of a structured learning sequence [30].

This study contributes to the existing literature by reinforcing the importance of frequency-based categorisation as a systematic, data-driven strategy. Unlike traditional methods, such as rote memorisation or undifferentiated lists, this approach ensures that vocabulary aligns with students' immediate learning needs and cognitive capacities. Additionally, the categorisation framework provides a structured pathway for students to move from high-frequency anatomical terms toward more complex terms—such as procedures and specialisms—by introducing vocabulary incrementally at appropriate learning stages. This method aligns with educational frameworks that emphasise scaffolding content according to students' evolving competencies [5], supporting the gradual expansion of their medical lexicon over time.

To further optimise learning, this study proposes splitting the Medical Terminology course into two modules: Medical Terminology and Advanced Medical Terminology. The first module focuses on high-frequency terms to establish a strong foundation, reducing cognitive overload and promoting retention [4]. The second module introduces more specialised, lower-frequency terms, such as procedures and rare conditions, allowing students to engage with complex content as they progress. This structure ensures a smooth transition between foundational and advanced knowledge, matching the curriculum with students' readiness and clinical experiences.

5. Conclusion

This research addressed the challenge of overwhelming medical terminology for pre-health science students, particularly those learning English as a foreign language (EFL). The study proposed a systematic, corpus-based approach to categorising and prioritising medical terms in the skeletal system chapter of a medical textbook. This approach used the relative frequency in the Medical Web Corpus to identify essential medical terms that should be introduced to students early in their learning. The next step was to categorise them according to the percentile-based approach into four categories: foundational, intermediate, advanced, and deferred. The first three categories were suggested to be integrated into the curriculum, while the deferred category was reserved for more advanced stages of health education.

The approach utilised offers a data-driven solution to managing students' cognitive load, allowing them to acquire essential terminology early in their education while preparing them for more specialised terms as they advance in their studies. Furthermore, focusing on the skeletal system as a case study provided a controlled framework that can be adapted and generalised to other medical domains, offering a scalable vocabulary instruction model.

However, the study's focus on a single chapter, while enhancing internal validity, limits the generalisability of the results to other medical domains. Systems such as the cardiovascular or nervous system may exhibit distinct frequency distributions, necessitating different categorisation models. Future research could expand the scope by including multiple chapters or comparing results with general medical corpora to provide broader insights into medical vocabulary acquisition. This broader perspective would not only strengthen the applicability of the findings but also offer a more comprehensive model for vocabulary

instruction across various fields of healthcare education.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data is included in the article.

Conflict of interest

The author declares no conflict of interest.

Appendix A

Single-Word Terms		
Acetabulum	Fibula	Osteoporosis
Acromion	Fontanelles	Osteotomy
Adhesions	Foramen	Patella
Amputation	Fracture	Pelvis
Ankles	Gout	Periosteum
Ankylosis	Hemarthrosis	Periostitis
Appendage	Hematopoietic	Phalanges
Appendicular	Humerus	Podiatrist
Arthritis	Ilium	Popliteal
Arthrocentesis	Immobilization	Prosthesis
Arthrodesis	Ischium	Pseudogout
Arthroplasty	Joints	Pubis
Arthrosclerosis	Kyphosis	Radiculopathy
Arthroscopy	Lamina	Radius
Bisphosphonates	Laminectomy	Rheumatologist
Bursa	Ligaments	Ribs
Bursitis	Lordosis	Rickets
Calcaneus	Lumbago	Sacroiliac
Callus	Malleolus	Sacrum
Carpals	Mandible	Scapula
Cartilage	Manubrium	Scoliosis
Chiropractor	Meatus	Skull
Chondromalacia	Meniscus	Spondylolisthesis
Clavicle	Metacarpals	Spondylosis
Clubfoot	Metatarsals	Sprain
Coccyx	Olecranon	Sternum
Costals	Orthopedist	Subluxation
Costochondritis	Orthotic	Synovectomy

Craniotomy	Ossification	Synovitis
Cranium	Osteitis	Talus
Crepitation	Osteoarthritis	Tarsal
Diaphysis	Osteochondroma	Tibia
Dislocation	Osteomalacia	Traction
Endosteum	Osteomyelitis	Ulna
Epiphyses	Osteopath	Vertebrae
emity	Osteopathy	
Femur	Osteophytes	

Appendix B

Multiple-Word Terms		
ACL reconstruction	Frontal bone	Psoriatic arthritis
Adhesive capsulitis	Greenstick fracture	Pubic bones
Allogenic bone marrow transplant	Hallux valgus	Pubic symphysis
Ankylosing spondylitis	Herniated disk	Pubic symphysis
Appendicular skeleton	Hinge joints	Red bone marrow
Arthroscopic surgery	Hip resurfacing arthroplasty	Revision surgery
Articular cartilage	Incomplete fracture	Rheumatoid arthritis
Auditory ossicles	Inferior conchae	Secondary bone cancer
Autologous bone marrow transplant	Internal fixation	Short stature
Avascular necrosis	Intervertebral disks	Shoulder replacement surgery
Axial skeleton	Juvenile idiopathic arthritis	Sphenoid bone
Baker's cyst	Lacrimal bones	Spina bifida
Ball-and-socket joints	Lower extremities	Spinal column
Bone density testing	Lumbar vertebrae	Spinal fusion
Bone grafting	Magnetic resonance imaging	Spinal stenosis
Bone marrow aspiration	Manipulative treatment	Spiral fracture
Bone marrow biopsy	Mastoid process	Spongy bone
Bone marrow transplant	Maxillary bones	Sternum body
Bone scans	Medullary cavity	Stress fracture
Buckle fracture	Multiple myeloma	Synovial capsule
Cartilaginous joints	Nasal bones	Synovial fluid
Cervical vertebrae	Nasal septum	Synovial joint
Closed fracture	Nucleus pulposus	Synovial membrane
Closed reduction	Oblique fracture	Temporo- mandibular joint
Colles fracture	Occipital bone	Thoracic cavity
Comminuted fracture	Open fracture	Thoracic vertebrae
Compact bone	Orthopedic surgeon	Total hip replacement
Compression fracture	Osteoblast cells	Total knee replacement
Computed tomography	Osteoporotic hip fracture	Transverse fracture
Cruciate ligaments	Paget's disease	True ribs
Decompressive craniectomy	Palatine bones	Ultrasonic bone density testing

Dual x-ray absorptiometry	Parietal bones	Upper extremities
Ethmoid bone	Partial knee replacement	Vertebra body
External auditory meatus	Pathologic fracture	Vertebral foramen
External fixation	Pectoral girdle	Vomer bone
False ribs	Percutaneous diskectomy	X-ray imaging
Fat embolus	Percutaneous vertebroplasty	Xiphoid process
Fibrous joints	Polymyalgia rheumatica	Yellow bone marrow

Appendix C

Foundational List		
Item	**Domain**	**Relative Frequency**
Arthritis	Pathology	105.3645
Joints	Anatomy	79.75458
Skull	Anatomy	64.46472
Fracture	Pathology	52.71792
Pelvis	Anatomy	41.47048
Adhesions	Pathology	33.86122
Cartilage	Anatomy	31.93512
emity	Anatomy	26.03794
Osteoporosis	Pathology	25.53859
Gout	Pathology	23.80272
Sternum	Anatomy	21.82907
Ribs	Anatomy	20.68768
Vertebrae	Anatomy	17.47752
Ligaments	Anatomy	17.45375
Femur	Anatomy	17.22
Traction	Treatment	17.12084
Amputation	Procedure	16.55015
Osteopathy	Treatment	15.17097
Osteomyelitis	Pathology	15.07585
Osteoarthritis	Pathology	14.7905
Meatus	Anatomy	13.00708
Dislocation	Pathology	12.03215
Tibia	Anatomy	12.00837
Rickets	Pathology	11.34256
Foramen	Anatomy	11.10477
Lower extremities	Anatomy	10.93831
Ankles	Anatomy	9.89204
Sacrum	Anatomy	9.72559
Periosteum	Anatomy	9.13111
Spinal column	Anatomy	8.0135
Scapula	Anatomy	7.58548
Humerus	Anatomy	7.58548

Clavicle	Anatomy	7.34769
Cranium	Anatomy	6.99101
Patella	Anatomy	6.80078
Bursa	Anatomy	6.34898
Pubis	Anatomy	6.2063
Scoliosis	Pathology	6.01607
Radius	Anatomy	5.96852
Prosthesis	Treatment	5.73073
Ossification	Process	5.58805
Arthroplasty	Procedure	4.92224
Coccyx	Anatomy	4.77957
Computed tomography	Procedure	4.66067

Appendix D

Intermediate List		
Item	**Domain**	**Relative Frequency**
Hematopoiesis	Anatomy	4.35155
Popliteal	Anatomy	4.25643
Osteotomy	Procedure	4.04242
Lamina	Anatomy	4.01864
Mandible	Anatomy	3.9473
Bursitis	Pathology	3.70952
Synovial Membrane	Anatomy	3.70952
Kyphosis	Pathology	3.59062
Ulna	Anatomy	3.54306
Osteomalacia	Pathology	3.44795
Epiphyses	Anatomy	3.44795
Ilium	Anatomy	3.42417
Tarsal	Anatomy	3.35283
Upper Extremities	Anatomy	3.30527
Periostitis	Pathology	3.23394
Immobilization	Treatment	3.21016
Ankylosis	Pathology	3.13882
Bisphosphonates	Treatment	3.13882
Fibula	Anatomy	3.01993
Chiropractor	Specialist	2.94859
Osteopath	Specialist	2.94859
Phalanges	Anatomy	2.87725
Osteitis	Pathology	2.82969
Acetabulum	Anatomy	2.78214
Sprain	Pathology	2.63946
Synovitis	Pathology	2.63946
Callus	Pathology	2.52057

Lordosis	Pathology	2.44923
Craniotomy	Procedure	2.33034
Lumbago	Pathology	2.33034
Frontal Bone	Anatomy	2.33034
Spondylolisthesis	Pathology	2.30656
Rheumatoid Arthritis	Pathology	2.30656
Appendage	Anatomy	2.28278
Laminectomy	Procedure	2.21144
Sacroiliac	Anatomy	2.18766
Cervical Vertebrae	Anatomy	1.97365
Magnetic Resonance Imaging	Procedure	1.97365
Crepitation	Pathology	1.92609
Subluxation	Pathology	1.73586
Meniscus	Anatomy	1.71208
Arthroscopy	Procedure	1.71208
Pubic Bone	Anatomy	1.69

Appendix E

Advanced List		
Item	**Domain**	**Relative Frequency**
Spondylosis	Pathology	1.66453
Malleolus	Anatomy	1.64075
Rheumatologist	Specialist	1.59319
Lumbar vertebrae	Anatomy	1.54563
Diaphysis	Anatomy	1.52185
Acromion	Anatomy	1.49807
Appendicular	Anatomy	1.47429
Podiatrist	Specialist	1.40296
Arthrodesis	Procedure	1.40296
Thoracic cavity	Anatomy	1.40296
Mastoid process	Anatomy	1.40296
Orthotic	Treatment	1.37918
Radiculopathy	Pathology	1.3554
Herniated disk	Pathology	1.30784
Olecranon	Anatomy	1.28406
Orthopedic surgeon	Specialist	1.23651
Total hip replacement	Procedure	1.21273
Fontanelles	Anatomy	1.11761
Parietal bones	Anatomy	1.07005
Nasal septum	Anatomy	1.04627
Ischium	Anatomy	1.02249
Nucleus pulposus	Anatomy	1.02249
Manubrium	Anatomy	0.9036

External auditory meatus	Anatomy	0.9036
Nasal bones	Anatomy	0.9036
Occipital bone	Anatomy	0.9036
Orthopedist	Specialist	0.87982
Osteophytes	Pathology	0.87982
Clubfoot	Pathology	0.85604
Calcaneus	Anatomy	0.78471
Pseudogout	Pathology	0.76093
Bone marrow transplant	Procedure	0.76093
Sphenoid bone	Anatomy	0.73715
Metacarpals	Anatomy	0.64203
Comminuted fracture	Pathology	0.64203
Stress fracture	Pathology	0.59447
Spina bifida	Pathology	0.54692
Compression fracture	Pathology	0.54692
Open fracture	Pathology	0.52314
Thoracic vertebrae	Anatomy	0.49936
Appendicular skeleton	Anatomy	0.47558

Appendix F

Deferred List		
Item	Domain	Relative Frequency
Spinal stenosis	Pathology	0.42802
Ankylosing spondylitis	Pathology	0.42802
Axial skeleton	Anatomy	0.42802
Medullary cavity	Anatomy	0.42802
False ribs	Anatomy	0.40424
Multiple myeloma	Pathology	0.40424
Pubic symphysis	Anatomy	0.40424
Talus	Anatomy	0.38046
Total knee replacement	Procedure	0.38046
Bone marrow biopsy	Procedure	0.35668
Ethmoid bone	Anatomy	0.35668
Hemarthrosis	Pathology	0.30913
Endosteum	Anatomy	0.30913
ACL reconstruction	Procedure	0.30913
Xiphoid process	Anatomy	0.30913
Synovectomy	Procedure	0.28535
Metatarsals	Anatomy	0.28535
Synovial joint	Anatomy	0.26157
Polymyalgia rheumatica	Pathology	0.26157
Transverse fracture	Pathology	0.23779
Closed fracture	Pathology	0.23779

Cruciate ligaments	Anatomy	0.23779
Psoriatic arthritis	Pathology	0.23779
Greenstick fracture	Pathology	0.23779
Avascular necrosis	Pathology	0.23779
Pathologic fracture	Pathology	0.23779
Arthrocentesis	Procedure	0.21401
Costochondritis	Pathology	0.16645
Bone density testing	Procedure	0.16645
Carpals	Anatomy	0.14267
Incomplete fracture	Pathology	0.14267
Bone scans	Procedure	0.14267
Synovial capsule	Anatomy	0.14267
Synovial fluid	Substance	0.14267
Floating ribs	Anatomy	0.11889
Pectoral girdle	Anatomy	0.11889
Osteoblast cells	Anatomy	0.11889
External fixation	Procedure	0.11889
Maxillary bones	Anatomy	0.09512
Decompressive craniectomy	Procedure	0.09512
Internal fixation	Procedure	0.09512
Bone grafting	Procedure	0.09512
X-ray imaging	Procedure	0.09512
Spiral fracture	Pathology	0.09512
Chondromalacia	Pathology	0.07134
Osteochondroma	Pathology	0.04756
Arthrosclerosis	Pathology	0.04756

References

[1] Ismayilli-Karakoç, A., 2020. Teaching medical terminology to speakers of English as a foreign language. In: Genç ZS. and Kaçar, IG. (Eds.), TESOL in the 21st Century: Challenges and opportunities. Peter Lang: Bristol, UK. pp. 235-252.

[2] Panocová, R., 2017. The Vocabulary of Medical English: A Corpus-Based Study. Cambridge Scholars Publishing: Newcastle Upon Tyne, UK. pp. 1-190.

[3] Kalyuga, M., Kalyuga, S., 2008. Metaphor awareness in teaching vocabulary. Language Learning Journal. 36(2), 249-257. DOI: https://doi.org/10.1080/09571730802390767

[4] Webb, S., Nation, P., 2017. How Vocabulary is Learned. Oxford University Press: Oxford, UK. pp. 1-336.

[5] Wang, X., Reynolds, B. L., 2024. Beyond the books: Exploring factors shaping Chinese English learners' engagement with large language models for vocabulary learning. Education Sciences. 14(5), 496.

[6] Sun, W., Park, E., 2023. EFL Learners' Collocation Acquisition and Learning in Corpus-Based Instruction: A Systematic Review. Sustainability. 15(17), 13242.

[7] Quero, B., Coxhead, A., 2018. Using a corpus-based approach to select medical vocabulary for an ESP course: The case for high-frequency vocabulary. In: Kırkgöz, Y. and Dikilitaş, K. (Eds.). Key Issues in English for Specific Purposes in Higher Education. Springer International Publishing: Cham, Switzerland. pp. 51-75.

[8] Chen, Q., Ge, G. C., 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). English for Specific Purposes. 26(4), 502-514. DOI: https://doi.org/10.1016/j.esp.2007.04.003

[9] Dang, T. N. Y., Webb, S., 2014. The lexical profile of academic spoken English. English for Specific Purposes. 33, 66-76. DOI: https://doi.org/10.1016/j.esp.2013.08.001

[10] Le, N. H., Ha, H. T., 2023. Lexical demands of academic written English: From students' assignments to scholarly publications. Sage Open. 13(4), 1-16.

DOI: https://doi.org/10.1177/21582440231216292

[11] Hsu, W., 2013. Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. Language Teaching Research. 17(4), 454-484. DOI: https://doi.org/10.1177/1362168813494121

[12] González-Fernández, B., Schmitt, N., 2017. Vocabulary acquisition. In: Sato, M and Loewen, S. (Eds.). The Routledge Handbook of Instructed Second Language Acquisition. Taylor & Francis: London, UK. pp. 280-298.

[13] Schmitt, N., 1998. Tracking the incremental acquisition of second language vocabulary: A longitudinal study. Language Learning. 48(2), 281-317.

[14] Schmitt, N., 2019. Understanding vocabulary acquisition, instruction, and assessment: A research agenda. Language Teaching. 52(2), 261-274.

[15] Najafi, M., Talebinezhad, M. R., 2018. The impact of teaching EFL medical vocabulary through collocations on vocabulary retention of EFL medical students. Advances in Language and Literary Studies. 9(5), 24-27. DOI: http://doi.org/10.7575/aiac.alls.v.9n.5p.24

[16] Coxhead, A., 2014. Vocabulary and ESP. In: Paltridge, B., Starfield, S. (Eds.). The Handbook of English for Specific Purposes. Wiley-Blackwell: Berlin, Germany. pp. 115–132.

[17] Liu, D., Lei, L., 2019. Technical vocabulary. In: Webb, S. (Ed.). The Routledge Handbook of Vocabulary Studies. Taylor & Francis: London, UK. pp. 111–124

[18] Barclay, S., Pellicer-Sánchez, A., 2021. Exploring the learning burden and decay of foreign language vocabulary knowledge: The effect of part of speech and word length. ITL-International Journal of Applied Linguistics. 172(2), 259-289. DOI: https://doi.org/10.1075/itl.20011.bar

[19] Ehrlich, A., Schroeder, C. L., Ehrlich, L., et al., 2021. Medical Terminology for Health Professions. Delmar Cengage Learning: Boston, United States. pp. 1-672.

[20] Kilgarriff, A., Baisa, V., Bušta, J., et al., 2014. The Sketch Engine: ten years on. Lexicography. 1(1), 7-36. DOI: https://doi.org/10.1007/s40607-014-0009-9

[21] Le, C. N. N., Miller, J., 2020. A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. English for Specific Purposes. 58, 102-121. DOI: https://doi.org/10.1016/j.esp.2020.01.004

[22] Leech, G., Rayson, P., Wilson, A., 2014. Word Frequencies in Written and Spoken English: Based on the British National Corpus, 2nd ed. Taylor & Francis: London, UK. pp 1-320.

[23] Cohen, J., 2013. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Taylor & Francis: London, UK. pp. 1-567.

[24] Wright, B. D., Masters, G. N., 1982. Rating Scale Analysis. MESA Press: Chicago, IL, USA. pp. 1-206.

[25] Hinton, P. R., 1995. Statistics explained: a guide for social science students. Routledge: London, UK. pp. 1-322.

[26] Biber, D., Conrad, S., Reppen, R., 2004. Corpus Linguistics: Investigating Language Structure and Use, 4th ed. Cambridge University Press: Cambridge, UK. pp. 1-340.

[27] Manning, C. D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Pres: Cambridge, United States. pp. 1-720.

[28] Drewry, H., Notterman, J., 2013. Psychology and Education: Parallel and Interactive Approaches. Springer: New York, United States. pp. 1-290.

[29] Baker, P., 2023. Using Corpora in Discourse Analysis, 2nd ed. Bloomsbury Publishing: London, UK. pp. 1-280.

[30] Bergman, E.M., de Bruin, A.B., Herrler, A., et al., 2013. Students' perceptions of anatomy across the undergraduate problem-based learning medical curriculum: a phenomenographical study. BMC Med Educ. 13, 152.