Xodabande *et al.*
Asian. J. Second. Foreign. Lang. Educ. (2023) 8:32
https://doi.org/10.1186/s40862-023-00205-5

Asian-Pacific Journal of Second and Foreign Language Education

RESEARCH Open Access

Developing and validating a mid-frequency word list for chemistry: a corpus-based approach using big data

Ismail Xodabande[1*], Mahmood Reza Atai[1], Mohammad R. Hashemi[1] and Paul Thompson[2]

*Correspondence:
ismail.kh.tefl@gmail.com

[1] Department of Foreign Languages, Kharazmi University, Tehran, Iran
[2] Department of English Language and Linguistics, University of Birmingham, Birmingham, UK

Abstract

Given the importance of specialized vocabulary in scientific communication and academic discourse, there is a growing need to create wordlists to address the vocabulary-learning needs of university students and researchers in different subject areas. The current study analyzed a corpus of chemistry research articles (with 278 million running words) to establish a mid-frequency vocabulary list for this field. Using frequency, range, and dispersion criteria, the study identified 560 lemmas in the fourth to the ninth British National Corpus/Corpus of Contemporary American English (BNC/COCA) lists that provided 6.4% coverage of all words in the corpus. The list was validated using specialized and general corpora, and the results confirmed the value and relevance of the items for chemistry. Moreover, for using the list for pedagogical goals, the vocabulary items were divided into five bands based on their coverage and importance. The 100 words in the first band were the most important mid-frequent vocabulary in chemistry, as they provided 3.05% coverage. The study highlights the significant contribution of mid-frequency words in research articles and the findings have implications for using large corpora as a big data source in identifying specialized and field-specific vocabulary.

Keywords: Wordlist, EAP/ESP vocabulary, Chemistry, Research article, Corpus linguistics, Mid-frequency vocabulary, Academic vocabulary, Big data

Introduction

In the current academic landscape, being able to read and publish Research Articles (RAs) in English has become a crucial factor for professional success of many university students (Flowerdew, 2015; Li & Flowerdew, 2020; Martínez et al., 2009). Relatedly, as there is an increasing demand for English proficiency in academic settings, students must prioritize improving their language skills to expand their opportunities for engaging with the global community of academics. Within this competitive culture of scholarly publishing in higher education, RA is considered as the "pre-eminent genre of the academy" being the most important site for creating and disseminating scientific knowledge (Hyland, 2009, p. 67). More than 90 percent of the top-tier international journals are published in English (Lillis & Curry, 2010), and scholars are required to publish in

those journals to attain academic achievements and recognition (Hyland, 2022). Given these considerations, the study of linguistic features of RAs has remained a significant area of inquiry within applied linguistics.

Against this background, it is widely recognized that non-native speakers of English (NNES) face significant linguistic challenges when attempting to publish in English (Corcoran, 2017; Flowerdew, 2019; Politzer-Ahles et al., 2020). One major obstacle in this regard is the insufficient vocabulary knowledge which makes it difficult for NNESs to write and read in English (Bazerman et al., 2012). This linguistic deficiency further hinders their engagement with research communities and impedes their academic and professional growth (Flowerdew, 2019). In this regard, research focusing on disciplinary and field-specific vocabulary is important for several pedagogical reasons. Firstly, vocabulary is a critical factor in language learning and proficiency development (Clenton & Booth, 2020) which correlates positively with scores in general proficiency tests (Alderson, 2006; Dodigovic & Agustín-Llach, 2020). Secondly, vocabulary size, measured as the number of words a person knows, is a strong predictor of writing quality and reading comprehension (Laufer, 1996; Morris & Cobb, 2004; Qian, 2002; Schoonen et al., 2011). In addition, specialized vocabulary is closely linked to the content knowledge in a specific field (Hyland & Tse, 2007; Woodward-Kron, 2008) and "constitutes a very important and required knowledge for those who work directly or indirectly in a subject area" (Liu & Lei, 2020, p. 111). Recognizing this indispensable role of vocabulary, creating discipline-specific word lists is gaining increased attention in English for Specific Purposes (ESP) education (Coxhead, 2018b; Coxhead & Demecheleer, 2018; Dang, 2019; Nation, 2016).

The current study aimed to contribute to this line of research by creating and validating a mid-frequency word list for chemistry RAs. The study is significant as it addresses vocabulary in chemistry, a field that received scant attention in the expanding literature on ESP vocabulary studies (Coxhead, 2018b). Moreover, the focus of the research is on mid-frequency vocabulary-learning needs which is conceptually different from traditional views on specialized vocabulary in terms of academic and technical words (Coxhead & Nation, 2001). The study also used Meso-level big data as the computerized writings of the experts in educational fields (Fischer et al., 2020) to identify field-specific words, that provide a more reliable resource for mining frequently used words in chemistry. The research is intended to inform English for Research Publication Purposes (ERPP) programs in chemistry by providing a resource for the frequently used vocabulary items in RAs. Additionally, chemistry students and researchers might find the developed wordlist instrumental in setting and planning their personal vocabulary-learning endeavors.

Review of the literature

Big data in education

Big data refers to large and complex datasets that exceed the processing capacity of traditional data management tools. More specifically, these high volume, fast-growing, and complex datasets has three key characteristics of volume (massive amounts of data), velocity (high speed at which data is generated and processed), and variety (heterogeneous data from diverse sources) (Ward & Barker, 2013). Big data provides numerous

affordances and opportunities for improvements in education. These affordances include monitoring, evaluating, and understanding learning processes to enhance educational effectiveness through informed decision-making (Baig et al., 2020; Fischer et al., 2020; Williamson, 2018). Recently, Fischer et al. (2020) outlined a three-level framework for using big data in education, encompassing *Micro* (e.g., clickstream data), Meso (e.g., text data), and Macro (e.g., institutional data) levels. The first level relates to the staggering amount of data generated when learners interact with digital learning environments or work with digital tools and resources. The most pertinent examples include intelligent tutoring systems, massive open online courses (MOOCs), simulations, and games. Meso-level big data accounts for the computerized writings of the learners and experts, which is available in the form of digital corpora. The analysis of available data at this level provides reliable information with respect to the actual uses of language in different contexts and discourse types. The third level, namely the Macro-level big data, encompasses the data collected at the institutional level (e.g. course enrolment data).

In line with these developments, the utilization of big data is also gaining increased interest in language teaching and learning (Godwin-Jones, 2017, 2021; Lee et al., 2019; Reinders & Lan, 2021; Thomas & Gelan, 2018). One of the recent developments in this area that impacted the field significantly is the availability of large and easily accessible corpora for language analysis (Chambers, 2019; Reinders & Lan, 2021; Römer, 2011). A corpus, defined as a principled collection of naturally occurring language (spoken or written) in a machine-readable format makes it possible to obtain valuable information concerning the patterns of language use, frequency of lexical items, collocations, and related statistics, which was unimaginable before (Baker, 2016; McEnery & Hardie, 2011). In recent years, a growing number of studies are using the affordances provided by big data in the form of large corpora for investigating the lexical profile of various discourse types (Ha, 2022a, 2022b; Nguy & Ha, 2022; Trang et al., 2023). For example, Ha (2022b) examined the lexical profile of informal spoken English using a 625 million-words corpus. The findings of the study revealed that a vocabulary size of 3000–5000 words in BNC/COCA frequency levels is needed for adequate level of comprehension of informal spoken English. Analyzing a much larger corpus containing 12 billion words from online newspapers and magazines, Ha (2022a) showed that 4000 most frequent word families in the BNC/COCA lists and familiarity with acronyms, marginal words, proper nouns, and transparent compounds seem to be necessary for gaining a 95% vocabulary coverage threshold for newspapers.

Moreover, the implementation of corpus-based pedagogy has been viewed as one of the most important developments in Computer Assisted Language Learning (CALL) in recent years that has great potential to revolutionize language education (Godwin-Jones, 2021). Accordingly, considering the long-standing interest in applied linguistics for analyzing language-related features in academic discourse, Meso-level big data also provides a more reliable source for studying vocabulary in academic discourse. More specifically, since the kind of texts that university students need to interact with in graduate or post-graduate studies are currently available in massive amounts in digital format, Meso-level big data as a corpus for language analysis provides insights regarding the most important vocabulary for using English for academic and research publication purposes. A recent study investigated the lexical profile of academic written English by analyzing a corpus

of 100,000 abstracts with 26 million words (Nguy & Ha, 2022). The findings indicated that there is a considerable variation in the lexical demands of academic language across different subject areas. Additionally, it was found that the comprehension of written academic English is significantly more demanding compared to spoken language and it required a much larger vocabulary size spanning from 7000 to 15,000 most frequent words in BNC/COCA base lists. Such findings point to the importance of building a large vocabulary base for university students that goes beyond the accepted thresholds of lexical knowledge for everyday language use.

Mid-frequency vocabulary

A commonly employed approach for teaching vocabulary involves grouping English words into four distinct categories based on their frequency of use and level of technicality. These categories consist of high-frequency, academic (semi-technical), technical, and low-frequency words (Coxhead & Nation, 2001; Nation, 2001). Language education programs designed for beginner English language learners are advised to focus on and prioritize the items in the first group (Nation & Waring, 1997). There are a number of resources for such words but the most notable and classic example is the General Service List (GSL) (West, 1953). These high-frequency vocabulary refers to the most basic English words that constitute a significant proportion of daily conversations and most words in all types of writing (Nation, 2001). On the other end of the continuum are the low-frequency words which are rarely used across different text types, and might be ignored in teaching if they are considered less important for comprehension (Laufer, 2013). Technical vocabulary constitutes subject-specific words used in different specialized fields. Academic or semi-technical vocabulary is viewed as "formal, context-independent words with a high frequency and/or wide range of occurrence across scientific disciplines, not usually found in basic general English courses; words with high frequency across scientific disciplines" (Farrell, 1990, p. 11).

Although this four-part categorization has been influential in many ESP vocabulary studies, recent research has indicated that it is untenable for a pedagogical description of vocabulary (Schmitt & Schmitt, 2014). More specifically, given the high demands of reading academic and authentic texts (Hsu, 2014, 2018; Laufer, 2013, 2020; Nation, 2006; Nguy & Ha, 2022; Schmitt et al., 2011), the four-part categorization leaves a considerable gap between high- and low-frequency words, that academic and technical vocabulary fail to cover appropriately (Vilkaitė-Lozdienė & Schmitt, 2019). Schmitt and Schmitt (2014) suggested a potential solution to address this issue by proposing two changes to the existing model. Firstly, they argued for expanding the high-frequency group to encompass the 3000 most frequently used word families based on the BNC/COCA lists (Nation, 2012). Secondly, they recommended introducing a new category referred to as *mid-frequency vocabulary*, which would encompass words ranging from the fourth to the ninth list in the BNC/COCA wordlists (3001–9000) (Nation, 2012). Although there are still ongoing debates around a suitable size for high-frequency words (Cobb, 2007; Dang & Webb, 2016; Dang et al., 2020; Nation, 2013), it has been argued that together with these words, mid-frequency words "represent the amount of vocabulary needed to deal with English without the need for outside support" (Nation, 2013, p. 18).

The importance of mid-frequency vocabulary is widely acknowledged in the literature (Masrai, 2019; Schmitt et al., 2011, 2017; Webb & Rodgers, 2009b), and the receptive knowledge of these items are considered to be necessary for 95% or minimal, and 98% or adequate comprehension thresholds (Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Moreover, it has been argued that after learning high-frequency vocabulary, further vocabulary knowledge developments are idiosyncratic and mostly related to one's job and interests (Nation, 2001). In this regard, Vilkaitė-Lozdienė and Schmitt (2019) believe that since mid-frequency words are more domain-specific, it is not possible to come up with general mid-frequency lists to serve the needs of a diverse group of learners in language education programs. Furthermore, given that mid-frequency words are not encountered frequently enough compared to high-frequency words (Cobb, 2007), there is no systematic attention to these words in textbooks or teacher talk in language education programs (Schmitt & Schmitt, 2014; Yang & Coxhead, 2020), which makes mastering them a daunting task for language learners at various levels. According to Vilkaitė-Lozdienė and Schmitt (2019), mid-frequency vocabulary can be dealt with by focusing on lists developed for specific purposes.

Previously compiled word lists based on RAs

In recent years, numerous studies have explored the vocabulary requirements for various types of communication. For instance, studies on daily conversations, television shows, and movies have indicated that individuals need to know approximately 3000 word families to participate effectively in these types of discourse (Nation, 2006, 2013; Rodgers & Webb, 2011; van Zeeland & Schmitt, 2012; Webb & Rodgers, 2009a, 2009b). Additionally, to comprehend academic spoken English, an extra 1000 word families are necessary (Dang & Webb, 2014). However, the vocabulary knowledge needed for understanding authentic texts such as academic books, novels, and newspapers goes beyond these thresholds. According to some estimates, for achieving sufficient comprehension of written discourse in English, a vocabulary size of around 8000–9000 word families seems necessary (Hsu, 2014, 2018; Nation, 2006; Schmitt et al., 2011). Accordingly, informed by the four-part categorization approach, a consistent concern in ESP vocabulary research has been the establishment of general academic (Browne et al., 2013; Coxhead, 2000, 2019; Farrell, 1990; Gardner & Davies, 2014) or field-specific academic vocabulary lists (Green & Lambert, 2018; Hsu, 2013; Lei & Liu, 2016; Valipouri & Nassaji, 2013; Wang et al., 2008; Ward, 2009; Yang, 2015). General academic lists aim to identify common academic words shared among different disciplines, while field-specific lists focus on the academic vocabulary used in specific fields (Dang, 2019). In recent years, the common core approach to academic vocabulary has been challenged in light of the accumulated evidence that provided a strong case for the specificity of academic literacy (Durrant, 2014, 2016; Hyland & Tse, 2007). Consequently, given the inadequacy of general academic word lists for vocabulary-learning needs across disciplines, developing field-specific word lists has attracted increased attention (Chen & Ge, 2007; Hyland & Tse, 2007; Konstantakis, 2007; Li & Qian, 2010; Martínez et al., 2009; Masrai & Milton, 2018; Valipouri & Nassaji, 2013).

In an early study within this field of research, Wang et al. (2008) analyzed a corpus of 288 medical research articles, containing 1093011 words. They applied three criteria

proposed by Coxhead (2000), including specialized occurrence, range, and word frequency, to develop a list of specialized academic vocabulary for medical writing. Specifically, the chosen words had to appear at least 30 times in the entire corpus (word frequency), be used in at least 16 of the 32 subject areas (range), and fall outside of the top 2000 most frequently used English words (specialized occurrence) according to the GSL (West, 1953). The study identified 623 word families that met the abovementioned criteria, and those items provided around 12.24% coverage of the corpus of medical RAs. Lei and Liu (2016) created a new academic wordlist for medical students. Their corpus contained medical RAs (760 articles, taken from 38 journals in medicine, 2.7 million words) and medical textbooks (3.5 million words). The study identified 819 lemmas as medical academic vocabulary list that consisted of 444 nouns, 133 verbs, 219 adjectives, and 23 adverbs. The findings indicated that the new list provided much more improved coverage in medical RAs (i.e., 19.44%) compared to the previously developed medical word list by Wang et al. (2008). Moreover, using an enhanced methodology for their data analysis and word selection criteria, the study filtered out some frequent but less valuable words related to medicine. Unlike Wang et al. (2008), who excluded the GSL items from their final list, this study investigated highly frequent and general service lemmas in the context of medical texts and identified 313 items with special meaning in medical fields, and subsequently included those items in their final list.

In another study, Valipouri and Nassaji (2013) investigated a 4-million words corpus, compiled with 1185 Chemistry RAs, to establish a Chemistry academic word list for EFL students. Following Coxhead (2000), the study identified 1577 word families that met the frequency, range and specialized occurrence criteria. However, it should be noted that although the researchers included specialized occurrence as a criterion in their word selection procedures, they did not exclude the GSL items and ended up with 1577 word families in their first round of data analysis. To establish the Chemistry academic word list, the researchers then excluded abbreviations, technical terms, and function words (such as pronouns, articles, and propositions) from the initial list, which resulted in a list with 1400 word families containing 683 GSL, 327 AWL, and 390 non-GSL/AWL word families. The coverage provided by GSL items in this study was 65.46% of the entire corpus, and the remaining AWL and non-GSL/AWL items provided additional coverage of 16.83%. The developed Chemistry academic word list provided total coverage of 81.18% in the 4-million words corpus of Chemistry RAs.

Yang (2015) established a nursing academic word list. This study investigated a corpus containing 1006934 words called Nursing Research Articles Corpus (NRAC) with RAs from different sub-disciplines within the nursing field. Similar to Wang et al. (2008), the criteria proposed by Coxhead (2000) were also employed in this study, and 676 word families were identified as Nursing Academic Word List (NAWL), which provided 13.46% coverage of NRAC. The top 100 word families in NAWL provided 6.75% coverage in the corpus, which is considerably high, given that the remaining 576 word families provided 6.89% coverage. Finally, a study by Liu and Han (2015) investigated an environmental science corpus of 200 RAs with 862242 running words. Using the three criteria proposed by Coxhead (2000), in addition to a dispersion criterion, the study identified 458 word families for the environmental academic word list (EAWL) and reported a coverage of 70.61% for the GSL items. The EAWL word families provided 15.43% coverage

in the corpus of environmental RAs. Liu and Han (2015) validated their EAWL using two additional corpora. The first validating corpus was created using RAs from ten environmental science academic journals (20 RAs, 99942 words). The second corpus was compiled with RAs from the same academic journals used in the study (20 RAs, 78,827 words). The 458 word families in the EAWL accounted for 14.92% of the tokens in the first corpus and 15.59% in the second corpus. These coverage figures were close to 15.43% coverage of the list in 200 RAs. The study concluded that the EAWL serves the environmental science academic learning better than other academic word lists such as the AWL (Coxhead, 2000).

The present study

The abovementioned studies contributed significantly to our understanding with respect to the specialized uses of vocabulary in RAs. It is now well established that there is a need to develop more restricted and discipline-related vocabulary lists to satisfy the needs of university students and researchers in different subject areas. Moreover, creating discipline-related vocabulary lists for university students is in line with the current emphasis on promoting disciplinary literacy, which highlights the close connection between language and a given discipline (Airey et al., 2017; Kuteeva & Airey, 2014). Nonetheless, some methodological issues in this line of research deserve further attention. First, most studies have used the GSL (West, 1953) and the AWL (Coxhead, 2000) as the lists for high-frequent and academic vocabulary in English, and both lists have been criticized recently for various shortcomings associated with them (Gardner & Davies, 2014; Hyland & Tse, 2007; Martínez et al., 2009). A related concern to this issue stems from the proposed distinctions among general service, academic, and technical words. In this regard, studies indicated a large overlap between academic and general vocabulary, making it difficult to define and operationalize the related constructs (Green & Lambert, 2018; Schmitt & Schmitt, 2014). The same difficulty arises when dealing with highly specialized and technical words, as there is a technicality gradient associated with those items (Ha & Hyland, 2017). These limitations in defining different types of words foreground adopting a more pedagogically functional classification of vocabulary in ESP studies.

The second issue of concern pertains to the limited size of the corpora examined in previous studies. For instance, the range varied from 862,242 running words for environmental science (Liu & Han, 2015) to four million words for Chemistry (Valipouri & Nassaji, 2013). Previous research suggests that a corpus of one million running words is sufficient to obtain a reliable list of highly frequent words (Brysbaert & New, 2009). However, for vocabulary items beyond this range, a corpus of around 20 million words is required (Nation, 2016; Sorell, 2013). As studies discussed above produced vocabulary lists that extended beyond highly frequent words, the relatively small sizes of the investigated corpora make it necessary to approach their findings with caution. To address this limitation, the present study used big data in the form of a large corpus of research articles to get more reliable results. Third, although validation of discipline-specific word lists is of crucial importance in their application (Coxhead, 2018b; Dang, 2019), it seems that this consideration has been neglected to a large extent in the existing literature, and only a few studies have built validating

corpora to test their findings against different databases (Liu & Han, 2015). Furthermore, given the importance of mid-frequency vocabulary in general, its role in ESP vocabulary studies remained less explored. The current study aimed to fill these gaps in the literature by answering the following research questions:

(1) What is the coverage of the mid-frequency vocabulary in chemistry RAs?
(2) What are mid-frequent vocabulary items used frequently in the corpus? What is the coverage of frequently used mid-frequency words?
(3) What is the coverage of frequently used mid-frequency words in chemistry in other specialized and general corpora?

Method

Corpus compilation

Following the criteria of representativeness, balance, and size (McEnery & Hardie, 2011; Sinclair, 1991), a corpus of around 50,000 chemistry RAs with 278,000,000 running words was created and analyzed. To create the corpus, the study employed the Ant Corpus Generator (AntCorGen) program, a freeware tool that creates subject-specific corpora utilizing the PLOS ONE database (Anthony, 2019). The corpus represented a wide range of sub-areas within chemistry, such as organic chemistry, inorganic chemistry, chemical thermodynamics, physical chemistry, chemical kinetics, spectroscopy, theoretical chemistry, and more. In the process of compiling RAs for vocabulary profiling, references and appendices were excluded from the corpus, while all other sections such as abstracts, introductions, materials and methods, findings, discussions, and conclusions, were included. Next, to make the database manageable for word profiling purposes (see the following subsection), the collected research articles were combined into a single text file, and then using the Ant File Splitter program (Anthony, 2017), the text was broken down into 278 sub-corpora, each containing one million running words. Given the enormous size of the corpus and a large number of files, creating smaller sub-corpora and balancing the number of words in 278 text files helped in establishing sound criteria for word selection, which was based on the guidelines proposed by Nation (2016) for analyzing very large corpora.

Lexical analysis software

The process of lexical profiling for chemistry research articles (RAs) was conducted using the Ant Word Profiler program (Anthony, 2021). This software, which is available as freeware, facilitates the assessment of vocabulary complexity levels in the text files that are uploaded to the program. By default, the program employs two word lists, the General Service List (GSL) and the Academic Word List (AWL) comprising 570 words. Nevertheless, it is feasible to evaluate the text against other vocabulary lists, which can be manually uploaded to the program. Once the data has been processed, the software produces a set of vocabulary statistics and comprehensive frequency information pertaining to the corpus. In the current study, the BNC/COCA lists (1st–34th) (Nation,

2012) freely accessible from the Ant Word Profiler website were used for profiling RAs. Given that the software could not provide the lexical profile output for 50,000 files, the 278 sub-corpora were given as the input texts to Ant Word Profiler.

Word selection criteria

Developing word lists involves a crucial decision regarding the unit of counting, with previous studies utilizing various units such as types, lemmas, and families. One widely adopted strategy has been to employ word families, which consist of the base word and its inflected forms as well as transparent derivations (Bauer & Nation, 1993). This approach is based on the assumption that familiarity with the base word in a family can aid in comprehending its derived and inflected forms (Coxhead, 2000; Webb & Nation, 2017; Xue & Nation, 1984). It is important to acknowledge that this perspective has been contested in recent years, as some researchers have called into question the effectiveness of using word families. Instead, they advocate for the use of lemmas, which comprise a headword along with its inflected forms which is more appropriate for creating pedagogically useful lists (Brown et al., 2020; Gardner & Davies, 2014; Lei & Liu, 2016). In a recent discussion, Nation (2016) argued that the choice for the counting unit should match the purposes of the list development. In this regard, it has been argued that the lower levels (types and lemmas) are suitable for productive purposes (Dang, 2019; Durrant, 2014), and higher levels for receptive uses of vocabulary items (Dang et al., 2017; Nation, 2016). Given this important consideration, and based on the intended uses of the list developed in the current study that aims to help university students and researchers in the chemistry field to read and write research articles in English, the present study used lemma as the unit for counting and analyzing words.

The data was processed through several steps. Firstly, the outputs obtained from Ant Word Profiler were transferred to Microsoft Excel documents to identify frequently occurring mid-frequency words across the entire corpus. To achieve this, Coxhead's (2000) three criteria of range, frequency, and specialized occurrence were utilized. In addition, due to the vast size of the corpus and its 278 sub-corpora, a fourth criterion of dispersion was also employed, as recommended by Egbert and Biber (2019). To meet the range criterion, vocabulary items had to appear in at least 200 sub-corpora (i.e., in over 70% of the source texts). For frequency and dispersion, the words needed to occur a minimum of 7923 times in the entire corpus and at least 28.5 times in each of the smaller corpora containing one million running words. Finally, the selected words had to fall outside of the 3000 most frequent word families in English, as determined by the BNC/COCA lists (Nation, 2012).

Results and discussion

Table 1 presents the lexical profile of chemistry RAs based on BNC/COCA lists. The 1000 most commonly used words in the English language make up a significant portion of the corpus, comprising 144,158,658 tokens and 5639 types, which account for 51.86% of the corpus. The next 1000 most frequent words account for 31,538,369 tokens and 52733 types, representing 11.34% of the corpus. The third BNC/COCA list provided 10.83% coverage, with 30,100,486 tokens and 5010 word types. Combined, these

Table 1 The lexical profile of chemistry RAs based on BNC/COCA lists

BNC-COCA lists	Token	Token%	CumToken%	Type	Group
1	144158658	51.86	51.86	5639	999
2	31538369	11.34	63.2	5273	1000
3	30100486	10.83	74.03	5010	1000
4	8785085	3.16	77.19	3839	996
5	5529036	1.99	79.18	3220	985
6	3410422	1.23	80.41	2971	976
7	3087538	1.11	81.52	2493	955
8	1873841	0.67	82.19	2236	923
9	1252030	0.45	82.64	1949	907
10–30	9895518	3.58	86.22	16562	10800
31–34	20637310	7.42	93.64	16409	15002
0	17731707	6.38	100.02	616952	616952
Total	278000000			682553	651495

three lists as the high-frequency vocabulary in English provided 74.03% coverage. As the first base list in the mid-frequency vocabulary, the coverage of the fourth list dropped to 3.16%, with 8,785,085 tokens and 3839 types. The coverage provided by lists 5 to 9 was 1.99%, 1.23%, 1.11%, 0.67%, and 0.45%, respectively. The mid-frequency words (i.e., base lists 4 to 9) provided a total coverage of 8.61%. Adding this coverage to 74.3% provided by high-frequency words, the high- and mid-frequency words accounted for 82.64% of the entire corpus. Moreover, the low-frequency words (i.e., base lists 10–30) provided 3.58% coverage. Finally, the last base lists that contain proper names, exclamations, alphabet letters, transparent compounds, and abbreviations provided a coverage of 7.42%. Totally, the BNC/COCA lists accounted for 93.64% of the corpus, and 6.38% of the words in the database were beyond BNC/COCA lists.

To generate a list of mid-frequency words relevant to chemistry RAs, we examined items listed in BNC/COCA lists 4 to 9. A total of 560 words (i.e., lemmas) satisfied the criteria proposed for this study, with these words comprising 17,796,194 tokens and 6.40% of the overall corpus. This indicates that the remaining mid-frequency words in the corpus accounted for just 2.21%. To clarify this finding further, data analysis revealed that 5742 mid-frequency word families that comprised 16,708 types or individual words were used in the corpus (with a total coverage of 8.61%). The selection criteria resulted in identifying 560 lemmas (expanding to 1074 word types) that provided around 6.40% coverage. Accordingly, the identified items are highly valuable for chemistry students and researchers who read academic articles in this field.

The study identified 560 mid-frequency words with a coverage of 6.40% in chemistry RAs, which is lower than the 12.24% coverage of 623 academic word families compiled for medical RAs (Wang et al., 2008), 13.46% coverage of 676 academic word families in nursing RAs (Yang, 2015), and 15.59% coverage of 458 word families in environmental RAs (Liu & Han, 2015). Moreover, Valipouri and Nassaji (2013) also identified 717 word families beyond the GSL that provided 16.83% coverage in their corpus of chemistry RAs, which is considerably larger than the coverage of 610 mid-frequency words identified in the current study. It is worth noting that the approach taken to define high-frequency vocabulary in this study was different from previous

studies. In comparison to studies that used the GSL as a reference for high-frequency words, this study utilized the first to third BNC/COCA lists and found a higher coverage of 74.03% for high-frequency words in the English language. For instance, previous studies found 65.46% coverage for the GSL in chemistry RAs (Valipouri & Nassaji, 2013) and 70.61% in environmental RAs (Liu & Han, 2015). Moreover, an analysis of 1400 academic vocabulary items in chemistry RAs identified by Valipouri and Nassaji (2013) revealed that 73.79% (about 1000 word families) of these items are high-frequency words when considering the 3000 most frequent word families in English. This suggests that several words previously categorized as academic vocabulary in research articles should be reclassified as high-frequency vocabulary when extending the range of high-frequency words to the 3000 most frequent word families.

Comparing the 1400 items identified by Valipouri and Nassaji (2013) as Chemistry academic word list with the current study findings, it was found that the two lists had only 11.18% shared items. As 267 out of 1400 word families in the list were mid-frequent vocabulary, the overlap of these items with the 560 items identified in this study was investigated. The findings showed that only 27.51% of the mid-frequent words in Valipouri and Nassaji (2013) were also present in the list of mid-frequency chemistry vocabulary developed in this study. This underscores the importance of replication studies in ESP vocabulary studies (Coxhead, 2018a, 2018b, 2019; Miller, 2022). Moreover, it should be noted that the current study investigated a much larger corpus, which resulted in identifying the words used frequently in chemistry RAs. Finally, compared to medical texts, 560 mid-frequency words in chemistry RAs provided less coverage than the 19.44% coverage of the 819 lemmas in the new academic wordlist for medical students (Lei & Liu, 2016). However, it should be noted that these lemmas contained some items from high-frequent English words in the GSL, which explains the observed differences.

In order to create a more pedagogically useful mid-frequency list, the 560 headwords were divided into five bands (see "Appendix A"). The following table provides additional information regarding the number of words in each band and their coverage in the corpus. As indicated below, the first 100 most frequent items accounted for 3.05% of the tokens in the corpus and should be regarded as very important words in chemistry RAs. The first 300 word families provided total coverage of 5%. Table 2 also provides some sample items from each band ordered by their frequency, and *acid,*

Table 2 mid-frequency words in chemistry RAs in 6 bands

Band	Coverage (%)	Sample words
1	3.05	Antibody, peptide, amino, receptor, substrate, buffer, mutation, subject, simulation, incubated
2	1.01	Inflammatory, interface, gradient, nucleus, helix, extracellular, catalytic, proliferation, duration, secretion
3	0.74	Cardiac, enrichment, deficient, inflammation, kinetic, specimen, aggregation, pulse, amplitude, pellet
4	0.53	Mineral, peripheral, gram, administered, analytical, cocaine, adjacent, scaffold, ammonium, deposition
5	0.6	Poly, aerobic, semi, micro, bulk, placebo, drought, pulmonary, horizontal, aligned

residue, glucose, mutant, assay, domain, induced, enzyme, membrane, and *parameter* were the ten most frequent mid-frequency words that accounted for 2,115,966 tokens and about 0.76% of the corpus.

Validating the list

After establishing the mid-frequency vocabulary list for chemistry RAs (Appendix A), the study used a number of specialized and general corpora to validate the list. To this end, the list was investigated against a 31-million-word corpus of RAs published in different disciplines. More specifically, this specialized corpus contained an equal number of RA (i.e., 600) in biology, medicine, earth sciences, ecology, computer sciences, engineering and technology, medicine and health, physical sciences, and social sciences. The selection of these research areas was based on the categorization used in Ant Corpus Generator software (Anthony, 2019). This corpus also contained 600 chemistry research articles from the Science Direct database. The coverage of the 560 mid-frequency items frequently used in chemistry RAs in different corpora is shown in Fig. 1. As shown below, the list provided 3.97% coverage in RAs published in biology, 2.63% in medicine, 2.34% in earth sciences, and 2.3% in ecology. One possible reason for the high coverage of the list in biology is that one main subarea of chemistry is organic chemistry. Accordingly, there might be considerable overlap between the fields. Additionally, chemistry has been regarded as the central science connecting life sciences and physical sciences (Henry & Malin, 2010). In this regard, the list has around 2.5% coverage in medicine, earth sciences, and ecology. The coverage of the list dropped considerably in other research areas, including people and places (1.74%), mathematics (1.58%), social sciences (1.45%), and computer sciences (1.23%). The lowest coverage was in technology (i.e., 1.1%). Finally, the list provided around 5.1% coverage in chemistry RAs collected from the Science Direct database. Although this is less than the 6.4% coverage of the list in the original corpus, it is considerably larger than the coverage provided for RAs in different research areas. Moreover, given the differences in the size of corpora, such variation is inevitable (Nation, 2016).

The second specialized validating corpus was compiled based on the classification of disciplinary groups into four major areas, including (1) Arts and Humanities (AH), (2) Life Sciences (LS), (3) Physical Sciences (PS), and (4) Social Sciences (SS). In

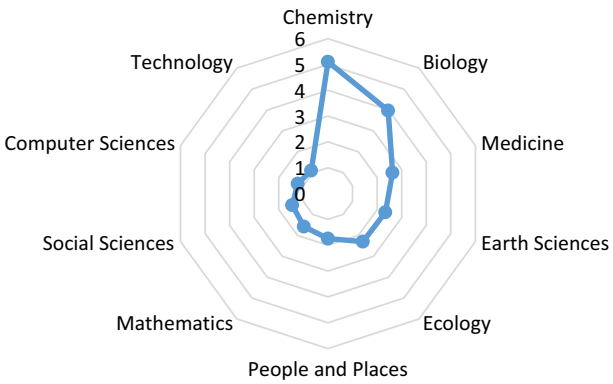

Fig. 1 Coverage (%) of the frequently used mid-frequency words in chemistry RAs across other research areas

Table 3 the coverage of 560 mid-frequent words in chemistry RAs in validating corpora

Statistics					
Level	**Token**	**Token%**	**Cumtoken%**	**Type**	**Group**
BNC/COCA 1st	55972999	59.03	59.03	5889	999
BNC/COCA 2nd	12706677	13.4	72.43	5499	1000
BNC/COCA 3rd	10935436	11.53	83.96	5341	1000
560 mid-frequency	2412420	2.54	86.5	1050	577
Off lists	12791597	13.49	99.99	314,422	314,422
Total	94819129				

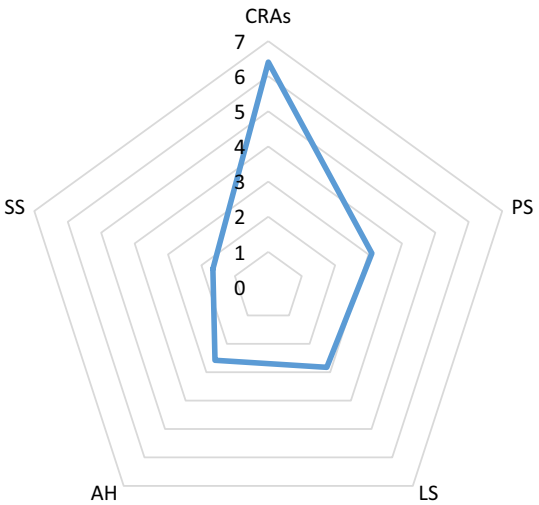

Fig. 2 Coverage (%) of the frequently used mid-frequency words in chemistry RAs across disciplines

doing so, an equal number of RAs were downloaded for each group (i.e., 4000) using Ant Corpus Generator tool. The inclusion of different fields under each category was based on the classification of the disciplines used in the British Academic Written English (BAWE) corpus (https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/). This large corpus contained 16,000 RAs and around 95 million words. The results of profiling the corpora against high-frequency vocabulary and frequently used mid-frequency words in chemistry are provided in Table 3. Overall, approximately 84% of the words in the second validating corpus were high-frequency words, which is 10% higher than the corpus of chemistry RAs. However, frequently used mid-frequency words in chemistry accounted for 2.54% of the corpus, which is around 4% lower than in chemistry alone.

Figure 2 shows the coverage of the mid-frequency word list for chemistry in different disciplines. The list provided around 3.1% coverage in PS. As chemistry is within PS, this high coverage indicates that the items are more frequent in this discipline. The coverage of the list was lower in LS; however, it provided 2.83% coverage which is close to PS. Taken together, the list provided around 3% coverage in PS and LS.

Additionally, the list accounted for 2.58% of the words in AH. The lowest coverage among the four disciplinary groups was for SS, as the list provided only 1.66% coverage of the words used in RAs published in this discipline. The results for validating the list against two balanced and specialized corpora (with 30 and 95 million words) indicated that the identified items are highly relevant for chemistry. Additionally, data analysis indicated that the list provides around 2.5% coverage in other disciplinary areas which is around 4% lower than chemistry.

After analyzing the list of frequently used mid-frequency words in chemistry against specialized corpora compiled by RAs, the coverage of the list was examined using general corpora. In this regard, the list was validated against different sections of the Corpus of Contemporary American English (COCA, https://www.english-corpora.org/coca/). The first corpus was a 37-million word sample from iWeb (https://www.english-corpora.org/iweb/), a 14-billion words database created from around 22 million web pages. The analysis indicated that the list developed in this study provided 0.72% coverage in the iWeb corpus. Additionally, analyzing the list against the COCA sample corpus (around 9 million words) pointed to 0.47% coverage. More specifically, the list provided 1.14% coverage in the academic sub-section (1.1 million words) of the COCA sample; however, the list coverage was 0.37% in other sections (8 million words) that included fiction, news, movies, blogs, and magazines. Finally, the coverage of the list was examined in relation to other general corpora samples available on the COCA website (https://www.corpusdata.org/formats.asp). This general corpus contained around 10 million words (Wikipedia, SOAP, movies, TV, corona, NOW, GloWbE, COHA), and the list provided 0.30% coverage of the entire text. Overall, data analysis indicated that the coverage of the list is considerably low in general corpora compared to specialized corpora created from RAs. Consequently, the validation process showed that the items in "Appendix A" (1) are highly relevant to field of chemistry, (2) contain a considerable proportion of mid-frequency words used in RAs generally, and (3) are specialized (or technical) terms with infrequent use in general English texts.

Conclusion

The current corpus-based study investigated a large corpus of chemistry RAs containing 278 million words to establish a list of mid-frequency vocabulary for researchers (and EAP students) in the field of chemistry. The study found that 560 mid-frequency words provide 6.4% coverage in chemistry RAs, and with the high-frequency words (the first 3000 most frequent words in English) based on the BNC/COCA list, the coverage reaches 80.43%. By adding the coverage of the 31–34 BNC/COCA lists that include proper names, exclamations, alphabet letters, transparent compounds, and abbreviations, the total coverage reaches around 88%. Accordingly, learning the 560 mid-frequency words in chemistry is a significant step for EAP students in chemistry and researchers in this field. Although understanding the text goes beyond the knowledge of the lexical items, such vocabulary knowledge significantly facilitates reading comprehension (van Zeeland & Schmitt, 2012). Moreover, the developed mid-frequency word list has been validated using specialized and general corpora. The low coverage of chemistry mid-frequency words in the general corpus (0.30%) and around 2.5% coverage in

other disciplines indicates that the list contains field-specific words that are more relevant to chemistry.

The findings of the study has implications for chemistry students, researchers, EAP teachers, and materials developers in this field. First, the list of mid-frequency items identified in this study could be an important vocabulary-learning goal for chemistry students and researchers. In order to get the most out of the time spent on learning these items, the words are divided into five bands depending on their importance, and any effort to learn the first 300 words should bring the most benefit. Second, given the importance of mid-frequency vocabulary, EAP teachers in chemistry can focus on these items in the classroom and spend more time focusing learners' attention on these items. This can be realized, for example, through using different approaches within focus on form in vocabulary instruction, such as task-embedded and task-related instruction (Laufer, 2005). Third, as mid-frequency words receive no systematic attention in the textbooks (Schmitt & Schmitt, 2014), there is a need for intentional focus on these items in EAP materials for chemistry. The study also has some methodological implications for vocabulary studies in ESP. First, given the recent developments in corpus analysis software, the study compiled and analyzed a large corpus of RAs. Given the importance of corpus size in creating word lists (Sorell, 2013), future studies can use the tools employed in the current study to create field-specific word lists. Even teachers, students, and researchers can create and use word lists if they receive relevant training in using such tools. Second, by comparing our list with a previous study on chemistry RAs (Valipouri & Nassaji, 2013), it was found that there is a considerable difference in the produced word lists, and only 11.18% of the items were the same in the final lists. Although the current study used a different corpus with a much larger size (70 times larger), this observation highlights the importance of replication research in ESP vocabulary studies (Miller, 2022) and using larger corpora to get more reliable results. Finally, the study further highlighted the need for validating the produced word lists to see their coverage in different text types. This will help users of the lists make informed decisions on setting vocabulary learning goals.

The current study had some limitations. First, texts from a single genre, i.e., RAs, were used to create a mid-frequency list for chemistry students and researchers. As these groups need to read other text types, such as textbooks, posters, and lab manuals, to name just a few, there is a need to investigate the role of mid-frequency words in such text types. Second, the study was concerned with identifying individual words; nonetheless, given the role of lexical bundles and chunks in academic discourse (Biber & Barbieri, 2007; X. Liu et al., 2023), there is also a need to create pedagogical resources containing such items. Moreover, incorporating the output from corpus studies into instructional programs has received far less attention in the literature. Such endeavors will bring the real benefits of word list research to stakeholders in EAP. In this regard, a promising direction for future research is integrating wordlist research with materials development for language teaching.

Appendix A: mid-frequency words in chemistry

Band 1 (3.05% coverage)

Acid, residue, glucose, mutant, assay, domain, induced, enzyme, membrane, parameter, antibody, peptide, amino, receptor, substrate, buffer, mutation, subject, simulation, incubated, activation, RNA, serum, insulin, inhibitor, oxygen, plasma, fluorescence, ion, spectrum, fraction, inhibition, metabolic, terminal, primer, metabolism, pre, genome, cellular, phosphate, leaf, gel, mediated, coli, vitamin, incubation, protocol, synthesis, liver, transcription, tumor, vector, linear, lipid, particle, neuron, loop, ethanol, purified, uptake, coefficient, affinity, hydrogen, conformation, sodium, mitochondrial, nitrogen, induce, intracellular, matrix, abundance, deviation, yeast, diabetes, blot, kit, sigma, nutrient, dynamics, degradation, calcium, regression, interval, intake, electron, motif, replicate, quantify, induction, flux, median, transcript, activated, physiological, deletion, inhibit, lung, threshold, strand, biomass.

Band 2 (1.01% coverage):

Inflammatory, interface, gradient, nucleus, helix, extracellular, catalytic, proliferation, duration, secretion, microbial, algorithm, diameter, zinc, microscopy, overnight, organism, tolerance, diffusion, solvent, axis, intermediate, differentiation, genomic, arrow, salinity, alignment, chronic, stimulus, viability, dilution, beta, transcriptional, diluted, saline, quantification, fusion, cholesterol, cohort, basal, kidney, UV, viral, putative, spatial, inhibited, sediment, antioxidant, carbohydrate, encoding, optimal, fluid, utilize, acute, aggregate, bead, soluble, tag, fluorescent, toxicity, embryo, centrifuge, chromosome, cleavage, parasite, biochemical, diabetic, inhibitory, donor, dynamic, purification, graph, electrode, deficiency, clone, polymer, differential, identical, replication, modulate, denote, template, docking, mitochondria, dissolved, microscope, respiratory, pathogen, trait, defect, alpha, renal, thermal, fasting, kinetics, antibiotic, prevalence, minimal, morphology, pore.

Band 3 (0.74% coverage):

Cardiac, enrichment, deficient, inflammation, kinetic, specimen, aggregation, pulse, amplitude, pellet, array, larva, lesion, enriched, velocity, centrifugation, voltage, trajectory, harvested, node, verify, conversion, tagged, equilibrium, validation, activate, amplified, residual, simulated, multi, magnitude, vascular, dye, abundant, nicotine, sensor, efficacy, coral, amplification, maximal, precursor, differentially, antigen, urine, epithelial, mediate, copper, lipids, toxic, hormone, therapeutic, synthetic, classification, impaired, neuronal, collagen, accession, synthesized, maternal, starch, onset, fungal, dependence, inoculate, diagnosis, chloride, mammalian, spike, transient, potassium, intact, filament, classified, overlap, nitrate, scenario, genus, temporal, cardiovascular, depletion, obesity, optical, dissociation, fetal, digestion, laser, blotting, ecosystem, cortex, locus, fluctuation, compartment, ethics, inclusion, homology, minimize, insect, saturation, spectral, upstream.

Band 4 (0.53% coverage):

Mineral, peripheral, gram, administered, analytical, cocaine, adjacent, scaffold, ammonium, deposition, dashed, precipitation, dysfunction, habitat, proton, cloned, complement, encode, null, accordance, oxide, toxin, robust, ex, vertical, trans, polar, aromatic, incidence, intestinal, peg, exponential, hybrid, wells, helice, decay, retention, coil, susceptibility, antagonist, dot, inactivation, wheat, diagram, mutated, acidic, encoded, fermentation, configuration, exclusion, virulence, calibration, adverse,

syndrome, hypertension, systemic, loci, lateral, fungi, neural, validated, digested, mid, skeletal, medication, annotation, saturate, potent, radius, optimized, duplicate, steroid, litter, decomposition, reef, ensemble, beneficial, modulation, cavity, consensus, capillary, crude, suppression, pancreatic, intrinsic, microorganism, urinary, sperm, isotope, prolong, cognitive, sterile, susceptible, obese, glut, resin, artery, batch, optimization, utilization.

Band 5 (0.6% coverage):

Poly, aerobic, semi, micro, bulk, placebo, drought, pulmonary, horizontal, aligned, grid, fixation, goat, molar, absent, helical, gut, dual, integrity, contamination, permeability, infusion, multivariate, secreted, homologous, morphological, resonance, amp, viable, diagnostic, caffeine, composite, inoculation, arterial, cumulative, spontaneous, buffered, indirect, gag, chemistry, tract, artificial, retinal, logistic, elongation, viscosity, ambient, germination, invasive, stationary, annotated, demographic, rotation, questionnaire, inducing, maize, broth, magnification, ammonia, impairment, nutritional, predominantly, offspring, turnover, activator, magnesium, clamp, sham, alkaline, inhibiting, bold, embedded, tobacco, sulfur, carbonate, comparative, photosynthetic, mesh, suppressed, arsenic, depleted, polymerization, proximal, displacement, lag, geometry, distilled, differentiated, posterior, cloning, numerical, abnormal, embryonic, irradiation, gastric, refinement, overlapping, distal, consecutive, biology, tuberculosis, ecological, dotted, cultivation, attenuated, cerebral, prism, inset, influenza, shear, serial, moisture, canonical, intestine, fibrosis, fused, inverse, larval, biotechnology, fret, phosphorus, mammal, colon, activating, vegetation, triple, cleaved, conversely, repression, binary, tandem, gamma, photosynthesis, automate, hybridization, penicillin, reproductive, anterior, ventilation, complementary, cadmium, relevance, qualitative, sensory, preliminary, diagnosed, immobilized, pathological, proximity, humidity, contrary, excretion, appendix, bleaching, longitudinal, mini, adherence, cultivated, asymmetric, dorsal.

Abbreviations

CALL	Computer assisted language learning
CEC	Cambridge English Corpus
EAP	English for Academic Purposes
EFL	English as a Foreign language
ESP	English for specific purposes
GSL	General Service List
AWL	Academic word list
BNC/COCA	British National Corpus—Corpus of Contemporary American English
RAs	Research Articles

Acknowledgements
Not applicable

Author contributions
All authors contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable for this study, as this is a corpus-based research with no human participants.

Competing interests
The authors declare no competing interests.

Received: 17 March 2023 Accepted: 20 July 2023
Published online: 01 October 2023

References

Airey, J., Lauridsen, K. M., Räsänen, A., Salö, L., & Schwach, V. (2017). The expansion of English-medium instruction in the Nordic countries: Can top-down university language policies encourage bottom-up disciplinary literacy goals? *Higher Education, 73*(4), 561–576. https://doi.org/10.1007/s10734-015-9950-2

Alderson, J. C. (2006). *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum.

Anthony, L. (2017). *AntFileSplitter* (1.0.0). Waseda University. https://www.laurenceanthony.net/software/antfilesplitter/

Anthony, L. (2019). *AntCorGen* (1.1.2). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antcorgen/

Anthony, L. (2021). *AntWordProfiler* (1.5.1w). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antwordprofiler/

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: A state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education, 17*(1), 44. https://doi.org/10.1186/s41239-020-00223-0

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics, 21*(2), 139–164. https://doi.org/10.1075/ijcl.21.2.01bak

Bauer, L., & Nation, I. S. P. (1993). Word Families. *International Journal of Lexicography, 6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Bazerman, C., Keranen, N., & Prudencio, F. E. (2012). Facilitated immersion at a distance in second language scientific writing. In M. Castelló & C. Donahue (Eds.), *University writing: Selves and texts in academic societies* (pp. 235–248). Brill. https://doi.org/10.1163/9781780523873_014

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*. https://doi.org/10.1093/applin/amaa061

Browne, C., Culligan, B., & Phillips, J. (2013). *The New Academic Word List*. http://www.newgeneralservicelist.org/nawl-new-academic-word-list

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching, 52*(4), 460–475. https://doi.org/10.1017/S0261444819000089

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. https://doi.org/10.1016/j.esp.2007.04.003

Clenton, J., & Booth, P. (Eds.). (2020). *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary*. Routledge. https://doi.org/10.4324/9780429285400

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology, 11*(3), 38–63.

Corcoran, J. (2017). The potential and limitations of an intensive English for research publication purposes course for mexican scholars. In M. J. Curry & T. Lillis (Eds.), *Global academic publishing: Policies, perspectives and pedagogies* (pp. 217–232). Multilingual Matters. https://doi.org/10.21832/9781783099245-021

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

Coxhead, A. (2018a). Replication research in pedagogical approaches to formulaic sequences: Jones & Haywood (2004) and Alali & Schmitt (2012). *Language Teaching, 51*(1), 113–123. https://doi.org/10.1017/S0261444815000221

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. Routledge.

Coxhead, A. (2019). Academic vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 97–110). Routledge. https://doi.org/10.4324/9780429291586-7

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

Coxhead, A., & Nation, I. S. P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252–267). Cambridge University Press. https://doi.org/10.1017/CBO9781139524766.020

Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge. https://doi.org/10.4324/9780429291586-19

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning, 67*(4), 959–997. https://doi.org/10.1111/lang.12253

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*(1), 66–76. https://doi.org/10.1016/j.esp.2013.08.001
Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL - International Journal of Applied Linguistics, 167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan
Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*. https://doi.org/10.1177/1362168820911189
Dodigovic, M., & Agustín-Llach, M. P. (2020). Introduction to vocabulary-based needs analysis. In M. Dodigovic & M. P. Agustín-Llach (Eds.), *Vocabulary in curriculum planning: Needs, strategies and tools* (pp. 1–6). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-48663-1_1
Durrant, P. (2014). Discipline and level specificity in University students' written vocabulary. *Applied Linguistics, 35*(3), 328–356. https://doi.org/10.1093/applin/amt016
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/10.1016/j.esp.2016.01.004
Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104. https://doi.org/10.3366/cor.2019.0162
Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the english of electronics and a study of semi-technical vocabulary* (No. 25; CLCS Occasional Paper).
Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education, 44*(1), 130–160. https://doi.org/10.3102/0091732X20903304
Flowerdew, J. (2015). Some thoughts on English for research publication purposes (ERPP) and related issues. *Language Teaching, 48*(2), 250–262. https://doi.org/10.1017/S0261444812000523
Flowerdew, J. (2019). The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. *Language Teaching, 52*(2), 249–260. https://doi.org/10.1017/S0261444819000041
Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. https://doi.org/10.1093/applin/amt015
Godwin-Jones, R. (2017). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning & Technology, 21*(1), 4–15.
Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology, 25*(1), 4–19.
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/10.1016/j.jeap.2018.07.004
Ha, A. Y. H., & Hyland, K. (2017). What is technicality? A technicality analysis model for EAP vocabulary. *Journal of English for Academic Purposes, 28*, 35–49. https://doi.org/10.1016/j.jeap.2017.06.003
Ha, H. T. (2022a). Lexical profile of newspapers revisited: A corpus-based analysis. *Frontiers in Psychology, 13*(1), 1–10. https://doi.org/10.3389/fpsyg.2022.800983
Ha, H. T. (2022b). Vocabulary demands of informal spoken English revisited: What does it take to understand movies, TV programs, and soap operas? *Frontiers in Psychology, 13*(1), 1–10. https://doi.org/10.3389/fpsyg.2022.831684
Henry, B., & Malin, J. M. (2010). Celebrating the international year of chemistry in 2011. *Chemistry in Australia, 77*(6), 16–17. https://doi.org/10.3316/ielapa.246887850785170
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. https://doi.org/10.1177/1362168813494121
Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*, 54–65. https://doi.org/10.1016/j.esp.2013.07.001
Hsu, W. (2018). The most frequent BNC/COCA mid- and low-frequency word families in English-medium traditional Chinese medicine (TCM) textbooks. *English for Specific Purposes, 51*, 98–110. https://doi.org/10.1016/j.esp.2018.04.001
Hyland, K. (2009). *Academic discourse: English in a global context*. Bloomsbury Academic Press.
Hyland, K. (2022). The scholarly publishing landscape. In C. Hanganu-Bresch, M. J. Zerbe, G. Cutrufello, & S. M. Maci (Eds.), *The Routledge handbook of scientific communication* (pp. 15–25). Routledge. https://doi.org/10.4324/9781003043782-3
Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x
Konstantakis, N. (2007). Creating a business word list for teaching business English. *Estudios De Linguistica Inglesa Aplicada (ELIA), 7*, 79–102.
Kuteeva, M., & Airey, J. (2014). Disciplinary differences in the use of English in higher education: Reflections on recent language policy developments. *Higher Education, 67*(5), 533–549. https://doi.org/10.1007/s10734-013-9660-6
Laufer, B. (1996). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge University. https://doi.org/10.1017/CBO9781139524643.004
Laufer, B. (2005). Focus on form in second language vocabulary learning. In S. H. Foster-Cohen, M. P. Mayo, & J. Cenoz (Eds.), *EUROSLA yearbook* (pp. 223–250). John Benjamins Publishing Company. https://doi.org/10.1075/eurosla.5.11lau
Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly, 47*(4), 867–872. https://doi.org/10.1002/tesq.140
Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*. https://doi.org/10.1002/tesq.3004
Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

Lee, H., Warschauer, M., & Lee, J. H. (2019). Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL, 31*(2), 135–149. https://doi.org/10.1017/S0958344018000162
Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42–53. https://doi.org/10.1016/j.jeap.2016.01.008
Li, Y., & Flowerdew, J. (2020). Teaching English for research publication purposes (ERPP): A review of language teachers' pedagogical initiatives. *English for Specific Purposes, 59*, 29–41. https://doi.org/10.1016/j.esp.2020.03.002
Li, Y., & Qian, D. D. (2010). Profiling the Academic word list (AWL) in a financial corpus. *System, 38*(3), 402–411. https://doi.org/10.1016/j.system.2010.06.015
Lillis, T., & Curry, M. J. (2010). *Academic writing in a global context: The politics and practices of publishing in English*. Routledge.
Liu, D., & Lei, L. (2020). Technical vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 111–124). Routledge. https://doi.org/10.4324/9780429291586-8
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/10.1016/j.esp.2015.03.001
Liu, X., Li, S., Fan, W., & Dang, Q. (2023). Corpus-based bundle analysis to disciplinary variations: Relocating the role of bundle extraction criteria. *English for Specific Purposes, 70*, 151–163. https://doi.org/10.1016/j.esp.2022.12.004
Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183–198. https://doi.org/10.1016/j.esp.2009.04.003
Masrai, A. (2019). Vocabulary and reading comprehension revisited: Evidence for high-, mid-, and low-frequency vocabulary knowledge. *SAGE Open, 9*(2), 2158244019845182. https://doi.org/10.1177/2158244019845182
Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes, 31*, 44–57. https://doi.org/10.1016/j.jeap.2017.12.006
McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method*. Cambridge University Press.
Miller, D. (2022). Replication as a means of assessing corpus representativeness and the generalizability of specialized word lists. *Applied Corpus Linguistics, 2*(3), 100027. https://doi.org/10.1016/j.acorp.2022.100027
Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a Second Language trainees. *System, 32*(1), 75–87. https://doi.org/10.1016/j.system.2003.05.001
Nation, I. S. P. (2012). *The BNC/COCA word family lists*. https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf
Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes, 63*(1), 59–81. https://doi.org/10.1353/cml.2006.0049
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656
Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.
Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 6–19). Cambridge University Press.
Nguy, N. L. Q., & Ha, H. T. (2022). Lexical profile of academic written English revisited: What does it take to understand scholarly abstracts? *SAGE Open, 12*(3), 21582440221126344. https://doi.org/10.1177/21582440221126342
Politzer-Ahles, S., Girolamo, T., & Ghali, S. (2020). Preliminary evidence of linguistic bias in academic reviewing. *Journal of English for Academic Purposes, 47*, 100895. https://doi.org/10.1016/j.jeap.2020.100895
Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193
Reinders, H., & Lan, Y.-J. (2021). Big data in language education and research. *Language Learning & Technology, 25*(1), 1–3.
Rodgers, M., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly, 45*(4), 689–717. https://doi.org/10.5054/tq.2011.268062
Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–225. https://doi.org/10.1017/S0267190511000055
Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212–226. https://doi.org/10.1017/S0261444815000075
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*, 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484–503. https://doi.org/10.1017/S0261444812000018
Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31–79. https://doi.org/10.1111/j.1467-9922.2010.00590.x
Sinclair, J. (1991). *Corpus, concordance, and collocation*. Oxford University Press.
Sorell, J. (2013). *A study of issues and techniques for creating core vocabulary lists for English as an international language*. Unpublished PhD thesis, Victoria University of Wellington, New Zealand.
Thomas, M., & Gelan, A. (2018). Special edition on language learning and learning analytics. *Computer Assisted Language Learning, 31*(3), 181–184. https://doi.org/10.1080/09588221.2018.1447723
Trang, N. H., Nguyen, D. T. B., & Ha, H. T. (2023). Vocabulary demands of academic spoken english revisited: A case of university lectures and TED presentations. *SAGE Open, 13*(1), 1–10. https://doi.org/10.1177/21582440231155334
Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*(4), 248–263. https://doi.org/10.1016/j.jeap.2013.07.001

van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479. https://doi.org/10.1093/applin/ams074
Vilkaitė-Lozdienė, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness: high-, mid-, and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge. https://doi.org/10.4324/9780429291586-6
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/10.1016/j.esp.2008.05.003
Ward, J. S., & Barker, A. (2013). *Undefined by data: A survey of big data definitions*. https://arxiv.org/pdf/1309.5821.pdf
Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170–182. https://doi.org/10.1016/j.esp.2009.04.001
Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.
Webb, S., & Rodgers, M. P. H. (2009a). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407–427. https://doi.org/10.1093/applin/amp010
Webb, S., & Rodgers, M. P. H. (2009b). Vocabulary demands of television programs. *Language Learning, 59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x
West, M. (1953). *A general service list of English words*. Longman, Green & Co.
Williamson, B. (2018). The hidden architecture of higher education: Building a big data infrastructure for the 'smarter university.' *International Journal of Educational Technology in Higher Education, 15*(1), 12. https://doi.org/10.1186/s41239-018-0094-1
Woodward-Kron, R. (2008). More than just jargon – the nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes, 7*(4), 234–249. https://doi.org/10.1016/j.jeap.2008.10.004
Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*, 215–229.
Yang, L., & Coxhead, A. (2020). A corpus-based study of vocabulary in the new concept English textbook series. *RELC Journal*. https://doi.org/10.1177/0033688220964162
Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes, 37*(1), 27–38. https://doi.org/10.1016/j.esp.2014.05.003

Publisher's Note