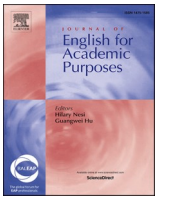
Journal of English for Academic Purposes 54 (2021) 101044

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap



When a bug is not a bug: An introduction to the computer science academic vocabulary list

David Roesler

Department of Applied Linguistics, Portland State University, 1825 SW Broadway, Portland, OR, 9720, USA

ARTICLE INFO

Keywords:
ESL
Word list
Computer science
Academic vocabulary
Language education

ABSTRACT

This article presents the Computer Science Academic Vocabulary List (CSAVL), a pedagogical tool intended for use by English-for-specific-purpose educators and material developers. A 3.5-million-word corpus of academic computer science texts was developed in order to produce the CSAVL. The CSAVL draws from the improved methodologies used in the creation of recent lemma-based word lists such as the Academic Vocabulary List (AVL) (Gardner & Davies, 2014) and the Medical Academic Vocabulary List (MAVL) (Lei & Liu, 2016), which take into account the discipline-specific meanings of academic vocabulary. The CSAVL provides specific information for each entry, including part of speech and CS-specific meanings in order to provide users with clues as to how each item is used within the context of academic CS. Based on the analyses performed in this study, the CSAVL was found to be a more efficient tool for reaching an minimal level of academic CS reading comprehension than the Academic Word List (AWL) (Coxhead, 2000), or the combination of the AWL with the Computer Science Word List (CSWL) (Minshall, 2013).

1. Introduction

Although academic vocabulary knowledge has been described as being vital to both reading comprehension (Corson, 1997; Jacobs, 2008; Nagy & Townsend, 2012) and academic success (Goldenberg, 2008), it may cause difficulties for learners because items of academic vocabulary often appear less frequently than general vocabulary items (Xue & Nation, 1984) and can also take on special meanings in the context of a specific academic discipline (Cohen et al., 1979; Lam, 2001). Academic word lists have been created to promote the acquisition of academic vocabulary and aid English for Academic Purposes (EAP) educators in setting vocabulary goals, selecting texts for instruction, and creating learning materials. Academic word lists can allow limited resources to be efficiently allocated in course design by focusing learners' attention on the vocabulary that will likely be most frequently encountered in academic reading and can also provide guidance for independent learners seeking an efficient path to developing the vocabulary knowledge necessary for reading comprehension (Laufer, 1989; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010).

1.1. The need for discipline-specific academic word lists

General academic word lists such as the Academic Word List (AWL) (Coxhead, 2000) attempt to represent academic vocabulary as a whole by identifying frequently used words within corpora of texts drawn from a variety of academic disciplines. Coxhead's AWL was designed as supplement to the General Service List (GSL) (West, 1953), one of the most widely-used general high-frequency word lists

E-mail address: roeslerdavidthomas@gmail.com.

https://doi.org/10.1016/j.jeap.2021.101044
Received 9 June 2021; Received in revised form 24 August 2021; Accepted 25 August 2021
Available online 28 August 2021

in English language education. Despite Coxhead's awareness of the problematic characteristics of West's GSL which have been described in more recent research (Browne, 2014; Gardner & Davies, 2014; Brezina & Gablasova, 2015; Lei & Liu, 2016), the AWL was intended to be combined with the GSL, allowing it to achieve something closer to the 95% lexical coverage required for the minimal comprehension of a text (Laufer, 1989; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010). In designing the AWL as a supplement, Coxhead excluded GSL items from the AWL, an approach that has been challenged for its problematic assumption that there is a clear distinction between the categories of high-frequency and academic words (Cobb, 2010; Gardner & Davies, 2014).

The AWL concept of a core 'general academic' vocabulary has also been challenged by Hyland and Tse (2007), who found that when testing the AWL on their own academic corpus, the coverage of the AWL was not evenly distributed across academic disciplines. Of the 570 word families of the AWL, only 36 were relatively evenly distributed across their three categories of academic disciplines. When examining the AWL for the distribution of homographs (e.g. *lead* as a noun versus *lead* as a verb) across disciplines, Hyland and Tse (2007) found that different disciplines "showed clear preferences for particular meanings and collocations" (p.244). A word list which represents a 'general academic' vocabulary ignores significant differences in meaning and usage of words as they occur in various disciplines. This can result in students devoting resources to studying word meanings that may never be encountered in their own discipline, or are used in completely different ways in that discipline.

For the discipline of computer science, Guo (2018), in an examination of English language learners studying computer programming, found that many terms are used with different meanings in the context of CS, including AWL words such as *instance* as well as high-frequency GSL words such as *for* and *while*. Lam (2001), in an examination of "sub-technical" vocabulary (words with both CS-specific meanings and general high-frequency meanings), also found that when reading CS texts, learners faced additional challenges which were linked to the transfer of general-usage meanings into a CS context, and when learners were provided with a technical dictionary of CS-specific meanings instead of a general-usage dictionary, they achieved significantly higher scores in a series of CS-based reading comprehension tasks which specifically focused on these polysemous terms.

The overlap between words with CS-specific meanings (such as "bug", "port", "tree", "string", "volume", and "mouse") and the contents of the GSL and AWL means that a CS vocabulary list which is built upon (and therefore excludes the contents of) the GSL and AWL word lists will be constrained by them and unable to indicate the special meanings found in the language of academic computer science. As studies of CS learners suggest (Bonar & Soloway, 1985; Guo, 2018; Lam, 2001), learners who have not examined the discipline-specific meanings of GSL and AWL items may be at a disadvantage and could face additional complications as a result of the transfer of general-usage meanings to discipline-specific contexts. In order to include words with CS-specific meanings, the Computer Science Academic Vocabulary List (CSAVL) was not built upon the GSL or AWL word lists, but instead uses other criteria to restrict the inclusion of general high-frequency words. Words with CS-specific meanings are explicitly marked in the CSAVL for two reasons: 1) so that list users are made aware of differences between the use of these items in a CS context and a general-usage context; and 2) so that list users can easily locate the appropriate CS-specific definitions for these terms.

In recent years, researchers have produced academic word lists that focus on describing the vocabulary of specific academic disciplines rather than a general academic vocabulary. Discipline-specific word lists have been created for business (Hsu, 2011), medicine (Wang et al., 2008; Lei & Liu, 2016), Science (Coxhead & Hirsch, 2007), agriculture (Martínez et al., 2009), applied linguistics (Khani & Tazik, 2013), engineering (Hsu, 2014), nursing (Yang, 2015), pharmacology (Fraser, 2007), and computer science (Minshall, 2013). The subject of this article, The Computer Science Academic Vocabulary List (CSAVL), continues in this trend of word lists that are increasingly focused on specific academic disciplines, but incorporates an improved methodology which makes it a more efficient tool for reaching a minimal level of academic CS reading comprehension than previously developed word lists.

1.2. The value of lemma-based word lists

The word list with the strongest influence on the design of the present study is Gardner and Davies' (2014) Academic Vocabulary List (AVL). The design of the AVL differed from the AWL (Coxhead, 2000) in that 1) the AVL was not designed as a supplement to the GSL and 2) the AVL was based on the organizational unit of the lemma rather than the word family (Bauer & Nation, 1993). Regarding design decision number one, Gardner and Davies found that many GSL high-frequency words appeared more frequently in academic registers than general ones (words such as *exchange*), which led them to argue for the inclusion of these high-frequency words in their own academic word list. For the creation of the Computer Science Academic Vocabulary List, I also adopted this position in order to prevent the full variation of the register represented by the list from being restricted by the GSL. If high-frequency words such as *memory* take on unique technical meanings within a CS context, they should also be considered for inclusion on a CS academic vocabulary list.

In reference to the second design decision, the present study also adopts the use of the lemma rather than the word family due to a number of problematic aspects of the word family counting unit that have been described by numerous researchers (Gardner, 2007; Gardner & Davies, 2014; Hyland & Tse, 2007; Wang & Nation, 2004). A word family is centered on a *headword*, and includes all inflected forms of that headword along with all derivations of the headword up to level six of Bauer and Nation's (1993) ranking system. The word family unit ignores parts of speech, so the family headed by *lead* will also contain both the verb *lead* and the noun *lead,* despite their distinctly different meanings. Wang and Nation (2004), in an analysis of Coxhead's word-family based AWL, found that one in ten of the word families in the AWL contained homographs such as *lead.* This finding, combined with that of Hyland and Tse (2007) regarding the preferences for particular meanings in specific disciplines, suggests that learners studying word family-based lists will be spending time studying meanings of a headword that could be infrequent in their chosen academic discipline. While *lead* as a noun may be highly relevant to students of chemistry, it may be less so for students of business. While word families are fairly inclusive groupings of inflections and derivations, the lemma (which only groups inflectional forms of a word within a single part of speech) is a

more specific unit, and lemma-based lists can provide a more precise description of the frequent items in a corpus (Brezina & Gablasova, 2015). The move to a lemma-based list mitigates some of the issues related to homographs since different parts of speech appear as different list entries (*lead* as a verb will appear separately from *lead* as a noun).

Gardner (2007) pointed out an additional problem with 6-level word families: inflectional knowledge comes sooner than derivational knowledge for English language learners. It may be problematic to assume that learners entering EAP programs will have enough derivational knowledge to recognize the relationship of a headword to derivational forms up to level 6 (e.g. the relationship between *react* and *reactor*). Furthermore, researchers have pointed out how learners from L1 backgrounds with different morphological systems may face additional challenges when analyzing these derivational forms (Corson, 1997; Ward, 2009). These derivational transparency issues for English language learners can be addressed by lemma-based lists, which group headwords only with inflected forms and list separate entries for items such as *react* and *reactor*.

Additional support for the development of lemma-based word lists has been provided by Brezina and Gablasova's new-GSL (2015), a lemma-based replacement for the GSL that provides similar coverage of general English corpora to West's GSL while containing substantially fewer lemmas. A lemma-based approach was also taken by Lei and Liu (2016) in the development of their discipline-specific Medical Academic Vocabulary List (MAVL), which provided more coverage of a medical corpus than the word family-based Medical Academic Word List (Wang et al., 2008), and did so with far fewer lemmas. A word list that achieves the same amount of coverage as another list, but with fewer words, could be described as a more *efficient* list because it will likely result in a shorter time investment for learners. The findings of the researchers who created the new-GSL and MAVL both provide evidence that, in addition to being more informative and user-friendly than word-family lists, lemma-based lists are also more efficient.

1.3. The computer science academic word list (Minshall, 2013)

A discipline-specific word list similar in purpose to the present study, Minshall's (2013) Computer Science Word List (CSWL), is based upon the word family unit and was created as a technical supplement to both the General Service List (West, 1993) and the Academic Word List (Coxhead, 2000). The CSWL was derived from a 3.6-million-word corpus which contained 408 texts of two text types: academic journal articles and conference proceedings. To filter general high-frequency and general academic words from the CSWL, Minshall excluded all GSL and AWL word families from his list. As a result of the exclusion of GSL word families, Minshall prevented high-frequency polysemous GSL words with CS-specific meanings (such as "memory" and "class") from being included in the CSWL. AWL academic vocabulary items such as "core" and "function" also take on discipline-specific meanings within a CS context and restricting items like these from the CSWL limits it from representing the variations of the register. Additionally, since Minshall based the CSWL on the word family unit, learners are assumed to be able to recognize the relationship of each headword to its derivational forms (e.g. the relationship between "serve" and a "server" in a computer network), an assumption that has been challenged by numerous researchers (Brezina & Gablasova, 2015; Corson, 1997; Gardner, 2007; Gardner & Davies, 2014; Ward, 2009).

Minshall's CSWL does not specify part of speech for list items and gives no indication of whether or not an item has a discipline-specific meaning. This lack of discipline-specific information reduces list efficiency and may create additional complications (as was observed in Lam, 2001) for learners who may spend time studying parts of speech or meanings of a headword that are infrequently used in academic computer science ("bug" as an insect, "port" as a city with a harbor) and mistakenly assume that those meanings can be applied to a CS context. Additionally, by not including frequency rank information, the CSWL does not fully allow list users to prioritize the study of vocabulary based on frequency of use, one of the primary purposes of a word list.

1.4. Research purposes of the CSAVL project

Although there is one existing academic computer science word list (Minshall's, 2013 CSWL), there has not been a lemma-based CS word list which attempts to represent the broader academic vocabulary of the discipline, identifies items with CS-specific meanings, identifies parts of speech for list items, or incorporates the technologies of part-of-speech tagging and statistical analysis. In order to develop a word list that can better meet the needs of ESL learners bound for study in the context of academic computer science, I propose the following list of ideal characteristics for a discipline-specific computer science academic vocabulary list (CSAVL):

(1) The list should be lemma-based in order to be as representative as possible of the register and maximize efficiency for its users by guiding them toward the specific parts of speech for each item that are most frequently found in academic CS texts.
(2) In order to more fully describe the variation of this register, the list should not completely exclude items appearing on a general service list, but instead use other criteria to exclude general high-frequency words in order to acknowledge that specialist vocabulary can be found within all frequency bands of vocabulary.
(3) The list should explicitly identify items with CS-specific meanings so that learners can focus on the meanings that are most relevant to their chosen discipline and the ways that they differ from their general-usage meanings.

After creating the Computer Science Academic Vocabulary List, the secondary purposes of this research project are as follows:

(1) To compare the coverage of the CSAVL with the Academic Word List (Coxhead, 2000), Academic Vocabulary List (Gardner & Davies, 2014), and the Computer Science Word List (Minshall, 2013).
(2) To determine the amount of coverage of an academic CS corpus that can be achieved by combining the CSAVL with a lemma-based general high-frequency word list such as the new-GSL (Brezina & Gablasova, 2015).

(3) If the combination of the CSAVL and new-GSL do not reach the 95% lexical coverage threshold suggested by Laufer (1989, 2010) for a minimal level of reading comprehension, to create a supplement to the CSAVL in order to reach this target.

2. Methodology

2.1. Corpus design overview

In order to develop the Computer Science Academic Vocabulary List (CSAVL), two corpora of academic CS texts were developed: The Computer Science Academic Corpus 1 and 2 (CSAC1 & CSAC2). Since a word list derived from a given corpus is to some degree a frequency measure of the items in that corpus, it will likely have an unfairly high lexical coverage percentage of that source corpus (Coxhead, 2000; Nation, 2016). Therefore, it is also important to develop a second corpus to make a more valid coverage test that might demonstrate its representativeness of texts across the discipline. In this study, the larger CSAC1 corpus was used to generate the CSAVL word list, and the smaller CSAC2 test corpus was used to evaluate the word list. A thorough description of the development of these corpora can be found in Roesler (2020) and is summarized here in the following steps:

1. Selection of text types

Both corpora are collections of academic computer science texts that were selected to represent the kinds of texts an undergraduate CS student studying at an American university might encounter. The vocabulary of the "core" CS classes, required courses that are taken in the initial years of study in a CS program, is represented by the CS textbook portion of the corpora. The vocabulary of the upper-division CS coursework, which is more topically varied and can include independent research projects, is represented by the content of the academic journal article portion of the corpora. In order to create topical sub-categories for journal articles in the corpus, 10 computer science sub-disciplinary categories were devised (listed in Table 1) based on the categories utilized by the Association of Computing Machinery's (ACM) Computing Classification System.

Table 1
Journal article categories.

Journal article categories (based on the ACM Computing classification system)	
1. Computer systems organization	6. Mathematics of computing
2. Computing methodologies	7. Networks
3. Hardware	8. Security and privacy
4. Human-centered computing	9. Software and its engineering
5. Information systems	10. Theory of computation

To develop sub-disciplinary categories for textbooks, I performed a systematic survey of CS textbooks currently used in core CS courses at 21 American universities. Based on these data, 10 categories of core CS textbooks were created (listed in Table 2) that were intended to mirror the categories used for journal articles as closely as possible.

Table 2
Textbook categories.

Core CS Textbook categories	
1. Computer systems organization	6. Mathematics of computing
2. Data structures*	7. Programming*
3. Algorithms*	8. Operating systems*
4. Human-centered computing	9. Software and its engineering
5. Probability and statistics*	10. Theory of computation

Categories that differ from journal article categories are marked with *.

2. Establishment of corpora sizes

For the CSAC1 source corpus, a size of 3.5 million tokens was chosen to match the size of the corpora used to derive both the Academic Word List (AWL) (Coxhead, 2000) and the Computer Science Word List (CSWL) (Minshall, 2013). The 3.5-million-token CSAC1 source corpus and the 700,000-token CSAC2 test corpus were designed to be as parallel in organization as possible. Each of the corpora were partitioned into 20 equally-sized sub-corpora (two text types divided into 10 topical categories).

3. Establishing text selection criteria

Three selection criteria were set for choosing the texts for inclusion in the CSAC1 and CSAC2 corpora. These criteria were

established for the purpose of selecting texts which best represented the population of texts that might be encountered by undergraduate CS students studying in American universities in 2019. For journal articles, the three criteria for selection were *topical relevance* (the topic of the text must match the category of the sub-corpus), *recency* (a preference for recent publications), and *impact and usage* (a preference for higher unique download count). For textbooks, the selection criteria were *topical relevance*, *institutional influence* (in use by a "top 20" US computer science program), and *current usage* (a preference for a higher usage count among the surveyed universities).

4. Balancing token counts within the sub-corpora

After selecting texts based on the above criteria, a mean token count for all of the collected texts in each sub-corpus was determined. Texts which greatly deviated from that mean count were removed when making the final selection. As shown in Table 3, a total of 12 textbooks and 142 journal articles were included in the CSAC1 source corpus, and chapters from 10 C S textbooks and 42 ACM journal articles were included in the CSAC2 test corpus.

Table 3
Overall makeup of corpora.

	CSAC1		CSAC2	
	No. of texts	Tokens	No. of texts	Tokens
Textbooks	12	1,773,522	10	356,679
ACM Journal Articles	142	1,758,934	42	357,557
Total:	154	3,532,486	52	714,236

2.2. Word selection criteria

Six word selection criteria were established for the creation of the CSAVL, based on those used by Lei and Liu (2016), Gardner and Davies (2014), and Coxhead (2000). These six criteria were applied to the CSAC1 source corpus in order to choose the final set of lemmas that were included in the CSAVL.

1. **Minimum Frequency:** This criterion requires that all items on the CSAVL occur with an overall minimum frequency of 100 times in the CSAC1 corpus. This was the minimum frequency set by Coxhead (2000) for the AWL, which was derived from a corpus of a similar size (3.5 million tokens). For the 3,532,486-token CSAC1, this equates to a normed frequency of 0.28 occurrences per ten thousand words.
2. **Discipline connection:** Candidate lemmas must demonstrate a special connection to the discipline by having either: 1) a frequency in the CSAC1 that is 150% of its frequency in a corpus of general English; or 2) a CS-specific meaning.
   In order to include those high-frequency lemmas with both general and CS-specific meanings that, as observed by Lam (2001), provide challenges for learners, this criterion is structured in two parts. The first part is equivalent to the *frequency ratio* criterion used in Gardner and Davies (2014) to eliminate general high-frequency words from their academic word list. However, when using frequency ratio alone, we eliminate terms that are used at similarly high frequencies in both general and academic CS corpora, but with different meanings. In a general English corpus, a term such as "gate_n" may often indicate "a hinged barrier covering an opening in a wall", whereas in an academic CS corpus it can often indicate "a device that implements a logic function". For a lemma such as "gate_n", the measure of frequency ratio does not account for meaning, and fails to recognize the discipline connection that is demonstrated by the metaphorical extension of the term to CS concepts. In consideration of this, I also allow lemmas with CS-specific meanings (identified using a technical dictionary) to remain as candidates.
   In order to prevent the CSAVL from being constrained by the scope of any specific technical dictionary, the *discipline connection* criterion is operationalized as an 'or' condition rather than an 'and'. Lemmas that are more relatively frequent in academic CS texts may still be useful to CS learners, regardless of whether they fall within the entries of any specific CS technical dictionary, and therefore, I do not eliminate them from the candidate lemmas at this stage.
   To implement frequency ratio, I use the non-academic portion of the BNC as a corpus of general English. The normed BNC frequency (per ten thousand words) of each lemma was compared to its normed frequency in the CSAC1. To identify candidate lemmas with CS-specific meanings, I make use of the Oxford Dictionary of Computer Science (ODOCS) (Butterfield et al., 2016).
3. **Range ratio:** Lemmas must occur with at least 20% of their expected frequencies in at least half of the 20 sub-corpora. Expected frequency is calculated by dividing overall frequency of a lemma by the number of sub-corpora (20 in the case of the CSAC1). In creating the AVL, Gardner & Davies required that lemmas occurred with 20% of expected frequency in 7 of their 9 disciplines (a range ratio of 78%). For the AWL, Coxhead required that words appear in at least half of her sub-corpora (50%). Lei & Liu adopted Gardner & Davies' minimum expected frequency requirement ($\geq$20%) and combined it with Coxhead's range requirement ($\geq$50%) since their corpus size was closer to the size of the AWL corpus. Lei & Liu's decision was also adopted for the creation of the CSAVL in order to limit the inclusion of lexical items that are specific to a smaller range of the CS sub-disciplines.
4. **Dispersion:** Dispersion was used to ensure that lemmas were evenly distributed throughout the corpus. Following Gardner and Davies (2014) and Lei and Liu (2016), Juilland's D was selected as the measure of dispersion. All lemmas must have a minimum

dispersion value of 0.3. The ideal target threshold is arbitrary and is a topic of debate (Gries, 2019). Gardner and Davies set this value at 0.8 for the creation of the Academic Vocabulary List (AVL), whereas Lei and Liu chose the value of 0.5 when creating the Medical Academic Vocabulary List (MAVL) from a much smaller corpus than the 120-million-word COCA academic corpus (Davies, 2008) used by Gardner and Davies.

After experimenting with a variety of values and examining their effects on list size, a final dispersion threshold of 0.3 was chosen for the CSAVL, matching the value selected in corpus-based research by Oakes and Farrow (2007).

5. **Discipline measure:** A lemma should occur no more than three times its expected frequency in any more than three of the sub-corpora. This also follows Lei & Liu's modifications of Gardner & Davies' criterion. This measure excludes lemmas which are clustered in a portion of the corpus, but are infrequent in the remaining portion. This criterion prevents technical items that are primarily used within a limited number of sub-disciplines from being included in the list.
6. **Additional meaning criterion for general high-frequency words:** Any remaining candidate lemmas that appear on a general high-frequency word list (Brezina and Gablasova's new-GSL) must also appear in a technical dictionary (the Oxford Dictionary of Computer Science). This criterion eliminated items such as "we_pron", which: 1) satisfied the frequency ratio portion of Criterion Two (due to the numerous multi-author journal articles in the CSAC1); 2) appeared in the new-GSL; and 3) did not have a CS-specific meaning. Because the new-GSL uses the lemma unit of counting, it was chosen as the general high-frequency list for this criterion.

2.3. Extracting candidate lemmas

Lancsbox (Brezina et al., 2015, 2018) was used to tokenize, POS tag, and lemmatize the corpora. The Lancsbox tools allowed for calculations of frequency, relative frequency, and dispersion to be made, all of which were necessary for applying the word list inclusion criteria. A set of programs was developed in C++ in order to apply the word selection criteria. Once all six criteria had been applied and the remaining candidate items for the CSAVL had been determined, manual edits were made to the remaining items. Multi-letter variables, numerals, initialisms and acronyms, proper nouns, and proper adjectives were removed from the word list. After manual adjustments and edits were made, a total of 904 lemmas remained on the final version of the CSAVL.

In addition to the CSAVL academic list, a supplement word list, the CSAVL-S, was created in an attempt to reach the 95% lexical coverage threshold suggested by Laufer (1989, 2010) for a minimal level of reading comprehension. While the CSAVL was intended to include the most frequently used items of CS academic vocabulary, the CSAVL-S was designed to include technical words that occurred more frequently within CS subdisciplines than they did relative to the discipline as a whole.

The lemmas of the CSAVL-S supplemental list were extracted in the following manner. First, since the CSAVL-S was a supplemental list, items already on the CSAVL were excluded. Next, the criteria which limited the inclusion of sub-disciplinary technical items were either modified or removed: criteria three (range ratio), four (dispersion), and five (discipline measure) were removed, and criterion one (minimum frequency) was lowered from 100 to 60. After applying criteria two (discipline connection) and six (special meaning), a list of 702 CSAVL-S lemmas was produced.

2.4. Removing undesired content from the test corpus

Before coverage tests could be performed on the CSAC2 test corpus, the CSAC2 was examined for items which would be ignored in the tests. Numbers and all non-alphabetic characters were removed from the CSAC2 using regular expressions.

As Nation (2016) argued, proper nouns are not necessarily unique to a single language, carry little meaning beyond a specific referent, and are usually omitted in word counts related to vocabulary lists. Vocabulary list developers (Coxhead, 2000; Konstantakis, 2007; Minshall, 2013) have also omitted proper nouns, acronyms, and abbreviations from their corpus token counts due to these terms often being recognizable across languages ("PC") and placing a low learning burden on a learner (Konstantakis, 2007). Following this reasoning, proper nouns, proper adjectives, acronyms, and initialisms were removed from the test corpus. These undesired items comprised a total of 3.99% of the CSAC2. The figure of 3.99% was comparable to the percentages of similar content found in the corpora used to generate the CSWL (Minshall, 2013) and the MAVL (Lei & Liu, 2016).

2.5. Comparing word lists based on different units of counting

In order to evaluate the CSAVL and CSAVL-S word lists, coverage tests were performed on the CSAC2 test corpus. By means of these coverage tests, comparisons were made with a number of academic and general high-frequency word lists. While the CSAVL and CSAVL-S are lemma-based lists, the majority of the comparison lists were based on the word family unit. To allow for the comparison of the lemma-based CSAVL and CSAVL-S with word family-based lists, lemma counts for the word family lists (the GSL, AWL, and CSWL) were produced and are listed in Table 1.

A list in word family form can be converted to lemma form without distorting the intended contents of the list since a word family is assumed to include all parts of speech for each headword. The word family headword "bank" is assumed to include "bank_v", "bank_n", "banking_n", "banked_adj", and "banker_n". However, when converting in the opposite direction (from lemma to word family), a different relationship is found. If one converts from the lemma "bank_v" to the word family headword "bank", then new information has been added ("banker", "bankers") which no longer reflects the contents of the original lemma-based word list. Converting from lemma to word family results in unfair coverage comparisons between lists of different units of counting (as noted by Gardner & Davies, 2014). For this reason, all comparisons in this study are made at the lemma level.

As Table 4 shows, the CSAVL is less than half the size of the AWL when comparing by lemmas. Because the lists being compared in this study are of substantially different sizes, normed coverage (coverage per 100 lemmas) was used a metric to determine each list's *efficiency* in covering a corpus of academic CS texts.

Table 4
Word list sizes.

word list	types	lemmas	families
GSL (West, 1953)	7822	5338	**2000**
AWL (Coxhead, 2000)	3082	1926	**570**
CSWL (Minshall, 2013)	1919	1225	**432**
new-GSL(Brezina & Gablasova, 2015)	5115	**2495**	
CSAVL	1853	**904**	
CSAVL-S	1398	**702**	

Bolded values indicate the original unit of list organization.

3. Results and discussion

3.1. Academic word list coverage comparisons

Nation (2016) recommends that when evaluating a word list, the list should be compared with other lists having a similar purpose. The Academic Word List (AWL) (Coxhead, 2000) and the Academic Vocabulary list (AVL) (Gardner & Davies, 2014), can be viewed as the primary standalone word list options for learners who hope to develop reading comprehension of academic computer science texts, and thus were chosen for the comparison made in this section.

By making this comparison, it was found that the CSAVL was able to provide substantially more efficient coverage of academic CS texts than the AVL (Gardner & Davies, 2014) and the AWL (Coxhead, 2000). Table 5 shows the results of the coverage tests performed on the CSAC2 test corpus using the lemma-based AVL, the word family-based AWL, and the CSAVL. The largest of the three lists, Gardner and Davies' AVL, covered 18.64% of the test corpus. In comparison, the CSAVL, which is less than one third of the size of the AVL, provided a coverage of 16.06%. Though the AVL exceeded the total coverage of the CSAVL by 2.58%, the large size of the AVL diminishes its efficiency for word list users. When the normed coverage per 100 lemmas values are compared to account for the difference in list size, the CSAVL covered 1.78% of the corpus per 100 lemmas, nearly tripling the efficiency the AVL's 0.62% coverage per 100 lemmas. In addition to the AVL's limitation in terms of efficiency, the AVL is also limited by its status as a standalone academic vocabulary list. The CSAVL can be supplemented by the CSAVL-S to provide additional coverage of an academic CS corpus while remaining notably more efficient than the AVL (as shown in Table 3).

When comparing the CSAVL with Coxhead's AWL, the CSAVL provided higher total coverage of the test corpus than the AWL (16.06% versus the AWL's 12.2%) while containing less than half the number of lemmas. When examining the normed coverage values, it can be seen that the CSAVL's normed coverage of 1.78% per 100 lemmas nearly tripled the 0.63% coverage per 100 lemmas provided by Coxhead's AWL. Coxhead's restriction of all GSL content (some of the most highly-frequent words in the English language) from the AWL was a likely factor in the AWL's limited ability to provide coverage of the CS corpus. The restriction of GSL items from the AWL resulted in the omission of items with CS-specific meanings such as "memory", "tree", and "server", which were some of the most frequent items in both the Computer Science Academic Corpus 1 and 2 (CSAC1 and CSAC2) and could be considered highly relevant to English learners intending to develop their reading comprehension of academic CS texts.

Table 5
CSAC2 coverage of the AVL, AWL, and CSAVL.

word list	lemmas	CSAC2 coverage	coverage per 100 lemmas
AVL (Gardner & Davies, 2014)	3015	18.64%	0.62%
AWL (Coxhead, 2000)	1926	12.20%	0.63%
CSAVL	904	16.06%	1.78%

3.2. Supplemental CS word list coverage comparisons

Because Minshall's (2013) CSWL is a supplement to the AWL (Coxhead, 2000), and does not attempt to independently describe the vocabulary of academic CS, it was not directly compared to the base CSAVL word list. However, a comparison was made of the combinations of the supplemental lists with their base lists; the combination of the AWL and CSWL was compared with the combination of the CSAVL and CSAVL-S. In making this comparison, it was found that the CSAVL/CSAVL-S combination provided higher total coverage and double the normed coverage of an academic CS corpus. Table 6 shows that the combination of the CSAVL/CSAVL-S covered 19.9% of the test corpus, slightly higher than the 17.26% covered by the AWL/CSWL combination. The CSAVL/CSAVL-S combination (nearly half the size of the AWL/CSWL combination) provided substantially higher normed coverage (1.24%) of the test corpus than the AWL/CSWL combination (0.55%), demonstrating the efficiency of the CSAVL/CSAVL-S in covering academic CS texts.

Table 6
CSAC2 coverage of the AWL, CSWL, CSAVL, and CSAVL-S.

word list	lemmas	CSAC2 coverage	coverage per 100 lemmas
AWL (Coxhead, 2000) & CSWL (Minshall, 2013)	3151	17.26%	0.55%
CSAVL & CSAVL-S	1606	19.90%	1.24%

3.3. General high-frequency word list coverage comparisons

The CSAVL and CSAVL-S were also combined with a general high-frequency word list, the new-GSL (Brezina & Gablasova, 2015), to determine whether the total coverage of this combination could reach the 95% coverage threshold suggested by Laufer (1989, 2010) for a minimal level of reading comprehension. A coverage test using the combination of the GSL (West, 1953), AWL (Coxhead, 2000), and CSWL (Minshall, 2013) was also performed to compare its efficiency in covering a corpus of academic CS texts to that of the new-GSL/CSAVL/CSAVL-S combination.

The results of these coverage tests demonstrated that the CSAVL and CSAVL-S word lists, when combined with a general service list, are able to approach the 95% coverage threshold in a corpus of academic CS texts, indicating that these word lists may be viable tools for developing reading comprehension of academic CS texts. It was also found that the new-GSL/CSAVL/CSAVL-S combination was able to provide similar coverage to the GSL/AWL/CSWL combination while containing less than half the total number of lemmas.

Table 7 lists the results of coverage tests performed with the three-list combinations: the GSL/AWL/CSWL and the new-GSL/CSAVL/CSAVL-S. The three-list combination of the GSL/AWL/CSWL covered 95.49% of the CSAC2, surpassing the 95% lexical threshold for minimal comprehension. This was slightly higher than the coverage results reported by Minshall (2013), who found this combination to cover 94.41% of his test CS corpus. The three-list combination of the new-GSL/CSAVL/CSAVL-S was found to cover 94.77% of the CSAC2, achieving coverage within 0.23% of the lexical threshold. The new-GSL/CSAVL/CSAVL-S combination, totaling 3918 lemmas, was able to achieve similar coverage to the 8489-lemma GSL/AWL/CSWL combination with less than half the number of lemmas, demonstrating the ability of the new-GSL/CSAVL/CSAVL-S combination to efficiently approach the 95% threshold.

Table 7
CSAC2 coverage of general frequency list combinations.

word list	lemmas	CSAC2 coverage	coverage per 100 lemmas
GSL, AWL & CSWL	8489	95.49%	1.12%
new-GSL, CSAVL & CSAVL-S	3918	94.77%	2.42%

3.4. Coverage across general, academic, and CS corpora

Following the evaluation methods of Gardner and Davies (2014) and Lei and Liu (2016), I also tested the CSAVL and CSAVL-S word lists against a variety of corpora types: a corpus of general English (the non-academic portion of the BNC), a general academic corpus (the academic portion of the BNC), and a second academic computer science corpus (the CSAC1). The results of these tests indicated that: 1) The CSAVL and CSAVL-S are more representative of academic vocabulary than the vocabulary of general English; and 2) The CSAVL and CSAVL-S are representative of academic CS English texts beyond those found in the corpus from which they were extracted.

The second column of Table 8 shows that the CSAVL covered only 2.96% of the BNC non-academic portion lemmas, somewhat similar to the 3.69% reported by Lei and Liu for the MAVL. For the academic portion of the BNC, the CSAVL covered 4.93%, which suggests that the items in the CSAVL appear more often in an academic context than a general one. The CSAVL-S supplemental list covered 0.62% of the BNC non-academic, which was comparable to the 0.39% reported by Minshall when testing his CSWL word list on a fiction corpus. When comparing the CSAVL-S to the academic portion, its coverage increased to 0.9%, suggesting that the contents of the CSAVL-S are more representative of academic than general vocabulary.

Table 8
Coverage of the CSAVL and CSAVL-S in the BNC academic and non-academic portions.

word list	BNC non-academic	BNC academic
CSAVL	2.96%	4.93%
CSAVL-S	0.62%	0.90%

To evaluate CSAVL and CSAVL-S performance in a second academic CS corpus, a coverage test was performed on the CSAC1 source corpus. Before performing this test, the CSAC1 corpus was processed in the same manner as the CSAC2 (as described in Section 2.4). It was found (as shown in Table 9) that the CSAVL covered a similar portion of both the CSAC1 (16.87%) and CSAC2 (16.06%). The CSAVL-S supplemental list also covered comparable portions of the CSAC1 (4.17%) and CSAC2 (3.84%). When the CSAVL and CSAVL-S were combined with the new-GSL, the total CSAC1 coverage of the combination was only 0.08% higher than its coverage of the CSAC2. The similarity of coverage values found in the two academic CS corpora provides evidence that the CSAVL/CSAVL-S lists are representative of academic CS texts beyond those found in their source corpus and are able to provide similar coverage of CS texts produced by a variety of authors.

Table 9
CSAVL coverage of academic computer science corpora.

word list	CSAC1	CSAC2
CSAVL	16.87%	16.06%
CSAVL-S	4.17%	3.84%
new-GSL, CSAVL & CSAVL-S	94.79%	94.71%

3.5. Usage of the CSAVL and CSAVL-S

Based on the comparisons and analyses performed in this study, the combination of the CSAVL and CSAVL-S can be considered a more efficient tool than the AVL (Gardner & Davies, 2014) or the pairing of the AWL (Coxhead, 2000) and the CSWL (Minshall, 2013) for developing academic computer science reading comprehension. The CSAVL and CSAVL-S are specifically representative of the vocabulary of academic computer science, contain a large portion of vocabulary that have CS-specific meanings (Author, 2020), and provide an indication of which parts of speech and derivational forms are most relevant to a CS context. By comparison, the AVL and the combined AWL/CSWL word lists are notably less efficient in their coverage of an academic CS corpus, provide no indication of CS-specific meanings, and provide no indication of which parts of speech and derivational forms are relevant to academic CS texts. Given these differences with the AVL, AWL, and CSWL, the CSAVL and CSAVL-S word lists are likely to better serve the needs of English language learners seeking the most direct route to English reading comprehension of academic CS texts by providing learners with an improved cost/benefit relationship and by describing the vocabulary of academic CS in a more detailed and context-specific manner. These lists can also serve as a valuable tool for English for Specific Purposes (ESP) course and material designers who need to identify the most frequently used terms that are specifically representative of academic CS in order to maximize the benefits of language instruction within a constrained period of time.

I provide the following recommendations for the use of the CSAVL/CSAVL-S:

1. The usage of the CSAVL/CSAVL-S in relation to technical dictionaries

Users of the CSAVL and CSAVL-S should also make use of technical dictionaries such as the Oxford Dictionary of Computer Science (ODOCS) (Butterfield et al., 2016) ODOCS and *the Oxford Concise Dictionary of Mathematics* (ODOM) (Clapham & Nelson, 2009) in conjunction with the word lists. To guide users of the lists toward the appropriate technical dictionary definitions, the items in the CSAVL/CSAVL-S are listed with a "*" to denote items that have ODOCS entries, and with a "#" to denote items that appear in the ODOM. As noted by Lam (2001), items with both technical and general meanings such as "pipe_n", "bug_n", or "mouse_n" may create more difficulties for learners than technical domain-specific terms such as "botnet_n". For this reason, list users should make special note of the terms that have been marked as having discipline-specific meanings, and consult a technical dictionary in order to find the meaning that is most appropriate for a CS context.

2. Suggested uses for educators and pedagogical material designers

The CSAVL/CSAVL-S can serve as a tool used by educators in the design of an English for Specific Purposes (ESP) course for students intending to study computer science, or by pedagogical material designers to create learning materials for students who have the intention of improving their reading comprehension of English academic CS texts. Educators using the CSAVL/CSAVL-S should prioritize the lemmas in order of their ranked frequency. The final versions of the lists include a rank number to the left of each item. Items with the lowest rank numbers were the most frequent items in the academic CS corpus that the lists were extracted from, and should be the first priorities for learners to acquire. In selecting terms for course or material design, users of the CSAVL and CSAVL-S lists should first focus on the 904 lemmas of the CSAVL, which were found to be the most frequent and evenly distributed terms in the corpus of academic CS texts. The 702 lemmas of the CSAVL-S, which were less frequent and were clustered in specific CS sub-disciplines, such as hardware or mathematics, should be turned to only after the main CSAVL list has been studied.

3.6. The limitations of the CSAVL project

The first limitation of this study was the size of the 3.5-million-word CSAC1 corpus. While the CSAC1 was of similar size to the corpus used for the creation of the AWL, the CSAC1 could have been expanded in order to include a greater sampling of academic CS texts. This would have allowed for the inclusion of a wider range of textbooks than the 1–2 textbooks that were sampled for each 10 textbook sub-corpora of the CSAC1. A larger and more varied sampling of authors within each sub-corpus would have helped to mitigate any possible issues of author bias within the CSAC1.

A second limitation was the choice of the non-academic portion of the BNC as the corpus of general English used to derive and evaluate the list. Given that the CSAVL was designed for use by American university students, a corpus such as COCA would have been a more ideal choice due to differences in American and British usage of English. In this study, the BNC was chosen for reasons of data accessibility, and this choice no doubt had some effect on the final contents of the wordlist.

A third limitation of this study is that although frequency counts were made of POS-tagged lemmas in the CSAC1, there was no

identification made of the intended meanings of the terms in each specific context that they were found. While "memory_n" was one of the most frequent terms in the CSAC1, it was unknown how many of the uses of "memory" were referring its CS-specific meaning and how many of the uses instead referred to its general usage meaning. Future word list research may be able to overcome this limitation through the use of semantic tagging systems that might be able to better identify intended word meanings in each specific context of use (Rayson et al., 2004).

3.7. Possibilities for future word list research

A complication that pointed toward future research came from the hyphenated forms in the CSAC1 and CSAC2 corpora. The removal of numbers from the CSAC1 source corpus resulted in the presence of hyphenated forms such as "real-time", but also a large number of split forms such as "-dimensional" and "-bit". In order to eliminate this inconsistency, all hyphens were removed from the corpora before tokenization. However, the decision to remove hyphens also meant that multi-word hyphenated forms such as "real-time" were not included in the final versions of the CSAVL/CSAVL-S. A comprehensive multi-word CS vocabulary list that also includes hyphenated forms and a list of common collocations may be a useful project for future research and could provide learners with an additional level of description of the language of academic CS.

4. Conclusion

Though a variety of authors have produced word lists representing a general academic vocabulary (Campion & Elley, 1971; Coxhead, 2000; Gardner & Davies, 2014; Xue & Nation, 1984), it has been only recently that researchers such as Lei and Liu (2016) have begun to develop discipline-specific academic vocabulary lists that take these three principles into account: 1) that academic and other specialist vocabulary can be found at all levels of word frequency; 2) that for many vocabulary items, certain derivational forms and parts of speech are more relevant than others to specific disciplines; and 3) that many polysemous vocabulary items have both general-usage meanings and meanings that are associated with specific disciplines.

These principals guided the development of the CSAVL project, which produced a discipline-specific computer science academic word list that built upon the methods of Gardner and Davies (2014) and Lei and Liu (2016). By allowing for the inclusion of high-frequency words with discipline-specific meanings (such as "mouse_n" and "tree_n") into the word list, the CSAVL more comprehensively represents the vocabulary related to the concepts and practices of computer science than previous lists which restricted general high-frequency items, such as the Computer Science Word List (CSWL) (Minshall, 2013). By explicitly indicating which terms have CS-specific meanings, and which parts of speech and derivational forms are most relevant to CS, the CSAVL reduces the burden of interpreting word list contents on list users and describes the vocabulary of CS more efficiently and in greater detail than was possible through previous methods. It is my hope that learners will be able to benefit from the efficiency and detail of the CSAVL, and that word list developers will be able to take advantage of and improve on the methods of this study when producing their own specialist word lists.

Declaration of competing interest

None.

Appendix A. Supplementary data

The CSAVL and CSAVL-S vocabulary lists may be found online at https://doi.org/10.1016/j.jeap.2021.101044.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(3), 253–279.
Bonar, J., & Soloway, E. (1985). Preprogramming knowledge: A major source of misconceptions in novice programmers. *Human-Computer Interaction, 1*(2), 133. https://doi.org/10.1207/s15327051hci0102_3
Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1–22. https://doi.org/10.1093/applin/amt018
Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139–173.
Brezina, V., Timperley, M., & McEnery, A. (2018). #LancsBox v. 4.x. [software package].
Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction, 3*(2), 1–10. https://doi.org/10.7820/vli.v03.2
Butterfield, A., Ngondi, G. E., & Kerr, A. (Eds.). (2016). *A dictionary of computer science*. Oxford University Press.
Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. New Zealand Council for Educational Research.
Clapham, C., & Nicholson, J. (2009). *The concise Oxford dictionary of mathematics*. Oxford University Press.
Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*(1), 181–200.
Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., & Ferrara, J. (1979). Reading English for specialized purposes: Discourse analysis and the use of student informants. *Tesol Quarterly, 13*(4), 551–564.
Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718. https://doi.org/10.1111/0023-8333.00025
Coxhead, A. (2000). A new academic word list. *Tesol Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée, 12*(2), 65. https://doi.org/10.3917/rfla.122.0065
Davies, M. (2008). *The corpus of contemporary American English*. Available at: http://corpus.byu.edu/coca/.
Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education, 10*, 127–143.
Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241–265. https://doi.org/10.1093/applin/amm010
Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. https://doi.org/10.1093/applin/amt015
Goldenberg, C. (2008). Teaching English Language Learners: What the research does—and does not—say. *American Educator, Summer*, 8–44.
Gries, S. T. (2019). Analyzing dispersion. In S. T. Gries, & M. Paquot (Eds.), *Practical handbook of corpus linguistics*. Retrieved from http://www.stgries.info/research/ToApp_STG_Dispersion_PHCL.pdf.
Guo, P. J. (2018). Non-native English speakers learning computer programming: Barriers, desires, and design opportunities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI, '18*, 1–14. https://doi.org/10.1145/3173574.3173970
Hsu, W. (2011). A business word list for prospective EFL business postgraduates. *Asian ESP Journal, 7*(4), 63–99.
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
Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.
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
Jacobs, V. (2008). Adolescent literacy: Putting the crisis in context. *Harvard Educational Review, 78*(1), 7–39. https://doi.org/10.17763/haer.78.1.c577751kq7803857
Khani, R., & Tazik, K. (2013). Towards the development of an academic word list for Applied Linguistics research articles. *RELC Journal, 44*(2), 209–232. https://doi.org/10.1177/0033688213488432
Konstantakis, N. (2007). Creating a business word list for teaching business English. *ELIA, 7*, 79–102.
Lam, J. (2001). A study of semi-technical vocabulary in computer science texts with special reference to ESP teaching and lexicography. In G. James (Ed.), *Research reports* (Vol. 3). Language Center, Hong Kong University of Science and Technology.
Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From humans to thinking machines* (pp. 316–323).
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.
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
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
Minshall, D. E. (2013). *A computer science word list (Master's thesis)*. Retrieved from https://www.baleap.org/wp-content/uploads/2016/03/Daniel-Minshall.pdf.
Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91–108. https://doi.org/10.1002/RRQ.011
Nation, I. S. P. (2016). *Making and using word lists for language learning and testing* (3rd ed.). Amsterdam, Netherlands: John Benjamins Publishing Company.
Oakes, M. P., & Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing, 22*(1), 85–99. https://doi.org/10.1093/llc/fql044
Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop* (pp. 7–12). Lisbon, Portugal, 2004.
Roesler, D. (2020). *A computer science academic vocabulary list [Master's thesis*. Portland State University]. https://doi.org/10.15760/etd.7414.
Wang, J., Liang, S., & Ge, G. (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
West, M. (1953). *A general service list of English words*. London, New York: Longman, Green.
Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215–229.
Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27–38. https://doi.org/10.1016/j.esp.2014.05.003