English for Specific Purposes 58 (2020) 77–89

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

How large a vocabulary do Chinese computer science undergraduates need to read English-medium specialist textbooks?

Jia Bi

Software College, Northeastern University, No. 3-11, Wenhua Road, Heping District, Shenyang, 110819, PR China

ARTICLE INFO

Article history:

Keywords:
Computer science
Word list
Lexical coverage
Academic vocabulary
Technical vocabulary

ABSTRACT

This study specifically addresses the needs for English-medium textbook reading comprehension of Chinese computer science undergraduates who have already mastered about 3,300 general English word families prescribed by the Ministry of Education before entering university. Thirty textbooks on various subjects of computer science were chosen to build a Computer Science Textbook Corpus (CSTC) containing 7.51 million running words. Based on criteria of range, frequency, and dispersion, 356 word families outside the 3,300 items within students' knowledge were extracted to form the Computer Science Vocabulary List (CSVL). The CSVL accounted for 4.79% of the tokens in the CSTC but only 0.39% in a fiction corpus. The CSVL, combined with students' lexical repertoire acquired from secondary education, provided 95.16% coverage of the corpus, reaching the minimum requirement for reading comprehension suggested by Laufer (1989). By analyzing the overlapping proportion of related word lists pairwise, this study further established that the development of specialized word lists achieved the best efficiency if targeting at a homogenous audience.

1. Introduction

With the globalization of China's higher education, an increasing number of major Chinese universities offer bilingual courses or courses in English for undergraduates of different majors to study specialist subjects. Students taking such courses are required to read English-medium textbooks to acquire academic knowledge and skills. This practice is particularly beneficial for students majoring in disciplines whose key technology initially derived from English speaking countries. Computer science is such a discipline whose core technology is based on English, and most of the classic and advanced technologies were first expressed in English. Therefore, many Chinese computer science faculties have introduced original textbooks written in English as teaching materials for students to read.

It has been generally accepted among scholars that readers' vocabulary size is a key factor in successful reading comprehension (Laufer, 1992; Nation, 2006; Schmitt, Jiang, & Grabe, 2011). For most Chinese first-year undergraduates, after at least six years of English study for general purposes, a significant obstacle to reading English-medium textbooks is their

E-mail address: bij@swc.neu.edu.cn.

https://doi.org/10.1016/j.esp.2020.01.001
0889-4906/

lack of knowledge about discipline-specific words and discourse since what they have learned is primarily general English (Cai, 2017).

To better equip computer science undergraduates with necessary discipline-specific vocabulary on top of what they have learned before entering university, this study aims to create a word list which will represent the core vocabulary used in the English-medium textbooks of computer science students need to read. It is anticipated that this word list combined with students' existing lexical repertoire, will help them better cope with specialist reading tasks.

2. Literature review

2.1. Lexical coverage

Lexical coverage is a fundamental concept in making and evaluating word lists. It refers to the percentage of tokens in the text a reader knows (Nation, 2006). Studies have found the lexical coverage threshold for successful reading comprehension. Laufer (1989) found that a reader's knowledge of 95% of the text's vocabulary was required to ensure reasonable reading comprehension of the text. Hu and Nation's (2000) study concluded that 98% was the necessary lexical coverage for adequate comprehension. Schmitt et al. (2011), in a more recent study, also suggested 98% as the optimal coverage for adequate comprehension of academic texts.

In a nutshell, the research to date would suggest two coverage points as a lexical threshold for minimum and optimal reading comprehension: 95% and 98%. As Minshall (2013) pointed out, setting 98% as the threshold is a problem of diminishing returns: "As coverage demands increase, vocabulary sizes become exponentially higher to meet them" (Minshall, 2013, p. 6). Hsu (2014) set 95% as the predetermined coverage percentage for her study to create an Engineering English Word List to bridge EFL students'lexical gap because 95% coverage seemed a more feasible goal for first-year EFL students. Ward (1999) and Valipouri and Nassaji (2013) both intended their investigated lexical resources to achieve 95% coverage threshold so that their students could have sufficient lexical knowledge to read textbooks or research articles in their areas of specialty. The reason why so many specialized academic word list studies have adopted 95% as the threshold lexical coverage value is that there is always a cost-benefit analysis involved in the decision on what vocabulary items to teach to EFL students. Nation (2006) discovered that 4,000 word families plus proper nouns provided 95% coverage of novels and newspapers, and 8,000–9,000 word families plus proper nouns provided 98% coverage; to achieve only 3% higher coverage, EFL students need to command twice as much vocabulary as they do with 95% coverage. Hence, while 98% coverage may be optimal, the goal of 95% is more achievable in this context, and so this study adopts 95% as the threshold value for Chinese computer science undergraduates to achieve reasonable reading comprehension of textbooks. In other words, the Computer Science Vocabulary List (CSVL) developed by this study combined with students' lexical repertoire acquired from secondary education would cover at least 95% of the tokens in the kinds of textbooks students are required to read.

2.2. Unit of counting

One of the critical issues in making word lists is how to define a word, which will make a big difference in the number of items. In previous studies, a lot of word lists were developed using word families as the unit of counting, such as West's (1953) General Service List (GSL), Coxhead's (2000) Academic Word List (AWL), Hsu's (2013) Medical Word List, Yang's (2015) Nursing Academic Word List, and Dang, Coxhead, and Webb's (2017) Academic Spoken Word List. A word family is the combination of a base word and its inflectional and derivational forms made up of affixes from Bauer and Nation's (1993) set of affixations for word families. For example, the members of the word *access* include both its inflectional forms *accessed, accesses, accessing,* and derivational forms *accessibility, accessible, inaccessibility, inaccessible*. This word family has altogether eight members which are counted as one item in the word list. The thinking behind choosing word families as the unit of counting is that once learners have mastered the stem word, with knowledge of basic word-building processes, they will infer the meanings of regularly inflected and derived forms of that word without much effort (Bauer & Nation, 1993). Other scholars used lemmas as the unit of counting (Brezina & Gablasova, 2015; Dang & Webb, 2016; Gardner & Davies, 2014). A lemma only includes the headword and its inflectional variants.

The counting approach in a word list study is usually dependent on the purpose of the word list. Generally speaking, for receptive uses (listening and reading), one word family is counted as one vocabulary item; for productive uses (speaking and writing), the lemma is a more sensible unit of counting (Brezina & Gablasova, 2015; Nation, 2016). The current study aims to provide a word list for computer science majors to achieve better reading comprehension of specialist books. The target audience of the word list are first-year undergraduates of major Chinese universities whose English language ability corresponds to level 4 described in *China's Standards of English Language Ability* (Ministry of Education, 2018a). Vocabulary competence described in level 4 includes knowledge about root words, prefixes, and suffixes. The students admitted to major universities are among the best of all Chinese university first-year students in the National College Entrance Examination, and their English morphological skills can help them infer the meanings of various forms of a base word. Therefore, the preferred unit of counting is word families rather than lemmas for the creation of the CSVL.

2.3. Previous studies on academic and technical word list development

2.3.1. Academic words and technical words

Nation (2013) classified vocabulary into three groups based on frequency levels: high-frequency words, mid-frequency words, and low-frequency words. Academic words and technical words are in focus for the current study. They may be found anywhere along the spectrum of frequency from high to low, but mainly within the mid-frequency range (Nation, 2013).

Although there is not a generally accepted definition for academic vocabulary, researchers (e.g., Dang, Coxhead, & Webb, 2017; Gardner & Davies, 2014; Lei & Liu, 2016) normally consider it to be a group of words with high frequency, wide range, and even distribution in academic texts but infrequent in other genres. Also known as sub-technical or semi-technical words, academic words play an organizational and rhetorical role in structuring the writer's arguments (Baker, 1988). Academic words account for a significantly higher proportion of running words in academic texts than in general texts.

Technical words are those used in specialized disciplines. There are several ways to decide whether a word is technical or not, such as an expert's opinion or reference to a specialized dictionary. Statistical methods are also used in corpus linguistics such as term frequency-inverse document frequency (tf-idf) which is a measure used to evaluate how important a word is to a document in a corpus (Rajaraman & Ullman, 2011). Widely cited by linguistic scholars (e.g., Coxhead & Demecheleer, 2018; Valipouri & Nassaji, 2013) is a rating scale for finding technical words designed by Chung and Nation (2003) (Figure 1) which can measure the strength of the relationship of a word to a particular specialized field. According to the scale, there are two kinds of technical vocabulary: words that are closely related to a specialized field but may also occur in general use at Step 3 and words that are unique to a specialized field and are not likely to be used in general language at Step 4. Chung and Nation (2003) analyzed the technical vocabulary in an anatomy textbook and an applied linguistics textbook. They found that 16.3% of the technical words at Step 3 in the anatomy texts and 50.5% of the words at Step 3 in the applied linguistics texts were general service words from the GSL or academic words from the AWL.

It seems that academic and technical words are only meaningful to a discipline-specific audience. Any existing academic word list will be of limited use because the development of that list has not considered the present target learners' specific situation.

The current study aims to extract a word list from a corpus of computer science textbooks excluding the words Chinese first-year undergraduates should have acquired from secondary education. As a pedagogical tool for English instruction for specific purposes, the word list will mainly consist of academic vocabulary and technical vocabulary which are frequently used in computer science.

2.3.2. Academic and technical word list development

The most influential collection of academic words is Coxhead's (2000) AWL which was extracted from a corpus of 3.5 million tokens composed of research articles, university textbooks, and manuals from four disciplines: arts, commerce, law, and science. The 570-word-family AWL provided approximately 10.0% coverage of the tokens in the academic corpus.

The AWL has made considerable contributions to EFL teaching, as well as set a paradigm for academic word list development. However, because of its intensive use, the AWL has come under scrutiny and received some criticism. Gardner and Davies (2014) pointed out the problems in the methodology of the AWL and created a new Academic Vocabulary List of 3,000 lemmas or 2,000 word families from a corpus of 120 million words of written academic materials.

One concern about a general academic word list is whether it has general applicability across different disciplines. Hyland and Tse (2007) questioned the usefulness of a general academic word list for all university majors by suggesting that individual lexical items often behave differently across disciplines in terms of range, frequency, collocation, and meaning. They recommended that "teachers help students develop a more restricted, discipline-based lexical repertoire" (Hyland & Tse, 2007, p. 235).

To address the specific vocabulary needs of EFL learners of different disciplines, researchers have developed many field-specific academic or technical word lists, largely following Coxhead's (2000) corpus analysis method. Hsu (2013) extracted 595 word families to form the Medical Word List from a corpus of 155 textbooks across 31 medical subject areas. Coxhead and Hirsh (2007) conducted a pilot study, identifying 318 word families with a considerably higher coverage over a science-specific corpus than over non-science corpora. There are also Ward's (2009) 299-word Basic Engineering List for less proficient engineering undergraduates, Yang's (2015) Nursing Academic Word List containing 676 word families in the nursing discipline, and Coxhead and Demecheleer's (2018) plumbing word list featuring 1,465 individual types.

Most relevant to the current study is a Computer Science Word List (CSWL) created by Minshall (2013). He obtained a list of 433 word families from a corpus of 3.66 million tokens compiled from journal articles and conference proceedings covering 10 sub-disciplines of computer science. The CSWL, in combination with the GSL and the AWL, accounted for 95.11% of all the tokens in his corpus, reaching the minimum requirement for reading comprehension set by Laufer (1989).

A predominant method used in specialized word list creation is that the researchers filter out the general high-frequency words such as those in West's (1953) GSL which are supposed to be already known by learners aiming for university study. However, some scholars (e.g., Dang et al., 2017; Gardner & Davies, 2014; Lei & Liu, 2016; Neufeld, Hancioğlu, & Eldrige, 2011) have questioned the rationality of creating academic word lists on top of a general service list by pointing out that there are many words in the GSL which have academic or technical meanings.

Step 1 Words such as function words that have a meaning that has no particular relationship with the field of anatomy, that is, words independent of the subject matter. Examples are: *the, is, between, it, by, 12, adjacent, amounts, common, commonly, directly, constantly, early,* and *especially.*
Step 2 Words that have a meaning that is minimally related to the field of anatomy in that they describe the positions, movements, or features of the body. Examples are: *superior, part, forms, pairs, structures, surrounds, supports, associated, lodges, protects.*
Step 3 Words that have a meaning that is closely related to the field of anatomy. They refer to parts, structures or functions of the body, such as the regions of the body and systems of the body. Such words are also used in general language. The words may have some restrictions of usage depending on the subject field. Examples are: *chest, trunk, neck, abdomen, ribs, breast, cage, cavity, shoulder, girdle, skin, muscles, wall, heart, lungs, organs, liver, bony, abdominal, breathing.* Words in this category may be technical terms in a specific field like anatomy and yet may occur with the same meaning in other fields and not be technical terms in those fields.
Step 4 Words that have a meaning specific to the field of anatomy and are not likely to be known in general language. They refer to structures and functions of the body. These words have clear restrictions of usage depending on the subject field. Examples are: *thorax, sternum, costal, vertebrae, pectoral, fascia, trachea, mammary, periosteum, hematopoietic, pectoralis, viscera, intervertebral, demifacets, pedicle.*

Figure 1. Steps in Chung and Nation's (2003, p.105) scale of technical vocabulary.

Among the previously mentioned studies, following Coxhead's (2000) practice, Hsu (2013), Minshall (2013), and Yang (2015) excluded either West's (1953) GSL or the top 3,000 most frequently-occurring words in the British National Corpus (Nation, 2004) when creating their discipline-specific academic word lists. On the contrary, Gardner and Davies (2014), Dang et al. (2017), Lei and Liu (2016), and Valipouri and Nassaji (2013) did not adopt a point of departure in their word list development.

In the studies with a departure point, there is either a specific audience, or an assumed audience, or an assumption that learners should have already mastered a large number of general words. For the present study, there is a specific target audience whose existing vocabulary can be properly measured. If this study had developed a word list without a departure point, it would have resulted in too many words that had already been mastered by the target students. Such a word list would increase students' learning burden and extend their learning time.

2.3.3. Chinese high school graduates' lexical repertoire

In China, the Ministry of Education published *The English Curriculum Standards for Compulsory Education* (2012) and *The English Curriculum Standards for High School* (2018b) to guide English language teaching before higher education. In the appendixes of the two official documents, vocabulary lists are given to specify the vocabulary items to be mastered by students. All the English textbooks "must pass an examination ahead of use by the committee commissioned by the Ministry of Education with a vocabulary check ensuring that vocabulary used in the teaching materials is covered by the curriculum vocabulary list" (Jin, Li, & Li,

2016, p. 957). Under the guidance of the national curriculum standards, students are required to understand the function and meaning of the words in context and be able to enlarge their vocabulary with word formation skills.

Using the online program Familizer/Lemmatizer (Cobb, n.d.), I converted the words in the word lists of curriculum standards into a list of 3,348 word families. This list was termed the Basic Word List (BWL) in this study, representing the lexical repertoire of the first-year students of major universities before they begin their tertiary study.

Although no detailed language descriptions could be found about the vocabulary lists, in an official document, *China's Standards of English Language Ability* (Ministry of Education, 2018a), which can be used as a yardstick for English teaching and learning in China, the requirement of high school students' vocabulary competence is described as understanding of basic vocabulary for daily topics relating to study, work, travel, and current affairs. Therefore, the BWL is assumed to be a general service list, and this study will also test its generality against several corpora.

Most of the existing word lists are set in a broad context, targeting a large or the whole population of non-native international students, without much consideration of the heterogeneous nature of the audience. The word list of this study is intended as a pedagogical tool for Chinese computer science undergraduates, and most of the university students commencing their studies have mastered about 3,300 words listed in the BWL. Therefore, the BWL will serve as the baseline for the development of a new specialized word list in computer science for Chinese learners of English.

2.4. Metaphor of computer science language

Dating back to only about 1960, computer science is a relatively new discipline deeply rooted in mathematics and electrical engineering (Belford, 2017). Therefore, much of its documentation naturally reflects the language of mathematicians and engineers (Chisholm, 1986). Different from the entities in physics or biology which have a foundation in the real world, the entities in computer science exist mainly in virtuality. One distinctive feature of computer science language is the pervasive metaphors in both formal and informal discourse. Although all scientific disciplines use metaphors, no field uses metaphor as idiosyncratically and pervasively as does computer science (Johnson, 1990). Computer scientists sense a need to compensate for the extremely abstract, or even ethereal nature of most computer realities by expressing them as metaphors. According to Johnson,

> Computer systems are sometimes portrayed as little universes in which traffic moves from one place to another, where there are ports, buses, and pipelines. And almost without transition, those very same realities are sometimes seen as structures and objects needing interfaces and platforms and hooks. (Johnson, 1991, p. 273)

Because of the exceedingly abstract nature of computer science, computer scientists name and describe various entities and activities using highly colorful and imaginative language. Many expressions, including both nouns and verbs in computer science, are derived from common English terms to describe the image and behavior of objects and the interactions between humans and machines as well as between components of computers. Metaphor as an important component of the fabric of computer language represents the shared reservoir of values of the discipline (Johnson, 1990).

The abstract nature and pervasive metaphors in computer science discourse may lead to the assumption that there are not many technical words exclusively used in this field, and instead, that computer science shares a large vocabulary with other disciplines, acquiring new abstract meanings specifically for computer science. This study will analyze the characteristics of the words extracted from computer science textbooks to identify the distinctive features of computer science vocabulary.

3. Research questions

This study aims to provide Chinese computer science undergraduates with a word list that can satisfy their needs for the comprehension of specialist textbooks. There are four research questions for this study:

1. What percentage of the running words in the computer science textbook corpus does the BWL cover?
2. Beyond the items in the BWL, how many more words do Chinese computer science undergraduates need to reach the level of 95% lexical coverage of the specialist textbooks?
3. Does corpus analysis confirm that the BWL is a general service list? Is the CSVL a reliable word list for Chinese computer science undergraduates?
4. What are the characteristics of the words in the CSVL? Are they academic or technical vocabulary?

4. Methodology

4.1. Building and preparing the corpus

To find the right materials for the corpus, I consulted two computer science professors. They presented the syllabuses of the courses the college offered to computer science majors. From the textbooks and reference books listed in the syllabuses, 30 books were chosen to build the Computer Science Textbook Corpus (CSTC) on the topics of computer architecture,

computer network, operating system, data structure, algorithm, software engineering, information security, cryptography, computer graphics, artificial intelligence, and several programming languages which are the compulsory or elective courses of computer science students. The list of all the textbooks in the corpus is provided online in the Supplementary Material.

Computer science literature provides a great challenge for core vocabulary extraction because it is full of mathematical, numerical, and programming data. These data do not create reading comprehension barriers to students, but a large number of non-linguistic characters will cause inaccuracy and confusion in the calculation of text coverage. To extract these items, the BNC/COCA 25,000 word family lists (Nation, 2017), developed from the British National Corpus and the Corpus of Contemporary American English (Davies, 2008), were used to compare against the CSTC; any items not in the BNC/COCA 25,000 word family lists were carefully examined, as described in what follows.

According to Nation (2016), formulae, numbers, and forms containing a mixture of letters and numbers, and non-words are normally not counted as words, so he suggests enclosing them in triangular brackets (<>). This is because, "This option preserves the original nature of the corpus and still allows the exclusion of some items." (Nation, 2016, p. 108). To do this, PowerGREP (Goyvaerts, 2016) was used to search and process those items. For example, the regular expression [0–9]+[a-zA-Z]+[0-9a-zA-Z]*|[a-zA-Z]+[0–9]+[0-9a-zA-Z]* matches any combination of letters and numbers, and therefore can be used to search for forms containing a mixture of letters and numbers like *s2*, *9cn*, and *temp0019*. These forms were enclosed in triangular brackets (<>) by replacing them with the regular expression <%MATCH:%>.

Other elements of the text which it was felt could be safely omitted for the purpose of this study including copyright information, figures, tables, and references were removed manually (Lei & Liu, 2016; Yang, 2015). This is because such information did not represent linguistic data and provided no burden on reading comprehension. After data cleansing, a corpus of 7,513,913 running words comprising 30 computer science textbooks was ready for analysis.

Following Coxhead and Demecheleer (2018), I used two useful word lists in the BNC/COCA word family lists to identify proper names and acronyms in the corpus. The two word lists are basewrd31 and basewrd34 which are lists of proper nouns and common acronyms respectively. New-found proper nouns and acronyms from the items outside the BNC/COCA 25,000 word family lists were added to the lists accordingly. For example, *GUI* for *Graphical User Interface* and *ACM* for *Association for Computing Machinery*, which are frequently used acronyms in computer science and provide no burden on reading comprehension were added to basewrd34.

Proper nouns in computer science literature mainly include names of scientists, places, enterprises, systems, applications, and programming languages. Acronyms are mostly multi-word proper nouns of computer components, systems, organizations, etc. reduced to the first letter of each word. Such information is not considered to be a barrier to reading comprehension. In this study, proper nouns and acronyms were put in separate lists when lexical coverage was calculated (Nation, 2006).

4.2. Word selection criteria

The first criterion for the selection of the CSVL words was specialized occurrence, which meant all the selected words had to occur outside the BWL which was within the first-year undergraduates' knowledge.

The second criterion was range. According to previous studies (e.g., Coxhead, 2000; Hsu, 2013; Wang, Liang, & Ge, 2008), candidate words should occur in at least half of the subject areas. There were 30 text files in the CSTC; therefore, for a word to be included in the CSVL, its family members had to appear in at least 15 of the 30 text files.

The third criterion was frequency: the number of occurrences of a word family that appeared in the corpus. With a corpus of 3.5 million tokens, Coxhead set the threshold of frequency at 100 times (28.46 times per million words) for multi-member word families and 80 times (22.77 times per million words) for one-member word families. Wang et al. (2008) decided on a threshold of 30 times (27.45 times per million words), a third of Coxhead's, because their corpus consisted of 1.09 million running words, a third of Coxhead's corpus. Many studies followed Coxhead (2000) to adopt 28 times per million words as the minimum frequency. This cutoff frequency is actually arbitrary because no study has explained the rationale for adopting this frequency threshold value. Hsu (2014) set the frequency threshold after repeated trials until the lexical coverage of her Engineering English Word List contributed to reaching 95%. The studies of Coxhead (2000) and Wang et al. (2008) did not set a lexical coverage for their word lists to contribute to, so they were not concerned about the total coverage they obtained with the GSL (West, 1953) and their specialized word lists. The desired coverage of this study was 95% text coverage. Therefore, following Hsu's (2014) practice, I experimented with several cutoff frequency points (150 times, 130 times, etc.) and finally decided on a frequency threshold of 100 times (13.31 times per million).

The last criterion was dispersion. Dispersion is a measure that shows how evenly a word is spread across the corpus. The adoption of a dispersion measure avoids the inclusion of words highly frequent in only a few parts of the whole corpus. Juilland's D (Juilland & Chang-Rodríguez, 1964) is the most commonly used dispersion measure in the studies of word list development (e.g.Dang et al., 2017; Gardner & Davies, 2014; Lei & Liu, 2016), and "It is a number between 0 and 1, with 0 signifying extremely uneven distribution and 1 perfectly even distribution" (Brezina, 2018, p. 51). Scholars have not reached a consensus on the threshold of Juilland's D value for determining the evenness of a word's distribution across the corpus. Oakes and Farrow (2007) decided on 0.3 as the threshold value for an evenly distributed word. Lei and Liu (2016) set 0.5 as the minimum Juilland's D value for developing the Medical Academic Vocabulary List. Dang et al. (2017) used 0.6 as the cut-off point for a word to qualify as being evenly distributed.

Experimenting with 0.6, 0.5, and 0.4 as cut-off points found that 0.6 and 0.5 values would exclude many frequent and useful words in computer science. It was finally decided that 0.4 was the most appropriate Juilland's D value for the CSVL to contribute to the minimum 95% text coverage. Compared with other studies, 0.4 is a rather low value. However, in the studies of Lei and Liu (2016) and Dang et al. (2017), researchers did not set a lexical coverage threshold value for their word lists to achieve. With an expected lexical coverage threshold, this study had to compromise on the dispersion value. As Oakes and Farrow (2007) argued, "whether we use range, D, or U, our cut-off point for discriminating between well and poorly dispersed words must be arbitrary" (p. 92). Hence, a Juilland's D of 0.4 may indicate a somewhat uneven distribution pattern, but it is approaching the half-way point between extremely uneven and perfectly even distributions.

In summary, a word family outside the BWL which appeared in at least 15 text files with at least 100 occurrences and a Juilland's D value of at least 0.4 in the CSTC would be included in the CSVL.

4.3. The software for analysis

Besides the text editing software PowerGREP, the tool used for vocabulary analysis in this study was AntWordProfiler (Anthony, 2014) since it was considered the best program for using BNC/COCA lists for the analysis of vocabulary in texts (Nation, 2016, 2017). It is a powerful corpus analysis tool that can generate vocabulary frequency and range data from a number of corpora by comparing the corpora with the loaded word lists. In this study, BNC/COCA word family lists were used to identify the items that might contain unwanted information. For the extraction of the CSVL, the BWL was loaded into AntWordProfiler to be compared against the CSTC.

As previously mentioned, an online program Familizer/Lemmatizer (Cobb, n. d.) was used to transform the investigated words into word families in a format that could be loaded into AntWordProfiler. Researchers just need to input all the investigated words into the website window; a list of word families is then generated.

5. Results and discussion

5.1. Text coverage of the CSVL and the BWL in the CSTC

Based on the criteria of specialized occurrence, range, frequency, and dispersion, a list of 361 word families was obtained from the 7.51-million-token CSTC. After being evaluated by computer science professors, five items were removed from the list: *hoc*, *millisecond*, *multi*, *semi*, and *inter*. *Hoc* always collocates with *ad* as in *ad hoc network*, and it does not occur singly in computer science textbooks. *Millisecond* is a transparent compound word without further implications. *Multi*, *semi*, and *inter* seldom convey meanings on their own, and they are mainly used as prefixes. The remaining 356 word families were included in the CSVL. A list of the headwords in the final version of the CSVL in alphabetical order can be found in the Appendix. Table 1 shows the coverage of the BWL and the CSVL in the CSTC with proper nouns and acronyms treated separately.

Of all the 3,348 word families in the BWL, 3,284 items occurred in the CSTC; those that did not were everyday English words rarely used in computer science, such as *anxiety*, *beggar*, *snack*, and *terrify*. The BWL's coverage in the CSTC was 89.20%, significantly higher than the coverage of the general service lists in other academic and technical word studies (e.g., 76.1% in Coxhead, 2000; 70.68% in Hsu, 2013; 76.43% in Minshall, 2013). The high coverage of the BWL in the CSTC suggests that the words in the BWL are quite representative in modern English. Another assumption we can make is that the literature of computer science is not very difficult in nature.

The 356 word families in the CSVL provided a further 4.79% coverage in the CSTC. The CSVL and the BWL, counted together with proper nouns and acronyms, covered 95.16% of the running words in the CSTC, reaching the minimum requirement for reading comprehension. The 356-item CSVL is obviously a manageable learning goal for the first-year undergraduates.

5.2. Validity test of the CSVL

According to Coxhead (2000), a validity test of an academic word list is performed to see how the list covers a different collection of texts on similar topics. Therefore, to check if the CSVL was a reliable list, I built a second computer science textbook corpus (CSTC2) with another 30 textbooks (the list included in online Supplementary Material) on subjects similar to the first 30 textbooks. After the same editing work was done, a corpus of 6,637,299 tokens, a little smaller than the CSTC,

Table 1
Coverage of BWL and CSVL in the CSTC with proper nouns and acronyms included.

Word List	Number of word families appearing in CSTC	Token coverage in CSTC
BWL	3,284	89.20%
CSVL	356	4.79%
Proper nouns (Basewrd31)	5,317	0.94%
Acronyms (Basewrd34)	630	0.23%
Total	9,587	95.16%

was ready for the test. When generating the CSWL, Minshall (2013) used journal articles to compile a corpus for word extraction; however, he used textbooks for a second corpus to test the viability of his word list. This research again used textbooks as test corpus materials because the purpose of the word list was to provide an essential vocabulary for computer science undergraduates to comprehend specialist textbooks. Therefore, only computer science textbooks were used as corpus materials.

AntWordProfiler was run again to compare the CSVL against the CSTC2. As can be seen in Table 2, the CSVL's coverage of the CSTC2 was 4.96%, which was 0.17% higher than its coverage in the CSTC, and all the 356 word families in the CSVL were present in the CSTC2. Given these results, reliability testing was found to be satisfactory. In comparison, Minshall's (2013) CSWL coverage was 1.32% less in the test corpus than in the observed corpus, and only 409 of the 433 word families of the CSWL appeared in the test corpus.

The validity test of an academic word list also includes a comparison against a general-purpose corpus. Following the example of Coxhead (2000), I collected 60 fiction texts from Project Gutenberg (http://www.gutenberg.org), an online repository that offers free eBooks of titles whose copyright has expired which are freely available for download. The fiction corpus contained 7,181,254 running words, which was a comparable size to the CSTC. It was loaded into AntWordProfiler to test the coverage of the CSVL.

From Table 3, we can see the CSVL's coverage in the fiction corpus was 0.39%, considerably lower than its coverage in the computer science textbook corpora (4.79% in CSTC, 4.96% in CSTC2). The huge difference supports the claim that the majority of word families in the CSVL are associated particularly with computer science literature and it is a valid word list for the targeted population of undergraduates of computer science.

As for the BWL, its occurrence was relatively stable across the three corpora. According to Gardner and Davies (2014), general high-frequency words appear with equally high frequency across all major registers, including the academic register. As can be seen from Table 3, neither the coverage nor the number of occurring items of the BWL showed much fluctuation across the three corpora. This indicates that the BWL is a robust general service list with stable occurrence across different genres.

5.3. Comparison of word lists pairwise

To further establish the applicability of the BWL and the CSVL to Chinese computer science undergraduates, the study made comparisons between the two lists and the GSL, BNC/COCA-3000, AWL, and CSWL to find out the similarities and differences between them.

As well as showing the overlap between the BWL/CSVL and other relevant word lists, Table 4 also tells how the BWL and the CSVL are different from other lists. The BWL is a group of general service word families supposed to be mastered by Chinese high school students. It indeed overlaps with a large percent of the GSL (90.0%) and the BNC/COCA-3000 (86.1%). However, it also shares items with Coxhead's (2000) AWL and Minshall's (2013) CSWL, which were claimed to be academic lists.

There are 413 items in the BWL which overlap with words in the AWL, such as *achieve, analyze, classify, communicate, conclude, contribute, interpret, phenomenon, undertake,* and *vary.* These words may be more frequent in academic texts than in fictional texts, but they may also occur in general texts and be known by Chinese high school graduates. Some words in the AWL, as Coxhead (2000) suggested, are candidates for a general service list. The AWL is a word list beyond the GSL, and the two lists combined consist of less than 2,500 word families since there are only 1,901 items in the GSL (obtained from http://

Table 2
Coverage of BWL and CSVL in the CSTC2 with proper nouns and acronyms included.

Word List	Number of word families appearing in CSTC2	Token coverage in CSTC2
BWL	3,237	88.93%
CSVL	356	4.96%
Proper nouns (Basewrd31)	4,363	0.90%
Acronyms (Basewrd34)	569	0.31%
Total	8,525	95.10%

Table 3
BWL and CSVL's coverage in the fiction corpus as compared with CSTC and CSTC2.

Corpus	Coverage of BWL	Number of BWL word families	Coverage of CSVL	Number of CSVL word families
Fiction corpus	89.28%	3,266	0.39%	313
CSTC	89.20%	3,284	4.79%	356
CSTC2	88.93%	3,237	4.96%	356

Table 4
Number and percentage of shared items between the BWL/CSVL and other word lists.

Word List	Number of shared items with BWL (3,348 items)	Percentage of overlap with BWL	Number of shared items with CSVL (356 items)	Percentage of overlap with CSVL
GSL (1,901 items) (West, 1953)	1,711	90.0%	10	0.5%
BNC/COCA-3000 (3,000 items) (Nation, 2017)	2,583	86.1%	116	3.9%
AWL (570 items) (Coxhead, 2000)	413	72.5%	98	17.2%
CSWL (433 items) (Minshall, 2013)	105	24.2%	144	33.3%

www.victoria.ac.nz/lals/staff/paul-nation). The above analysis may well explain the huge overlap between the BWL and the AWL.

Between the 356-word-family CSVL and Coxhead's (2000) 570-word-family cross-disciplinary AWL, there is an overlap of 98 words, such as *adjacent, coordinate, enforce, proportion,* and *offset.* Words like these are used in various disciplines, playing an organizational role in showing the relationship between different agents or components.

Although both are computer science word lists, the CSVL and the CSWL only share 144 items (40.45% of the CSVL). Among these shared words, words like *algorithm, array, asynchronous, batch, binary, cache, configure, default, encrypt, execute, intersect, optimize, query, retrieve,* and *stack* are typical of computer science literature. They may appear in other contexts, but they are a lot more frequent in computer science texts. However, the CSVL's 40.45% overlap with the CSWL is rather low, considering they were both extracted from computer science literature. One reason behind this is that the CSWL was developed by filtering out the GSL and the AWL while the CSVL was developed on top of the BWL. The CSVL shared 108 items with the CSWL's departure points: the GSL and the AWL. The CSWL shared 105 items with the CSVL's departure point: the BWL.

The second plausible explanation is that the two word lists were created from different sources of computer science literature for different purposes. The CSWL was intended as a supplement to both the GSL and the AWL in the instruction of non-native English speakers who were studying computer science in UK universities. Journal articles and conference proceedings in computer science were chosen as corpus materials. The present study aimed to provide a pedagogical tool for Chinese computer science undergraduates who have already mastered the vocabulary in the BWL and adopted textbooks as corpus materials. Therefore, with different departure points and different types of literature, studies aiming at specialized vocabulary in the same field may yield quite different results. This again may facilitate the argument that word lists are more effective and efficient if they are targeted at a homogenous learner group.

5.4. Characteristics of the vocabulary in the CSVL

After examining the items in the CSVL and the texts of computer science books, I did not see many words exclusive to this field like *sternum*, *pectoral*, and *viscera* in the field of anatomy. Measured by the four-step rating scale (Chung & Nation, 2003), there are mainly two types of vocabulary in the CSVL. A large number of words from the CSVL are used to describe the features, movements, and interactions of various entities which may belong to the subset of academic vocabulary that plays an organizational and rhetorical role in structuring the writer's arguments. For example, *implement*, *optimize*, *hierarchy*, *adjacent*, *iterate*, *underlie*, *simultaneous*, *incorporate*, *fraction*, *conjunction*, *consistency*, and *transition*. These words are also used in other disciplines and even in general English, but acquire special meanings when used in computer science. The second type of vocabulary are words with narrow meanings at Step 3, used mainly in mathematics and computing science, such as *algorithm*, *exponent*, *theorem*, *polynomial, integer*, *encrypt*, *coefficient*, and *binary*. About 50 of 356 items (14.04%) in the CSVL fall into this category. With narrow senses, they are within easy mastery for computer science students. An overwhelming proportion (85.96%) of words fall into the category of academic vocabulary. Their senses in computer science come from the transferred meanings of those general academic terms which project abstract relationships in computer science. Words of this type provide a real challenge since they may have different meanings in other contexts (Hyland & Tse, 2007). Students will learn these words better as they delve into the specialist subject; and in turn, a good command of these words will help them better cope with the study of specialist subjects.

5.5. Answers to the research questions

From the above discussion, we can give clear answers to the four research questions.

1. The BWL, which is a list of 3,348 word families Chinese university first-year students have already mastered before entering university, covers about 89% of the running words in computer science textbook corpora.
2. Beyond the word families in the BWL, first-year computer science undergraduates need to command 356 word families which are frequently used in computer science textbooks to achieve 95% lexical coverage for minimum comprehension of the kinds of specialist textbooks they need to read.

3. Although 413 items in the BWL are also included in the AWL, some of them are general service words while some are general academic words which also occur in general-purpose language. Its high coverage in the fiction corpus (89.28%) indicates it is basically a general service list. The test against a second computer science textbook corpus and comparisons between the CSVL and other relevant word lists show the CSVL is a reliable and tailored word list for Chinese computer science undergraduates.
4. The words in the CSVL are mostly academic words and some are technical words. Due to the special characteristics of computer science language, only a small proportion of words in the CSVL are used exclusively in this field. For most of the word families, while they do have special meanings in computer science, they also occur in other contexts.

6. Pedagogical implications

The CSVL can help EFL teachers and syllabus designers to focus their attention on the most frequent words in the field when developing teaching materials for Chinese computer science undergraduates. For the Chinese computer science undergraduates, the list can set the priority for their vocabulary learning.

As discussed above, a large number of words on the list also occur in other areas and general language. Therefore, learners should notice how the use and meaning of the words in computer science are different from those in the general-purpose language. As computer scientists project the schema of ordinary life onto their technical understanding of computation processes (Colburn & Shute, 2008), students need to know how the words are used in the physical world to better understand their use in the virtual world.

Another pedagogical implication is that teachers should raise students' awareness of how the words in the CSVL typically collocate with other words in the context of computer science. Entries in *A Dictionary of Computer Science* (Butterfield, Ngondi, & Kerr, 2016) and *Dictionary of Computer and Internet Terms* (Downing, Covington, Covington, & Covington, 2009) mainly consist of common collocations in the discipline. This shows how important it is to understand the meaning of collocations as a whole.

7. Limitations and implications for future research

This study has several limitations. First, it only produced a list of single words in computer science. Various studies (e.g., Martinez & Schmitt, 2012; Simpson-Vlach & Ellis, 2010) have confirmed the importance of multi-word units for EFL learners to improve their English proficiency. Future research should investigate the common collocations in computer science. Second, the study only used information from corpora to develop the CSVL. Such a method may leave out some words which occur frequently in other texts not included in the observed corpus. Although designated as a computer science vocabulary list, it may not be a list that can satisfy all EFL computer science professionals. Compared with other studies, the size of the observed corpus of this study is large enough for the extraction of core vocabulary, but only choosing textbooks as corpus materials may make it a biased list given that it does not draw upon journal articles which have been an important source of corpus texts for many studies. For example, some important words in computer science listed in Minshall's (2013) CSWL are not present in either the CSVL or the BWL, such as *acyclic*, *calibrate*, *instantiate*, and *intrusion*. Since the purpose of this study was to provide a vocabulary list for computer science undergraduates to better comprehend specialist textbooks, journal articles which are not supposed to be the reading materials for most of them were not included in the corpus. This drawback provides a possibility for future research to generate a computer science word list applicable to all levels of students as well as to researchers and practitioners in this field.

8. Conclusion

The huge differences between the BWL/CSVL and other widely used word lists justify compiling more targeted word lists for specific audiences. The GSL, BNC/COCA-3000, AWL, and various discipline-specific word lists have a larger target audience in a broader context. Nevertheless, it is sensible and advisable for local EFL teachers to make more focused word lists to address their students' needs more efficiently. Especially in the creation of a discipline-specific word list as a pedagogical tool for teaching English at tertiary level, understanding the target students' background knowledge of vocabulary and extracting words beyond their existing lexical repertoire will produce a more effective and efficient word list.

Funding

This work was supported by the China Scholarship Council [grant number 201606085039].

Acknowledgments

I would like to thank the guest editors, Averil Coxhead and Oliver Ballance, and all the anonymous reviewers, for their insightful and constructive comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.esp.2020.01.001.

Appendix. The Computer Science Vocabulary List (Headwords in alphabetical order)

adjacent	administrator	aggregate	alert	algorithm
alias	align	alpha	alternate	ambiguity
analog	analogous	analogy	annotate	anonymous
append	applicable	archive	array	assembly
assert	asymmetry	asynchronous	attribute	augment
authorize	automate	automobile	axis	baseline
batch	bias	bin	binary	boolean
bounds	brace	bracket	browse	buffer
bug	bulk	byte	cache	canonical
chunk	classification	clause	closure	cluster
coefficient	coherent	collaborate	commutative	compatible
compile	complement	composite	compound	compress
concatenate	concise	concurrent	configure	conjunction
consecutive	consequent	considerable	consistency	constitute
constrain	contiguous	converge	conversion	convert
coordinate	copyright	correlate	curve	customize
debug	decimal	decompose	decrement	dedicate
default	denote	dense	dependency	depict
deploy	derive	descend	deviate	diagonal
digit	discard	discrete	dominant	dual
dummy	duplicate	effective	elementary	eliminate
embed	empirical	encode	encrypt	enforce
entity	equivalence	equivalent	essence	evolve
exceed	exclusive	execute	exhaust	expertise
exponent	exponential	extensive	extract	factor
feasible	feedback	filter	finite	flaw
flip	font	formula	forth	fraction
fragment	frame	furthermore	generator	genre
graphic	grid	handbook	hardware	hash
header	heterogeneous	heuristics	hexadecimal	hierarchy
hint	homogeneous	horizontal	hybrid	hypothesis
implement	implicate	implicit	incorporate	increment
index	indirect	induct	infinite	informal
infrastructure	inherent	inherit	install	integer
integral	intercept	interface	intermediate	intersect
intuition	intuitive	invalid	inventory	inverse
invert	invoke	iterate	kernel	keyword
layer	legacy	legitimate	lever	lightly
likewise	linear	locality	login	loop
lowercase	magnet	magnitude	malicious	manipulate
manual	manufacture	margin	matrix	maximize
mechanism	merge	migrate	minimal	minimize
modify	module	mutual	naïve	node
nonetheless	notate	notion	null	numeric
numerous	offset	optimal	optimize	oriented
overlap	overload	override	overview	pad
paradigm	parameter	parenthesis	parse	participant
partition	password	patch	peer	penalty
persist	perspective	pipeline	pixel	polynomial
precede	predecessor	preface	prefix	presence
presume	prime	privacy	probe	propagate
property	proportion	proprietor	protocol	prototype
pseudo	publication	query	rand	ratio
rational	recipient	recursive	redundant	refine
relevance	remainder	repository	reset	retain
retrieve	revenue	reverse	robust	rotate
scalar	scenario	scheme	scope	script
segment	semantic	sensible	sensor	sequence
serial	simulate	simultaneous	sketch	slot
sole	sorted	spam	span	sparse
spatial	specify	spectrum	stack	static
stock	strip	subscript	subtle	succession
suffix	summate	symmetry	symposium	synchronize

syntax	tab	tag	taxonomy	teller
template	temporal	terminate	terminology	theorem
thereby	thesis	threshold	token	tolerance
transact	transfer	transition	traverse	trigger
triple	trivial	tuple	underlie	unify
upgrade	uppercase	usage	utilize	utility
validate	vector	vendor	verify	versa
versus	vulnerable	whereas	whereby	wireless
workshop				

References

Anthony, L. (2014). *AntWordProfiler (version 1.4.1) [computer software]*. Retrieved from http://www.laurenceanthony.net/software/antwordprofiler/.

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language, 4*(2), 91-105.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Belford, G. (2017). *Computer science in encyclopedia britannica online*. Retrieved from https://www.britannica.com/science/computer-science.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1-22.

Butterfield, A., Ngondi, G. E., & Kerr, A. (2016). In *A dictionary of computer science* (7th ed.). Oxford: Oxford University Press.

Cai, J. (2017). Debates around the orientation of TEFL in Chinese tertiary education. In H. Reinders, D. Nunan, & B. Zou (Eds.), *Innovation in language learning and teaching: The case of China* (pp. 115-153). London: Palgrave Macmillan.

Chisholm, R. M. (1986). Selecting metaphoric terminology for the computer industry. *Journal of Technical Writing and Communication, 16*(3), 195-220.

Chung, T. M., & Nation, I. S. P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103-116.

Cobb, T. (n.d.) Familizer/lemmatizer (version 2.0) [computer software]. Retrieved from https://www.lextutor.ca/familizer/.

Colburn, T. R., & Shute, G. M. (2008). Metaphor in computer science. *Journal of Applied Logic, 6*(4), 526-533.

Coxhead, A. (2000). A new academic word list. *Tesol Quarterly, 34*(2), 213-238.

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes, 51*, 84-97.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée, 12*(2), 65-78.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning, 67*(4), 959-997.

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153-167). Amsterdam: John Benjamins.

Davies, M. (2008). *The corpus of contemporary American English: 560 million words, 1990-present*. Retrieved from http://corpus.byu.edu/coca/.

Downing, D., Covington, M., Covington, M. M., & Covington, C. A. (2009). *Dictionary of computer and internet terms* (10th ed.). Hauppauge: Barron's Educational Series.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.

Goyvaerts, J. (2016). *PowerGREP (version 4.7.3) [computer software]*. Retrieved from http://www.powergrep.com/.

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.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes, 33*(1), 54-65.

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.

Jin, T., Li, Y., & Li, B. (2016). Vocabulary coverage of reading tests: Gaps between teaching and testing. *Tesol Quarterly, 50*(4), 955-964.

Johnson, G. J. (1990). Computer science: A philosophical primer. In *Proceedings of the 28th annual southeast regional ACM conference* (pp. 70-72).

Johnson, G. J. (1991). Agents, engines, traffic, objects and illusions: Paradigms of computer science. *Journal of Technical Writing and Communication, 21*(3), 271-283.

Juilland, A. G., & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren, & M. Nordman (Eds.), *Special language: From human thinking to thinking machines* (pp. 316-323) Clevedon: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud, & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132) London: Macmillan.

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.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320.

Ministry of Education China. (2012). *English curriculum standards for compulsory education*. Beijing: Beijing Normal University Press.

Ministry of Education China. (2018a). *China's standards of English language ability*. Beijing: Higher Education Press.

Ministry of Education China. (2018b). *English curriculum standards for high school*. Beijing: People's Education Press.

Minshall, D. (2013). *A computer science word list*. Unpublished master's thesis. Swansea: Swansea University.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3-13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists (Version 1.0.0)*. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation.

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System, 39*(4), 533-538.

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*, 85-99.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge: Cambridge University Press.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

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.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442-458.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language, 12*(2), 309-323.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170-182.
West, M. (1953). *A general service list of English words*. London: Longman, Green, & Co.
Yang, M. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27-38.

Jia Bi is a lecturer of English in Software College of Northeastern University in China. Her research interests include corpus-based studies in vocabulary, discourse analysis, second language writing, and English for specific purposes.