INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 8 No. 3, January 2019, pp. 657-667

Available online at:
http://ejournal.upi.edu/index.php/IJAL/article/view/15269

doi: 10.17509/ijal.v8i3.15269

The development of science academic word list

Todsaporn It-ngam* and Supakorn Phoocharoensil

Language Institute of Thammasat University, 2 Prachan Road, Bangkok 10200, Thailand

ABSTRACT

Knowledge of specialized academic vocabulary is important for the academic success of EFL natural science students. Specialized words outside the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000) are necessary for comprehending scientific text. The existing lists of words do not cover all sub-disciplines of natural science. The present study aims to explore the specialized academic words across 11 sub-disciplines of natural science. To identify the words, a corpus-based approach and an expert-judged approach were used. A 5.5-million-word corpus called the Science Academic Journal (SAJ) Corpus was created for this study. Applying the established word selection criteria, 513 word families were selected. The potential list was reviewed by a panel of experts in order to remove the overly-technical words from the list. The Science Academic Word List (SAWL) was established with 432 word families and provided 5.82% coverage of the running words in the SAJ corpus. To validate the word list, the SAWL was tested against two independent corpora. The findings revealed that the SAWL contains 432 word families that are useful for reading journal articles in natural science disciplines. In addition, it was also found that the SAWL performed better on an independent corpus compared to the Science World List (Coxhead & Hirsh, 2007). It is expected that the SAWL established in this study will be a useful source for learning and teaching vocabulary in natural science disciplines.

Keywords: Corpus linguistics; English for Specific Purposes; natural science; specialized word list

First Received:	Revised:	Accepted:
19 September 2018	19 December 2018	26 December 2018
Final Proof Received:		**Published:**
30 January 2019		31 January 2019

How to cite (in APA style):
It-ngam, T. & Phoocharoensil, S. (2019). The development of science academic list. *Indonesian Journal of Applied Linguistics, 8*, 657-667. doi: 10.17509/ijal.v8i3.15269

INTRODUCTION

Academic vocabulary knowledge is crucial for academic success. Educators and language experts are calling for explicit instruction on academic vocabulary, including lists of academic vocabulary (Brezina & Gablasova, 2015). The development of academic vocabulary lists can be traced back to the most influential and widely used word list – West's *General Service List* (GSL) from 1953. The 2,000-word families of the GSL provide approximately 80% to 90% coverage of most written texts (Gilner, 2011; Matsuoka, 2012). In response to the GSL, pioneering scholars attempted to explore academic texts to see words which are not in the GSL but frequently occur across academic disciplines. During the 1970s, according to Gardner and Davies (2014), several word lists of general academic vocabulary were developed based on small corpora of academic materials thanks to the technology at that time. A more robust academic vocabulary list called the *University Word List* (UWL) was published by Xue and Nation (1984). The developers built the UWL on the four different word lists. As a result, the UWL contains over 800-word families and has 8.5% coverage in a corpus of academic texts. However, the UWL lacked consistent selection principles because it was made from different word lists. This inconsistency has made Coxhead's (2000) *Academic Word List* (AWL) the new standard word list since 2000, replacing the UWL.

Coxhead's AWL consists of 570 words based on a 3.5-million-word corpus of academic English texts

* Corresponding Author
Email: itngam@gmail.com

across four disciplines: Arts, Science, Law, and Commerce. Each group consists of seven sub-disciplines. The 570 words were chosen based on the criteria that they occurred in all four disciplines, in 22 of the 28 sub-disciplines, and at least 100 times in total. The words were then compared with a 3.5-million-word corpus of novels to distinguish the words that were truly academic and were not in the GSL. As a result, Coxhead (2000) claims that the new word list provides 10% coverage of the running words in an academic corpus, which is superior to that of UWL. Up to the present time, Coxhead's AWL has served as an important source for vocabulary learning in English language education.

Even though Coxhead's AWL is influential and widely used, the list has been criticized for several issues. Gardner and Davies (2014), for example, point out that there are two main problematic issues: the use of word families for initial word counts and the relationship of AWL to GSL. The use of word families has been criticized because members of some word families might not share the same core meaning. In addition, the AWL was built on the GSL, which is an old list containing more general, high-frequency words. Yet, it is found that 79% of the AWL word families are still among the high-frequency words. That is to say, the good coverage of the AWL in academic texts is the direct result of high-frequency words in the list instead of its academic representativeness. As a result, Gardner and Davies introduced a new *Academic Vocabulary List* (AVL) in 2014. One of the key characteristics of the AVL is that it represents contemporary English. The text coverage of the AVL is reported to have twice as much as the AWL, but Nation (2013) found that 40% of the top 500 words of the AVL are also in the GSL. This means the AVL includes high-frequency words which students most likely know (e.g. 'study,' 'use,' 'group,' 'level,' 'however'). Webb and Nation (2017) suggest that, as the AVL contains about 3,000 academic words, it is too big to be used in a language course. The AVL might be a good resource for researchers rather than for learners or teachers.

A specialized word list, also known as *technical word list, field-specific academic vocabulary list, discipline-specific academic word list,* and *discipline-based lexical repertoires* refers to the list of academic words that are closely related to particular disciplines (Liu & Han, 2015). Experts have drawn considerable attention to this type of word list because several studies have shown that not all words in the interdisciplinary academic word lists (e.g., Coxhead's AWL) are equally important to learners with highly specific needs. The usefulness of the AWL varies significantly across disciplines in terms of range, frequency, collocation, and meaning (Lei & Liu, 2016). Coxhead and Hirsh (2007) indicate that there is a certain amount of specialized academic vocabulary consisting of words outside the GSL and AWL. Yang (2015) also suggests that each specific discipline has its own conventions. It is, therefore, necessary to develop academic vocabulary lists for specific disciplines, which have beneficial effects on language instruction and academic vocabulary research (Liu & Han, 2015; Valipouri & Nassaji, 2013). Nation (2016) suggests that making a specialized word list will help those working with academic texts to understand the size of the vocabulary of a technical area. It will also suggest paths towards dealing with such vocabulary from a curriculum perspective. Specialized word lists can guide the development of appropriate vocabulary learning strategies and help in developing subject matter materials for English for Academic Purposes courses. Finally, making the word list will help teachers to examine the role of technical vocabulary in specialized texts and its possible effects on comprehension and in developing tests of previous topic knowledge.

In scientific disciplines, corpus linguistics has been employed to develop specialized word lists for pedagogical purposes. For example, the *Science Word List* (SWL) (Coxhead & Hirsh, 2007) has been developed based on a Pilot Science Corpus of Written Academic English, which includes 14 sub-disciplines (agricultural science, biology, chemistry, computer science, ecology, engineering and technology, geography, geology, horticultural science, mathematics, nursing and midwifery, physics, sport and health science, and veterinary and animal science). These disciplines were included in the word list because they are the disciplines of science degrees offered at Massey University and the University of Sydney – the two universities where the study was carried out. The SWL consists of 318 word families and covers 3.79% of the running words of the Pilot Science Corpus.

However, the fact that the SWL was drawn from 14 sub-disciplines of science makes this word list too broad. The sub-disciplines can be divided into three branches: natural science, technological science, and health science. According to Biber (2006), the specialized vocabulary in natural science (i.e., biology, chemistry, mathematics, and physics) is different by nature from other scientific branches. This implies that many words in the SWL might not be equally valuable and may become a burden of vocabulary learning for science students who are not majoring in the disciplines related to engineering or medical science. In contrast, other existing word lists in science are too specific to a certain discipline, e.g., *Chemistry Academic Word List* (Valipouri & Nassaji, 2013), *Microbiology Academic Word List* (Boonyos, 2014), and *Environmental Academic Word List* (Liu & Han, 2015). Hence, it is necessary to develop a new specialized academic word list for natural science disciplines, the *Science Academic Word List* (SAWL).

To make an academic word list for science disciplines, special characteristics of scientific English need to be taken into account. The language of science is different from that of several other academic disciplines. Reeves (2005) describes that scientific language is a simple, descriptive system. The language in the scientific reports must be "as free as possible from connotations that reflect or create cultural biases and emotional attachment" (p.10) because the goal of scientists is to

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

report facts carefully. Scientists need to be very careful when dealing with words that may have other meanings because scientists from different disciplines may define the same terminology in different ways. For example, the word *homology* in the fields of evolutionary biology and biochemistry has different technical meanings. In evolutionary biology, the word *homology* means similarity between organisms based on genetics, while similarity based on similar adaptation to a common function is called an *analogy*. According to Halliday and Webster (2004), scientific English has many technical features developed over time by experts. These features could cause difficulty for non-English speaking science students. This implies that distinguishing general words and specialized words cannot be done solely through a corpus approach, despite the objective nature. The polysemous words that have specialized meanings can be differentiated among others by using an expert's judgment.

Chung and Nation (2004) suggest four approaches to identify technical words: using expert's judgment, using clues, using a technical dictionary, and using corpora. The expert-judged approach, in which a panel of experts is given a four-point Likert scale to measure the strength of the relationship of a word to the discipline, is the most thorough way of identifying specialized words. This laborious approach is commonly used to overcome the limitation of the corpus-based approach. In scientific disciplines, the expert's judgement approach was applied in some projects to create word-lists in some disciplines, such as in chemistry (Valipouri & Nassaji, 2013), plumbing (Coxhead & Demecheleer, 2018), and finance (Tongpoon-Patanasorn, 2018). The present study also used the expert-judged approach to distinguish specialized words and complement the corpus approach.

The purpose of this study is to make a new academic word list for science disciplines. This research focuses on the academic words that are not found in the GSL (West, 1953) or the AWL (Coxhead, 2000). Drawing on journal articles of science disciplines, the new word list will help teachers design an appropriate syllabus and allow science students to use it as a guideline for self-study. With the appropriate instructions based on the developed word list, it is expected that the students will be able to read academic texts more effectively. Related to the creation of science academic word list in this study, the following research questions were formulated: (1) What are the academic words frequently found in journal articles of science disciplines?; and, (2) How does the present science academic word list differ from the SWL (Coxhead & Hirsh, 2007)?

METHOD

The compilation of the corpus

The corpus created for the present study is the Corpus of Scientific Academic Journal (hereafter SAJ corpus). The SAJ is a corpus of 5.5 million running words from 1,062 journal articles in science disciplines. Located in the eastern region of Thailand, the university where the current study was carried out has a Faculty of Science with 11 subject areas: applied physics, aquatic science, biochemistry, biology, biotechnology, chemistry, food chemistry, mathematics, microbiology, physics, and statistics. These natural science disciplines are commonly taught in many universities across the country. The present study has included research articles and reviews articles as science students are required to read both text types. To make the SAJ corpus for these 11 subject areas, 1,062 journal articles were chosen equally according to the number of journals and running words.

The process of selecting journal articles for building the corpus involved three main steps. First, 11 professors from the different disciplines of natural science were requested to recommend five major journal titles in their disciplines, the articles of which are written in English by international authors and frequently assigned to their students. Table 1 shows the selected journal titles in each discipline. The corpus comprises 11 sub-corpora. Second, each sub-corpus was expected to contain approximately 500,000 running words from the five recommended journals in each discipline (as shown in Table 1), each of which was expected to contain approximately 100,000 running words. Finally, the researchers selected both research articles and review articles published from October to December 2017 and downloaded them from online databases. The number of the articles was not fixed because the length of articles varied among different disciplines. However, the articles were downloaded and included in the corpus until each sub-corpus comprised approximately 500,000 running words. The irrelevant sections in the articles such as acknowledgements, references, appendices, and biographies were excluded from the analysis. The SAJ corpus eventually contains 5.5 million running words and was divided equally into 11 sub-corpora, as presented in Table 2.

Research tools

To analyze the corpus, two concordance programs were used: AntWordProfiler (Anthony, 2014) and AntConc (Anthony, 2016). These programs are comprehensive and freely available for corpus linguistic research. They are recommended by Nation (2016) and are widely used for making many word lists (e.g., Chanasattru & Tangkiengsirisin, 2016; Pugsee, Limgomolvilas, Wudthayagorn, & Janpugdee, 2017).

In this study, AntWordProfiler was used to generate word lists from the SAJ corpus and to compare the lists against reference word lists: West's (1953) GSL, Coxhead's (2000) AWL, and Coxhead and Hirsh's (2007) SWL. AntWordProfiler was also used to evaluate the SAWL by analyzing its text coverage rate on other corpora. AntConc was used to examine the words in the SAWL. The concordance function was used to investigate the SAWL words in the SAJ corpus. The results from this program were given to the experts in the following steps to support their judgement.

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

Table 1. Selected journal titles for the Scientific Academic Journal (SAJ) Corpus

Disciplines	**Journals**
1. Applied Physics	1.1 Applied Surface Science Journal
	1.2 Journal of Alloys and Compounds
	1.3 Surface and Coatings Technology Journal
	1.4 Thin Solid Films Journal
	1.5 Wear Journal
2. Aquatic Science	2.1 Aquaculture Journal
	2.2 Coral Reefs Journal
	2.3 Hydrobiologia Journal
	2.4 Marine Biology Journal
	2.5 Zoological Studies Journal
3. Biochemistry	3.1 Biochemical Journal
	3.2 Biochemistry Journal
	2.3 Journal of Biochemistry
	3.4 Journal of Biological Chemistry
	3.5 PLOS One Journal
4. Biology	4.1 Cell Stem Cell Journal
	4.2 Nature Protocols Journal
	4.3 Nature Reviews Microbiology Journal
	4.4 The FEBS Journal
	4.5 Translational Research Journal
5. Biotechnology	5.1 Applied Microbiology and Biotechnology Journal
	5.2 Bioresource Technology Journal
	5.3 Biotechnology and Bioengineering Journal
	5.4 Current Opinion in Biotechnology Journal
	5.5 Plant Biotechnology Journal
6. Chemistry	6.1 Analytica Chimica Acta Journal
	6.2 Analytical Chemistry Journal
	6.3 Journal of Chromatography A
	6.4 Talanta Journal
	6.5 The Analyst Journal
7. Food Chemistry	7.1 Food Chemistry Journal
	7.2 Food Microbiology Journal
	7.3 Journal of Food Science
	7.4 Journal of the Science of Food and Agriculture
	7.5 Meat Science Journal
8. Mathematics	8.1 International Journal of Mathematical Education in Science and Technology
	8.2 Mathematics Magazine
	8.3 Mathematics Teacher Journal
	8.4 The American Mathematical Monthly
	8.5 The College Mathematics Journal
9. Microbiology	9.1 Biocontrol Journal
	9.2 Biological Control Journal
	9.3 Mycologia Journal
	9.4 Phytopathology Journal
	9.5 Plant Disease Journal
10. Physics	10.1 ACS Nano Journal
	10.2 Journal of Computational Physics
	10.3 Nature Physics Journal
	10.4 Physics of Life Reviews
	10.5 Physics Reports
11. Statistics	11.1 Computational Statistics & Data Analysis Journal
	11.2 Journal of Statistical Planning and Inference
	11.3 Open Journal of Statistics
	11.4 Statistics and Computing Journal
	11.5 Statistics and Probability Letters Journal

To ensure that the words in the SAWL are useful for most science students, the expert-judged approach was used. According to Chung and Nation (2004), the expert-judged approach is the most reliable method for identifying technical words. The main tool for the expert-judged approach is a rating scale. The scale used in the present study was adapted from Chung and Nation (2004). In Chung and Nation (2004), words graded at Levels 3 and 4 were judged as technical words. Valipouri and Nassaji (2013) employed a similar scale. However, words at Level 4 were not included in their Chemistry Academic Word List (CAWL) because the purpose of their study was to develop an academic word list applicable to all four areas of chemistry. The words at Level 4 were considered too technical and specific to only one of the subject areas. They were not included in

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

the final CAWL. Tongpoon-Patanasorn (2018) explored the technical words in financial disciplines and used Chung and Nation's (2004) 4-point rating scale. The scale was reduced to three levels because the original Levels 1 and 2 could be viewed as non-technical words and the 3-point scale was easier for the raters. Similarly, Coxhead and Demecheleer (2018) employed Chung and Nation's (2004) 4-point rating scale and modified it. They also reduced the scale to three levels. The original Levels 2 and 3 were combined because they were slightly different and the scale with three levels allowed for a focus on technical words alone.

Likewise, Chung and Nation's (2004) 4-point rating scale was changed for the present study. The original Level 1 was removed from the modified rating scale because general words had been excluded from the potential list in the earlier step. The modified rating scale consists of three levels (shown in Table 3). As the present study aims to make a word list for 11 disciplines of natural science, the words rated at Level 3 by at least two of three experts were excluded from the list because they were considered to be too technical or very specific to few subject areas. The words classified at Levels 1 and 2 were included in the final SAWL.

Table 2. The size of the SAJ Corpus

Subject areas	Articles	Running words
1. Applied Physics	98	507,044
2. Aquatic Science	88	508,337
3. Biochemistry	89	501,808
4. Biology	66	508,004
5. Biotechnology	99	505,450
6. Chemistry	89	505,922
7. Food Chemistry	96	503,609
8. Mathematics	146	502,157
9. Microbiology	91	507,532
10. Physics	76	505,310
11. Statistics	124	507,772
Total	1,062	5,562,996

Table 3. The rating scale for the present study (adapted from Chung & Nation, 2004)

Level 1 Words that have a meaning that is **minimally related** to the 11 subject areas of science
Level 2 Words that have a meaning that is **closely related** to the 11 subject areas of science. The words are also used in general language but may have some restrictions of usage depending on the subject fields.
Level 3 Words that have a meaning specific to one or some of the 11 subject areas of science and are not likely to be known in general language. The words have clear restrictions of usage depending on the subject fields.

Word selection criteria

To make the SAWL from the SAJ corpus, the word selection criteria were established. This study adapted the word selection criteria in the AWL (Coxhead, 2000). According to the AWL, words were selected based on three criteria: specialized occurrence, range, and frequency.

Specialized occurrence refers to the occurrence of the words in specialized manners. Coxhead (2000) did not include general words from West's (1953) GSL. Many specialized academic word lists developed after the AWL also follow this rule and some researchers insist that the specialized words should not be listed in the AWL either. Coxhead and Hirsh's (2007) SWL focuses on specialized words occurring outside the GSL and AWL. However, some specialized word lists allow words in the GSL and AWL (e.g., Valipouri & Nassaji, 2013), while other word lists may include words in the AWL (e.g., Boonyos, 2014; Liu & Han, 2015; Yang, 2015). In the present study, both GSL and AWL were considered essential for science students. They should know these words prior to learning specialized academic words. Therefore, for the creation of the SAWL, the words occurring in the GSL and AWL were removed.

The range of a word refers to the occurrence of the word in each of the sections (or sub-corpora) of the corpus (Nation & Webb, 2011). The AWL was developed from a large corpus divided into four faculty divisions where each division comprises 875,000 running words from eight disciplines (or 28 discipline divisions in total). To be included in the AWL, the words have to occur at least 10 times in each of the four faculty divisions (i.e.1 time in every 87,500 running words) and in at least 15 of the 28 discipline divisions (53.6%). The SAJ corpus contains 11 sub-corpora. By applying Coxhead's (2000) principle to the present study, the words to be included in the SAWL occurred at least six times (500,000 ÷ 87,500 = 5.71) in at least six of the 11 subject areas (54.5%).

The frequency of a word in a corpus was the third condition. According to the AWL, each word in the list had to occur with a frequency of at least 100 times in the whole corpus of 3.5 million running words. That is equal to approximately 28.6 times in every one million running words of the corpus. This principle was adopted for many specialized word lists. For example, Coxhead and Hirsh's (2007) SWL was derived from a 1.7 million-word corpus. The cut-off frequency rate was 50 times in the corpus (28.6 x 1.7 = 48.6). Valipouri and Nassaji's (2013) CAWL was based on a four million-word corpus. The words in the list must occur at least 114 times in the corpus (28.6 x 4 = 114.4). Liu and Han's (2015) EAWL was developed from a 0.86 million-word corpus. The frequency rate for EAWL was 30 times in the corpus (28.6 x 0.86 = 24.6). In the present study, the corpus contains around 5.5 million running words. Hence, the appropriate frequency rate for the SAWL was 155 times in the whole corpus (28.6 x 5.5 = 157.3).

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

In summary, the word selection criteria for the SAWL had three conditions. (1) *Specialized occurrence*: The first 2000 most frequent words in the GSL and the 570 academic words in the AWL were removed. (2) *Range*: A word family included in the SAWL had to occur at least six times in six or more of the 11 subject areas. (3) *Frequency*: A word family included in the SAWL had to occur with a frequency of at least 155 times in the whole CAJ corpus.

Data analysis

Creating the SAWL involved two methods: a corpus-based approach and expert-judged approach. The corpus-based approach consists of four major steps. First, the SAJ corpus was loaded into the AntWordProfiler program. The SAJ corpus comprises 11 text files. Each file contains around 500,000 running words derived from research articles and review papers published in selected scientific, academic journals. An overall list of word families occurring in the SAJ corpus was created. Second, the word families in the list were refined and compared with West's (1953) GSL and Coxhead's (2000) AWL. The word families coinciding in the GSL and AWL were removed. Next, the remaining words were further investigated. Words like transparent compounds, proper names, non-words, foreign words, and abbreviations were removed from the results. Finally, the words that met all selection criteria were kept. The potential SAWL was generated based on this result. At this stage, AntConc program was employed to closely explore some words in detail to make a decision whether they should be counted as a word or not.

In the expert-judged approach, a panel of three experts was invited to review whether the words in the potential SAWL should be included in the final list from a scientific point of view. In the present study, the panel of three experts consisted of three experienced lecturers from the Faculty of Science who volunteered to participate in the study. A detailed written summary of the scope and objectives of this study was sent to all the lecturers. They also received the questions and rating scale, which was modified from Chung and Nation (2004). Each of the experts was asked to make an independent judgment based on the question of whether the word was specific to any discipline of natural science. The words were excluded in the SAWL if they were rated too specific by two of the three raters. The inter-rater reliability test (the Kappa statistic) was applied to the analysis. The reliability test showed a high rate of agreement among the experts: 0.93, or 93%.

FINDINGS AND DISCUSSION

The science academic words

The first objective of the study was to identify science academic words frequently used by academic writers. The 5.5-million-word SAJ corpus was compiled for the study. The words in the corpus were divided into four levels: GSL-K1, GSL-K2, AWL, and others (lower frequency words). Table 4 shows the proportion in the SAJ corpus.

The proportion in the SAJ corpus reflects the notion that scientific English has special characteristics. In general, the GSL covers around 70% to 95% of most text (Gilner, 2011; Nation & Hwang, 1995). However, as the SAJ corpus is the corpus of scientific academic text, the GSL provides only 63% coverage. In other words, the SAJ corpus contains fewer general words than corpora of general texts. It is worth noting that 108 GSL words were not found in the SAJ corpus, especially those with connotative or emotional meaning such as *absolutely*, *ashamed*, *laughter*, *loyal*, and *polite*. The findings are in line with the characteristics of scientific language. Halliday and Webster (2004) and Reeves (2005) propose that the English language used in science has many technical terms and it avoids general words with connotative or emotional meanings.

Table 4. The proportion of word types in the SAJ corpus

	Word Levels	Running Words		Groups	
		No. of running words	Percent	No. of Groups	Percent
1	GSL K-1	3,239,363	58.23%	994	0.95%
2	GSL K-2	285,525	5.13%	898	0.86%
3	AWL	561,119	10.09%	568	0.54%
4	Others	1,476,989	26.55%	102,259	97.65%
	Total	5,562,996	100.00%	104,719	100.00%

The SAJ corpus also comprises a significant proportion of AWL. As a corpus of academic text, the coverage of the AWL in the SAJ corpus was 10%, in which 568 AWL word families were detected. This figure is at the same level of Coxhead's (2000) study that the 570 words of AWL cover 10% of the academic corpus. The GSL and AWL altogether brought coverage of the SAJ corpus up to 73%. To identify science academic words that are worth learning, the Level-4 words were further investigated.

Science academic words were selected from SAJ corpus based on the three criteria of specialized occurrence, range, and frequency. Altogether, 513 word families met the word selection criteria. Then, the possible science academic words were rated by a panel of three experts using the 3-level rating scales adapted from Chung and Nation (2004). From 513 word families, the experts agreed to remove 81 words from the list. Most of the eliminated words were scientific names, e.g., *Bacillus*, *cerevisiae*, *Drosophila*, and *necrosis*. Some words were those usually occurring together with specialized words, e.g., *efficiently*, *favorable*, and *mapping*. This is in line with Chung and Nation (2004), which noted that this

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

method could not detach collocations of technical words from the list.

The final SAWL list comprises 432 word families (see Appendix A for the alphabetical list of 432 headwords). Table 5 shows the coverage of the SAWL in the SAJ corpus. The whole list covers 5.82% of the corpus. The combination of the GSL, the AWL, and the SAWL provides up to 79.43% coverage of the running words in the SAJ corpus. However, Nation (2013) points out that 95% - 98% coverage is sufficient for comprehending reading text.

Excerpt 1 provides an example of text from the SAJ corpus (136 running words). The high-frequency words (GSL-K1, GSL-K2) are unmarked, the AWL words are in *italics*, the SAWL words are in bold, and the other lower frequency words and abbreviations are underlined. Twenty-seven SAWL words occur in this text and account for 20% of the running words. The four lists (GSL-K1, GSL-K2, AWL, and SAWL) brought text coverage up to 90%. In other words, only one word in every ten words is not in the four lists.

To aid vocabulary selection, Coxhead (2000) divided the AWL into 10 sub-lists based on frequency, each of which contains 60 word families. This method has been applied in the SWL (Coxhead & Hirsh, 2007) and the CAWL (Valipouri & Nassaji, 2013). The first sub-list of 60 most frequent word families in the SAWL was also created, shown in Table 6.

The coverage of the first 60-word sub-list was 2.52%, while the whole list covers 5.82% of the SAJ corpus. The figures imply that this sub-list should be learned before learning the words with less coverage because it provides a better return for learning effort.

Table 5**.** The coverage of different base word lists over the SAJ corpus

Word lists	Running words	% of SAJ	Headwords
1st GSL	3,239,363	58.23%	994
2nd GSL	285,525	5.13%	898
AWL	561,119	10.09%	568
SAWL	323,611	5.82%	432
Off-list	1,155,034	20.76%	100,888
Total	5,562,996	100.00%	103,780

Excerpt 1. An example of text on biology from the SAJ corpus

With the development of life science and **biomedical** science, the *detection* of low-**abundance proteins** and the *acquisition* of ultra-weak **biological** signals have become a bottleneck of these fields. We *predict* a bright future for **nanoparticle**-based immunoassays owing to their *unique physical* and *chemical* properties. Moreover, recently *published* reports also *indicate* that **nanoparticles conjugated** with various *targeting* **molecules** or **antibodies** can be used to *target specific* **substrates** in **vitro**. Possibly, upcoming work will be performed by *coupling* **functionalized** nanomaterials with **molecular biological** *techniques*. By introducing the **functionalized** nanomaterials, **novel** technologies such as rolling circle **amplification** (RCA), *target-induced* repeated **primer** extension, **hybridization** chain *reaction*, **loop**-*mediated* **amplification** and *target* DNA recycling **amplification,** including endonuclease, exonuclease and **polymerase**-based circular **strand**-replacement **polymerization** have been applied to **amplify** the electrochemical, **optical** and *visual* signals.

Table 6. The 60 most frequent words in the SAWL

***Note:* Headword *(Range, Coverage%)* / (*) = also in SWL Sublist 1**

1.	protein *(11,0.163)**	21.	assay *(10,0.038)*	41.	spatial *(11,0.028)*
2.	species *(11,0.158)**	22.	carbon *(10,0.037)**	42.	incubate *(9,0.028)*
3.	acid *(9,0.127)**	23.	column *(11,0.036)**	43.	membrane *(10,0.026)**
4.	gene *(11,0.104)*	24.	correlate *(11,0.035)*	44.	fraction *(11,0.026)*
5.	mathematics *(11,0.071)*	25.	composition *(11,0.034)*	45.	magnet *(10,0.026)**
6.	molecule *(11,0.062)**	26.	synthesis *(11,0.034)*	46.	organic *(9,0.026)**
7.	strain *(10,0.061)*	27.	lipid *(9,0.034)*	47.	peptide *(8,0.026)*
8.	matrix *(11,0.061)*	28.	fluorescent *(10,0.032)*	48.	coefficient *(11,0.025)*
9.	ion *(10,0.056)**	29.	residue *(11,0.031)*	49.	receptor *(9,0.025)*
10.	dense *(11,0.052)**	30.	fungus *(9,0.031)*	50.	buffer *(10,0.024)*
11.	activate *(11,0.051)*	*31.*	amino *(9,0.03)*	51.	laboratory *(11,0.024)**
12.	linear *(11,0.049)**	*32.*	cancer *(11,0.03)*	*52.*	nanoparticle *(7,0.024)*
13.	infect *(11,0.048)**	33.	genetic *(10,0.029)*	53.	abundant *(11,0.024)*
14.	tissue *(11,0.045)**	34.	genome *(9,0.029)*	54.	transcript *(10,0.024)*
15.	coating *(10,0.044)*	35.	muscle *(10,0.029)**	*55.*	reference *(11,0.024)*
16.	bacterium *(10,0.043)**	36.	plasma *(11,0.028)*	56.	virus *(9,0.023)*
17.	enzyme *(9,0.042)**	37.	pathogen *(9,0.028)*	57.	diffuse *(11,0.023)*
18.	pathway *(11,0.04)*	*38.*	spectra *(10,0.028)*	58.	microscope *(11,0.*023)
19.	cellular *(11,0.039)*	*39.*	electron *(11,0.028)**	*59.*	optic *(10,0.023)*
20.	peak *(11,0.039)*	*40.*	imaged *(11,0.028)*	60.	absorb *(11,0.023)**

To prove that the SAWL is appropriate for the learning of natural science disciplines, the validity of SAWL was tested. Nation and Webb (2011) suggest that a good word list should work well on the corpus from which it was made and work poorly on another independent corpus. The coverage of the SAWL on the SAJ corpus was 5.82%. It was cross-checked against two different corpora – a corpus of English news (EN) and a

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

corpus of science academic texts (SAT). The performance of the SAWL on the EN corpus was very poor (0.51% coverage) while it worked well on the SAT corpus (5.72% coverage). This indicates that the SAWL contains specialized academic words of natural science disciplines.

Comparing the SAWL and SWL

The present study also explored the distinguishing features of the SAWL that make it different from the SWL (Coxhead & Hirsh, 2007) in order to claim that the SAWL better serves the needs of EFL science students. The findings reveal two aspects to support the claim.

First, all word families in the SAWL are more specific to natural science disciplines than the SWL. Of its 432 word families, the SAWL shares 176 (41%) with the SWL, while 256 (59%) are different. In other words, the majority of word families in the SAWL are different from SWL. It was found that words related to health science and technological science, which are in the SWL, are not included in the SAWL (e.g., *anatomy, glad, hormone, insulin, cylinder, fuel,* and *propel*). Moreover, some of the SWL words are the words removed from the SAWL during the rating process, including *calcium, capture, carbohydrate, carbon, cavity, chamber, chloride, chronic, climate, cluster,* and *defense.* These words have been removed from the final SAWL because the experts found that their meanings are specific to only a few disciplines of natural science. As a result, SAWL contains more word families that are useful for the EFL students majoring in natural science disciplines.

Second, the SAWL words families are more frequently used in natural science research articles, which implies that science students could have more opportunities to encounter them. The SWL claims that it has 3.79% coverage, which means one word in every 25 words. The coverage of the SAWL is 5.82% or one in every 17 words. As the aforementioned coverage rates are the result of performing on different corpora, the SAWL and the SWL were tested again on the same corpus – the 1.1-million-word SAT corpus. As shown in Table 5, this method also yields almost similar results (5.72% and 3.91% coverage respectively). These findings confirm that the SAWL, which has been developed for the 11 subject areas of natural science, is more useful for the science students.

Table 7. The coverage of SAWL and SWL over the SAT corpus

Word lists	Running Words	% of SAT	Headwords
1st GSL	649,071	58.41%	923
2nd GSL	52,258	4.70%	674
AWL	106,978	9.63%	554
SAWL	**63,522**	**5.72%**	**435**
SWL	**43,406**	**3.91%**	**313**

CONCLUSION

The present study developed the specialized academic word list (SAWL) for 11 natural science disciplines. The corpus-based approach and the expert-judged approach were used to identify specialized academic words to make a list. The SAJ corpus, the corpus used for this study, was derived from 1,062 articles published in international academic journals recommended by 11 professors from different natural science disciplines. The SAWL was then reviewed by the panel of three professors as the experts in the field. The final list contains 432 word families and covers 5.82% over the SAJ corpus. Moreover, the SAWL performed better than the SWL (Coxhead & Hirsh, 2007).

The findings confirm previous studies (e.g., Ackermann & Chen, 2013; Coxhead & Demecheleer, 2018; Tongpoon-Patanasorn, 2018; Valipouri & Nassaji, 2013) in that making technical word lists should involve more than the corpus-based approach. The weakness of the corpus approach is that it cannot detach the collocates of technical words from the list (Chung & Nation, 2004). Therefore, the expert-judged approach was also applied in this study. Decisions from experts in the field are beneficial for selecting useful items into specialized word lists. In addition, the rating scale used in this study was reduced from four levels to three levels, similar to the method used by Coxhead and Demecheleer (2018) and Tongpoon-Patanasorn (2018). It seems that the modified rating scale helped the experts make decisions more effectively.

The results of this study suggest several pedagogical implications. As the SAWL provides high coverage of science English in research articles, it should be a good resource for students and teachers of science English, syllabus designers, and material developers. There are three specific suggestions for using the SAWL. First, attention should be given to collocations used together with the SAWL words. Teachers should introduce how the SAWL words are used in the correct context. Second, apart from reading, teachers should encourage EFL students to use the SAWL words in their academic writing and speaking. Finally, the SAWL was built on the notion that the science students are familiar with the most commonly used words in GSL (West, 1953) and general academic words in AWL (Coxhead, 2000). However, for low proficiency students, teachers might design their ESP courses that are accompanied by GSL, AWL, and SAWL.

There are some limitations to this study. Although the corpus used for this study included 5.5 million running words, it is from only one text type – journal articles. Particular attention should be given when using the SAWL with other text types such as textbooks or technical documents. Second, this study covers 11 subject areas of natural science disciplines. They are the disciplines of science offered at the university where this study was carried out. Other universities may not offer the same disciplines, and this can limit the replication of this study. In addition, the present study does not offer

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

lexical information (e.g., part of speech, specific meaning, or collocations) of each item.

Future research of specialized academic word lists can be conducted to address the following issues. As noted earlier, one of the limitations of this study is that the SAJ corpus was compiled from only one text type, i.e., journal articles. Other text types such as textbooks, conference papers, reports, and theses could be explored. Second, more research could be done on a list of multiword units or collocations in specialized academic fields, e.g., the Academic Collocation List (Ackermann & Chen, 2013). Finally, future research could be done on effective methods for integrating these specialized academic word lists into professional practice.

REFERENCES

Ackermann, K., & Chen, Y. H. (2013). Developing the academic collocation list (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. doi: 10.1016/j.jeap.2013.08.002

Anthony, L. (2014). AntWordProfiler (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University, Retrieved from http://www.laurenceanthony.net/.

Anthony, L. (2016). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Boonyos, K. (2014). *Establishment of a microbiology academic word list.* (Unpublished master thesis). Thammasat University, Thailand.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics, 36*(1), 1-22. doi: 10.1093/applin/amt018

Chanasattru, S., & Tangkiengsirisin, S. (2016). Developing of a high frequency word list in Social Sciences. *Journal of English Studies, 11*, 41-87.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System, 32*(2), 251-263. doi: 10.1016/j.system.2003.11.008

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

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

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue française de linguistique appliquée, 12*(2), 65-78.

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

Gilner, L. (2011). A primer on the general service list. *Reading in a Foreign Language, 23*(1), 65-83.

Halliday, M. A. K., & Webster, J. J. (2004). *The language of science*. London: Bloomsbury Academic.

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. doi: 10.1016/j.jeap.2016.01.008

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

Matsuoka, W. (2012). Searching for the right words: Creating word lists to inform EFL learning. In D. Hirsh (Ed.), *Current perspectives in second language vocabulary research* (pp. 151-177). Bern: Peter Lang.

Nation, P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

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

Nation, P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35-41.

Nation, P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Pugsee, P., Limgomolvilas, S., Wudthayagorn, J., & Janpugdee, P. (2017). *A framework for generating ICT word lists.* Paper presented at the 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), Pattaya, Thailand.

Reeves, C. (2005). *The language of science*. Oxon: Routledge.

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

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. doi: 10.1016/j.jeap.2013.07.001

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

West, M. (1953). *A general service list of English words with semantic frequencies*. London: Longman.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication, 3*, 215-229.

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27-38. doi: 10.1016/j.esp.2014.05.003

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

Appendix A

The 432 headwords of the Science Academic Word List (SAWL) – in an alphabetical order.
***Note:* (*) = also in SWL (Coxhead & Hirsh, 2007)**
The full SAWL is available at https://sites.google.com/go.buu.ac.th/sawl.

1. absorb*
2. abundant*
3. acetate
4. acetic
5. acetone
6. acid*
7. activate
8. acute*
9. additives
10. adhesion
11. adverse*
12. affinity
13. agar
14. agarose
15. albumin
16. alcohol*
17. align*
18. alkaline
19. allele
20. ambience
21. amino*
22. ammonia*
23. amplify
24. amplitude
25. anaerobic
26. anneal
27. antimicrobial
28. apoptosis
29. aqueous
30. architecture
31. aromatic
32. array*
33. assay
34. atmosphere*
35. atom*
36. bacillus
37. bacterium*
38. barrier*
39. basal*
40. baseline
41. batch
42. bead
43. bioactive
44. biochemical
45. biology
46. biomass
47. biomedical
48. biosynthetic
49. biotechnology
50. bovine
51. breakdown*
52. breast
53. buffer*
54. calcium*
55. calibrate*
56. cancer
57. candidate
58. capillary
59. capture*
60. carbohydrate*
61. carbon*
62. cardiac
63. cardiovascular
64. cascade
65. catalyse
66. cavity*
67. cellular
68. cellulose
69. centrifuge
70. cerevisiae
71. chamber*
72. chemistry
73. chip
74. chloride*
75. cholesterol
76. chromatography
77. chromosome
78. chronic*
79. climate*
80. clinic
81. clone
82. coating
83. coefficient*
84. colon
85. column*
86. composition
87. configure*
88. confocal
89. conserve*
90. contaminate*
91. correlate*
92. covalent
93. crude
94. crystal
95. cumulative*
96. cysteine
97. dash
98. database
99. dataset
100. decompose*
101. defect*
102. deficiency*
103. degrade*
104. dehydrogenased
105. dense*
106. dependence*
107. depict
108. deplete
109. deposit*
110. deposition
111. developmental
112. diagram*
113. diet
114. differential*
115. diffract
116. diffuse*
117. digest*
118. digital
119. dilute*
120. discrepancy
121. disperse*
122. disrupt
123. dissolution
124. dissolve*
125. distil
126. diverge*
127. donor
128. downstream
129. droplet
130. drug*
131. dual
132. dye
133. ecological
134. ecosystem
135. efficiently
136. electrode*
137. electron*
138. electrophoresis
139. electrostatic
140. elemental
141. elevate*
142. elongate*
143. embed*
144. emit*
145. encode
146. endogenous
147. engineered
148. enrich
149. enzymatic
150. enzyme*
151. epithelial
152. ester
153. ethanol
154. evaporate*
155. excitation
156. exogenous
157. exponential*
158. fabricate
159. favorable
160. favorably
161. feasible
162. feedback*
163. feeding
164. ferment
165. fibre*
166. filter*
167. fluid*
168. fluorescent
169. flux*
170. fraction*
171. fragment*
172. frequencies
173. functionalize
174. fungus*
175. fuse*
176. gel
177. gene
178. genetic
179. genome
180. genotype
181. genus*
182. geography
183. geometry
184. germ
185. glucose
186. glycerol
187. graph*
188. grid*
189. gut
190. height*
191. hybrid*
192. hydrogen*
193. hydrolysis
194. hydrophilic
195. hydrophobic
196. hydroxyl
197. imaged
198. immobilize
199. immune
200. incubate*
201. infect*
202. inflame
203. infrared*
204. inject*
205. inset
206. intact
207. intake
208. interestingly
209. interface*
210. interior*
211. intestine*
212. invasive
213. inverse
214. ion*
215. kernel
216. kidney*

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468

217.kinase
218.kinetic
219.kit
220.laboratory*
221.lactic
222.laser
223.latent*
224.lateral*
225.lattice
226.linear*
227.lipid*
228.liver*
229.localise
230.locus
231.longitudinal*
232.loop*
233.lysine
234.magnesium*
235.magnet*
236.magnify
237.magnitude*
238.mapping
239.marine*
240.mathematics
241.matrix
242.maximal
243.median*
244.membrane*
245.mesh*
246.metabolic*
247.metabolism
248.metabolite
249.methanol
250.methionine
251.micro
252.microbe*
253.microorganism
254.microscope*
255.mitochondria
256.mobile*
257.molar
258.molecule*
259.morphology*
260.mortal*
261.mount
262.muscle*
263.mutant
264.mutate
265.nanoparticle
266.negligible
267.neural
268.nitrogen*
269.node*
270.novel
271.nucleotide
272.nucleus*
273.null
274.nutrient*
275.nutrition
276.online
277.onset*
278.optic
279.optimal
280.optimise
281.optimum*
282.oral
283.organic*
284.organism*
285.oven
286.overnight
287.overview*
288.oxidant
289.oxide*
290.oxidise
291.oxygen*
292.pathogen
293.pathogenic
294.pathway*
295.patients
296.peak*
297.penetrate*
298.peptide
299.periphery*
300.peroxide
301.pharmaceutical
302.phenotypic
303.phosphate*
304.phylogenetic
305.physiological*
306.plasma*
307.plastic
308.platform
309.plot*
310.polar*
311.poly
312.polymer
313.polymerase
314.polynomial
315.pooled
316.pore*
317.posterior
318.potassium*
319.potent
320.precipitate*
321.precursor
322.prevalent
323.primer
324.probe
325.profile*
326.progression
327.proliferate
328.proline
329.propagate*
330.protease
331.protein*
332.pulse*
333.purify*
334.purity
335.putative
336.quantify
337.reagent
338.receptor
339.redox
340.reference
341.regress
342.replicate*
343.reservoir*
344.residue
345.resonance
346.resuspend
347.robust
348.rotate*
349.routine*
350.saline*
351.saturate*
352.scaling
353.scan
354.score
355.seasonal
356.secrete*
357.segment*
358.sensing
359.sensor
360.serine
361.serum*
362.setup
363.silica
364.silicon
365.simultaneous*
366.skeletal*
367.sodium*
368.software
369.soluble*
370.solvent
371.spatial*
372.species*
373.spectra
374.spectre
375.spontaneous
376.static*
377.storage
378.strain*
379.strand
380.subset
381.superior*
382.supernatants
383.suppress
384.susceptible*
385.switch*
386.symmetry*
387.symptom*
388.synergistic
389.synthesis*
390.synthetic*
391.tank
392.taxonomy
393.template
394.temporal*
395.tertiary
396.therapeutic
397.therapy
398.thermal*
399.threshold
400.tissue*
401.tolerance
402.toxic*
403.toxin
404.tract*
405.transcript
406.transient*
407.triangle
408.triple
409.triplicate
410.tumour
411.tyrosine
412.unclear
413.untreated
414.uptake*
415.urea*
416.vascular
417.velocity
418.verify*
419.versus*
420.vertical*
421.viable*
422.video
423.virus*
424.vitamin*
425.volatile
426.volt*
427.wavelength*
428.weighted
429.worldwide
430.yeast
431.zinc*
432.zone*

Copyright © 2018, *IJAL*, e-ISSN: 2502-6747, p-ISSN: 2301-9468