Academic and Technical Vocabulary in the Corpus of Chemistry Research Articles

Pinrada Nuamjapho, Khon Kaen University, Thailand
Kornwipa Poonpon, Khon Kaen University, Thailand

The Asian Conference on Education & International Development 2017
Official Conference Proceedings

Abstract
Many researchers have mentioned the importance of vocabulary in language learning , in particular, vocabulary for specific disciplines. However, there are some problems with the vocabulary selected to teach students because some teachers strongly believe in their intuition for choosing word families to teach. Therefore, Schmitt and Schmitt (2005) claimed that the best way to determine vocabulary frequency is by using frequency lists compiled from vocabulary databases called a corpus or corpora**.** This study is a corpus-based study which aims to develop lists of high frequency words and to identify the proportions of words from the General Service List (GSL), the academic word list (AWL), and technical words in a chemistry research articles corpus in order to know the proportion of each word types compared with previous studies**.** A corpus of about 1.2 million token words was compiled from 210 chemistry research articles, derived equally from seven sub fields of chemistry. The RANGE Program was used in this study to identify the first 1000 and second 1000 GSL, AWL, and rare words. Then**, a** Rating Scale adapted from Chung and Nation (2003) was used to identify technical vocabulary. From the analysis, it was found that the proportion of GSL, AWL, technical words, and other words found in the present study were 62.5, 9.2, 7.0 and 21.3, respectively**.**

Keywords: Academic vocabulary, Technical Vocabulary, Corpus

Introduction

Vocabulary plays an important role for learning a language because it is always viewed as a crucial tool for language learners to achieve successful communication. Many researchers have mentioned the importance of vocabulary in language learning. Wilkins (1972) said that "while without grammar very little can be conveyed, without vocabulary *nothing* can be conveyed" (pp. 111–112). Moreover, Lewis (1993) claimed that vocabulary is the core or heart of language. He also mentioned that "language consists of grammaticalised lexis, not lexicalized grammar" (p. 51). Nation (2001) suggested that knowing vocabulary is considered useful for language learners. Meara (1996) also stated that learners who know more vocabulary are proficient in language skills than those who know a smaller amount of vocabulary. Therefore, it can be said that vocabulary is an essential factor for teaching and learning a language.

Due to the significant role of vocabulary, many approaches have been used and investigated for teaching and learning vocabulary. To answer a question about how vocabulary should be taught and learned, Schmitt and Schmitt (2005) suggested that the major approach for English vocabulary teaching is a *frequency* perspective due to the large amount of English vocabulary (Schmitt & Marsden, 2006). However, some teachers trustin their own intuition for choosing word families to teach. Therefore, Schmitt and Schmitt (2005) claimed that the best way to determine vocabulary frequency is by using frequency lists compiled from vocabulary databases called a corpus or corpora. From this perspective, Nation (2001) and others divided the entirety of vocabulary into four levels based on frequency. Firstly, the general service list or GSL (West, 1953) is the list of the highest frequency words. The GSL list contains the 2,000 most frequently used word families (70-80% of texts) including function words. Secondly, the academic word list or AWL (Coxhead, 2000), and the university word list or UWL (Xue and Nation, 1984) are the common academic vocabularies in academic texts and context. The third level, technical or specialized vocabulary, is distinct in each discipline. Finally, the fourth level is the list of low frequency words or unusual words that may not occur in general texts, such as certain symbols, abbreviations etc. Based on the high coverage of the GSL in different texts, some researchers suggest that GSL words are significant for all English language learners and that students should primarily pay attention to GSL words when studing a second language (Coxhead & Nation, 2001; Nation & Kyongho, 1995; Nation & Waring, 1997). For academic contexts, such as university education, academic words and technical words are very important. Students should know academic and technical vocabulary because they will need to employ them during their academic courses or disciplines.

Due to the importance of academic vocabulary, there have been a number of studies focusing on academic words in different disciplines. Coxhead (2000) made an attempt to create the academic word list (AWL) by collecting words comprising texts in law, arts, commerce, and science. The total size of her corpus was 3.5 million. The AWL consists of 570 word families in total. The AWL in Coxhead's study accounted for about 10% of the total words in that corpus. According to Coxhead (2011), the aim of the academic word list is to help teacher of English for Academic Purposes (EAP) "set goals for their student's vocabulary learning" (p.357). The study by Coxhead (2000) was an important innovation in several fields of vocabulary learning and teaching. Based on her study, several other vocabulary studies investigated academic

words in specific perspectives from different disciplines such as engineering by Mudraya (2006), electrical engineering by Wasuntarasophit (2008), medicine by Wang, Liang, and Ge (2008), agriculture by Martinez, Beack, and Panza (2009), applied linguistics by Vongpumivitch, Huang, and Chang (2009), finance by Li & Qian (2010), chemistry by Valipouri & Nassaji (2013), education by Mozaffari & Moini (2014), nursing by Yang (2015), environmental science by Liu & Han (2015), and medical science by Lei & Liu (2016). As exemplified by previous studies, the AWL can be regarded as the most widely cited academic word list from different fields (Yang, 2015). Although Nation (2001, p.12) mentioned that Coxhead's AWL was "the best list", more recently, some researchers have focused on the development of AWLs found in a specific discipline and some other specific vocabularies appeared in each field called *technical vocabulary*.

Apart from GSL and AWL, technical words occurring across different disciplines are also beneficial for learners, as these types of words can help them understand discipline- specific texts. Nation (2001) and Strevens (1973) suggested that students who study in the scientific area may have problems with technical terms. This means knowledge of the specific terms is essential in learning the technical language of that field. However, a few studieshave been conducted regarding technical vocabulary and the criteria for deciding which words are technical words and which are not, depending on different disciplines of vocabulary. There have also been some previous studies focusing on technical vocabulary. Chung and Nation (2003) investigated technical vocabulary from anatomy textbooks with a corpus of 450,000 words and linguistics textbooks with a corpus of 93,445 words. Wasuntarasophit (2008) explored technical vocabulary from electrical engineering textbooks with a corpus of 120,000 words, and Lessard-Clouston (2010) studied technical vocabulary from theology lectures with a corpus of 10,470 words. The results of these studies show the proportion of vocabulary. Nation (2001), claimed that academic vocabulary covers 10% and technical vocabulary covers 5 % of running words in academic texts, whereas other researchers discovered different vocabulary group proportions in their corpus. Chung and Nation (2004) found that technical vocabulary appears with higher frequency in anatomy texts, at 37.6 % and only 16.3 % in applied linguistics texts. Wasuntarasophit (2008) argues that technical vocabulary should not be abandoned because it covers 20.6% of all words in electrical engineering texts. However, several of the studies above focused on vocabulary in textbooks from that field. Therefore, another way to identify technical vocabulary is through research articles themselves, as they are authentic written texts from several writers around the word.

As mentioned earlier, Nation (2001) and Strevens (1973) suggested that students in scientific areas may have difficulty in technical terms. One challenge among learners who study particular disciplines especially chemistry, is the difficulty to read and understand chemistry research articles that regularly consist of many types of vocabulary. Varipouri & Nessaji (2013) made an attempt to compile a large chemistry research articles corpus in order to investigate the frequency of academic words and general words based on AWL of Coxhead (2000) and GSL of West (1953). However, the study of Varipouri and Nessaji (2013) did not mention technical vocabulary which is one of the most important word types for students who study in very specific discipline such as chemistry. The researcher in this present study is well- qualified with a bachelors degree in chemistry, and experience with research in inorganic

chemistry, which may raise the researcher's passion to study and conduct research about different kinds of vocabulary from the chemistry corpus.
To increase the reliability of the method used in this study, the researcher hope that the present study can address some limitations found in the studies mentioned above, particularly the study about technical vocabulary in chemistry based on the framework of Chung and Nation (2003). The present study aims to identify and classify vocabulary types from the chemistry research articles corpus and hope that the vocabulary list from the study will provide guidelines or materials to help students develop greater vocabulary knowledge in order to improve comprehension of chemistry research papers.

(1) There are three research questions in this study: What are the proportions of GSL and AWL in the chemistry research articles corpus based on the RANGE program?
(2) What are the high-frequency academic words in the chemistry research articles corpus?
(3) To what extent are the chemistry technical words used in the chemistry research articles corpus?

Methodology

This corpus was a collection of 210 research articles in chemistry. Based on the SCIMAGO Institutions Ranking (SJR) the chemistry research articles were divided into seven subject areas: analytical chemistry, chemistry (miscellaneous), electrochemistry, inorganic chemistry, organic chemistry, physical and theoretical chemistry, and spectroscopy**.** Therefore, the research articles were selected equally from seven sub fields of chemistry.

After the chemistry research articles were downloaded, each article was prepared in the IMRaD research paper format, Swales (1990). The IMRaD format is a common organization structure of scientific research articles formatted to consist of four parts as follows: introduction, methodology, results and discussion. Then, the data was saved into a text file (*.txt) and analyzed by the RANGE program. The RANGE program (Nation and Heatley, 2002) was the software used in this study for basic classification of three main vocabulary groups: 1) GSL 2) AWL and off lists. However, it was found that some of GSL and AWL can be technical word, therefore the Rating Scale was used in order to identify technical vocabulary and it is presented below.

Group 1 Function words that have meanings with no particular relation to the field of chemistry. This group also includes grammatical words such as articles, preposition, modal verbs, auxiliary verbs, conjunction, pronoun, and some adverbs. Examples are: *the, is, and, between, always, become, its, with* and so on. The vocabulary of this group appears in GSL (West, 1953).
Group 2 Content words including nouns, verbs, adjective and adverbs that are used in the academic field, but whose meanings have minimal relation to the field of chemistry. Examples are: *storage, complete, major, enormous, cycle, similar* and so on. The

vocabulary from group 2 is not considered technical chemistry vocabulary.
Group 3 Content words that have two different meanings. They contain meaning both in the field of chemistry and in other, non-related fields. These words are found in the GSL (West, 1953) and AWL (Coxhead, 2000). This group of words is considered technical chemistry words when they are used or appear in chemistry research articles. Examples are: *absolute, abundance, lake, daughter, habit, lead* and so on.
Group 4 Content words or specific words used to explain results or, principle of theory related to the field of chemistry. They can be units, materials, equipment, derivatives, process technique, machines, theory names or chemist names. They are used to describe the experiment and theory. The vocabulary in this group is considered technical vocabulary. Examples are: *quartz, radiation, isomer, octahedral, spectroscopy, unsaturation, reagent, Fahrenheit*, and so on.
Group 5 Content words or very specific words that are used in chemistry field and rarely found in other fields. These words are the name of chemical elements, chemical substances and chemical materials used in the laboratory experiments. Examples are: *cadmium, palladium, pyruvic, sulfuric, methanol* and so on.

The rating scale in the present study was trained before being used and it was used by three experts in Chemistry, as well as one researcher. The purpose of this is so that four raters can identify words with the same criteria. Furthermore, a Chemistry Dictionary was used in this study in order to double check the words that have specific meanings in the field of chemistry. The vocabulary in groups three, four and five according to the Rating Scale were considered technical vocabulary.

Results and Discussion

This corpus was a collection of 210 chemistry research articles from seven sub fields of chemistry (30 articles from each sub-field). These sub-fields included analytical chemistry, chemistry (miscellaneous), electrochemistry, inorganic chemistry, organic chemistry, physical theoretical chemistry and spectroscopy. The total number of running words in the Chemistry Research Articles Corpus used in this study was around 1,129,000 words.

Table 1 Proportion of four kinds of words list in the Corpus of Chemistry Research Articles

Word List	Running Words (Tokens)	Types
1st GSL	655,705 (58.09%)	2,574 (6.13%)
2nd GSL	54,010 (4.78%)	1,374 (3.27%)
AWL	118,056 (10.47%)	1,965 (4.69%)
Off list	300,910 (26.66%)	36,065 (85.91%)
TOTAL	**1,128,681 (100.00%)**	**41,987 (100.0%)**

From Table 1, we can see that GSL (1st GSL and 2nd GSL) has the highest proportion of words in the corpus, covering approximately 62.9% of the words in the corpus, AWL covers around 10.5% and the last group, which contained both technical and rare words, cover around 26.7% of the total words in the corpus. It is clear that words in the GSL makeup the biggest part of the corpus. In other words, on average, for every 100 running words in the corpus, we will see words from the GSL approximately 60 times. Furthermore, it can be inferred that, for every 100 words of the text in the corpus, we will see around 10 AWL word appearances.

Research question 2: *What are the high-frequency academic words in the chemistry research articles corpus?*

The analysis of high-frequency academic words shows that there were 162 academic words found in the corpus of chemistry research articles. They are shown in Table 2.

Table 2 High-frequency used AWL words in the corpus of Chemistry Research Articles

Rank	Word	Frequency	Rank	Word	Frequency
1	Structure	1,725	16	Section	569
2	Data	1,655	17	Ratio	547
3	Chemical	1,616	18	Via	511
4	Method	1,075	19	Potential	506
5	Complex	1,015	20	Shift	498
6	Energy	996	21	Bond	484
7	Range	814	22	Site	447
8	Transfer	804	23	Source	899
9	Function	759	24	Significant	445
10	Approach	749	25	Mechanism	444
11	Phase	721	26	Specific	433
12	Nuclear	678	27	Scheme	416
13	Sequence	676	28	Region	399
14	Similar	596	29	Available	395
15	Process	593	30	Constant	370
Rank	**Word**	**Frequency**	**Rank**	**Word**	**Frequency**
31	Mode	340	62	Primary	210
32	Research	340	63	Role	205
33	Compound	334	64	Design	200
34	Domain	328	65	Furthermore	198
35	Initial	324	66	Core	196
36	Target	310	67	Final	192
37	Technique	291	68	Image	191
38	Factor	290	69	Occur	190
39	Overall	274	70	Error	188
40	Area	272	71	Theory	181
41	Transport	272	72	Environment	178
42	Layer	268	73	Indicate	178
43	Major	263	74	Subsequent	177
44	Obtain	260	75	Achieve	176

45	Internal	255	76	Prior	172
46	Individual	252	77	Identify	168
47	Whereas	248	78	Require	164
48	Previous	246	79	Detect	163
49	Network	245	80	Medium	163
50	Volume	245	81	Device	162
51	Component	241	82	Stress	158
52	Constants	239	83	Cycle	155
53	Dynamic	238	84	Alternative	152
54	Period	231	85	Appropriate	146
55	Stable	224	86	Strategy	141
56	Hence	222	87	Unique	136
57	Series	219	88	Focus	134
58	Dimension	213	89	Generate	133
59	Positive	213	90	Sufficient	129
60	Parameter	212	91	Feature	125
61	Access	211	92	Rigid	120
Rank	**Word**	**Frequency**	**Rank**	**Word**	**Frequency**
93	Exhibit	119	124	Distinct	71
94	Investigate	115	125	Sphere	70
95	Despite	113	126	Element	69
96	Demonstrate	112	127	Obvious	69
97	Technology	111	128	Apparent	68
98	Affect	103	129	Locate	68
99	Estimate	103	130	Construct	67
100	Principal	103	131	Author	64
101	Uniform	103	132	Induce	63
102	Issue	102	133	Interact	63
103	Correspond	97	134	React	61
104	Identical	96	135	Summary	61
105	Bulk	95	136	Monitor	60
106	Thereby	94	137	Code	60
107	Minimum	93	138	Concept	59
108	Normal	93	139	Confirm	59
109	Index	92	140	Capable	58
110	Random	92	141	Extract	58
111	Enhance	91	142	Facilitate	55
112	Challenge	89	143	Compatible	54
113	Contact	88	144	Intense	54
114	Remove	87	145	Visible	54
115	Enable	85	146	Convert	52
116	Media	84	147	Create	51
117	Crucial	81	148	Assess	50
118	Label	79	149	Evident	50
119	Evaluate	78	150	Approximate	48
120	Transform	78	151	Maintain	48
121	Ensure	76	152	Formula	47
122	Intrinsic	74	153	Framework	47
123	Reveal	72	154	Display	45

Rank	Word	Frequency	Rank	Word	Frequency
155	Reverse	45	159	Select	40
156	Proportion	44	160	Media	40
157	Vary	44	161	Contribute	40
158	Rely	43	162	Phenomenon	40

From the word selection criteria of high-frequency used AWL words in this study, the framework of Coxhead (2000) was applied. The words have to occur at least 40 times in the entire corpus and occur at least 4 times in each sub field.

From table 2, it is clear that there are 162 academic words in the corpus of chemistry research articles that meet the word selection criteria. The words that occur with the highest frequency are “structure”, “data”, “chemical”, “method”, “complex”, and “energy”, respectively.

However, there were some words that did not meet the word selection criteria because they were not covered in the seven sub fields of chemistry but they should not be ignored because they occurred with high frequency in the corpus. These were words such as “sensitivity”, “progress”, “space”, “plant” and so on.

Research Question 3: to what extent are the technical chemistry words used in the chemistry research articles corpus?

From the analysis, it was found that the technical vocabulary was derived from three group of words (GSL, AWL, and Off lists). It is surprising that some general word such as “exciting” can have specific meaning in chemistry field. The word “exciting” can be found in contexts such as this example from the chemistry corpus, “are relatively inefficient in *exciting* the vibrational and rotational modes.” Moreover, this word, when it occurs with other words, can have different meanings and contain more detail about chemistry, such as “photo exciting”.

The results revealed that there were 649 technical chemistry words that can be classified into two main groups by meaning: 1) academic chemistry words, or words used to describe results, principle of theory, machines or techniques related to chemistry and 2) technical words used in the field of chemistry, such as chemical substance or compoundsThese words are rarely found in other fields.

Examples of the first group of technical vocabulary are “spectral”, “magnetics”, “spectroscopy”, and “resonance” which both appeared with high frequency in the corpus and appeared in more than four out of seven sub fields of chemistry research articles. Eamples of the second groups of technical vocabulary are “carbon”, “oxygen”, “butanol”, “oxide”, and“sodium”.

Furthermore, it was found that about 96% of technical words were academic chemistry words used to describe things like results, principle of theory, machine or techniques related to chemistry. The results describing the proportions of words confirmed that teachers should focus on teaching these types of vocabulary more than

focusing only on chemical substances. These words may not appear with high frequency and students may not have ample chances to encounter them.

Moreover, the study of technical vocabulary reveals that the real proportion of technical vocabulary in the corpus of chemistry research articles is about 7%, whereas the percentage of other words or words with low frequency in the corpus is about 21.3%. The reason for this high number is that there are many words that appear with low frequency or it can be assumed that it depends on the nature of the research articles, the reason being this data was brought from several chemistry research articles with various topics of study.

Examples of low frequency words found in the corpus include “B-hydroxysteroid”, “terphynyl”, “B-cyclodextrin” and other specific jargon used only in research articles.

To summarize, it can be said that the research instrument used to identify technical vocabulary or rating scale can still reliably validate different field of technical vocabulary.

Conclusion

In this study, it is clear that some general words can be considered technical chemistry vocabulary. As mentioned before, the proportion of GSL words in this study is 62.5% which is lower than the first results, which were identified by only using the RANGE program.

Moreover, it was found that the number of high frequency AWL words in the whole corpus is 162. After the analysis of Academic words bythe RANGE Program, the results of the study revealed that the proportion of academic words is around 10.5%. After this, the process to identify technical vocabulary showed that the proportion of AWL in this corpus was just 9.2%.

The total technical vocabulary in the chemistry research articles corpus was 649 distinct words or 78,641 running words which is approximately 7% of the entire corpus.

Implication of the Study

According to an analysis of vocabulary in the corpus of chemistry research articles bythe RANGE program, words were shown as “type” forms, for example, state, states, yielded, and yielding. However, more attention should be paid to the collapse of types into concerned word families when applying vocabulary for pedagogical purposes.

In addition, the results of the present study of the chemistry research articles corpus can be used for ESP pedagogy in the field of chemistry. The present study provided vocabulary of high frequency used in the chemistry corpus. Apart from knowing the popular vocabulary lists, teachers should raise awareness of vocabulary from the GSL and AWLwhich have more than one meaning in their specific field.

There are two main limitations in this study: the limitation of research articles and the limitation of single words. The data was drawn from chemistry research articles, so

results can vary from different sub fields of chemistry because of the research trends in each sub field during the given time period. One suggestion for future research about analyzing the vocabulary in a chemistry corpus is that researchers should collect the data from textbooks from which students learn so that the data can provide different results.

As mentioned earlier, this study is mainly focused on single words. However, it may have overlooked some technical words formed by more than one word, AKA compound words. A suggestion for further study is to study collocations of technical words or noun phrases to help students attain a more dynamic sort of knowledge and to help students have more advanced comprehension of academic texts.

References

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

Chung, T. M. and Nation, P. (2004). Identify technical vocabulary. *System, 32*, 251-263.

Coxhead, A. (2000). A New Academic Word List. *TESOL QUATERY, (34)*2, 213-218.

Coxhead, A. and Nation, P. (2001). The specialized vocabulary of English for specific purposes. In J. Flowerdew and M. Peacock (Eds.), *Research Perspectives on English for Academic Purposes* (pp. 252-267). Cambridge: Cambridge University Press.

Lessard-Clouston, M. (2010). Theology lectures as lexical environments: A case study of technical vocabulary use. *Journal of English for Academic Purposes, 9*, 308-321.

Lewis, M. (1993). *The lexical approach: The states of ELT and a way forward*. Hove, England: Language Teaching Publications.

Martinez, I., Beck, S., Panza, C. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*, 183-198.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, et al. (Eds.), *Performance& Competence in second language acquisition* (pp. 37). Cambridge: Cambridge University Press.

Mozaffari, A. and Moini, R. (2014). Academic Words in Education Research Articles: A Corpus-Based Study. *Procedia Social and Behavioral Science, 98*, 1290-1296.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*(2), 235-256.

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

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Downloadable from http://www.victoria.ac.nz/lals/about/staff/paul-nation.

Schmitt, D., and Schmitt, N. (2005). *Focus on vocabulary: Mastering the Academic Word List*. White Plains, NY: Pearson Education.

Strevens, P. (1973). Technical, technological, and scientific English. *ELT Journal, 27*, 223–234.

Swales, J.M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Valipouri, L., Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*, 248-263.

Vongpumivitch, V. Huang, J & Chang, Y. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, 28, 33-41.

Wasuntarasophit, S. (2008). *Technical and academic vocabulary in electrical engineering textbooks*. Nakhon Ratchasima, Thailand: Institute of Social Technology Suranaree University of Technology.

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

Wilkins, D. (1972). *Linguistics in language teaching.* London: Arnold.

Xue, G. & Nation, I.S.P. (1984). A university word list. *Language learning and communication, 3*(2), 215-229.

Contact email: camera_14@hotmail.com