Corpus-based Data for Determining the Vocabulary Found in the ICLH Research Articles Written by Myanmar Researchers

Sandar Htay*

Abstract

This study explores the use of the lexical words in the academic articles written by the Myanmar scholars in the English language. The lexical words, the vocabulary, is regarded as one of the main features in academic writing. The purpose of the study is to develop a list of the most frequent vocabulary used by the Myanmar researchers in the field of language teaching and social sciences and compared it with the lists of West's (1953) General Service (GSL) and Coxhead's (2000) Academic Word Lists (AWL). The corpus approximately 170,000 running words was build using the 39 academic research articles, written by Myanmar scholars, presented at the 1st International Conference on Languages and Humanities (ICLH) which was hold at Yangon University of Foreign Languages in 2020. The lists of the frequent words found in respective levels of GSL and AWL were developed. The analysis revealed that the top 20 frequent words included 18 basic words and only two academic words. The results also show that, amongst 2000 word families of GSL, 731 word families occurred frequently in the articles with the coverage of 79% of the tokens in the corpus. For academic words, among 570, only 96 word families with the coverage of 9 % of the all tokens were found as the frequent words in the research articles. It is hoped that the construction of the lists will definitely provide the Myanmar researchers the knowledge on the use of vocabulary in academic writing of the research articles in the English language.

Keywords: Corpus-based Lexical Analysis, Academic Words, General Service Words, Myanmar researchers, Language and Humanities

I. INTRODUCTION

Research articles are one of the platforms for the exchange of new knowledge and new discoveries between academic and speech community. In this way, it is not surprised that research has become a standard marker of a scholastic establishment. In many institutions, it is required for the researchers to complete the investigation so as to share their discoveries and results to the others. Whilst sharing their knowledge and findings in the English language, somewhat they may have the troubles in picking or utilizing the suitable words. Therefore, it is believed that the study of the language used in the article has revealed many valuable results, which are helpful to enhance the development of the researcher's knowledge of the writing of the research article.

It is also recorded that vocabulary is significant for expressing and sharing knowledge by the users, the writers. What's more, a large amount of vocabulary can make it effectively simpler for readers to comprehend texts and dialogues, and more effectively use various technical skills in English (Meara, 1996). As pointed out by Coady (1997), Donley & Reppen (2001) and Nation (2001), Vocabulary has also played a major role in one's success not only in learning a language but in academic lives. Koda (2005, 2007) and Laufer (1997, 2003) stated that without sufficient vocabulary, the one might not comprehend the texts or express their thoughts successfully. Therefore, these findings led to the idea of developing the vocabulary most commonly used by researchers in their research.

It is not surprising that the use of the academic vocabulary has become the prominent factor in writing the English written research. As indicated by Coxhead (2001), the academic vocabulary are the words which frequently occurred in academic works and texts, yet rarely

* Lecturer, Department of Linguistice, Yangon University of Foreign Languages

found in other text types. The other words that appeared frequently in all the other texts were identified as General Service words which the "greatest general" use of English language users (West, 1953).

Furthermore, there here have been increasing number of researches conducted by the native as well as non-native researchers. For the specific, it was interested to study the academic vocabulary and general service words used frequently by the Myanmar researchers in the field of language teaching and social sciences as the researcher is in the field of language teaching and linguistics. So, the particular corpus has built using the articles published in the conference proceeding of the 1st International Conference on Languages and Social Sciences held at Yangon University of Foreign Languages.

On the other hand, there has been no such kind of research which studied the use of language of the Myanmar researchers. As the contribution to this filed, this study aimed to develop a list of the vocabulary used frequently by the Myanmar researchers in the field of language teaching and social sciences and compared it with the lists of West's (1953) General Service (GSL) and Coxhead's (2000) Academic Word Lists (AWL). It is hoped that the results, the vocabulary lists, would be worthwhile for the researchers in order to develop their research article writing ability in the English language since it has been driven by the needs of the researchers.

II. LITERATURE REVIEW

Academic Vocabulary is one of the Nation's (2001) four levels of vocabulary division which consists of high frequency words, academic vocabulary, technical and low frequency words. Coxhead (2000) developed the Academic Word List (AWL) by analysing and using the 3.5 million corpus of written academic English texts in four prominent areas: Arts, Commerce, Law and Science. In addition, West (1953) created the General Service List (GSL) referenced to the high frequency words of Nation (2001). The GSL includes 2000 most frequent word families, but the Academic Word List (AWL) constructed by Coxhead (2000) comprises 570 word families. Nation and Waring (1997) recommended that those who are in the academic context have to know the first 2000 words in GSL followed by the large amount of academic vocabulary which appeared in almost all the academic disciplines.

There have been numerous studies that developed the academic vocabulary list for the academic disciplines in different context. Campion et al. (1971) built up the lists of most frequent vocabulary used by the university students from different specializations. As mention in the previous part, Coxhead (2000) built the AWL which incorporates 3112 tokens barring the GSL vocabulary of West (1953). In view of thoughts of the previous scholars, in sociology field, Kwary and Arta (2017) determine the academic-article vocabulary list and in medical field, Lei Lei at al. (2016) build an academic vocabulary list comparing the specialized word lists. Wang (2008) additionally made a list of vocabulary used in the medical field which includes non-GSL or AWL, yet they are the most frequent words found in the research articles. In addition, Chen et al. (2007) called attention to in their investigation that only around 50% of the AWL were found in the medical research articles. Valipori et al. (2013) build the AWL list in science. They stated that there were 1400 word families in the 400,000 running words in the science research articles. However, there has been no such study conducted using the Myanmar scholars' research articles.

In the present study, it is aimed to develop the lists of the frequent vocabulary used by the Myanmar researchers and determine the AWL and GSL in the present corpus by comparing the standardized AWL and GSL. The academic vocabulary and general service words used

frequently by the Myanmar researchers in the field of language teaching and social sciences would be investigated by using corpus-based approach.

Research Questions

Based on the review of the previous studies and practical needs for the particular context, for this corpus-based study, the research questions have been set as follows:

1. What are the most frequent words used in the ICLH research articles written by Myanmar scholars?
2. What are the most frequent academic words and general service words in a corpus of ICLH research articles?
3. Are there any differences of English research in the academic writing between Myanmar scholars and the others in terms of the frequency of the general service words (i.e. basic words) and academic words?

III. METHODOLOGY

This study is a non-experimental research which utilized corpus-based strategies to assess the data on the words used in the studies in ICLH proceedings.

The Corpus

A corpus is a group where language texts are gathered in. It gives a large number of occurrences of language, which vary in purposes. The texts assembled in the corpus are either written or spoken, but transcribed the spoken in written form. There are two kinds of corpus: sample and monitor. The sample corpus, which is also called the balanced corpus, is built to present a particular kind of language at a specific timeframe. The sample corpus contains a restricted measure of the texts. After it is assembled, it will not be amended or changed. Rather than the sample corpus, the monitor corpus is dynamic as its size continues growing.

The particular corpus in this study was created as the sample corpus which was aimed to represent the frequent vocabulary found in the Proceeding papers of the 1st International Conference on Languages and Humanities hosted by Yangon University of Foreign Languages. The corpus would be named as the International Conference on Languages and Humanities Corpus (hereby ICLHC). In this proceeding, the three main themes were included: Oriental Languages, European Languages and Humanities for Diverse Society.

In the present corpus, the samples were perused from the selected research articles of the three themes. It was structured after the models proposed by Sinclair (1991, 2005) and Barnbrook (1996), considering "representativeness, balanced, specificity of corpus, use of whole documents, and availability in electronic form". The articles from all the three themes were chosen as the themes cover the whole main theme of the conference, languages and humanities. In terms of the articles written by the Myanmar scholars, the 9 articles from the theme Oriental Languages; 13 from the European Languages; and 17 from Humanities for Diverse Society were chosen. So, the numbers of the articles seemed not to be much differences. These would be the evidence for the representative and balance which have been the important features in the corpus-based research. In addition, the ICLHC also obtained the specificity of the corpus since the articles which were aligned with the main theme, language and humanities. Moreover, as reviewed by the editorial board of the conference proceeding,

most of the articles would be included in the corpus had "identifiable Abstract, Introduction, Method, Result and Discussion sections (IMRD)" Swale (1990). So, a total of 39 articles had chosen for the ICLHC. The final version of the ICLHC contains around 169,362 running words. As the study focuses on one single register, the language used in the three areas in the proceeding of ICLH, the size of the corpus seems to be adequate for its purpose. The Details of the ICLHC would be seen in Table 1.

Table 1

Description of the ICLHC

Corpus	Number of Articles	Number of Tokens
ICLHC	39	169,362

Instrumentations

The instruments used in this study were Lexical Frequency program by Heatley et al. (2002) and AntWordProfiler software by Anthony, L. (2014). The first one was used to create the frequency list of the ICLHC and the second software was used to compares the files with a set of vocabulary level lists of GSW and ASL. The software is quite modernized and can create the lexical statistical analysis and frequency lists of the corpus. Though the software can give the frequency information, the data given by the Lexical Frequency program would be much better to see the results readably. As both of the programs work with the plain text format, the electronic data gathered could be smoothly processed.

Procedures of Analysis

Firstly, all the articles were collected in their electronic version as the word files. The researcher removed "the reference lists, appendices, captions, footnotes, and acknowledgments" (Swales, 1990). Then, the modified electronic version was converted into the plain text format. Firstly, since it had been needed to create the frequency list, the data, kept in plaintext format, were uploaded into Lexical Frequency program. After receiving a command, the software processed the data. Then, all the types were ranged in terms of frequency occurred in the text accompanied with the tokens for each type. Finally, the results found were manually noted to response to the first research question, the most frequent used words in a corpus of ICLH research articles. In investigating the most frequent academic words and general service words in ICLHC, the Antwordprofiler was used to compare the ICLHC with the West's GSL two levels and Coxhead's AWL lists to categorize the words and find the frequency range of each word in the ICLHC. This can give the response to the second research question what the academic and basic words were most frequently used in the articles. Finally, the researcher compared the results in the study with the previous ones done by the other scholars to examine the research paper writing in academic English between Myanmar scholars and the others in terms of the frequency of the general service words (i.e. basic words) and academic words.

IV. RESULTS AND DISCUSSION

The most frequent words in the ICLHC

The ICLHC corpus comprises 885 word families with 169,362 running words. The most frequent words found in the corpus would be defined based on the criteria of range and

frequency of Coxhead (2000). A word which appeared in the corpus more than 50 times and occurred 5 times in the sub-corpora of the whole corpus were determined as the most frequent words in the ICLHC.

Among the 996 word families in the corpus, 1260 met the criteria. However, 82 were the technical word families and 29 were the numbers and figures. So they were excluded in the list. So 885 word families, which comprises 68488 token, were included in the most frequent vocabulary list found in the ICLHC. The top 20 word families found in the ICLHC can be seen in Table 2.

Table 2

The top 20 most frequent words in the ICLHC

Rank	Word	Raw Frequency (Tokens)	% of the ICLHC
1	language	4562	2.69
2	develop	4258	2.51
3	learner	4121	2.43
4	word	4072	2.40
5	student	3987	2.35
6	teacher	3939	2.33
7	find	3886	2.29
8	describe	3871	2.29
9	consist	3806	2.25
10	mean	3728	2.20
11	different	3665	2.16
12	write	3420	2.02
13	study	3353	1.98
14	focus	3342	1.97
15	consider	3213	1.90
16	show	2954	1.74
17	analyse*	2552	1.51
18	understand	2547	1.50
19	society	2375	1.40
20	participate*	2331	1.38
Total		*69982*	*41.3*

* The word “analyse” and “participate” have been included in Coxhead’s AWL. The other words are under the 1st Level GSL.

This study was aimed to investigate the most frequent words found in ICLHC. Regarding the first research question, altogether, the 69982 of 169,362 tokens, a coverage of 41.3 % of the ICLHC were found as the 20 most frequent words in the corpus. The present

frequent words included the words occurred in GSL or AWL, as previous researchers (Billuroglu et al, 2005 and Ward, 2009) stated that AWL words oppose the eliminating the academic words in the frequent word list. As indicated by the data presented in the list, it was found out that, amongst the most frequent 20 word families found in the ICLHC, only two words, "analyse" and "participate" included the AWL of Coxhead (2000). The other words are under the GSL of West (1953). The word "language" placed as the top role and it is not surprising that the corpus was constructed over the articles in the fields of language and humanities. It can also be said that the Myanmar researchers frequently use the basic words, yet only a few academic words have been used in their academic writing.

General Service Words and Academic Words found in the ICLHC

Based on the data obtained from comparing with the standardized vocabulary lists of AWL and GSL, it was found that the 420 word families with 110,459 tokens (65% of the 169,362 running words in the ICLHC) would be occurred in the first level GSL list (i.e. within the first most frequent 1000 word families). Within the second most frequent 1000 word families in GSL, altogether 311 word families with 32,891 tokens represent the 19% of the whole ICLHC. As for the Academic Words, of the 570 word families in standardized list, 96 word families containing 14,787 frequency tokens occurred in the ICLHC. It was only 9% of the ICLHC. In addition, the 58 word families with 11,225 tokens (7% of the ICLHC) were defined as the non-GSL/AWL words. The results can be seen in Table 3 and it was also illustrated in Figure 1.

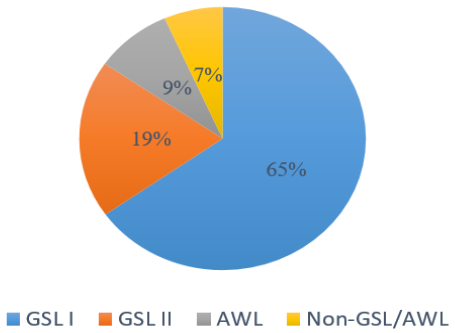
Table 3

The coverage of different word lists over ICLHC

Word Lists	Word Families	Tokens	% of the ICLHC
GSL 1st Level	420	110459	65
GSL 2nd Level	311	32891	19
AWL	96	14787	9
Non-GSL/AWL	58	11225	7
Total	885	169362	100

Figure 1

The coverage of the different vocabulary lists found in ICLHC



To response to the second question, the most frequent academic words and general service words used by Myanmar researchers in a corpus of ICLH research articles were investigated and listed. By looking at the data presented in Table 3, of 2000 GSL, only 731

word families occurred with high frequency in the present corpus. However, it could be said that the Myanmar scholars over use the words appeared in 1st Level GSL, indicated by 65% and in the 2nd Level GSL, 9% of the whole ICLHC. Only 9% of the whole ICLHC occurred as the academic words and the other 7% were in the non GSL/AWL. A comparison of the words with standardized lists showed that not many of AWL words were frequently used, so the word used were have high frequency. In addition, the Myanmar researcher have used much vocabulary only on the basic words appeared in the GSL of West (1953). Moreover, some words excluded in the GSL or AWL were also found in the corpus. It was shown in the following table, Table 4.

Over ICLHC, the 30 most frequent words in each different word list compared with the standardized lists can be seen in Table 4. In this table, the lists of the words which are non-GSL/AWL were also inserted.

Rank	GSL 1st Level		GSL 2nd level		AWL		Non-GSL/AWL	
	Word	**Tokens**	**Word**	**Tokens**	**Word**	**Tokens**	**Word**	**Tokens**
1	language	4562	skill	2312	analyse	2552	linguistic	210
2	develop	4258	sentence	2013	participate	2331	humanities	160
3	learner	4121	express	812	approach	1021	adverbs	135
4	word	4072	compare	462	structure	823	oriental	131
5	student	3987	perform	452	research	612	sociolinguistics	131
6	teacher	3939	collect	452	process	598	morphemes	122
7	find	3886	purpose	442	survey	551	reduplication	122
8	describe	3871	improve	398	data	436	affixation	122
9	consist	3806	apply	391	similar	383	morpheme	121
10	mean	3728	discuss	373	design	339	grammatical	120
11	different	3665	opportunity	364	method	278	vocabulary	116
12	write	3420	material	345	classify	205	affixes	110
13	study	3353	believe	329	feature	198	lexical	106
14	focus	3342	combine	327	available	161	usage	102
15	consider	3213	examine	324	concept	122	suffix	97
16	show	2954	recommend	321	significant	104	modifier	93
17	understand	2547	measure	311	individual	92	syllable	91
18	society	2375	opposite	231	item	95	loanwords	89
19	discuss	2341	distinguish	226	indicate	86	prefix	89
20	introduce	2314	frequent	215	intensive	86	multilingual	85
21	include	2282	inform	214	source	86	rhyme	85
22	foreign	2243	necessary	213	function	85	reduplicated	83
23	class	2236	international	212	culture	84	derivational	79
24	relation	2213	argue	211	outcome	84	prestige	78
25	same	1432	intend	210	randomly	69	semantic	75

Rank	GSL 1st Level		GSL 2nd level		AWL		Non-GSL/AWL	
	Word	**Tokens**	**Word**	**Tokens**	**Word**	**Tokens**	**Word**	**Tokens**
26	produce	1347	altogether	203	separately	68	supernatural	73
27	follow	1264	especial	125	category	68	legend	71
28	change	1253	instrument	123	positive	66	civilization	71
29	form	1243	comprise	109	emphasis	64	mood	61
30	give	1145	framework	78	constitute	62	slot	59
Total		86412		12798		11809		3087

Regarding the third question, it was found that there would be the academic writing of Myanmar scholars in English-written research would be different from other scholars in terms of the frequency of the general service words (i.e. basic words) and academic words. Chen et al. (2007) found that 50% of the AWL were found in the medical research articles, but in the present study, it was found that only 9% of the whole corpus in language and social science field were found as the academic words. The result of the present study seemed to be aligned with the findings of Valipori et al. (2013). In their study, it was found that only around 11% were found as academic words in their study of the physiotherapy research works. So, it could be said that, despite the differences in terms of the frequency of the words, the Myanmar researchers can use the academic words in their writing the research articles. They could have the enough vocabulary occurred in GSL. However, it is need for them to develop their knowledge on academic vocabulary as the percentage was quite low.

V. CONCLUSION

The present study developed the lists of vocabulary used by the Myanmar researchers in their academic-articles in the field of linguistics and social sciences. It was found that the 20 most frequent words listed above comprises 18 basic words and only 2 academic words. So, it can be said that the Myanmar scholars could have the much ability in using the general service words than the academic words. It might be because of their knowledge on vocabulary or their English proficiency level. In addition, it could be said that, by comparing the work of the researchers in other contexts, the use of the academic words by the Myanmar scholars was quite low (determined by 9% of the ICLHC). In the same way, based on the data, the coverage of the words in ICLHC, it could be stated that Myanmar scholars used a large amount of basis words (indicated by 84% of the whole corpus) and they needed to develop their writing in terms of using the academic words as well as technical words in their specific fields.

In short, the empirical study attempted to develop the most frequent word lists, including basic words and academic words, found in the academic articles written by the Myanmar scholar in the field of linguistics and social sciences. However, there were some limitations in the study, the corpus should be built using the research articles written by Myanmar researchers in the various disciplines. Yet, according to the time constraint, the researcher listed the frequent words used in the articles. In future studies, it would be better to include the analysis of collocation occurred in the corpus in order to help the researchers develop their knowledge of vocabulary. Except some limitations mention above, the results of the present study can definitely provide the Myanmar researchers with strong evidence and beneficial implication in the development of their knowledge of the writing of the research article in terms of the use of the vocabulary in their academic writing.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Professor Dr. Kyi Shwin, Rector of Yangon University of Foreign Languages, and the Pro-rector, Dr. Mi Aung for giving me a chance to study corpus linguistic analysis. I would like to give my gratitude to Professor Dr. Yin Myo Tint, Head of the Linguistics Department, Yangon University of Foreign Languages for her encouragement and her suggestion in doing this study. Next, I would like to express my thanks to Prof. Dr. Khin Thida, Head History Department and Daw San San Cho, Assistant Director of the Computer Department, Yangon University of Foreign Languages, for their supports in collecting the data for this study. My special thanks to due to the Assistant Professor, Dr Sukhum Wasuntarasophit, an assistant professor, faculty of humanities and social sciences, Khon Kaen University, Thailand for sharing me his invaluable knowledge on corpus linguistics during his lectures. Last but not least, I am grateful to the researchers presented at ICLH as I had a chance for analyzing the vocabulary used by the Myanmar researchers in their academic research articles.

REFERENCES

Anthony, L. (2013). AntWordProfiler (version 1.40) [computer program software]. Tokyo, Japan: Waseda University. Available from: http://www.laurenceanthony.net/ software.

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

Coxhead, A. et al. (2001). The specialized vocabulary of English for academic purposes. *Research perspectives on English for academic purposes*, 252-267.

Coxhead, A. (2017). Vocabulary and English for Specific Purposes Research: Quantitative and Qualitative Perspectives. Routledge.

Donley, K., & Reppen, R. (2001). Using corpus tools to highlight academic vocabulary in sustained content language teaching. *TESOL Journal*, *10*, 7-12.

Dudley-Evans, T. et al. (2011). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

Farrell, P. (1990). Vocabulary in ESP: A lexical analysis of the English for electronics and a study of semi-technical vocabulary. *Centre for Language and Communication Studies CLCS Occasional Paper, 25.* Dublin: Trinity College.

Jamalzadeh, M. et al. (2019). A Corpus-based Study of Academic Vocabulary in Psysiotherapy Research Articles. *Language teaching Research Quaterly*, *9*, 69-82.

Khamis, N. et al. (2020). Corpus-based Data for Determining Specialized Language Features. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(1), 36-41.

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

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston: Heinle Cengage Learning.

Nation, I. S. P. et al. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge : Cambridge University Press.

Tongpoon-Patanasorn, A. (2017). Constructing an academic word list of business English: A corpus-based approach. *Humanities & Social Sciences, 34*(2), 1-31. Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical word list.

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

Zi, W. (2017). Corpus-based Research on Lexical Features of English Abstract in Postgraduates' Thesis of Fashion Majors. *International Journal of English Language Education,* 5(2), 141-149.