International Journal of English Linguistics; Vol. 8, No. 4; 2018
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Academic Vocabulary Use in Doctoral Theses: A Corpus-Based Lexical Analysis of Academic Word List (AWL) in Major Scientific Disciplinary Groups

Habibullah Pathan[1], Rafique A. Memon[2], Shumaila Memon[3], Syed Waqar Ali Shah[4] & Aziz Magsi[5]

[1] Mehran University of Engineering & Technology, Sindh, Pakistan

[2] Institute of English Language & Literature, University of Sindh, Sindh, Pakistan

[3] Institute of English Language & Literature, University of Sindh, Sindh, Pakistan

[4] Mehran University of Engineering & Technology, Sindh, Pakistan

[5] Mehran University of Engineering & Technology, Sindh, Pakistan

Correspondence: Habibullah Pathan, Mehran University of Engineering & Technology, Sindh, Pakistan.

Received: November 1, 2017 Accepted: April 9, 2018 Online Published: May 2, 2018

doi:10.5539/ijel.v8n4p282 URL: https://doi.org/10.5539/ijel.v8n4p282

Abstract

Since the development of academic word list (AWL) by Coxhead (2000), multiple studies have attempted to investigate its effectiveness and relevance of the included academic vocabulary in the texts or corpora of various academic fields, disciplines, subjects and also in multiple academic genres and registers. Similarly, this study also aims at investigating the text coverage of Coxhead's (2000) AWL in Pakistani doctoral theses of two major scientific disciplinary groups (Biological & health sciences as well as Physical sciences); furthermore the study also analyses the frequency of the AWL word families to extract the most frequent word families in the theses texts. In order to achieve this goal, a pre-built corpus of Pakistani doctoral theses (PAKDTh) (Aziz, 2016) comprises of 200 doctoral theses from two major scientific disciplinary groups was used as textual data. Using concordance software AntConc version 3.4.4 (Anthony, 2016), computer-driven data analysis revealed that in total 8.76% (496839 words) of the text in Pakistani doctoral thesis corpus is covered by the AWL words. Further distributing the analysis per sub-lists, shows that the first three sub-lists of AWL accounted for almost 57% of the whole text coverage. An attempt was made to further analyze the AWL text coverage by considering the frequency of occurrences in terms of word families. The findings showed that among 570- word families of Coxhead's (2000) AWL, 550-word families with the sum of 96.49% are found to occur more than 10 times in PAKDTh corpus, which are taken as word families used in the corpus. This study concludes that Coxhead's (2000) AWL is proved effective for the writing of theses. On the basis of the findings, further possible academic implications are discussed in detail.

Keywords: academic vocabulary, AWL, doctoral thesis, corpus-based, scientific disciplines, word frequency

1. Introduction

Pakistani students, being non-native speakers of English and belonging to ESL context, at all academic levels are assumed to have very limited vocabulary knowledge which might be a factor influencing their proficiency in academic discourse (Mozaffari & Moini, 2014). There might have been less focus on teaching vocabulary by language teachers which could possibly be the main factor for students' lack vocabulary knowledge. Coady (1997) reports that these kinds of practices by ESL teachers are only because of the traditional language teaching practices (with negligence of vocabulary) which they have experienced during their earlier learning period. Similarly, According to Macaro (2003), language teachers from ESL context often neglect this area (vocabulary) of language and they must be provided proper research-based practices to incorporate the component of vocabulary in their teaching. Another most important factor is learning resources and curriculum (Fan, 2003; Warsi, 2004) which hinder language teachers to do so.

Despite all the facts, students learning English also find vocabulary as one of the most important areas to be achieved. Leki & Carson's (1994) survey also provided evidence for students' serious attitude towards vocabulary learning. As students advance to upper academic levels, by the expansion of more subjects and

textbooks they experience more vocabulary (Nagy & Anderson, 1984; Stahl, 1998; Schmitt, 2000; Biemiller, 2005; Stahl, 2005) feel themselves surrounded by a vast variety of texts containing specific vocabulary. Thus, they mainly focus on learning the vocabulary which is specified and specialized to their courses and subjects. Hence, it is suggested by Nation "… to direct vocabulary learning to more specialized areas when learners have mastered the 2000...3000 words of general usefulness in English" (2001, p. 187); but it is not always effective for learners whether they are native or non-native speakers of English. There might be chances for students to master vocabulary in general or specifically but become less acquainted with the vocabulary that they may require for better achievement in academics and for the effective understanding of academic discourse at higher levels. Subsequently, academic vocabulary, occurring less frequently than general vocabulary items (Worthington & Nation, 1996; Xue & Nation, 1984), seemed difficult for learners (Cohen et al., 1988) because of their more familiarity with technical or specialized vocabulary in comparison with that of academic. Thus, it is very crucial to take academic vocabulary development into consideration while teaching English to the students at any academic level (such as primary, elementary, secondary, higher secondary or tertiary).

The multitudinous advancement in technology and its role in linguistics cannot be unappreciated. So that, the recent development of corpus linguistic research, particularly in English for specific purposes (EAP) and English for academic purposes (EAP) are widespread. ESP or EAP practitioners and researchers find it the most valuable in linguistic research which helps them develop better and explicit knowledge about language. Corpus linguistics is generally defined as an approach and research method in linguistics rather than a branch of linguistics which empirically examines natural languages through corpus-based techniques using computers (McEnery & Wilson, 2001). The use of corpus linguistics is also widespread in vocabulary research studies. Coxhead (2000) used an academic corpus of 3.5 million words and attempted to construct the list of the academic word list (AWL).

The current study aims at investigating the use of AWL words in Pakistani doctoral thesis. It also attempts to examine AWL words use distinctively between two major disciplinary groups of Engineering & Technological, Biological, and Health Sciences.

The current paper attempts to answer the question given below:

1) What is the text coverage of AWL words in the corpus of Pakistani doctoral theses (PAKDTh)?
2) What are word families of AWLfrequently used in Pakistani doctoral theses?

2. Review of Literature

The corpus, compiled by Coxhead (2000) for making AWL, comprises texts from diverse academic sources (such as academic journal articles, academic web articles, textbooks, course books, scientific texts and laboratory manuals) of 28 subject areas from four major academic disciplines of arts, law, science, and commerce. Since the development of AWL, various research attempts have been made to determine its effectiveness across various academic fields and disciplines. The AWL contains 570-word families in total. The word families are referred to the root word which has different word forms such as assume, assumed, assumes, assuming, assumption and assumptions.

The review of the literature indicates that there has been fewer research studies focusing the use of AWL in particular disciplines and fields. Such as, Chung & Nation (2003) comparatively studied the use of Coxhead's (2000) AWL & West's (1953) General service list (GSL) in applied linguistics and anatomy books, Mudraya (2006) analysed AWL use in Student Engineering English corpus of 2 million words, Chen & GE (2007) did a lexical analysis on medical research papers corpus, Vongpumivitch et al. (2009) analysed the frequency of Coxhead's (2000) AWL word families in the corpus of Applied linguistics research articles, and Martinez (2009) critically investigated Coxhead's (2000) AWL word families in agriculture research articles.

The importance and effectiveness of Coxhead's (2000) AWL have been brought under discussion by the various researchers (e.g., Chung & Nation, 2003; Mudraya, 2006; Chen & Ge, 2007; Vongpumivitch et al., 2009). While some studies (Martinez, 2009; Mozaffari & Moini, 2014) do not consider the AWL as an effective source in terms of text coverage in the specific fields and disciplines. This study mainly tries to explore the usefulness of the AWL word forms in education research papers as well as an attempt is made to extract the non-AWL words frequently appear in education research. To this end, the relatively large corpus of education research papers was compiled.

This study attempts to analyze the use of academic vocabulary through the AWL in the doctoral theses of two distinct scientific disciplinary groups and also tries to extract the most frequently used AWL word families in the theses. In order to achieve this aim, a pre-built corpus of Pakistani Doctoral theses (PAKDTh) is used as textual data representing the written English language of theses as an academic genre and a non-native variety of written

academic English.

3. Pakistani Doctoral Thesis (PAKDTh) Corpus

An existing corpus PAKDTh (Aziz, 2016) comprises of 200 texts of Pakistani doctoral thesis from 17 disciplines categorized into two sub-corpora of major disciplinary groups PHSc and BHSc. PAKDTh contains 200 theses, 100 from each group. The size of PAKDTh corpus is approximately 5.6 million words. The exact number of words and disciplines included in the corpus are shown in the table given below:

Table 1. PAKDTh corpus description

Disciplinary Group	Discipline	Number of Theses	Tokens (words)
Physical Sciences	Chemistry	46	1,460,924
	Earth Sciences	9	217,610
	Mathematics	5	130,340
	Physics	40	940,504
		Σ 100	**Σ 2,749,378**
Biological and Health Sciences	Applied Biological Sciences	2	48,875
	Biochemistry	1	18,048
	Biotechnology	7	258,202
	Clinical Medicine	8	162,297
	Health Sciences	2	114,695
	Human Physiology	1	20,678
	Microbiology	15	385,539
	Molecular Biology	15	371,971
	Pathology	3	86,120
	Pharmacy	17	569,131
	Plant Sciences	8	248,625
	Zoological Sciences	5	134,197
	Biological Sciences	16	504,059
		Σ 100	**2,922,437**

Note. Adapted from "Linguistic Variation across Major Disciplinary Groups of Pakistani Academic Writing: Multidimensional Analysis of Doctoral Theses" by Aziz, Pathan, & Ali (2016), ARIEL-An International Research Journal of English Language and Literature, 27, 27-60.

4. Data Analysis

Coxhead's (2000) AWL word families, which are categorized into 10 sub-lists on the basis of frequency, were retrieved from the internet. The sub-lists were saved separately in notepad files including all the word families and their forms. The lists were modified to create lemma list files of the sub-lists so that, they may be used in AntConc 3.2.4 a concordancing program) for generating lemma search result lists for frequency counting based on AWL word families (head words) and their forms (lemmas).

The 570 AWL word families comprise of 3111 lemmas (word forms). The sub-corpora of PAKDth corpus for both the major disciplinary groups Biological & health sciences (BHSc) and Physical Sciences (PhSc) were separately loaded into the concording program (AntConc). Lemma search feature of Antconc 3.2.4 was employed to generate lemmatized frequency lists of both the sub-corpora. The search results were transferred to separate Microsoft excel (spreadsheet) files for both the groups and frequency counts were calculated to generate results to analyze the use of AWL in sub-corpora as well as in PAKDTh corpus. The results and findings are discussed in the next sections in detail.

4.1 Coverage of AWL in PAKDTh Corpus

The analysis of AWL words in PAKDTh corpus reveals that in total 8.76% of the text in Pakistani doctoral thesis corpus is covered by the AWL words. As shown below in Table 2, the occurrences of 496839 words were found the whole PAKDTh corpus. Similarly, in each of disciplinary groups' sub-corpora (BHSc & PHSc) the AWL's coverage is almost similar to the accumulative text coverage percentage with 8.62% (251879 words) of BHSc texts and 8.91% (244960 words) of PHSc texts. These findings of the current analysis show relative effectiveness of AWL words in both disciplinary groups of science and the written academic genre of doctoral theses. According to Coxhead's (2000) analysis, the text coverage of AWL in Science sub-corpus, which included texts from the subject areas of biology, chemistry, computer science, geography, geology, mathematics, and physics, was 9.1%. So, the use AWL words in PAKDTh corpus and its sub-corpora is almost close to that of Coxhead's

(2000). It is worth notable, that the counts for AWL coverage analyzed in this study include the occurrences of all the AWL word families and their forms, but the counts are not filtered on the basis of range and frequency criteria which is employed by Coxhead (2000) for the development of AWL. Following such criteria, the results for AWL coverage of PAKDTh corpus may vary suggestively.

Table 2. Coverage of AWL words in PAKDTh Corpus

Total Words in PAKDTh Corpus	**BHSc (Sub-Corpus)**	**PHSc (Sub-Corpus)**
	2,922,437	2,749,378
Frequency count for AWL words in PAKDTh Corpus	251879	244960
Percentage of AWL Text Coverage in each sub-corpora	8.62	8.91
Overall Percentage of AWL Text Coverage	4.44	4.32

The text coverage of AWL words in PAKDTh corpus distributed per sub-list (from sub-list 1-10) is provided in table 3. The results, distributed per sub-list, show that the coverage of the AWl words (included in sub-lists 8, 9 and 10) is significantly less than those which are included in sub-lists 1 to 7. The greater part of AWL is covered by sub-lists 1, 2 and 3 with 28.38%, 15.74% 12.72% respectively.

Table 3. Coverage of AWL words in PAKDTh Corpus (Per sub-list 1-10)

AWL Sub-Lists	AWL words coverage in PAKDTh corpus (Total Words: 5.67 million)	AWL coverage %	Token Types	Token Types %
Sub-list 1	141021	28.38	363	14.86
Sub-list 2	78220	15.74	278	11.38
Sub-list 3	63212	12.72	289	11.83
Sub-list 4	47308	9.52	253	10.36
Sub-list 5	43542	8.76	231	9.45
Sub-list 6	29797	6.00	279	11.42
Sub-list 7	43803	8.82	226	9.25
Sub-list 8	22737	4.58	235	9.62
Sub-list 9	22196	4.47	209	8.56
Sub-list 10	5003	1.01	80	3.27
Total	**496839**	**100**	**2443**	**100**

Observing the AWL text coverage in terms of token type (word forms), sub-lists 1-9 covers 96.73% of 2443 word forms/token types of AWL found in PAKDTh corpus which is 78.55 % of 3110 the total token types included in AWL word families and 1.79 % of 135789 the total token types of PAKDTh corpus.

4.2 Frequency of AWL in PAKDTh Corpus

The second objective of this study was to analyze the frequency of AWL word families in Pakistani doctoral theses. So, an attempt was made to further analyze the AWL text coverage by considering the frequency of occurrences in terms of word families. To answer the research question 2, the frequencies of the AWL word families in the entire PAKDTh corpus were calculated and arranged on the basis of the frequency of occurrences in the corpus which are given in Table 4.

The analysis on the basis of the frequency of occurrences shows that among 570- word families of Coxhead's (2000) AWL, 550 word families with the sum of 96.49% are found to occur more than 10 times in Pakistani doctoral theses corpus (PAKDTh), which are taken as word families used in the corpus. Whereas, only 19-word families with 3.33 % were found occurring less than 10 and between 1-9 times and only 1-word family was found with 0 occurrences, both of these word families can be regarded as the word families not frequently used in PAKDTh corpus.

Table 4. Frequency of occurrence for AWL Word families in PAKDTh Corpus

Frequency of occurrences	No. of Word Families	%	Accumulative %
≥ 1000	136	23.86	23.86
500~999	88	15.44	39.30
400~499	37	6.49	45.79
300~399	33	5.79	51.58
200~299	42	7.37	58.95
100 ~ 199	81	14.21	73.16
50~99	66	11.58	84.74
20~49	43	7.54	92.28
10~19	24	4.21	96.49
1~9	19	3.33	99.82
0	1	0.18	100
Total	570	100.00	

In this study, the word "analyze" was found to be the most frequently used AWL word family with 10442 frequency in PAKDth Corpus. Other AWL word families such as significant, react, method, found, extract, concentrate, data, differ, conflict and positive were also observed with high frequency in the corpus. Most importantly, the majority of the AWL word families 136 (23.86%) are found with the frequency more than 1000 times in PAKDTh corpus, which shows the importance of the academic vocabulary included in Coxhead's (2000) AWL in the texts of doctoral theses. The top 100 most frequently used AWL word families found in Pakistani doctoral theses are listed in Appendix A.

It is important to note that there is a significant difference between the results of this study and Coxhead's (2000) arrangement of AWL word families into the sub-lists (1-10) on the basis of the frequency of occurrences. There are various AWL word families which are positioned as high-frequency words in Coxhead's (2000) sub-lists of AWL were not found to occur with the relevant frequency in this study in comparison with Coxhead's analysis and vice versa. For instance, such words as authority, contract, export, finance, labour, legal, legislate were included in sub-list 1 of Coxhead's (2000) AWL, because they were found with high frequency in Coxhead's (2000) academic corpus, but the frequency of occurrences of these words in PAKDTh corpus is highly less ranging from 9 to 86 occurrences. However, certain word families which are infrequent in Coxhead's (2000) analysis, such as detect, exhibit, induce, intense, nuclear, radical, found, mature, medium, and so-called, which are included in the sub-lists 8, 9 and 10 of AWL, seemed to be from the topmost frequent word families (Appendix A) in the analysis of PAKDTh corpus with frequency of occurrences ranging from 1211 to 7335.

Only one AWL word family of the word compound does not occur in PAKDTh corpus. Taken all together, it can be assumed that this difference might be due to the texts included in PAKDTh corpus which has only been taken from two scientific disciplinary groups of (BHSc and PHSc) rather than the inclusion of other disciplinary groups such as arts, humanities, and social sciences.

5. Conclusion

The current study was an attempt to analyze the frequency and coverage of academic vocabulary in scientific doctoral theses texts using a corpus. For this purpose, a pre-built corpus of Pakistani doctoral theses (PAKDTh) (Aziz, 2016) was taken, which comprises of 200 Pakistani doctoral theses from two major scientific disciplinary groups of (biological & health sciences) and (physical sciences) covering 17 distinct disciplines and subject areas. The study reveals that the text coverage of AWL word families in the scientific doctoral theses corpus was 8.76% which indicates the effectiveness and importance of Coxhead's (2000) academic word list in the academic genre of theses and also in the two disciplinary groups (sub-corpora) of science. Furthermore, the findings of the analysis also revealed that the first three sub-lists of AWL accounted for almost 57% of the whole text coverage. Simply, it can be concluded that the word families included in the first three sub-lists of AWL play very important role in the coverage of AWL in the doctoral theses of sciences or PAKDTh corpus.

The results of this study also reveal that 550 world families (96.50%) among the total 570-word families of Coxhead's (2000) AWL are found to be frequently used in the doctoral theses of scientific disciplinary groups. On the basis of the findings of this study, all the individuals concerned with academic and scientific writing learning and instructions (e.g., novice researchers, EAP learners & teachers, research writers, writing instructors and course books and material designers) are suggested to rely upon the use and effectiveness of vocabulary included in Coxhead's AWL. The AWL can highly be considered as one of the most reliable sources for the development, learning, and teaching of academic vocabulary, specifically at higher secondary and tertiary level.

References

Aziz, A., Pathan, H., & Ali, S. (2017). Linguistic Variation across Major Disciplinary Groups of Pakistani Academic Writing: Multidimensional Analysis of Doctoral Theses. *ARIEL-An International Research Journal of English Language and Literature, 27*, 27-60.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In A. Hiebert & M. Kamil (Eds.), *Teaching and learning vocabulary: Bridging research to practice* (pp. 223-242). Mahwah, NJ: Erlbaum.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching, 36*, 81-109. https://doi.org/10.1017/S0261444803001903

Coady, J. (1997). L2 vocabulary acquisition: A synthesis of the research. In J. COADY & T. HUCKIN (Eds.), *Second Language Vocabulary Acquisition* (pp. 273-290). Cambridge: Cambridge University Press.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1988). Reading English for specialised purposes: Discourse analysis and the use of standard informants. In P. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 152-167). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524513.017

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

Macaro, E. (2003). *Teaching and learning a second language*. New York: Continuum.

McEnery, A., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

Mozaffari, A., & Moini, R. (2014). Academic words in education research articles: A corpus study. *Procedia-Social and Behavioral Sciences, 98*, 1290-1296. https://doi.org/10.1016/j.sbspro.2014.03.545

Nagy, W., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304-330. https://doi.org/10.2307/747823

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Stahl, S. A. (1998). *Vocabulary development*. Cambridge, MA: Brookline.

Stahl, S. A. (2005). Four problems with teaching word meanings (and what to do to make vocabulary an integral part of instruction).

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. https://doi.org/10.1016/j.esp.2008.08.003

Warsi, J. (2004). Conditions under which English is taught in Pakistan: An applied linguistic perspective. *Sarid Journal, 1*(1), 1-9.

Appendix A

Top 100 Word Families Most Frequently Used in Pakistani Doctoral Theses (PAKDTh) Corpus

No.	Word Family	Freq	No.	Word Family	Freq
1	analyse	10442	51	major	2605
2	significant	10138	52	volume	2587
3	react	9136	53	generation	2582
4	method	7623	54	inhibit	2577
5	found	7335	55	involve	2490
6	extract	6929	56	individual	2438
7	concentrate	6872	57	equate	2376
8	data	6757	58	derive	2374
9	vary	6390	59	culture	2374
10	conflict	6247	60	induce	2357
11	positive	6076	61	research	2292
12	range	5573	62	medium	2258
13	isolate	5214	63	remove	2255
14	indicate	5073	64	negate	2252
15	process	5013	65	require	2239
16	sequence	5000	66	proceed	2220

17	obtain	4625	67	site	2193
18	area	4513	68	role	2173
19	release	4432	69	period	2166
20	function	4400	70	estimate	2152
21	structure	4395	71	external	2121
22	ratio	4216	72	complex	2079
23	region	4201	73	reveal	2068
24	chapter	4190	74	occur	2060
25	enforce	4165	75	so-called	2019
26	maximise	4149	76	available	2006
27	parameter	4066	77	statistic	1972
28	detect	4021	78	consist	1955
29	normal	4010	79	mechanism	1938
30	factor	3984	80	enhance	1891
31	source	3871	81	mature	1874
32	similar	3754	82	layer	1852
33	technique	3647	83	component	1846
34	evaluate	3608	84	confirm	1844
35	identify	3581	85	media	1794
36	percent	3436	86	constant	1765
37	select	3413	87	exhibit	1764
38	chemical	3234	88	image	1735
39	interact	3180	89	correspond	1689
40	affect	3039	90	previous	1664
41	formula	3025	91	final	1651
42	investigate	2968	92	modify	1608
43	distribute	2947	93	target	1604
44	stable	2897	94	initial	1602
45	phase	2827	95	shift	1583
46	environment	2770	96	whereas	1540
47	respond	2767	97	nuclear	1505
48	section	2767	98	conclude	1467
49	specific	2705	99	link	1460
50	stress	2632	100	cycle	1427