Sci.Int.(Lahore),26(4),1557-1561,2014 ISSN 1013-5316; CODEN: SINTE 8

FEATURES OF ENGINEERING RESEARCH ARTICLES

Nurul Farahin Musa[1], Noorli Khamis [1]
[1] Center for Languages and Human Development,
Universiti Teknikal Malaysia Melaka (UTeM)
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka
Email: farahin.musa@gmail.com

ABSTRACT : *Research articles is one of the widely used medium in communicating new knowledge to the academic community. There have been many investigations into the language use of different text types, including research articles. Today, corpus-based investigations into language features offer more varied and systematic description of language use, especially in genre studies. This paper is a corpus-based study of the engineering research articles. The corpus for this study comprises the journal articles from an engineering discipline, retrieved from the Scopus Website. A preliminary investigation into the word frequency of the corpus was carried out with the Wordsmith Tool 6.0 program. This paper provides the findings of the frequency wordlist analysis which informs the features of the engineering research articles. Understanding the characteristics of the research articles can better equip the writers in preparing their drafts for a journal submission.*

KEYWORDS: Corpus-based Analysis, Frequency Wordlist, Engineering Research Articles, Scopus-indexed Journals

1.0 INTRODUCTION

Research articles (RA) is one of the genres in academic writing; it is regarded as a medium to exchange and communicate new knowledge to the academic community members[1,2,3]. The study on language use in texts has been an interest for many researchers. Some of them study the language use in dissertations [4,5], research articles [6,7], letters [8], weblogs [9] and other text types. Some of the studies focus on the grammar used in RAs, and some on verbs, pronouns and nouns [10], while others on wordlists [11].

The corpus-based approach has been the current tool in identifying the characteristics of the language in RAs [1]. Using a corpus to investigate the language use in a text type or genre has been proven useful in identifying its grammar, vocabulary and words variance [12]. Corpus-based studies allow the discovery of the language behaviour in a text; thus, the identification of the characteristics of a corpus [13]. With this in mind, corpus-based genre analysis would be an appropriate method of establishing a clear picture on how RA writers of a discipline write and present their new findings to the academic community [1].

Hence, this paper discusses the findings from the frequency wordlist analysis of the Scopus-indexed engineering RAs corpus. The Scopus-indexed journals are the concern for the study due to the arising needs of the academicians and researchers to publish their work in prestigious journals. The findings offer interesting information on the features of the corpus. Understanding the features of the research articles can better equip the writers in preparing their drafts for a journal submission.

2.0 METHODOLOGY

For this study, a corpus of 60 engineering RAs, retrieved from the Scopus Website http://www.scopus.com, was created. Then, the frequency wordlist was retrieved using the Wordsmith Tool 6.0 software [14]. The British National Corpus (BNC) was used as a reference corpus. Another program, RANGE, was employed for a further analysis of the wordlist.

2.1 Engineering Research Articles Corpus (ERAC)

The corpus used in this study is the Engineering Research Articles Corpus (ERAC), which consists of 60 engineering RAs from the high impact journals: *Biomaterials*, *Biomechanics and Modelling*, and *Biosensors and Bioelectronics*, which were retrieved from the Scopus Website. The journals selected for this study are from the Biomedical Engineering field. The RAs for this corpus are mainly selected based on two characteristics proposed by Nwogu [15]: *accessibility* and *reputation*. *Accessibility* refers to the ease of the text to be collected for the creation of a corpus. For this corpus, only the articles that can be retrieved online were selected. *Reputation* refers to the esteem which members of an assumed readership hold for a particular publication [15]. As for the corpus of this study, the selection of the research articles was based on the impact factors of the journals.

2.2 British National Corpus (BNC)

This corpus consists of 100 million tokens, which were collected from written and spoken British English. It represents the English used from the 20th century onwards. The sample of words in BNC were taken from academic journals, newspapers, books, memoranda, published and unpublished letters and conversation from various gender and ages. In this study, BNC is taken as a reference corpus and as the general English to obtain a statistical comparison between ERAC and BNC.

2.3 Wordsmith Tool 6.0

The Wordsmith Tool 6.0 is software that offers programs in investigating the language behaviour of a text or corpus. This tool has been used in several studies as a means for describing various textual characteristics of different genres [16]. This software offers 3 main programs that are *Wordlists*, *Keyword* and *Concordance*. However as for this paper, only the *Wordlists* program was employed to identify the most frequent words of ERAC.

2.4 RANGE

This is a vocabulary analysis program developed by Paul Nation, which is available at http://www.vuw.ac.nz/lals/staff/Paul_Nation. An easy-to-use program, RANGE allows the user to analyse a number of

1558 ISSN 1013-5316; CODEN: SINTE 8 Sci.Int.(Lahore),26(4),1557-1561,2014

vocabulary features in multiple texts simultaneously [17]. This program is especially useful for this study because it allows the comparison of wordlists in order to identify which words are and are not in the other list.

3.0 RESULTS AND DISCUSSION

To have meaningful interpretation of the results, the statistical information of ERAC was compared with the information of BNC, the reference corpus. Also, the comparison between both corpora is relevant to investigate any possible similarities or differences between BNC, which can be regarded as the general English, and ERAC, the engineering journal articles English – a specialized language.

3.1 General Statistics

Table 1: Statistical Data for ERAC and BNC

Statistical Details	ERAC	BNC
Tokens used for wordlists	246,695	97,860,872
Types (distinct words)	13,217	512,588
Standardized TTR	35.45	43
Mean word length (in characters)	5.20	5
Ratio of 1-4 letter words	56%	58%

Table 1 shows that ERAC consists of 60 texts with a total of 246,695 words or *tokens*, and 13,217 different words or *word types*. In comparison, BNC comprises 97,860, 872 tokens and 512,588 word types. A valid comparison can be observed from the standardized token ratio (Standardized TTR or STTR) values of both corpora. The STTR is obtained by computing the token ratio for the first 1000 words in the corpus, and for the following sets of 1000 words to the end of the corpus. A running average is computed, and the standardized token ratio is obtained. A high value means that the corpus consists of a variety of words, and a low value means that the corpus is using the same words repeatedly [11]. Thus, STTR suggests the variation of words or diversity of the corpus [11]. Table 1 shows that the STTR value of ERAC is 35.45, lower than BNC (43). It suggests that ERAC has lesser word variation than BNC; there are more repeated words in ERAC. This difference promotes a possibility that there are distinctive language features between general English and the engineering journal articles English, which worth to be discovered and investigated. Nevertheless, the finding may also be accounted by the characteristics of ERAC as a specific domain corpus – an engineering discipline; the specific areas or topics allow more specific and lesser words to be used [18].

The mean word length presents the value of the difficulty and stylistics of the text. It has been suggested that word-length can be a useful index to investigate the difficulty of a text; the higher the value of the mean word length, the more difficult the text to be read [18]. The employment of longer words suggest that the target texts may have many difficult words from a solely empirical perspective [18]. Table 1 content words, and to determine their coverage in the corpus. reveals that ERAC has almost the same word-length average as BNC, that is 5.2 (ERAC) to 5 (BNC) characters. This suggests that generally, ERAC has the same level of readability as BNC from the empirical point of view. ERAC is generally not made up of longer words, which suggests the same text difficulty or complexity level with any other general English (BNC) texts.

This same notion is also suggested by the ratio of 1-4 letter words, which reveals a relatively small difference in both corpora (56% for ERAC and 58% for BNC). A lower value of 1-4 letter words ratio represents a more difficult text. Therefore, the ratio values imply that the difficulty level of ERAC is quite similar to general English. The small difference (2%), which suggests that ERAC could be slightly difficult than general English, can be accounted by the use of its technical and/or sub technical words.

3.2 High Frequency Words of ERAC

Table 2 shows the top 50 words in ERAC and BNC, including the frequency of the words. It is found that the top 23 most frequent words in ERAC are function words, which cover about 37.3% of the corpus. The first content word *cells* ranks as the 24th most frequent word. The content words in this top 50 list indicate that the words are predominantly from the technical and/or sub-technical vocabulary: *bone*, *cell*, *model*, *tissue*, *surface*, *collagen*. Function words are closed-class words, which include prepositions, pronouns, determiners, conjunctions, modal verbs, auxiliary verbs and particles [19]. On the other hand, content words include the opened-class words, such as nouns, verbs, adverbs and adjectives. It should also be noted that the most frequent content words in the list are nouns; the nouns suggest the subjects mostly discussed in this field.

A comparison with the top 50 frequent words from BNC reveals the difference in the nature of the words between ERAC and general English. The top 50 frequent words in BNC are all function words. Table 2 also reveals that with their top 50 words, the corpus coverage of ERAC and BNC are about 44% and 38.7% respectively. The results show that with the coverage of the 50 words, the words in BNC are more general. The coverage of its top 50 words (38.7%) is close to the coverage of the 23 function words in ERAC (37.3%). More function words are identified in BNC with almost the same coverage.

This initial observation calls for a further look into the distribution of function and content words in ERAC. The RANGE program was used to extract the lists. This program allows the comparison of several word lists and the extraction of words which overlap from the lists. 216 functions words were obtained from the Brown Function Words, which can be retrieved online at http://web.simmons.edu/~veilleux/fw_project/bcfw_list.htm. These function words constitute the most frequently occurring words in any texts. RANGE is employed to categorise the ERAC frequency wordlist according to function and

Sci.Int.(Lahore),26(4),1557-1561,2014 ISSN 1013-5316; CODEN: SINTE 8

Table 2 : Frequency list of ERAC and BNC

	ERAC				BNC			
N	**Word**	**Freq.**	**%**	**Cum. %**	**Word**	**Freq.**	**%**	**Cum. %**
1	THE	18936	7.17	7.17	THE	6011078	6.04	6.04
2	#	17569	6.65	13.81	OF	3039337	3.05	9.09
3	OF	9981	3.78	17.59	AND	2606316	2.62	11.71
4	AND	7149	2.71	20.30	TO	2590123	2.60	14.31
5	IN	6142	2.32	22.62	A	2166581	2.18	16.48
6	TO	5185	1.96	24.58	IN	1935991	1.94	18.43
7	A	4468	1.69	26.27	#	1599061	1.61	20.03
8	FOR	2684	1.02	27.29	THAT	1108198	1.11	21.14
9	WAS	2668	1.01	28.30	IT	1025494	1.03	22.17
10	WITH	2663	1.01	29.31	IS	968175	0.97	23.15
11	IS	2465	0.93	30.24	FOR	876133	0.88	24.03
12	WERE	1981	0.75	30.99	WAS	860488	0.86	24.89
13	BY	1841	0.70	31.68	S	815075	0.82	25.71
14	THAT	1808	0.68	32.37	I	798208	0.80	26.51
15	AL	1745	0.66	33.03	ON	727508	0.73	27.24
16	AS	1740	0.66	33.69	WITH	657293	0.66	27.90
17	ET	1688	0.64	34.33	AS	650256	0.65	28.55
18	ON	1563	0.59	34.92	BE	649442	0.65	29.21
19	AT	1561	0.59	35.51	YOU	639390	0.64	29.85
20	THIS	1321	0.50	36.01	HE	629101	0.63	30.48
21	BE	1156	0.44	36.45	AT	521031	0.52	31.00
22	FROM	1122	0.42	36.87	BY	511048	0.51	31.52
23	ARE	1001	0.38	37.25	Â€™	492665	0.49	32.01
24	CELLS	990	0.37	37.62	ARE	454804	0.46	32.47
25	FIG	971	0.37	37.99	THIS	447973	0.45	32.92
26	AN	930	0.35	38.34	HAVE	446268	0.45	33.37
27	WHICH	770	0.29	38.63	BUT	433220	0.44	33.80
28	BONE	765	0.29	38.92	NOT	425108	0.43	34.23
29	WE	762	0.29	39.21	FROM	424343	0.43	34.65
30	CELL	734	0.28	39.49	THEY	413396	0.42	35.07
31	MODEL	728	0.28	39.77	HAD	411753	0.41	35.48
32	OR	714	0.27	40.04	HIS	408710	0.41	35.89
33	TISSUE	699	0.26	40.30	OR	367577	0.37	36.26
34	AFTER	615	0.23	40.53	WHICH	364372	0.37	36.63
35	USING	608	0.23	40.76	SHE	346185	0.35	36.98
36	NOT	577	0.22	40.98	AN	336235	0.34	37.31
37	IT	573	0.22	41.20	WE	332301	0.33	37.65
38	USED	569	0.22	41.41	T	328185	0.33	37.98
39	CAN	537	0.20	41.62	THERE	310618	0.31	38.29
40	SURFACE	482	0.18	41.80	WERE	307271	0.31	38.60
41	HAVE	470	0.18	41.98	HER	302513	0.30	38.90
42	THESE	467	0.18	42.15	ONE	292214	0.29	39.19
43	BETWEEN	454	0.17	42.33	ALL	273366	0.27	39.47

44	COLLAGEN	436	0.16	42.49	Â€	269844	0.27	39.74
45	RESULTS	432	0.16	42.65	BEEN	259698	0.26	40.00
46	STUDY	420	0.16	42.81	THEIR	254054	0.26	40.26
47	HAS	402	0.15	42.97	HAS	251748	0.25	40.51
48	MM	401	0.15	43.12	WILL	249773	0.25	40.76
49	H	399	0.15	43.27	IF	245564	0.25	41.01
50	BEEN	391	0.15	43.42	CAN	239281	0.24	41.25

The analysis with RANGE program reveals that there are 150 function words found in ERAC. These function words cover almost 37.7% of the corpus. Figure 1 and Figure 2 show the distribution of function to content words in ERAC based on types (distinct words) and tokens (words) respectively. Both reveal that the function words are highly repetitive in ERAC. Though the RANGE program identifies these words as function words, the occurrence of these words has to be examined with caution because some of these words do not behave as function words all the time, for example the word *is*, which can also be a verb. However, for the purpose of this study, all the identified functions words are not edited, and all are treated as words that match the function words from the Brown Corpus Function Wordlist. This findings warrant further analysis on the function words

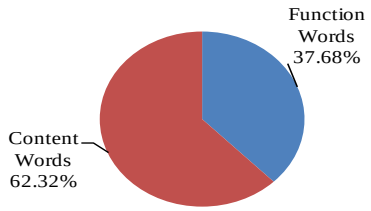

Figure 1 : The distribution of Function and Content Words (Tokens)

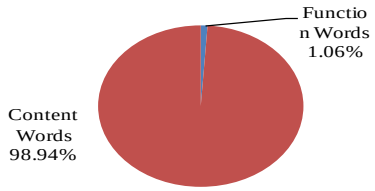

Figure 2 : The distribution of Function and Content Words (Types).

to be carried out on the nature of the function words in ERAC to discover possible features that significantly distinguish it from general English. Function words are unique because many of the members display a quality that joins grammar and lexis, such as the word *from*, which has 26 definitions in the COBUILD dictionary [20]. Empirical observations of the function words also may lead to significant findings about rhetorical functions in specialised texts.

4.0 CONCLUSION

This paper demonstrates the characteristics of the engineering RAs from the investigation of its most frequent words. The results show that there are distinct features between a specialised text (or a genre) and the general English. Hence, the findings also suggest the needs of specific wordlists in writing an engineering RA [21]. A specific wordlists in writing an engineering RAs is highly encouraged for the writers to understand the nature of writing in their specific disciplines. A good understanding of the linguistic features of the engineering RAs will help the novice researchers, especially the NNES (Non-Native English Speaker) writers, to produce clear and impactful RAs. A list of specific words for the specific disciplines should be a part of the whole rhetoric organization of a text, especially RAs [22].

ACKNOWLEDGEMENTS

This work is funded by Universiti Teknikal Malaysia Melaka short grant (PJP/2013/PBPI(4D)S01157).

REFERENCES

[1] W. Amnuai and A. Wannaruk, "Investigating Move Structure of English Applied Linguistics Research Article Discussions Published in International and Thai Journals," vol. 6, no. 2, pp. 1–13, 2013.

[2] L. Flowerdew, "An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies," *English Specific Purposes*, vol. 24, no. 3, pp. 321–332, Jan. 2005.

[3] B. Kanoksilapatham, "Facilitating Scholarly Publication : Genre Characteristics of English Research Article Introductions and Methods," vol. 18, no. 4, pp. 5–19, 2008.

[4] N. A. Manan and N. M. Noor, "Analysis of Reporting Verbs in Master's Theses," *Procedia - Social Behaviourial Science.*, vol. 134, pp. 140–145, May 2014.

[5] J. Ward, "A basic engineering English word list for less proficient foundation engineering undergraduates," *English Specific Purposes.*, vol. 28, no. 3, pp. 170–182, Jul. 2009.

[6] A. Mozaffari and R. Moini, "Academic Words in Education Research Articles: A Corpus Study," *Procedia - Soc. Behav. Sci.*, vol. 98, pp. 1290–1296, May 2014.

[7] I. a. Martínez, S. C. Beck, and C. B. Panza, "Academic vocabulary in agriculture research articles: A corpus-based study," *English Specifi Purposes*, vol. 28, no. 3, pp. 183–198, Jul. 2009.

[8] V. B. M. P. dos Santos, "Genre analysis of business letters of negotiation," *English Specific Purposes.*, vol. 21, no. 2, pp. 167–199, Jan. 2002.

[9] S. C. Herring and E. Wright, "Bridging the Gap : A Genre Analysis of Weblogs," vol. 00, no. C, pp. 1–11, 2004.

[10] A. Koyalan and S. Mumford, "Changes to English as an Additional Language writers' research articles: From spoken to written register," *English Specific Purposes*, vol. 30, no. 2, pp. 113–123, Apr. 2011.

[11] N. Khamis and I. H. Abdullah, "SOCIAL SCIENCES & HUMANITIES Wordlists Analysis : Specialised Language Categories," vol. 21, no. 4, pp. 1563–1581, 2013.

[12] K. Chujo and K. Oghigian, "Selecting Level-Specific Kyoto Tourism Vocabulary Using Statistical Measures," 2006.

[13] C.-F. Chang and C.-H. Kuo, "A corpus-based approach to online materials development for writing research articles," *English Specific Purposes.*, vol. 30, no. 3, pp. 222–234, Jul. 2011.

[14] M. Scott, "WordSmith Tools," 2012.

[15] Kevin Ngozi Nwogu, "The Medical Research Paper : Structure and Functions," vol. 16, no. 2, pp. 119–138, 1997.

[16] T. Berber-sardinha, R. M. Alegre, and S. P. Sp, "Comparing corpora with WordSmith Tools : How large must the reference corpus be ?"

[17] A. H. A. Coxhead, "Range (Computer Program)," 2002. [Online]. Available: http://www.victoria.ac.nz/lals/staff/paul.nation.aspx. [Accessed: 24-Jul-2014].

[18] Noorli Khamis, "A Lexical Investigation of Engineering English: A Corpus-Based Approach," Universiti Kebangsaan Malaysia, 2011.

[19] C. Chung and J. Pennebaker, "The Psychological Functions of Function Words," pp. 343–359, 2007.

[20] L. Flowerdew, "academic writing," vol. 54, no. October, 2000.

[21] I. a. Martínez, S. C. Beck, and C. B. Panza, "Academic vocabulary in agriculture research articles: A corpus-based study," *English Specific Purposes.*, vol. 28, no. 3, pp. 183–198, Jul. 2009.

[22] Q. Chen and G. Ge, "A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs)," *English Specific Purposes.*, vol. 26, no. 4, pp. 502–514, Jan. 2007.