ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse07912020.pdf
https://doi.org/10.30534/ijatcse/2020/07912020

Corpus-based Data for Determining Specialised Language Features

Noorli Khamis [1], Nurul Farahin Musa [2]
[1]Centre for Languages and Human Development, Centre for Technopreneurship Development CTeD, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia
noorli@utem.edu.my
[2]Jabatan Pembangunan Perniagaan, Syarikat Pembangunan Pertanian Melaka Sdn. Bhd. Ladang Lembah Kesang, Jln Jasin-Semujuk, Merlimau, Melaka Malaysia
farahin.musa@gmail.com

ABSTRACT

Corpus-based data have been used extensively to describe language use. Studies into specialised languages have adopted this approach to describe the English language used in different disciplines, such as Engineering English and Business English. Corpus-based analysis has also been used to determine the characteristics of specialised languages used by writers in RAs. Serving to contribute to the body of knowledge on characteristics of language use in RAs, this study presents the word lists analysis of research articles (RA) from two different disciplines - engineering and, business technology and innovation. The findings provide insights into the distinct features of these specialised languages in RAs. The RAs for both corpora were obtained from the Scopus database, and the frequency word lists for both were generated using the Wordsmith Tool 6.0. This study demonstrates not only the different word lists, but also empirical evidences in describing the two specialised languages. To do this, the analyses of the corpora involve the comparison of the general statistical details, the high frequency word lists, and the function vs. content word distributions. Insights into the characteristics of the specialised languages, such as provided in this study, are helpful in assisting students, researchers, writers, language practitioners to be more well-informed and more effective in using the specialised language.

Key words : corpus-based analysis, language features, research articles, specialised language, word list analysis

1. INTRODUCTION

A corpus is a body of data, collected based on specific criteria set by a corpus designer to serve certain purposes, among others are to create a dictionary [1] and to identify the rhetorical organisation of specific documents [2]. Corpus work has made significant contribution in the analysis and description of many English for Specific Purposes (ESP) [3]-[5]. Corpus-based analyses prove that language use is greatly determined by the concepts of tendencies and probabilities, as opposed to the knowledge of rules as advocated by earlier linguists. This suggests that a word (or a string of words) is considered significant or regular, if it is used frequently in a language. Therefore, if these frequent words can be identified in a specialised language, such as an Engineering English or Business English, useful information about the language can be formed; the characteristics of the language can be established. More well-informed decision-makings for ESP teaching and learning activities can be achieved.

This paper demonstrates how a corpus data provides useful insights into the features of a specialised language. This work attempts to identify the distinctive features of research articles (RA) of two different disciplines: electrical engineering and, business innovation and technology. Findings of language use in RAs, thus far, have revealed the moves RA writers use to express ideas on their new discoveries [6]. The findings also prove that different disciplines possess different features in the RA writing. Therefore, this paper serves to contribute to the body of knowledge on characteristics of language use in RAs of more different disciplines.

Research article (RA) is a platform to communicate new knowledge and new findings among academics and discourse community [7]. Studies on language use in RAs have revealed many valuable results, which have been the impetus for the development of many ESP courses on research article writing. Research activities is the main performance indicator of an academic institution. In most universities, it is a requirement for the researchers to carry out studies and discover new knowledge or innovation; and they are also required to share their findings and discoveries with their professional community through publications [8]. Several universities even make it a requirement for their academics to publish research papers for career advancement.

The different cultures in all fields of specialisation are depicted in the use of words and structures in the language [9]. As such, different specialisations have different lists of most frequent specific words. Studies on specialised language words in academic texts have gained more interest because of the increasing demand for the researchers of different contexts and backgrounds to publish [10]. Knowing and understanding the linguistic features of the specialised language can assist especially the early-career non-native English speaker (NNES) writers in writing clear, coherent and impactful research articles [11]. There have been many studies on the features of RAs from various disciplines. These studies focused on, among others, the language use [12], word lists [13], keywords [14] and rhetorical organisations [15].

Various word lists have been developed to assist the NNES writers in learning the vocabulary, as well as the many types of academic texts for a discipline. Such word lists include the Academic Word List (AWL), Basic Engineering Word List and General Service List of English (GSL). Studies on word lists for different disciplines reveal that some words mean and behave differently according to the fields, as well as to the genres. It is also established that some words from the AWL and GSL can carry technical meanings in other specific corpora [16]. As such, there is a need to develop a specific word list for every specialisation to represent the expressions of the discipline. Every discipline has its own voice and way to present their findings and to form their arguments on certain subjects [17]. There are more differences than similarities exist across disciplines. Hence, there is a need for the writers to know the technical, sub technical or general English words to be proficient in the specialised language.

Apart from the word lists, the characteristics of a field is also important for writers to be familiar with. The knowledge offers insights into the nature of language use of the field. The specialised word list assists the writers to understand the terms as used by the community of the field. As such, it is crucial for writers to identify the technical and sub technical terms, which characterise the language use of their specific disciplines.

This work employs the frequency word list of a corpus to determine the features of the specialised language by investigating the distribution of the function and content words [16]. It has been found that function words occupy a larger portion of texts. Closed-class words, such as pronouns, modal and auxiliary verbs, prepositions, determiners and conjunctions, are classified as function words [18]. These words are commonly used to form grammatical sentences. Function words express or represent the connection between the content words, which shows relationships between actions, activities, entities and verbs. Hence, by observing the behaviours of these function words in the specialised RAs, significant information on their rhetorical functions in the texts can be formed. In contrast, content words are open-class words, which include verbs, nouns, adjectives and adverbs; these words help writers to deliver a picture, ideas and content in readers' mind [19]. Thus, by using the corpus data, this work aims to identify the distribution of function vs. content words for both engineering and business technology innovation disciplines to determine the characteristics of these specialised languages.

2. METHOD

Table 1 provides the two corpora created for this study - the Electrical Engineering Research Articles Corpus (EERAC) and the Business Innovation and Technology Corpus (BITC). 60 RAs were randomly selected from the Scopus website at http://www.scopus.com for each corpus. The articles selected for EERAC were obtained from three journals: Solid State Electronics, Microelectronic Reliability and Microelectronic Engineering, while for BITC, also from three journals: Technovation, Information and Organization, and Technological Forecasting. The RAs for these corpora were selected mainly based on two characteristics - accessibility and representativity [20]. Accessibility takes into account the ease of texts collection in creating a corpus. Hence, for this study, only the articles which can be obtained online were included in both corpora. It is important to ensure that the RAs were selected systematically to ensure the representation of RAs from both disciplines.

Table 1: Composition of EERAC and BITC

	EERAC	**BITC**
No. of RAs	60	60
No. of journals	3	3

The British National Corpus (BNC) was used as the reference corpus; it is also used to represent the general English. A reference corpus is required to allow statistical comparisons between the specialised corpora (EERAC and BITC) and general English. BNC comprises the written and spoken British English, totalling to 100 million words.

The word lists and language use were analysed using the Wordsmith Tool 6.0 [21] and RANGE32 software [22]. Wordsmith Tool 6.0 has been employed in many studies as a tool to analyse textual features and language behaviours of a corpus or genre. It provides 3 functions for language investigation: Word lists, Keyword and Concordance. However, for this study, only the Word lists program was employed to generate the most frequent word lists from the corpora. The RANGE program was used to compare the most frequent word lists and extract the words which overlap from the lists

3. RESULTS AND DISCUSSION

The first stage of analysis involves the comparison of the statistical information between the specialised languages (EERAC and BITC) and the general English (BNC). The comparison is useful to investigate any similarities or differences between the general English and the specialised languages.

3.1 General Statistics of EERAC, BITC and BNC

Table 2 shows the general statistics comparison between EERAC, BITC and BNC, obtained using the Wordsmith Tool 6.0 software. With 60 texts, EERAC has a total of 170,078 tokens (words), and 8,198 word types (different words). BNC, on the other hand, has 97,860, 872 tokens and 512,588 word types. BITC, on the other hand, has a total of 509,307 words or tokens, and 19,516 different words or word types. Generally, RAs of BITC have more words than EERAC.

Table 2: Statistics of EERAC, BITC and BNC

Statistical details	EERAC	BITC	BNC
Tokens used for word lists	170,078	509,307	97,860,872
Types (distinct words)	8,198	19,516	512,588
Standardized TTR (STTR)	32.49	38.79	43
Mean word length (in characters)	4.82	5.45	5
Ratio of 1-4 letter words	56%	50.65%	58%

A more valid observation on the differences of these corpora can be made from the standardized token ratio (Standardized TTR or STTR) values. The STTR value is generated with the computation of the token ratio for the first 1000 words in the corpus. Next, the computation of the token ratio continues to be generated for the following sets of 1000 words until the end of the corpus. Then, a running average is calculated to determine the STTR value. The STTR value indicates whether the corpus comprises a variety of words. Otherwise, a low STTR value signals that there are many repetitive words in the corpus. In other words, STTR suggests the word range of the corpus [16]. Table 2 shows that the STTR value of EERAC is 32.49, lower than BITC (38.79) and BNC (43). This suggests that EERAC has more repeated words than BITC. The finding also shows that both specialised languages have lesser variation of words in comparison to general English.

Based on this finding, it is evident that the language features of general English are different from the language features of the RAs of both specialised languages. Similarly, there are distinctive differences for the RAs of both specialisations. These features worth to be discovered and investigated. Nevertheless, the finding may also be accounted by the characteristics of EERAC and BITC as specific domain corpora; the specific topics discussed in the RAs allow more specific and lesser words to be used.

The mean word length indicates the text difficulty and stylistics. Word-length has been used to inform the level of text difficulty. It is suggested that a high mean word length means that the text has high level of difficulty; thus, low level of readability. In other words, long word length may suggest the analysed texts have more difficult words.

Table 2 reveals that BITC (5.45) has a higher word-length average than EERAC (4.82). This suggests that generally, BITC has a higher level of readability than EERAC, and even BNC (5); BITC may be made up of longer and more complex words. Interestingly, from the empirical point of view, EERAC has almost the same level of readability as BNC. EERAC is generally made up of lesser long words, which suggests the same text difficulty or complexity level with general English (BNC) texts.

The same notion is also suggested by the ratio of 1-4 letter words results. A lower value of 1-4 letter words ratio represents a more difficult text. Table 2 shows that there is a relatively small difference between EERAC (56%) and BNC (58%). The ratio values also imply that the difficulty level of EERAC is quite similar to general English. The small difference (2%), which suggests that EERAC could be slightly difficult than general English, can be accounted by the use of its technical and/or sub technical words. However, there is a difference in the values of BITC (50.65%) in comparison to EERAC and BNC. The difference suggests higher difficulty level of BITC as compared to EERAC and even, the general English

3.2 High Frequency Words

Table 3 shows the top 50 words in EERAC, BITC and BNC, including the frequency of the words. From the table, it shows that 25 most frequent words in EERAC and 24 in BITC are function words. The first most frequent content word in both specialised corpora are nouns: EERAC – LAYER (26) and BITC – KNOWLEDGE (25). In fact, most of the content words in both top 50 lists are nouns. These nouns suggest the subjects mostly discussed in the respective fields. The content words from ERACC suggests that the words of this Engineering specialisation are from the technical and/or sub-technical vocabulary: *layer*, *gate*, *temperature*, and *current*. As for the nouns of BITC, the words are from both academic and sub-technical vocabulary: *knowledge*, *research*, *technology*, *firms*, and *study*. Interestingly, one adjective appears in this top 50 frequent words of BITC – *new*, highlighting the core focus of this specialisation. BNC, on the other hand, displays all function words in its 50 most frequent word list, suggesting that there are no specific and prominent subjects discussed in general English. The preliminary findings from the wordlists warrant an investigation into the function and content words distribution in both specialised corpora. To carry out the task, this study employs another software, the RANGE program, to categorise the function and content words from EERAC and BITC.

First, a function word list was identified; the Brown Functions Words was obtained at http://web.simmons.edu/~veilleux /fw_project/bcfw_list.htm. The list has 216 function words, which are words with high occurrence in any texts. The list was used to extract function words from both corpora. Next, the text coverage of both function and content words in the specialised corpora was determined.

Table 3: Top 50 words for EERAC, BITC and BNC

EERAC						BITC						BNC					
No	Words	Freq.	No	Words	Freq.	No.	Words	Freq.	No.	Words	Freq.	No.	Words	Freq.	No.	Words	Freq.
1	THE	15735	26	LAYER	591	1	THE	31558	26	NOT	1888	1	THE	6,055,105	26	FROM	431,075
2	#	11756	27	WHICH	578	2	OF	19161	27	S	1836	2	OF	3,049,564	27	HAD	425,987
3	OF	6965	28	GATE	573	3	AND	18307	28	WERE	1823	3	AND	2,624,341	28	HIS	413,144
4	AND	4449	29	SI	566	4	#	14911	29	HAVE	1655	4	TO	2,599,505	29	THEY	410,294
5	IN	4137	30	AN	555	5	TO	14363	30	THEY	1566	5	A	2,181,592	30	OR	376,289
6	TO	3631	31	TEMPERATURE	518	6	IN	11934	31	WHICH	1565	6	IN	1,946,021	31	WHICH	370,166
7	A	3444	32	WE	511	7	A	9162	32	MORE	1545	7	THAT	1,604,421	32	AN	366,196
8	IS	3100	33	CURRENT	494	8	THAT	6555	33	RESEARCH	1497	8	IS	1,052,259	33	SHE	338,743
9	FOR	2041	34	NM	492	9	IS	5026	34	CAN	1494	9	IT	974,293	34	WERE	325,351
10	WITH	1755	35	RESULTS	470	10	FOR	4844	35	WORK	1449	10	FOR	922,687	35	HER	308,363
11	AT	1445	36	NOT	465	11	AS	4543	36	THESE	1428	11	WAS	880,848	36	WE	304,311
12	AS	1370	37	HAVE	438	12	ON	3914	37	BETWEEN	1305	12	I	863,917	37	ONE	300,833
13	THAT	1344	38	INTERFACE	422	13	WITH	3342	38	NEW	1299	13	ON	732,523	38	THERE	290,466
14	BY	1311	39	DIFFERENT	417	14	THIS	3335	39	TECHNOLOGY	1283	14	WITH	731,319	39	ALL	285,870
15	ON	1263	40	HAS	417	15	ARE	3133	40	AT	1271	15	AS	659,997	40	BEEN	277,566
16	BE	1166	41	OR	413	16	BY	2991	41	ALSO	1263	16	BE	655,259	41	THEIR	260,360
17	WAS	1138	42	THAN	411	17	BE	2639	42	AL	1232	17	HE	651,535	42	IF	254,603
18	THIS	1133	43	USED	402	18	IT	2602	43	ET	1231	18	YOU	593,609	43	HAS	253,804
19	ARE	1131	44	OBSERVED	391	19	WE	2333	44	SUCH	1160	19	AT	588,503	44	WILL	252,703
20	FIG	944	45	BEEN	389	20	THEIR	2324	45	OUR	1151	20	BY	524,075	45	SO	251,179
21	FROM	856	46	SURFACE	388	21	FROM	2253	46	OTHER	1149	21	ARE	513,444	46	WOULD	239,549
22	C	758	47	SOLDER	386	22	OR	2189	47	USE	1136	22	THIS	458,368	47	NO	229,699
23	WERE	736	48	DEVICES	379	23	WAS	2166	48	FIRMS	1111	23	HAVE	454,419	48	WHAT	229,618
24	CAN	653	49	ALSO	378	24	AN	2155	49	STUDY	1097	24	BUT	448,684	49	CAN	225,524
25	IT	627	50	AFTER	371	25	KNOWLEDGE	2085	50	HAS	1048	25	NOT	446,783	50	WHEN	211,093

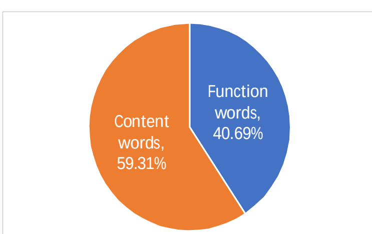

Figure 1(a): The Distribution of Function and Content Words in EERAC (tokens/words)

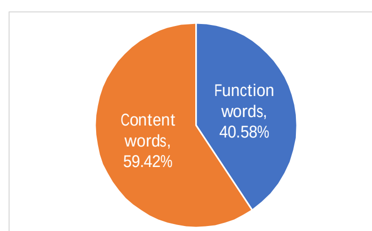

Figure 2(a): The Distribution of Function and Content Words in BITC (tokens/words)

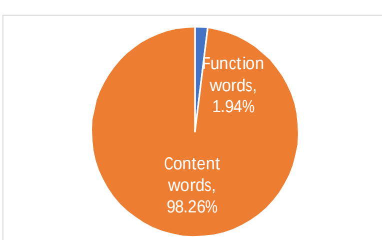

Figure 1(b): The Distribution of Function and Content Words in EERAC (types/distinct words)

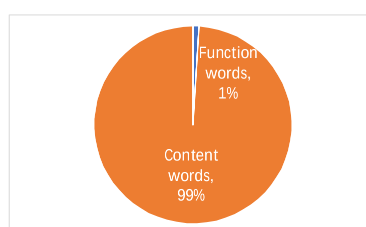

Figure 2(b): The Distribution of Function and Content Words in BITC (types/distinct words)

The analysis with RANGE program reveals that there are 150 function words (2%) in EERAC, and 186 (1%) in BITC. The ERRAC function words cover almost 41% of the corpus. Figure 1 shows the distribution of function to content words in EERAC based on types (distinct words) and tokens (words) respectively. The function words in BITC also cover almost 41% of the corpus. Figure 2 shows the distribution of function to content words in BITC based on types (distinct words) and tokens (words) respectively. The distributions of the function words in Figs. 1 and 2 reveal that the occurrence of function words in EERAC is almost similar to BITC, despite the fact that EERAC has more percentage of distinct function words to content words than BITC. This notion suggests that the functions words in EERAC, though lesser, are used repeatedly.

An implication that can be derived from this discovery is on the teaching and learning of function words to the learners of each specialised language. For EERAC, for example, the identified most frequent function words should be further explored in terms of their neighbouring words and usage contexts. The mastery of these function words is apparently crucial since they are used repeatedly in the field. Not only it assists the learners to better understand the language expressions in the texts, but also to write more effectively as expected by the community of the specialisation.

It should be noted that the use of RANGE program in extracting the function words from the specialised corpora should be exercised with caution. The program extracts the function words as their prototypical forms, instead of their functions in the texts. This means that the extracted words may not behave as function words all the time in the texts, such as the word *is,* which can function as either the auxiliary verb (a function word) or a verb (a content word). Nonetheless, the function words in this study were taken as their prototypical forms, as to determine the preliminary distributions of the words.

These findings, indeed, warrant further investigations into the nature of the function words in EERAC and BITC to discover possible features that significantly distinguish the specialised language from each other, as well as from general English.

4. CONCLUSION

This paper demonstrates a corpus-based comparison of general language features between the electrical engineering and business innovation and technology research articles (RAs). The results show that there are distinct features not only between the specialised languages and the general English, but also between the different specialised languages (EERAC vs. BITC). Hence, these differences warrant the needs for specific word lists in writing RAs for different disciplines. The frequency word lists analysis assists in understanding the nature of writing in the specific disciplines. Understanding the linguistic features of these RAs from different fields can better assist the novice researchers, especially the NNES writers, in producing RAs which are meaningful to the community of the domain. It also promotes the fact that the writing of the RAs needs to include the discipline-specific words to achieve the expected rhetorical organisation of the text in the discipline [23].

Another highlight from this study is the pedagogical implications that can be derived from the discoveries. Understanding the language characteristics of a specific discipline, such as the word length and text complexity, discussion topics, and distribution of technical or academic vocabulary, promises a more well-informed decision-making in selecting, planning and applying the language information for ESP learners of that discipline. Hence, the teaching and learning of ESP is more directed and meaningful.

Finally, this work demonstrates only several analyses of corpus-based data in describing a specialised language, which are general statistics, high frequency word lists and function vs. content word distributions. More analyses can be conducted in relation to corpus-based data to inform language users about a specialised language, in comparison to the general English. The knowledge of specialised language use assists language users to be more accurate in expressing ideas within their community, by using the accurate word choices and frequent structures.

Thus, this work provides yet more evidence to corpus-based specialised language investigations for more effective ESP considerations.

ACKNOWLEDGMENT

This paper is funded by a grant GLuar/MPM/2017/ PBPI-CTED/I00028 from Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

1. N. S. A. A. Bakar. **The development of an integrated corpus for Malay language**, in *Computational Science and Technology. Lecture Notes in Electrical Engineering,* vol. 603, R. Alfred, Y. Lim, H. Haviluddin, C. On, Ed. Singapore: Springer, 2020.
2. D. M. Ponton. ***Understanding Political Persuasion: Linguistic and Rhetorical Analysis***, Delaware: Vernon Press, 2020.
3. C. Suganthan, and R. RLN. **Classification of Errors in the Project Reports of Engineering Aspirants**, *International Journal of Advanced Trends in Computer Science and*

Engineering, vol. 8, no. 3, pp. 598-600, 2019.

4. O. Al-Omari, and N. Omari. **Enhanced Document Classification Using Noun Verb (NV) Terms Extraction Approach**, *International Journal of Advanced Trends in Computer Science and Engineering ,* vol. 8, no. 1, pp. 85-92, 2019.
5. M. Syamala, and N.J.Nalini. **A Deep Analysis on Aspect based Sentiment Text Classification Approaches**, *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 8, no. 5, pp. 1795-1801, 2019
6. S. Y. Uba. **Metadiscourse in research article genre: A cross-linguistic study of English and Hausa**, *English Language Teaching,* vol. 13, no. 2, pp. 57-62, 2020.
7. E. Sheldon. **Dialogic spaces of knowledge construction in research article conclusion sections written by English L1, English L2 and Spanish L1 writers**, *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE),* vol. 35, pp. 13-40, 2018.
8. M. Khedri, S. J. Ebrahimi, and C. S. Heng. **Interactional metadiscourse markers in academic research article result and discussion sections**, *3L : The Southeast Asian Journal of English Language Studies,* vol. 19, no. 1, p. 65–74, 2013.
9. L. Bercuci, and M. Chitez. **A corpus analysis of argumentative structures in ESP writing**, *International Online Journal of Education and Teaching,* vol. 6, no. 4, pp. 733-747, 2019.
10. P. Ventura, and E. Martin-Monje. **Learning specialised vocabulary through Facebook in a Massive Open Online Course**, in *New Perspectives on Teaching and Working with Languages in the Digital Era*, A. Pareja-Lora, C. Calle-Martínez, and P. Rodríguez-Arancón, Eds. Dublin: Research-publishing. net, 2016, pp. 117-128.
11. N. A. Manan, and N. M. Noor. **Analysis of reporting verbs in master's theses**, *Procedia - Social and Behavioral Sciences,* vol. 134, pp. 140–145, 2014.
12. H. Su, and L. Zhang. **Local grammars and discourse acts in academic writing: A case study of exemplification in Linguistics research articles**, *Journal of English for Academic Purposes,* vol. 43, pp. 1-11, 2020.
13. T. Lau. **Noun phrase construction in academic research articles**, *International Journal Online of Humanities,* vol. 3, no. 6, p. 17, 2017.
14. A. Duvvuru, S. Radhakrishnan, D. More, S. Kamarthi, and S. Sultornsanee. **Analyzing structural & temporal characteristics of keyword system in academic research articles**, *Procedia Computer Science,* vol. 20, p. 439–445, 2013.
15. N. F. Musa, and N. Khamis. **Research article writing: a review of a complete rhetorical organisation**, *Pertanika Journal of Social Sciences & Humanities,* vol. 23, no. S, p. 111 – 122, 2015.
16. N. Khamis, and I. H. Abdullah. **Wordlists analysis : specialised language categories**, *Pertanika Journal of Social Sciences & Humanities,* vol. 21, no. 4, p. 1563–1581, 2013.
17. S. Y. Uba. **Semantic categories of reporting verbs across four disciplines in research articles**, *English Language Teaching,* vol. 13, no. 1, pp. 89-98, 2020. https://doi.org/10.5539/elt.v13n1p89
18. C. Chung, and J. Pennebaker. **The psychological functions of function words**, *Social Communication*, pp. 343–359, 2007.
19. R. Schmauder, R. K. Morris, and D. V. Poynor. **Lexical processing and text integration offunction and content words: Evidence from priming and eye fixations**, *Memory & Cognition,* vol. 28, no. 7, pp. 1098–1108, 2000.
20. K. N. Nwogu. **The medical research paper : structure and functions**, *English for Specific Purposes,* vol. 16, no. 2, pp. 119–138, 1997.
21. M. Scott. ***Wordsmith Tools***. Oxford: Oxford University Press, 2006.
22. A. Heatley, I. S. P. Nation, and A. Coxhead. (2002). **Range: A program for the analysis of vocabulary** [Computer software]. Available http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx
23. 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 for Specific Purposes,* vol. 26, no. 4, pp. 502–514, 2007. https://doi.org/10.1016/j.esp.2007.04.003