International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-12, October, 2019

Extracting Significant Words in Engineering Texts for Specialised Language Descriptions

Noorli Khamis, Imran-Ho Abdullah

Abstract*: *The academic discourse of a specialised language is characterised by specialised and technical vocabulary, and lexicogrammar. Studies on language description suggest the need to explore and determine the specific characteristics of the academic discourse of each specialised language, to serve the language needs of the learners. This study demonstrates an exploration of this discipline specificity by looking at the nouns used in a specialised language - an Engineering English. It attempts to integrate a multivariate technique, i.e. the Correspondence Analysis (CA), as a tool to extract significant nouns in a specialised language for any further language use scrutiny. CA allows visual representations of the word interrelationships across different genres in a specialised language. To exemplify this, an Engineering English Corpus (E2C) was created. E2C is composed of two sub-corpora (genres): Engineering reference books (RBC) and online journals articles (EJC). The British National Corpus (BNC) was used as the reference corpus. 30 key-key-nouns were identified from the E2C, and the frequency lists of the words were retrieved from all the corpora to run the CA. The CA maps of the nouns display how these corpora are different from each other, as well as, which words characterise not only E2C from a general corpus (BNC), but also the different genres in E2C. Thus, CA proves to be a potential tool to display words which characterise not only a specialised corpus from a general corpus, but also the different genres in that specialised corpus. This study promises more informed descriptions of a specialised language can be made with the identification of specific and significant vocabulary for any academic discourse investigations.

Keywords: academic discourse, Correspondence Analysis, ESP, nouns, specialised corpus

I. INTRODUCTION

The knowledge of specialised language characteristics provides a platform for language practitioners to devise strategies to extend the language and knowledge to learners [1]. The most common way to analyse the language characteristics of different specialisations is from the written or spoken texts of the domains. The texts reflect specific purposes, contexts, and characteristics of the language used in the domain, and thus, reveal different lexical properties of the specialised language. corner of the paper.

The academic discourse of a specialised language is characterised by specialised and technical vocabulary, and lexicogrammar. It is reported as being conceptual, lexically dense, extensively elaborated (nominal groups) and objective [2]. Studies on language description suggest the need to explore and determine the specific characteristics of the academic discourse of each specialised language, to serve the language needs of the learners [3]-[6]. It is crucial to see more than what is taken as a generalised view of academic discourse to describe the usage of a specialised language to learners. All the various disciplines have distinctive ways to express ideas, and the members of these disciplines can understand their discourses accordingly. Therefore, discourse analyses of different specialised languages are of great importance, and the analyses will continue to be relevant with the emergence of every new specialisation.

This study demonstrates an exploration of this discipline specificity by looking at the nouns used in a specialised language - an Engineering English. The paper attempts to integrate a multivariate technique, i.e. the Correspondence Analysis (CA), as a tool to extract significant nouns in the specialised language for any further language use scrutiny.

II. LITERATURE REVIEW

A. Nouns in Academic Texts

Universities have been giving emphasis on equipping the students with language skills for various academic genres, such as academic writing and reading. Lan and Sun [2] highlight that syntactic complexity in academic register indicates the proficiency levels of the learners in the language. Their study suggests that academic writing courses should focus on complex noun phrase constructions to assist the learners in mastering the academic texts. The complexity of noun phrases in professional legal texts is also discussed by Ariwibowo and Tedjasuksmana [7]. Noun phrases are found significantly functioning as subject, object, subject complement, or complement of a preposition in legal texts.

There have been many other studies which suggest that analyses of nouns in academic texts provide useful insights into not only developing understanding in reading, but also cohesion in writing [8]. Işık-Taş [9] reports on a list of investigations into related concepts and functions of nouns in academic texts, which include general nouns, anaphoric nouns, carrier nouns, shell nouns, signalling nouns, and stance nouns. Some grammatical observations from noun groups in academic genres are also identified to assist language practitioners to comprehend specialised languages [10]-[13]. Findings from these studies offer valuable information on how nouns behave in various academic registers. These nouns are found to appear recurrently in academic texts.

Revised Manuscript Received on October 10, 2019

* Correspondence Author

Noorli Khamis*, Pusat Bahasa dan Pembangunan Insan (PBPI), Centre for Technopreneurship Development (CTeD), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia. Email: noorli@utem.edu.my

Imran-Ho Abdullah, School of Language Studies and Linguistics, Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia. Email: imranho@ukm.edu.my

Thus, nouns are identified to be a prominent landscape in academic texts [14]. In fact, the use of complex nouns, which reflect the level of sophistication in structures, as well as knowledge, makes up the feature of academic genres [15]-[16]. In addition, it is found to be a strong association with L2 learners' academic level; exposures to complex noun constructions are proven to boost the academic writing performance of L2 learners [2].

However, the complexity of noun groups in academic genres poses problems to L2 learners [16]. The different and specific nature of academic discourse of various disciplines also provide challenges in learning specialised languages. Due to this quality nouns possess in academic texts, it is crucial to identify and highlight significant nouns used for specialised language description, as well as for ESP classrooms.

This study aims to make a relevant contribution in extracting significant nouns from a specialised language for any further language investigation. Hence, the study adopts a multivariate technique, i.e. the Correspondence Analysis, in identifying and highlighting all the significant nouns, which are statistically found to characterise a specialised language.

B. Multivariate Statistics and Genre Analysis

With the emergence of big data, Graham, Kim, DeVasto, & Keith [17] posit that features of genres can be accurately identified not only with investigations across a genre, but also across massive genre collections. It is due to the fact that features which are not significant in a study of a genre, can be found significant when investigated across a huge number of texts. This approach looks at wealth of information in extracting the language features of a genre.

The ability of computers to handle enormous datasets has made what regarded as challenging language investigations before, possible. Not only it allows numerous interesting investigations about a language to be carried out, but also offers many scopes of visual representations to be displayed [18]. As more variables of language features to be studied, more complex interrelationships of the variables can be conceptualized, from either one or several sets of data. One approach of comprehending complex interrelationships of datasets is through a computational application of statistical tools - the multivariate analysis.

There are many techniques the multivariate analysis offer, and they are mainly categorised as:

a) **exploratory** – an analysis which looks at the consistency of data occurence which can lead to the contruction of hypotheses about a genre, and the data structures are frequently represented in the form of graphics.
b) **confirmatory** – an analysis which identify any significant associations of several selected independant variables against one or more dependant variables.

Thus, these two types of analysis are used together to validate the hypotheses formed about a genre.

The employment of the multivariate analysis in genre studies has received attention by many [19]-[21]. Hall-Millsa and Apela [22] employ the exploratory factor analyses to study the interrelationship of microstructure and macrostructure features in different written genres across different grade levels. Doring and Poeschl [23] use the exploratory statistics as one of the approaches to investigate the representation of relationships between humans and robots in the media. Of all the techniques, the multi-dimensional analysis has been commonly adopted in genre-based investigations. Jin [24] uses the multidimensional analysis to describe the different linguistic characteristics in the discussion sections between low- and high-impact Engineering research articles. Wu [25] employs Biber's multi-dimensional (MD) approach and Halliday's Systemic Functional Linguistics (SFL) to look into the linguistic variations present in corporate blogs.

The principle of textual studies is in the words. Thus, many lexical investigations of specialised languages look at different types of word lists, such as by their frequency, keyness and key-keyness, to describe the features of the language [5]-[6], [26]. Though generally these different word lists reveal the same features of a language, they offer different lists of words to look at. Out of these word lists, the key-keyword list includes the most words of different word classes in the top list [27]. On this note, this study employs the key-keyword list from which the significant nouns of the specialised will be extracted.

Key-keywords are words that are extracted from the keyword database. They represent the notion that the more texts they are 'key' in, the more 'key-key' they are. In other words, key-keywords suggest the range of the words - how many texts in the corpus does the word occur in. Hence, the adoption of the key-keyword lists in this study is to extract the most significant nouns that give the identity to the specialised language.

This study attempts to visually show the interrelationships of the nouns across genres of a specialised language. Thus, the multivariate technique adopted is the Correspondence Analysis (CA). CA allows the analysis of both dependant and independent variables to take place simultaneously [28]. An interesting feature of the CA is it allows cross-tabulation analysis of sets of data. With regard to lexical analysis, CA enables the examination of word frequencies across sets of text types and displaying of their relationships in graphical representations. Thus, CA offers not only clear, but also fast understanding of the word interrelationships [29]. Mealand [30] uses CA to extract words which contribute the most to genre differences, before further statistical tests are conducted to verify the findings. Deshors [31] employs the covarying collexeme analysis to precede the CA to prove the complex interactions of more than 6000 progressive constructions in five corpora.

As such, this study intends to contribute to the body of knowledge by analysing significant nouns, extracted from the key-keywords lists of different genres of the specialised language, and showing their associations in the CA. The observation defines the nouns in the specialised language more precisely, highlighting those which can be further scrutinised to inform the discipline specificity in terms of its linguistics features, rhetorical functions, and pedagogical implications.

III. METHOD

A.The Corpora

Two main corpora were used in this study: the Engineering English Corpus (E2C), and, a reference corpus - the British National Corpus (BNC).

Representing the Engineering English for the study, E2C compiles 102 texts, with 677,993 words. E2C comprises two genres (sub-corpora) of the Electronics and Computer Engineering English. They are written texts of suggested reference books (for Electronics and Computer Engineering) and online journals articles. These two sub-corpora are labelled as Reference Books Corpus (RBC) and Engineering Journals Corpus (EJC) respectively.

RBC contains suggested reference books, selected from a handbook of an Electronics and Computer Engineering faculty of a local university. Only two books were considered for the study to ensure manageability. The chapters from the books made up 34 texts, with 425,854 tokens. Next, EJC are articles, collected from online engineering journals, which were retrieved from four databases: ASME Online journals, ScienceDirect, IEEE Xplore and Wilson Applied Science & Technology. The articles were selected based on the chapter titles of the reference books. The total of articles for this sub-corpus is 68, with 252,139 tokens. Table I presents the distribution of the articles retrieved according to the databases.

Hence, there are 102 articles in E2C, with 677,993 tokens. Table II provides the composition of E2C based on the two sub-corpora.

For this study, BNC is the reference corpus, used to identify any distinctive lexical behaviors in E2C.

Table I: The distribution of EJC retrieved from databases

Databases	Book 1	Book 2	Total
ACME Online journals	9	8	17
ScienceDirect	8	9	17
IEEE Xplore	9	8	17
Wilson Applied Science & Technology	8	9	17
Total	34	34	68

Table II: The composition of E2C

Sources	No. of texts	Running Words
Reference Books	34	425,854
Journal Articles	68	252,139
Total	102	677,993

B. Software

The *Wordsmith* software was used in this study to extract the key-keywords from E2C. Another software used in the study was XLSTAT; a statistical analysis software which is compatible with the Microsoft Excel program. All the tools and functions of this software can be accessed from the Excel toolbars and menus. XLSTAT was employed particularly to perform the CA. The CA output includes hypotheses tests, data analysis models and data visuals. The visualisation of the results facilitates the profiling of E2C by displaying words which characterise the corpus the most.

IV. DISCUSSION OF FINDINGS

A. The Key-Key-Nouns of E2C

Using the Wordsmith software, 30 nouns were extracted from the E2C key-keyword list. The raw frequencies of the nouns were derived from the keyword list. Table III gives the key-key-nouns including their frequencies and number of texts in which they occur in E2C.

Table III: Frequency and no. of texts occurrence of 30 key-key-nouns in E2C

NOUNS	Freq	%	Texts	(%)
CIRCUIT	4064	0.68	82	80.39
VOLTAGE	6051	1.01	85	83.33
CURRENT	4347	0.72	86	84.31
OUTPUT	3753	0.62	80	78.43
TRANSISTOR	2391	0.40	67	65.69
CIRCUITS	1209	0.20	76	74.51
AMPLIFIER	1474	0.25	59	57.84
INPUT	2822	0.47	80	78.43
SIGNAL	2169	0.36	71	69.61
GAIN	2224	0.37	66	64.71
TRANSISTORS	969	0.16	60	58.82
BIAS	1090	0.18	66	64.71
EMITTER	1178	0.20	49	48.04
VOLTAGES	552	0.09	64	62.75
DEVICE	1348	0.22	76	74.51
RESISTANCE	1693	0.28	65	63.73
DIODE	1371	0.23	55	53.92
SOURCE	1570	0.26	95	93.14
RESISTOR	673	0.11	57	55.88
FIGURE	2252	0.37	77	75.49
FREQUENCY	1470	0.24	70	68.63
LOAD	1461	0.24	62	60.78
CAPACITOR	660	0.11	55	53.92
PARAMETERS	760	0.13	64	62.75
CHARACTERISTICS	902	0.15	72	70.59
VALUE	1090	0.18	83	81.37
DESIGN	946	0.16	83	81.37
POWER	1515	0.25	83	81.37
DEVICES	675	0.11	67	65.69
VALUES	672	0.11	68	66.67

Further relationships of all the 30 nouns in all the four corpora (including BNC) in this study is performed using the correspondence analysis (CA).

B. CA of Nouns

To run the CA, the frequencies of the same words were obtained from BNC, EJC and RBC. The frequency lists of nouns from all the corpora are as shown in Table IV.

The visual representations or CA maps of the key-key-nouns are in Fig. 1. Table V provides the inertia values of both axes in Fig. 1. It shows that the inertia values of the nouns are 97.7% along the F1 axis and 2.3% along the F2 axis. This means that the CA of these nouns is of good quality with 100%.

The values also indicate that the differences between the corpora are mainly along the F1 axis. Therefore, the main differences among the corpora can be described from the information or words along the F1 axis.

Table IV: Frequency of nouns for CA

Nouns	E2C	BNC	EJC	RBC
CIRCUIT	4,064	2,615	686	3,378
VOLTAGE	6,051	918	870	5,181
CURRENT	4,347	14,212	821	3,526
OUTPUT	3,753	6,077	553	3,200
TRANSISTOR	2,391	254	281	2,110
CIRCUITS	1,209	620	245	964
AMPLIFIER	1,474	333	273	1,201
INPUT	2,822	3,684	307	2,515
SIGNAL	2,169	3,127	282	1,887
GAIN	2,224	5,155	366	1,858
TRANSISTORS	969	114	224	745
BIAS	1,090	1,397	250	840
EMITTER	1,178	38	209	969
VOLTAGES	552	90	98	454
DEVICE	1,348	2,861	581	767
RESISTANCE	1,693	3,641	276	1,417
DIODE	1,371	106	262	1,109
SOURCE	1,570	9,038	336	1,234
RESISTOR	673	127	172	501
FIGURE	2,252	17,214	378	1,874
FREQUENCY	1,470	2,803	349	1,121
LOAD	1,461	3,096	193	1,268
CAPACITOR	660	206	137	523
PARAMETERS	760	1121	193	567
CHARACTERISTICS	902	3,749	128	774
VALUE	1,090	17,758	210	880
DESIGN	946	12,852	356	590
POWER	1,515	31,627	656	859
DEVICES	675	2172	394	281
VALUES	672	7575	180	492

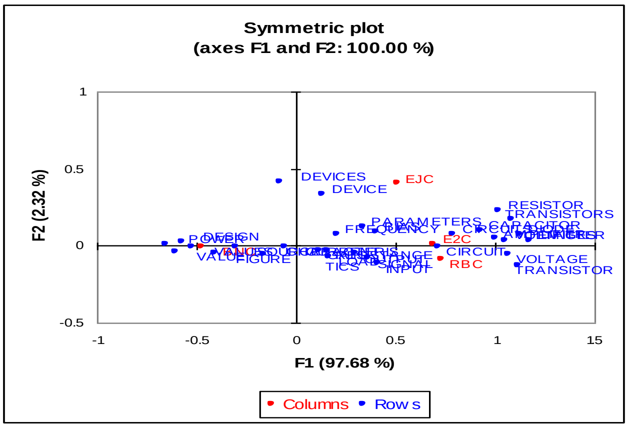

Fig. 1(a): CA map of nouns (columns and rows)

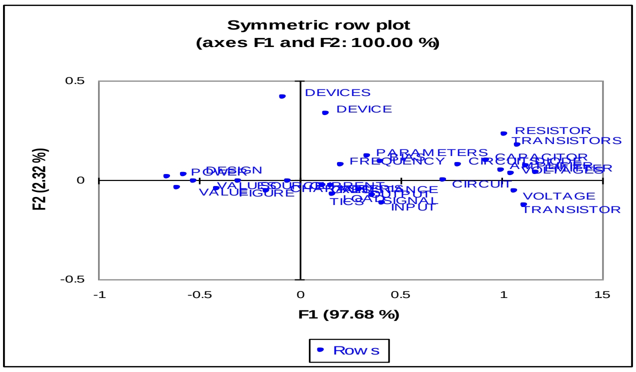

Fig. 1(b): CA map of nouns (rows)

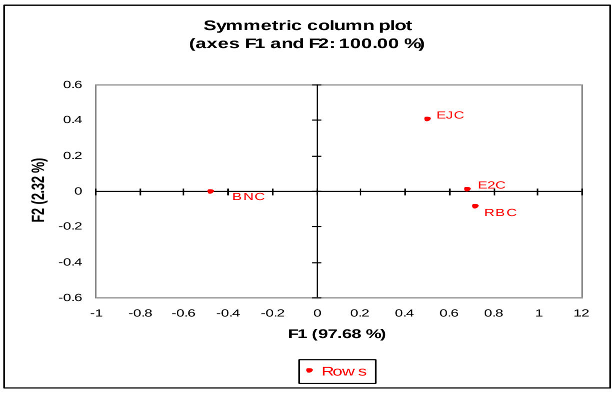

Fig. 1(c): CA map of nouns (columns)

Fig. 1: CA maps of nouns

Table V: Eigenvalues and percentages of inertia

	F1	F2
Eigenvalue	0.325	0.008
Inertia (%)	97.681	2.319
Cumulative (%)	97.681	100.00

As seen in Fig. 1(a), the 30 nouns offer more interesting information on the distribution of the words on the map. The coordinates of both the columns (corpora) and rows (nouns) on the map are respectively listed in the following Tables VI and VII.

Table VI: Coordinates of corpora

	F1	F2
E2C	0.685	0.007
BNC	-0.473	-0.005
EJC	0.510	0.404
RBC	0.726	-0.088

Retrieval Number L25171081219/2019©BEIESP
DOI: 10.35940/ijitee.L2517.1081219

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table VII: Coordinates of nouns

Nouns	F1	F2	Nouns	F1	F2
CIRCUIT	0.71	-0.002	RESISTANCE	0.152	-0.03
VOLTAGE	1.066	-0.055	DIODE	1.126	0.073
CURRENT	-0.058	-0.008	SOURCE	-0.307	-0.005
OUTPUT	0.297	-0.05	RESISTOR	1.015	0.231
TRANSISTOR	1.112	-0.125	FIGURE	-0.407	-0.042
CIRCUITS	0.785	0.076	FREQUENCY	0.206	0.079
AMPLIFIER	0.996	0.049	LOAD	0.162	-0.071
INPUT	0.409	-0.115	CAPACITOR	0.924	0.099
SIGNAL	0.357	-0.078	PARAMETERS	0.333	0.122
GAIN	0.114	-0.029	CHARACTERISTICS	-0.166	-0.057
TRANSISTORS	1.081	0.176	VALUE	-0.607	-0.041
BIAS	0.404	0.091	DESIGN	-0.573	0.028
EMITTER	1.171	0.038	POWER	-0.655	0.015
VOLTAGES	1.05	0.032	DEVICES	-0.079	0.417
DEVICE	0.134	0.335	VALUES	-0.525	-0.004

Fig. 1(c) clearly shows the difference between the specialised corpora and General English (BNC). RBC is found to be the most specific among the specialised corpora from the use of nouns. However, with these 30 nouns, the BNC point (-0.473) is seen further from the other specialised corpora. This implies that the use of the nouns further distinguishes the three corpora (E2C, EJC and RBC) from General English.

Nevertheless, there is not much difference with the corpora points on the F2 axis. EJC and RBC undoubtedly represent two different genres in the Engineering English. The EJC point (0.404) is further apart from RBC (-0.088). The use of nouns, apparently, further distinguishes even the corpora of the same specific domain.

An interesting observation on the F2 axis is that RBC (-0.088) has a closer coordinate to BNC (-0.005); they also share the same negative quadrant. This suggests that on the F2 axis, the RBC shows closer association to BNC than EJC. It implies that RBC has more general nouns than EJC. However, as shown in the earlier inertia values, the main difference among the corpora is largely contributed by the information along the F1 axis.

The contribution values of the corpora on both axes are displayed in Table VIII. A contribution value is the percentage of variance (inertia) of an axis, which is explained by the point. Table VIII shows that on the F1 axis, BNC has the highest contribution value of 40.7% or 0.407, followed by E2C 29.4% and RBC 26.8%. This means that these three corpora account for the most total of information on F1 axis, and EJC has the lowest contribution value at 3.1%.

On the other hand, the F2 axis sees a higher contribution value of EJC (83.3% or 0.833). RBC has a lower contribution on this axis (16.4% or 0.164).

Table VIII: Contribution values of corpora

	F1	F2
E2C	0.294	0.001
BNC	0.407	0.002
EJC	0.031	0.833
RBC	0.268	0.164

Nonetheless, the value is sufficient to show the existing differences between EJC and RBC on the F2 axis in terms of the 30 nouns. E2C and BNC have very little contribution on this axis with only 0.1% or 0.001 and 0.2% or 0.002. These low values further signify the difference of EJC and RBC on the F2 axis.

Next, the contribution values of the corpora on the axes direct the analysis on significant words that characterise each corpus in the study. Table IX lists the 16 nouns which contribute to the differences between the specialised corpus (E2C and sub-corpora: EJC and RBC) and General English. The values indicate that *power* has the highest contribution (17.5%), followed by *voltage* (17.4%), *value* (8.7%), *transistor* (7.3%), *circuit* (6.4%) and *design* (5.7%). The rest of the nouns have contribution values lower than 0.050 or 5%: *circuits*, *amplifier*, *emitter*, *voltages*, *diode*, *source*, *figure*, *capacitor*, and *values.* However, *current* shows equal low contribution on both axes (0.001 or 0.1%).

Table IX: Nouns that contribute to F1

NOUNS	F1	F2	NOUNS	F1	F2
CIRCUIT	0.064	0	DIODE	0.042	0.007
VOLTAGE	0.174	0.02	SOURCE	0.013	0
CURRENT	0.001	0.001	FIGURE	0.042	0.019
TRANSISTOR	0.073	0.039	CAPACITOR	0.015	0.007
CIRCUITS	0.022	0.009	VALUE	0.087	0.017
AMPLIFIER	0.038	0.004	DESIGN	0.057	0.006
EMITTER	0.039	0.002	POWER	0.175	0.004
VOLTAGES	0.016	0.001	VALUES	0.029	0

Fig. 1(b) and Table IX, however, reveal that *voltage* (1.066), *transistor* (1.112), *emitter* (1.171), *voltages* (1.050) and *diode* (1.126) are differentiated from *figure* (-0.407), *source* (-0.307), *value* (-0.607), *design* (-0.573), *power* (-0.655) and *values* (-0.525) along the F1 axis. The big gap between these two groups of nouns suggests that *voltage*, *transistor*, *emitter*, *voltages* and *diode*, which have coordinates more than 1.0, distinctly mark the differences between the specialised corpora and General English. On the other hand, *figure*, *source*, *value*, *design*, *power* and *values*, which have coordinates more than -0.4, provide the profiling of General English.

In contrast, Table X gives the 14 nouns which distinguish EJC from RBC. The nouns which are arranged according to the contribution values include *device* (30.9%), *devices* (30.3%), *input* (6.1%), *resistor* (3.9%), *transistors* (3.2%), *signal* (2.2%), *parameters* (2%), *frequency* (1.8%), *output* (1.7%), *bias* (1.5%), *load* (1.5%), *characteristics* (0.9%), *gain* (0.4%) and *resistance* (0.3%).

Interestingly, Fig. 1(b) and Table X provide another perspective on the form of contribution of the nouns along the axis. From Fig. 1, it appears that *device* (0.335) and *devices* (0.417) scatter very closely to EJC (0.404); thus, they are differentiated from *voltage* (-0.055) and *transistor* (-0.125) on the other side of the quadrant, which are closer to RBC (-0.088). It is also found that there are more nouns scattering around E2C (0.007) and RBC than there are around EJC. This, again, seems to support the assumption that RBC has greater influence on E2C than EJC.

Other nouns which coordinates are closer to RBC include *circuits* (-0.076), *signal* (-0.078), *output* (-0.050), *input* (-0.115), *load* (-0.71) and *characteristics* (-0.057).

Table X: Nouns that contribute to F2 axis

Nouns	F1	F2
OUTPUT	0.014	0.017
INPUT	0.018	0.061
SIGNAL	0.011	0.022
GAIN	0.001	0.004
TRANSISTORS	0.028	0.032
BIAS	0.007	0.015
DEVICE	0.001	0.309
RESISTANCE	0.002	0.003
RESISTOR	0.018	0.039
FREQUENCY	0.003	0.018
LOAD	0.002	0.015
PARAMETERS	0.003	0.02
CHARACTERISTICS	0.002	0.009
DEVICES	0	0.303

The results thus far strongly show that E2C appears very closely to RBC on both axes, thus, suggesting that E2C is composed more of the features of RBC than EJC.

V. CONCLUSION

The findings from this paper reveal that in the Engineering English (E2C), different genres possess different sets of nouns that strongly make up the features of one genre to the other. Using the CA, the key-key-nouns in E2C prove to differentiate the specialised language from General English (BNC). The technique also reveals how the nouns specifically characterise the genres even in E2C by looking at the coordinates of the nouns on the axes and their contribution values. Most interestingly, these differences are presented visually, thus, making the understanding of the differences easier and faster. In addition, the specialised language can be described more precisely.

As such, the CA proves to be another potential tool to display significant words which characterise not only a specialised corpus from a general corpus, but also the different genres in that specialised corpus. The complex interrelationships of the words in all the corpora are clearly presented in the graphical presentation of the words. CA also allows the identification of specific composition of a specialised corpus in terms of groups of words.

This study promises more informed descriptions of a specialised language can be made with the identification of specific and significant vocabulary for any academic discourse investigations. Future studies may benefit from this knowledge by focusing on the lexical properties of the words and their behaviours in a specialised language. As such, more meaningful descriptions of a specialised language can be made for the benefits of ESP classrooms.

ACKNOWLEDGMENT

We would like to thank the Centre for Technopreneurship Development (CTeD), Universiti Teknikal Malaysia Melaka (UTeM) and Malaysian Examination Council (MEC) in providing support for this paper. This paper is funded by the grant, **GLuar/MPM/2017/PBPI-CTED/I00028**.

REFERENCES

1. Z. Fang, M. J. Schleppegrell and B. E. Cox, "Understanding the Language Demands of Schooling: Nouns in Academic Registers," *Journal of Literacy Research,* vol. 38, no. 3, p. 247–273, 2006.
2. G. Lan and Y. Sun, "A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course," *Journal of English for Academic Purposes,* Vols. 38,, pp. 14-24, 2019.
3. D. Biber and B. Gray, "Grammatical change in the noun phrase: The influence of written language use.," *English Language and Linguistics,* pp. 223-250, 2011.
4. P. Peters and T. Fernández, "The lexical needs of ESP students in a professional field," *English for Specific Purposes,* vol. 32, p. 236–247, 2013.
5. W. Lu, S.-M. Lee and S. Jhang, "Keyness in maritime institutional law texts.," *Linguistic Research,* vol. 34, no. 1, pp. 51-76, 2017
6. L. Lei and D. Liu, "A new medical academic word list: A corpus-based study with enhanced methodology.," *Journal of English for Academic Purposes,* vol. 22, no. L, pp. 42-53, 2016.
7. S. Ariwibowo and H. Tedjasuksmana, "The Variations of the English Noun Postmodifiers in The Undergraduate Students' Compositions," *Beyond Words,* vol. 6, no. 1, pp. 52-61, 2018.
8. E. E. Işık-Taş, "Nominal stance construction in IELTS tests.," *Journal of English for Academic Purposes,* vol. 34, pp. 1-11, 2018.
9. M. Charles, "Argument or evidence? Disciplinary variation in the use of the Noun that pattern in stance construction," *English for Specific Purposes,* pp. 203-218, 2007.
10. J. Dong and L. Buckingham, "The collocation networks of stance phrases.," *Journal of English for Academic Purposes,* vol. 36, pp. 119-131, 2018.
11. Jalilifar, P. White and N. Malekizadeh, "Exploring nominalization in scientific textbooks: A cross-disciplinary study of hard and soft science," *International Journal of English Studies,* vol. 17, no. 2 , p. 1–20, 2017.
12. M. Currie, K. Sinwongsuwat and K. Nicoletti, "Non-Conventional English Language Use in the Writing and Speech of Thau Academic Writers: A Preliminary Study.," *E.S.P. Today,* vol. 4, no. 2, pp. 251-264, 2016.
13. C. E. Snow, "Academic Language and the Challenge of Reading for Learning About Science.," *Science,* vol. 328, pp. 450-452., 2010
14. M. K. Tåqvist, ""A wise decision": Pre-modification of discourse-organising nouns in L2 writing," *Journal of Second Language Writing,* vol. 41, pp. 14-26, 2018.
15. M. A. Halliday and J. R. Martin, Writing science: Literacy and discursive power., London: The Falmer Press, 1993.
16. M. Ryshina-Pankova, "A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor.," *Journal of Second Language Writing,* vol. 29, pp. 51-63, 2015.
17. S. Graham, S.-Y. Kim, D. DeVasto and W. Keith, "Statistical Genre Analysis: Toward Big Data Methodologies in Technical Communication.," *Technical Communication Quarterly,* vol. 24, p. 70–104, 2015.
18. S. C. Herring and J. C. Paolillo, "Gender and genre variation in weblogs," *Journal of Sociolinguistics,* vol. 10, no. 4, 2006.
19. H. Moisl, "Exploratory Multivariate Analysis," in *Corpus Linguistics. An International Handbook*, L. A. and K. M., Eds., Berlin, Mouton de Gruyter, 2009, pp. 874-899.
20. S.-w. Chan, "The Human Factor in Machine Translation. Terminology.," *International Journal of Theoretical and Applied Issues in Specialized Communication,* vol. 25, no. 1, pp. 121 - 127, 2018.
21. B. Shao, Y. Cai and G. Trousdale, "A Multivariate Analysis of Diachronic Variation in A Bunch of noun: A Construction Grammar Account," *Journal of English Linguistics,* vol. 47, no. 2, pp. 150-174, 2019.
22. N. Döring and S. Poeschl, "Love and Sex with Robots: A Content Analysis of Media Representations," *International Journal of Social Robotics,* pp. 1-13, 2019.
23. B. Jin, "A Multidimensional Analysis of Research Article Discussion Sections in the Field of Chemical Engineering," *IEEE Transactions on Professional Communication,* vol. 61, no. 3, pp. 242-256, 2018.

24. M. Richarme. (2002). Decision Analyst. [Online]. Available: https://www.decisionanalyst.com/whitepapers/multivariate/.
25. K. Noorli and A. Imran-Ho, "What Do Different Word Lists Reveal about the Lexical Features of a Specialised Language? .," *3L: The Southeast Asian Journal of English Language Studies,* p. 26 – 42, 2018.
26. O. Mudraya, "Engineering English: A lexical frequency instructional model," *English for Specific Purposes,* vol. 25 , pp. 235-256, 2006.
27. Y. Wu, *A Corpus-based Register Analysis of Corporate Blogs – Text Types and Linguistic Features.,* Liverpool.: University of Liverpool., 2016.
28. M. Greenacre, Correspondence Analysis in Practice, Boca Raton: Taylor & Francis Group, 2017.
29. D. L. Mealand, "Measuring genre differences in Mark with correspondence analysis," *Literary and Linguistic Computing,* p. 227–245, 1997.
30. S. C. Deshors, "Zooming in on Verbs in the Progressive: A Collostructional and Correspondence Analysis Approach," *Journal of English Linguistics ,* vol. 45, no. 3, p. 260 –290 , 2017.
31. S. Hall-Millsa and K. Apela, "Linguistic Feature Development Across Grades and Genre in Elementary," *Language, Speech, and Hearing Services in Schools,* vol. 46, p. 242–255, 2015.

AUTHORS PROFILE

Noorli binti Khamis, PhD, is a Senior Lecturer at Universiti Teknikal Malaysia Melaka (UTeM). She has 22 years of experience in teaching English at different institutions in Malaysia. Her academic achievements include a B.Ed. TESL, M.Ed. TESL and PhD in English Studies. Her research interests are corpus linguistics and ESP. Her study areas thus far revolve around lexical investigation of specialised languages. Her teaching background in a local technical university has triggered her interest in discovering more about the features of specialised languages to better understand the languages and, hence, prepare better for its teaching and learning. She has presented papers in several national and international conferences, and written research articles in international journals.

Imran Ho Abdullah, PhD, is a Professor of Cognitive and Corpus Linguistics at the School of Language Studies and Linguistics, UKM. His research interests are corpus linguistics and cross-cultural semantics/cognitive semantics and the natural extension of these interests to translation and the use of corpus methodologies in Translation Studies. Currently, he is the Deputy Vice-Chancellor (Industry and Community Partnership) at Universiti Kebangsaan Malaysia (UKM). His has vast and significant publications in national and international journals.