

# Developing a Corpus of Entrepreneurship Emails (COREnE) for Business Courses in Malaysian University Using Integrated Moves Approach

Nur Baiti Ismail Shauki<sup>\*</sup>, Manvender Kaur Sarjit Singh

*School of Languages, Civilisation & Philosophy, Awang Had Salleh Graduate School of Arts and Sciences,  
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia*

\*Corresponding author: nur\_baiti@ahsgs.uum.edu.my

**Article history:** Received: 3<sup>rd</sup> August 2021 | Received in revised form: 24<sup>th</sup> September 2021 | Accepted: 9<sup>th</sup> December 2021 | Published online: 1<sup>st</sup> January 2022

## Abstract

This paper describes the processes that went into the design and development of the Corpus of Entrepreneurship Emails (COREnE), a specialised corpus that contains pedagogical sales emails written by undergraduates at a Malaysian public university. It describes in detail the processes involved in data collection and selection, as well as raw data processing, which includes data coding and tagging. The coding and tagging utilises online softwares; CLAWS Tagger and AntConc. The study adopted Integrated Moves Approach (IMA) which combines both structural and syntactical analysis. This paper also discusses the evolution of corpora in Malaysia and their relevance to the field of English for Specific Purposes (ESP). COREnE was created with the goal of providing corpus linguistic researchers with a place to explore the field, as well as ESP practitioners in Malaysia with resources for developing local-based ESP curriculum and teaching and learning materials. COREnE has approximately 18 thousand words from 89 electronic documents retrieved from eCampus website and the corpus also provides lexical analysis on students CEFR (Common European Framework of Reference) level. The finding suggests that student's proficiency level is at the target level of CEFR descriptor. At present, users can make queries to access COREnE from the researcher for free.

**Keywords:** Corpus analysis, CEFR, Corpus of Entrepreneurship Emails, English for Specific Purposes, Integrated Moves Approach (IMA), Sales emails

## Abstrak

Kajian ini menerangkan proses dan perkembangan Corpus of Entrepreneurship Emails (COREnE), iaitu sebuah korpus khusus yang mengandungi e-mel perniagaan yang ditulis oleh pelajar-pelajar di sebuah universiti awam Malaysia. Ia menerangkan secara terperinci tentang proses yang terlibat dalam pengumpulan dan pemilihan data, serta pra-pemrosesan data mentah, yang merangkumi pengekodan dan penandaan data. Proses pengekodan dan penandaan menggunakan perisian dalam talian; CLAWS Tagger dan AntConc. Kajian ini dijalankan menggunakan Integrated Moves Approach (IMA) yang menggabungkan analisa struktur dan sintaks. Kajian ini juga membincangkan evolusi korpora di Malaysia dan kaitannya dengan bidang English for Specific Purposes (ESP). COREnE dibina dengan tujuan untuk menyediakan penyelidikan linguistik korpus tempat untuk menerokai bidang ini dengan lebih mendalam, serta menyediakan sumber bagi pengamal ESP di Malaysia untuk mengembangkan kurikulum dan bahan pengajaran dan pembelajaran ESP berasaskan analisa lokal. COREnE mempunyai 18 ribu perkataan dari 89 dokumen elektronik yang diambil dari laman web eCampus dan korpus ini juga memberikan analisa leksikal tahap CEFR (pelajar). Dapatan kajian merumuskan pencapaian pelajar adalah menepati tahap CEFR. Pada masa ini, pengguna boleh memohon untuk mengakses COREnE dari penyelidik secara percuma.

**Kata kunci:** Analisa Korpus, CEFR, Corpus of Entrepreneurship Emails, English for Specific Purposes, Integrated Moves Approach (IMA), Email Perniagaan

© 2022 Penerbit UTM Press. All rights reserved

## 1.0 INTRODUCTION

Corpus linguistics has been a major research subject for the past 40 years due to its natural occurrences compiled from large samples. Biber and Reppen (2015, p. 2) in the introduction of *The Cambridge Handbook of English Corpus Linguistics* state that many linguists believe corpus linguistics emerge in the 1980s although its dominant practice has started as early as in the 1960s. That is if the pre-computer corpora are to be excluded from the discussion. Samuel Johnson's *Dictionary of the English Language* published in 1755 is of those pre-computer corpora pioneers.

Leech (1998, p. xvi) mentions that the computer corpus linguistics era has gathered its momentum over the past 30 years in congruent with the growth of the usage of computers in storing and processing data. He went on further by describing the slow progress of computerised corpus in the education sphere and the currents that steer away researchers from the field. One of the current that Leech (1998) specifically stated is the emphasis of computerised corpus on the process of communicative teaching and learning, rather on the product of teaching and learning which is what learners write and say. Despite many debates on the history of corpora development,

researchers mutually agree that corpus linguistics is one of the comprehensive methods in data compilation in the field of contemporary linguistics (Hunston, 2006; Davies, 2007; Gries, 2009; Nesselhauf, 2011; Subtirelu & Baker, 2017; Grundmann, 2021).

There are various perspectives of learner corpus formation given by the experts of linguistics, however, this study adhered closely to Biber's et. al. (1998) four significant principles which are:

- 1) exploring patterns of language used in natural texts
- 2) the use of a major original collection of texts (corpus) as the fundamental basis for analysis
- 3) the basis is to describe the quantitative results and to enable the application of language patterns in the context of real life
- 4) the use of the corpus tool that accommodates such an analysis can contribute a comprehensive explanation of the textual collection

Furthermore, previous practises of specialised corpus formation have aided the work. Much of the design framework, particularly in data compilation (i.e., setting external criteria and text categories), adheres to the framework established by Sadjirin et al. (2018) in developing Malaysian Corpus of Financial English (MaCFE). Nonetheless, some design changes were required as needed; for example, the text categories finalised in COREnE did not include some of the text categories used in MaCFE due to confidentiality and accessibility concerns.

Corpus linguistics provides various benefits to the study of language. Corpora bring an empirical dimension to language studies which were not possible before due to the constraint of technology and time, giving an opportunity for researchers to fortify elucidations, emphasize important points, and discuss language with an improved understanding and knowledge (Romer & Wulff, 2010; Gries, 2015; Almutairi, 2016). A large collection of texts is needed to develop a corpus with varying purposes and contents. It is used for many kinds of investigations and usually carries information such as details of the text (Bowker & Pearson, 2002), part of speech tags (McEnery, Xiao & Tono, 2006), and analysis information (Anthony, 2016).

Although corpus has long been debated on its relevancy in linguistics field and researched worldwide, its development in the Malaysian educational setting is rather late. The earliest corpus study in Malaysia can be traced back to the year 1996 with the integration of a corpus study of the preposition *by* in the Fiction Section of the Wellington Corpus of Written New Zealand English (NZE) and a corpus of Malaysian Short Stories (ME) and since then, there are only 42 studies on corpus found among others are MEN Corpus and EMAS. This is according to Joharry and Rahim (2014), as mentioned in their bibliographic analysis study, "The process of ploughing through the 100-page Google Scholar citations found only 42 corpus-related studies on either Malay or English in Malaysia (p. 2)". From the discussion in the study, there are only two genre-specific learner corpora mentioned, which are the Business and Management English Language Learner Corpus (BMELC) compiled by Singh, Zaid and Shamsudin in (2012a & 2012b) and the legislative corpus of Kamariah and Su'ad (2011). Moreover, the reason why a genre-specific learner corpus is much needed to be developed is due to "different types of writing skills and strategies are needed to perform different types of writing tasks (Joharry and Rahim, 2014, p. 3)." Thus, the significant effect that a genre specific corpus imposed on different types of pedagogical writing strategies ensures the need for this study in Business courses in the Malaysian tertiary education field.

Having sufficient vocabulary and the vocabulary used is of the required frequency is essential in ascertaining successful communication especially in pedagogic writing and later on benefit the students as well as the society as a whole. Sufficient vocabulary according to Schmitt (2000), (as quoted in Kaur, 2013, p. 113) is "about 3,000 spoken word families are required to cover about 96% of one's daily communication lexicon use or about 2,000 words to maintain conversations". According to Tostlebe (2011), communication needs context and intended audience and thus this sets the premise on any writing tasks. On the other, frequency of vocabulary used plays a significant part in moderating linguistic behavior and sociolinguistics variations (Calude and Pagel, 2014). Therefore, when learners' vocabulary knowledge is at the standard level, they would have a positive attitude towards learning and strive for language proficiency and they can explicitly excel in various language functions be it writing or any other linguistics skills. The vocabulary profiles that the current study wished to compile are hoped to aid teachers monitor vocabulary of suitable frequency to be introduced in writing class. Apart from that, materials and writing activity selection for the lesson can be devised accordingly once the word level of the students is identified.

In the field of corpus, there is a serious lack of learner corpora covering the field of languages for specific purposes (Granger, 2008; Vicic, 2008; Myles, 2005). Besides its lacking, there are multiple concerns on learner corpora analysis. Kotani and Yoshimi (2015) claim that there are lacking of learner corpora focusing on speaking skills while Tono (2003) disputes that there is issue with the standardisation of corpus annotation. These arguments over the functions and benefits of learner corpora however, could possibly be resolved through understanding the definition of learner corpora provided by the prominent figures in this field. Researchers nevertheless defines learner corpus as a systematic computerized textual database of the language produced by language learners (Leech, 1998; Nesselhauf, 2004; Meurers, 2015). A more recent study on learner corpora by Cirillo (2020, p. 150) added, reliable information on frequency and lexical distribution are among the advantages of learner corpora. These definitions provided by influential figures in linguistics field show that there is standardization in terms of corpus annotation as the corpora are collected and organized using reliable technology and software. The AntConc software for example has nearly 90% error free rate and high reliability. The lacking of speaking corpora which is due to the difficulty in transcribing data can relatively be reduced using the advancement of technology and devices available nowadays. Prior to deciding whether to build a spoken or written corpus, researcher needs to determine what is their focus and how the corpus will be utilised in the English for Specific Purposes (ESP) classroom.

In general, ESP is concerned with the lessons applied in higher learning institution that contributes to the preparation of work-related elements such as the professional vocabulary and its community. Hutchinson and waters (1987) regarded ESP as "an approach to language teaching which all decisions as to content and method are based on the learner's reason for learning". While Bennett (2010, p.14) describes learner corpora in ESP as having two specific characteristics which are (1) comprising of transcripts of written and/or spoken language utilized by students who are learning the language, and (2) are often tagged and can be analysed, for instance, to see common errors the students made.

This study is developed from data compiled from first year and second year students enrolled in a degree program in Universiti Malaysia Kelantan (UMK). To accommodate the sample of data taken from this institution, where the core value instilled in the undergraduates is entrepreneurship skills, the term entrepreneurship genre is interwoven with business genre definition. Furthermore, research in this field fulfils the current trend of higher education in which the changes definitely asked for certifying workplace-based research to accommodate the cycle. Thus, as the country is moving towards empowering entrepreneurship values in all levels of education, the current study aids in providing a comprehensive structural analysis of sales email and developing vocabulary profile for an inclusive

entrepreneurship corpus. Hence, in preparing the students with adequate digital and data literacy, it is also the responsibility of the teachers to improve the methods and the tools they are using. Rather than depending on the traditional dictionary and non-standardised writing structure to be applied in their lesson, educators must be equipped with the latest advancement of technology which is what this study implores to produce, a learner corpus which was built specifically for the business course utilisation.

As mentioned earlier, the dearth of genre-specific learner corpora in Malaysia especially in pedagogical setting of business and entrepreneurship courses is appalling. Genre theories generally have three perspectives which are the New Rhetoric, the Systemic Functional Linguistics and the English for Specific Purposes. The similarity that these genre theories share, in various scholars' points of view, is each of them relate its definition with communicative purposes (Bhatia, 2004; Chandler 1997; Martin, 1997; Swales, 1990).

In this study, the type of text and its function in Corpus of Entrepreneurship Emails (COREnE) shows the sole communicative event and purpose, integrating entrepreneurship traits within the student's writing. The realization of the genre as a more extended approach and the highlights given to social functions are the significant elements in shaping a genre-based methodology. As divulged by Myskow & Gordon (2010, p. 284), the first step of teaching a genre-based writing is to form a social context with the texts then the focus can shift to linguistics and rhetorical features. Hence, as a promotional genre, the pedagogical sales email obliges to catch readers' attention, to build the writer's ability to speak to their lecturer's needs and expectations, and to bridge the link between the author and the field of entrepreneurship.

Electronic mail, or simply known as email, has revolutionized the way the business world functions and companies operate (Rice & Bair, 1984; Safayenj, Lee & MacGregor, 1988). Nonetheless, it has gained little scientific observation in English for Specific Purposes (ESP) researches and is rarely addressed in studies of Business English (Kankaanranta, 2006; Evans, 2012). Lack of research dwelling in the application of email in global business settings could be the contributing factor behind the gap that separates offices and classes despite the fact that office workers are generated from the classroom. Therefore, this study will examine the style of writing and sentence structure used by the entrepreneurial students in improving their writing skills to be applied in professional fields.

Despite the common need for the students to produce a substantial range of written genres which can incorporate business writing of various genres – complaint, persuasive or negotiation – the gap in the research literature with respect to pedagogical sales email to support the teaching and learning of Business English is still broad. Despite the scarcity of genre specific researches especially in pedagogical sales emails, studies of textual analysis have seen an interesting trend of different types of business correspondence by various researchers; for instance, Naksawat & Akkakoson (2015) on scam emails, Mehrpour & Mehrzad (2013) on business letters, Cheung (2011) on sales promotional letters and Gimenez (2006) on professional business emails.

Upton & Cohen (2009, p. 601) mention, 'Once the move structure for each of the letters in the corpus was identified, it was then possible to look at what the typical move structure patterns are for these types of letters.' For the current research, move investigation is chosen as a tool for corpus building due to its capacity to perceive distributional patterns of students' sentence structures. Move examination as indicated by Ding (2007, p. 370), "is a helpful tool in genre studies since moves are semantic and functional units of texts, which can be identified because of their communicative purposes and linguistic boundaries". Numerous researchers utilize the terms 'genre analysis' and 'move analysis' reciprocally in their writing. For instance, Nodoushan, (2014, p. 112) notes "More recently, some researchers reported on studies that implement the findings of genre-/move-based text analysis investigations in the teaching of writing".

This research reports a study exploring the linguistic features of sales inquiry email and sales reply email of undergraduate students writing. The features examined were based on Bhatia's (1993) sales promotion letter approach that listed seven (7) moves which are explained in detail in next section.

The reason for adopting Bhatia's (1993) framework is that it is comprehensive and easy to use as an analysis framework because each move is clearly defined from moves 1 to moves 7. Bhatia has provided examples for each move to help audience identify moves more concisely. Furthermore, the framework was created using business sales letter in a professional setting, thus it will be a great writing guidance for Business course students. Using this comprehensive model, a list of moves and steps were identified using a representative corpus of Business courses undergraduates.

## ■ 2.0 METHODOLOGY

This study is using purposive sampling throughout, that means purposive sampling is not limited to the respondents but it is also utilised when choosing the industry informants and subject specialists. Creswell & Clark (2011) note that purposive sampling deals with the selection of individuals or groups of individuals who are especially knowledgeable about or experienced with a phenomenon of interest.

All of the students have enrolled in Business English class or the subject code name is English II, which is a required course for them to complete before registering themselves in English III class in the following semester. Business English or English II is a requirement subject for the intermediate level students who have completed English I or have obtained a Band 3 and above in the Malaysian English University Test (MUET). This course focuses on workplace interaction and the focused skills are aimed to polish students' English proficiency in the workplace setting. Among the topics covered in this subject are email writing, vocabulary for describing people and objects, and some basic grammar. The ultimate reason for selecting students of English II subject is due to the noticeable lack of description in the assessment and topics covered that is supposed to enhance students' entrepreneurship traits.

The features analysed were based on Bhatia's (1993) sales promotion letter framework that consists of seven (7) moves which are:

- 1) Establishing credentials,
- 2) Introducing the offer,
- 3) Offering incentives,
- 4) Enclosing documents,
- 5) Soliciting response,
- 6) Using pressure tactics,
- 7) Ending politely.

Figure 1 below describes clearly the research procedure of this study. Research procedure consists of certain structural processes or steps to carry out research effectively. Hence, in the process of building COREnE, the first step carried out by the researcher is to gain access to the data. She then collected the sales emails written by the undergraduates and analysed the text. Noticing the problems in the writing, the researcher then conducted needs analysis by using Munby's Communication Needs Processor (CNP) (1978) methodology. Three industry informants and four subject specialists participated in a series of interviews to determine what is required in the process of teaching and learning of pedagogical sales emails. From the information and interview responses, the next process was done by manually tagging the moves and coding each of them.

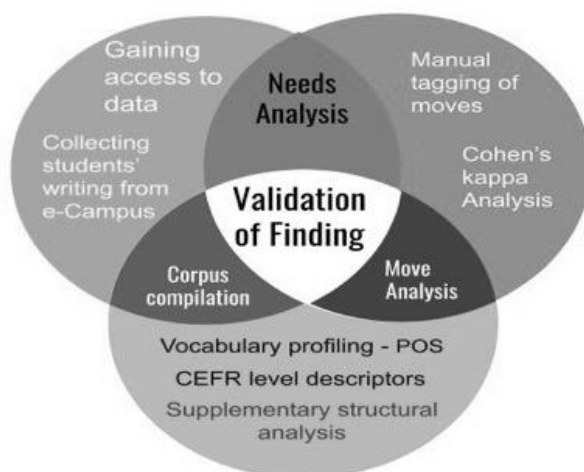
Cohen's kappa analysis was performed with two well-trained interraters using Bhatia's sales promotional letter moves (1993) as the base of the training. In the training, the interraters were provided with three checklists for each rater (Sales Inquiry Emails (SIE), Sales Reply Emails (SRE) and Bhatia's sales moves) and instruction of the rating procedure. Once the move types are clearly defined and all the documents in the dataset coded (and checked by the raters), the next step is to look for the patterns in how and when the different moves are used in order to help explain the specific role of each move more broadly within the genre. The goal is to have a full understanding of the communicative purposes and functions that the different parts of the genre have and how they work together to accomplish the overall communicative aim of the genre.

After the researchers understand the needs in pedagogical sales email writing, the corpus is then ready to be developed. The corpus development is divided into four major processes namely; 1) data collection, 2) data structuring (moves analysis), 3) parts-of-speech (POS) tagging and 4) CEFR descriptor level.

The text was coded manually and individually for the moves analysis following Integrated Moves Approach which integrates structural analysis (moves analysis) and syntactical analysis (Parts-of-speech and CEFR analysis). Using Bhatia's 1993) sales promotional letter framework, the researchers analysed the text and code the moves found as M1, M2 and further and then thematic analysis was applied to identify each moves appear in the text.

The text then uploaded on CLAWS Tagger for vocabulary profiling of Parts-of speech (POS) tags. Once the text was tagged, the researcher used AntConc, a software developed by Laurence Anthony of Waseda University, to organise the words alphabetically and sort its frequency. A supplementary structural analysis to determine the text's CEFR level is the final step of this study's research procedure.

**Figure 1** The Current Study's Research Procedure



### ■ 3.0 RESULTS

For sales reply emails, the analysis is quite direct following the structure of Bhatia's (1993) communicative purpose of sales promotion letter. It is a letter written from a company to address the concern of its potential customers. However, for sales inquiry emails, the communicative purpose is a little deviated from Bhatia's (1993) because the email comes from the customers to the company, even though the customer is asking on behalf of their company.

Thus, considering the need to understand the kind of information that a customer seeks in making the decision to purchase a product or service, this study performs the analysis on inquiring and responding to business information using the same structure of Bhatia's. Hopefully, it provides insights on how the customers interact and communicate their needs to the company which will be useful in the future when the graduates are asked to write in these types of genres.

Table 1 and Table 2 below highlight the moves that are missing from the student's writing compared to Bhatia's (1993) and also the moves that are added by the students (if there are any). In this analysis, Sales Reply Email will be coded as SRE and Sales Inquiry Email will be referred to as SIE. In Bhatia's promotional letter genre, there are seven (7) steps involved and the number of steps were recorded in SIE while 9 moves were present in SRE.

**Table 1** Sales Inquiry Emails Moves (SIE)

Code of SIE moves	Types of moves
M1: Designation	Indicate designation to establish trust and respect.
M2: Purpose	Introduce the purpose of writing the email.
• M2.1	Provide essential detailing of product of interest.
• M2.2	Show initiative of getting the info about the company and the product.
• M2.3	
M3: Incentives	Inquire for incentives: discount or after-sales service that should come with the product/ service offered.
M4: Inquiry	Inquiry for extra detailing: brochure/ flyers/ payment info
M5: Networking	Solicit further networking: future connection
M6: Pressure	Pressure tactics
M7: Polite ending	Ending politely.

**Table 2** Sales Reply Emails Moves

Code of SRE Moves	Types of moves
M1: Addressing interest	Addressing customer's interest in the product.
M2: Establishing credentials	Establishing company's credentials
M3: Responding to the request	Responding to customer's requests.
• M3.1	• Providing essential detailing.
• M3.2	• Indicating value of the offer.
M4: Incentive	Offering incentives.
M5: Attachment	Including methods of payment.
M6: Networking	Soliciting networking
M7: Product/ service credentials	Establishing credentials of the product.
M8: Pressure tactics	Using pressure tactics.
M9: Polite ending	Ending politely.

In order to make COREnE beneficial to a wide range discourse community of specialised corpus, the study adopted Integrated Moves Approach (IMA) that combines structural and syntactical analysis with structural analysis is the analysis of moves and syntactical analysis is the analysis of vocabulary of the text. COREnE underwent three stages of vocabulary profiling which are; 1) POS Tagging, 2) Frequency Analysis and 3) CEFR Descriptor Levels. Each of the stages is explained in the following subsections.

#### 1) POS Tagging

The tagger application allows up to 100 000-word limit per analysis with free subscription. This tool analyses the text according to its parts of speech such as adjectives, nouns, and verbs. There are 60 tagset used in this tagger among others are *DT* for determiner, *PNP* for personal pronoun and *VVB* for verb, base form. The simplified tagset is in Table 3 below.

**Table 3** Simplified Tagset

Code	Keyword	Code	Keyword
AJ	adjectives	PNP	personal pronoun
AT	article	PRP	preposition
AV	adverb	PUN	punctuation
DT	determiner	UNC	unclassified
NN	noun	VBB	base form of verb
ORD	ordinal	VM	modal

## 2) Frequency Analysis

Gries (2014) concurs the main use of frequency in the second language teaching and learning is to aid in selecting and prioritising the aspects of language to focus on (be they words, expressions, or grammatical structures). Based on the analysis, both genres recorded *nn* that represented the noun form with the highest frequency 2180 hits in SIE and 3873 in SRE. With 1001 (SIE) and 1981 (SRE) hits, *pun* (punctuation) is the second in the list; however, for this study, punctuation is excluded from discussion for it is not taught in the syllabus. The reason why punctuation is more frequent in SRE is because in the writing, students have to indicate specific information in bullet points, for example venue, time and date of the event. The detail of the analysis is presented in Figure 2 below.

Figure 2 Frequency analysis of sales emails

Word Types: 1079		Word Tokens: 22204
Rank	Freq	Word
1	2180	nn
2	1001	pun
3	821	prp
4	804	pnp
5	773	to
6	631	at
7	568	sent
8	564	vvi
9	444	av
10	437	the
11	436	dps
12	371	np
13	367	unc
14	361	aj
15	306	vm
16	294	vvb

## 3) CEFR level

Out of 10 000-word counts from SIE, 5549 tokens come from A1 vocabulary (Figure 3). The highest frequency of A1 level words is *the* with 427 tokens. It is followed by the word *to* (382), the pronouns *I* (287) and followed closely by another pronoun which is *your* (259). Among other words of A1 level used in SIE are both, good, meet, soon and writing. The second highest word frequency comes from A2 level with 1600 tokens counted. The word *company* is used for (261) times in SIE, followed by *printing* (181) and *machine* (111). The third highest frequency comes from B1 level with 832 tokens.

Figure 3 SIE CEFR analysis

Word List	Types	Tokens
A1	264 (26.01%)	5549 (59.85%)
A2	198 (19.51%)	1600 (17.26%)
B1	195 (19.21%)	832 (8.97%)
B2	163 (16.06%)	480 (5.18%)
C1	43 (4.24%)	58 (0.63%)
C2	10 (0.99%)	35 (0.38%)
Unlisted	142 (13.99%)	718 (7.74%)

For SRE, Figure 4 below shows the types and tokens of CEFR word list from the six (6) descriptors A1, A2, B1, B2, C1 and C2. The highest tokens of 5640 made up half of overall percentage of the analysis came from A1. 363 word-types are found under this level. As for A2 level, the tokens found are 983 while for B1 is 1016.

Figure 4 SRE CEFR analysis

Word List	Types	Tokens
A1	363 (25.17%)	5640 (57.76%)
A2	233 (16.16%)	983 (10.07%)
B1	255 (17.68%)	1016 (10.41%)
B2	172 (11.93%)	472 (4.83%)
C1	46 (3.19%)	353 (3.62%)
C2	13 (0.90%)	13 (0.13%)
Unlisted	360 (24.97%)	1287 (13.18%)

#### 4.0 DISCUSSION AND RECOMMENDATION

Sales email move analysis seeks to identify the components of a genre by the communicative purposes they serve, in this case, the moves. These communicative purposes must be identified within the context of the genre as well as the social context in which the genre resides. The two types of genres found in this study are Sales Inquiry Email and Sales Reply Email. This analysis enables ESP practitioners to investigate the ability of students in writing sales emails, investigate their vocabulary level, and also allows a systematic and detailed study of the learner's linguistic behaviour from the point of view of overdone (what linguistic features does the learner use more than a native speaker?) and underdone (what features does the learner use less than a native speaker?). Data analysis supported the comprehension of linguistic behavior of the students through the full use of COREnE for the extraction of linguistic information and different kinds of annotation (error tagging, lexical frequency counts, POS tagging). Ironically, the utilization of corpus in ESP researchers has only been widely recognized before it comes to be optimally useful. The current approaches seem to discard the traditional concept of error analysis and to continue to collect and store corpus data indefinitely for future reference.

Furthermore, another aspect of sales email analysis that many considered as retrograde is the tendency to concentrate on writing rather than speaking skills. The truth is that although speaking skills are more common in worldwide research, computers and technology are more practical and efficient in the current era. Even the spoken discourse collected as corpus has been transcribed into written form.

The analysis of the distributional patterns on CEFR level vocabulary illustrates the dominating frequency of the common noun forms in most of the sentences written by the undergraduates. The common noun appears in the above discussion with 11 out of the words above are nouns while the rest are DPS (possessive determiner form), VVB (base form of lexical verb) and PRP (preposition). The lack of adjective usage in this analysis supports the result of the previous research question in adjacent discussion above. Although the inadequacy of adjectives does not affect the communicative intention of the text, its positive and powerful impact in professional business setting is inevitable. Singh (2014, p. 221) described clearly on the importance of adjectives in writing, "(it) provides clarity and emphasized the embedded meaning of the sentence helping the reader to clearly understand the instruction (if it is) used with specific intentions within the ultimate communicative goal of each sentence structure." Hence, there is a necessity for a decent manifestation of adjectives to inculcate the undergraduates in business communication education and practice.

Aside from the variety utilisation of CEFR level frequency research, the finding supported the realistic target level for tertiary education in Universiti Malaysia Kelantan. The framing of the result is heavily influenced by the notion of business discourse as texts intended for learning and assessment and not in the context of real-world business. Hence, the lexis was identified and analysed within the framework of these communicative purposes and merely reflected their competence at appropriate level of their business specialist field. There is a minor but eminent learner progress in writing at the C1 and C2 level of CEFR which suggests that students have achieved the desired result that corresponds to the global scale of proficiency.

One important contribution of this study is identifying the moves of sales inquiry email and sales reply email in the data and describing the main characteristics that distinguish them from Bhatia's (1993) using Integrated Moves Approach (IMA). The analysis of the data indicates the main objective of the present study was to identify and describe the structural linguistic patterns used in the sales email in terms of the moves used. The research questions were answered using a combination of qualitative and quantitative research methods. The research instruments used consist of a corpus built from written pedagogical sales emails collected from the undergraduates of Business Courses and a set of semi-structured interview questions for the Industry informants and subject specialists. The corpus identified as COREnE, short for Corpus of Entrepreneurship Emails, was then used to conduct a genre analysis consisting of two complementary approaches namely a structural analysis and a syntactical analysis.

One of the most noticeable features that distinguish this framework of sales emails to other research of the same topic is the validation of the findings using interviews with the industry informants and the subject specialists. The responses elicited from the interviews were analysed using a thematic approach which informed the pedagogical and professional needs of the undergraduates. This has led to the major findings of this particular study which is a framework of sales email that suits pedagogical content-based ESP courses in the tertiary learning institutions in this country.

This study presented the use of a corpus-based genre analysis which consisted of a move analysis and a complementary syntactical analysis in order to propose a guided framework for designing suitable ESP courses. The gap between the two major stakeholders of ESP course designs; the learners and the industries, can be closed if ESP course designers and developers are encouraged to cross the borders of

academically- based ESP course designs. The course designers and developers should be encouraged to base their ESP course designs on research findings which are based on the authentic use of materials.

## 5.0 CONCLUSION

The importance of moves in writing sales emails implies that writers should be aware of the structure of content and the pattern of language use. Further analyses should indicate various moves and linguistic patterns on the basis of wide scale genre research. This study is an attempt to contribute to such research, but due to the limitations imposed upon the present data, the present study also suggests that further research with larger and more varied samples can be done to arrive at more conclusive results or generalizations.

The present study identified POS used in the corpus known as COREnE, which is a compilation of pedagogical sales emails and describes the forms, functions and types of relations at the sentence level. Furthermore, the identification of CEFR levels from the vocabulary that the students use in their writing indicates that there is a need to look at other text types of pedagogical writings. Future research could examine persuasive vocabulary in sales emails to describe their function, the type preferred by text-type, and their frequency and distribution at both the sentence and paragraph level. Comparing the use of POS and CEFR levels in different business means of communication could provide a more comprehensive understanding of the role that these items play in the production and perception of business emails in general.

## References

- Abdul Haseeb, Md. A. S. (2018, January 10<sup>th</sup>). Higher education in the era of IR 4.0. *New Straits Times*. <https://www.nst.com.my/education/2018/01/323591/higher-education-era-ir-40>
- Almutairi, N. D. (2016). The Effectiveness of Corpus- Based Approach to Language Description in Creating Corpus-Based Exercises to Teach Writing Personal Statements. *English Language Teaching*, 9(7), 103-111.
- Anthony, L. (2016). Introducing corpora and corpus tools into the technical writing classroom through Data-Driven Learning (DDL). In J. Flowerdew and T. Costley (Eds.), *Discipline Specific Writing* (pp. 162-180). Abingdon, UK: Routledge.
- Bhatia, V. K. (2004). *Interdiscursivity in critical genre analysis*. 391-400.
- Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Harlow, UK: Pearson Education Limited.
- Biber, D., & Reppen, R. (2015). Introduction: In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of Corpus Linguistics*. (pp. 1-8). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139764377.001
- Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routledge. ISBN: 0-415-23698-3 (hb); 0-415-23699-1 (pb)
- Calude, A. & Pagel, M. (2014). Frequency of Use and Basic Vocabulary. In *Multilingual Cognition and Language Use: Processing and Typological Perspectives*. M. Putz and L. Filipovic (Eds.) Publisher: John Benjamins. doi:10.1075/hcp.44.02cal.
- Chandler, D. (1997). An introduction to genre theory. *The Media and Communications Studies Site*, 1-15.
- Cirillo, L. (2020). Learner Corpora and Embedded Assessment of Undergraduate EFL Writing: The Case of Metadiscourse Markers. In *Approaches to English for Specific and Academic Purposes: Perspectives on Teaching and Assessing in Tertiary and Adult Education* by M.J. Ennis & J. Prior (Eds.) (pp.147-172). Publisher: bu.press
- Creswell, J., & Clark, V. P. (2011). *Designing and conducting mixed methods research*. (2<sup>nd</sup>Ed.). Thousand Oaks, CA: Sage.
- Davies, A. (2007). *An Introduction to Applied Linguistics: From Practice to Theory*. Edinburgh: Edinburgh University Press. Retrieved July 14<sup>th</sup>, 2020, from <http://www.jstor.org/stable/10.3366/j.ctt1g09wtk>
- Evans, S. (2012). Designing email tasks for the business English classroom: Implications from a study of Hong Kong's key industries. *English for Specific Purposes*, 31(3), 202-212. doi:<https://doi.org/10.1016/j.esp.2012.03.001>
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring Learner Language Through Corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130-154. doi:<https://doi.org/10.1111/lang.12226>
- Granger S. (2008). Learner corpora. In A. Ludeling & Kyto, M. (Eds.), *Corpus Linguistics: An International Handbook*. Berlin & New York: Walter de Gruyter, 259-275.
- Granger, S. (1998). *Learner English on Computer*. London, New York: Addison Wesley Longman.
- Gries, S. T. (2009). What is Corpus Linguistics?. *Language and Linguistics Compass*, 3(5), 1225-1241. doi:<https://doi.org/10.1111/j.1749-818X.2009.00149.x>
- Gries, S. T. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 160-181). Cambridge, UK: Cambridge University Press.
- Grundmann, R. (2021). Using large text news archives for the analysis of climate change discourse: some methodological observations. *Journal of Risk Research*. 1-13. 10.1080/13669877.2021.1894471.
- Hunston, S. (2006). Corpus Linguistics. In *Encyclopedia of Language & Linguistics* by K. Brown (Ed.), (pp. 234-248). Elsevier Publications. doi:<https://doi.org/10.1016/B0-08-044854-2/00944-5>
- Hyland, K. (2004). Graduates' gratitude: The generic structure of dissertation acknowledgements. *English for Specific Purposes*, 23(3), 303-324. doi:[https://doi.org/10.1016/S0889-4906\(03\)00051-6](https://doi.org/10.1016/S0889-4906(03)00051-6)
- Joharry, S. A., & Rahim, H. A. (2014). Corpus research in Malaysia: A bibliographic analysis. *Kajian Malaysia*, 32(1), 17-43.
- Kamariah, Y., & Su'ad, A. (2011). Collocational Competence among Malaysian Undergraduate Law Students. *Malaysian Journal of ELT Research*, vol. 7(1), p. 151 – 202.
- Kaur, N. (2013). A case study of tertiary learners' capability in lexical knowledge. *GEMA Online Journal of Language Studies*, 13(1), 113-126.
- Kankaanranta, A. (2006). "Hej Seppo, Could You pls Comment on This!" - Internal email communication in lingua franca English in a multinational company. *Business and Professional Communication Quarterly*, 69(2), 216-225. doi:<https://doi.org/10.1177/108056990606900215>
- Kotani, K. & Yoshimi, T. (2015). *Design of a Learner Corpus for Listening and Speaking Performance*. 9th Pacific Asia Conference on Language, Information and Computation pages 351 – 358.
- Leech, G. (1998). Preface. In Granger, S. (Ed.), *Learner English on Computer*, pp. xiv-xx. London and New York: Longman.
- Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 65-100). doi:<https://doi.org/10.1075/scl.59.06loz>
- Martin, J. R. (1997). Analysing genre: functional parameters. In J. R. Martin, F. Christie & J. Martin (Eds.), *Genre and Institutions: Social Processes in the Workplace and School* (pp. 3-39). London: Cassell: Continuum.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies. An Advanced Resource Book* (Routledge Applied Linguistics). London, New York: Routledge. xx, 386 S.
- Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 537-566). Cambridge, UK: Cambridge University Press. doi:<https://doi.org/10.1017/CBO9781139649414.024>
- Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.



- Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, 21(4), 373-391. Retrieved October 30, 2017, from <https://hal.archives-ouvertes.fr/hal-00572085/document>
- Myskow, G., & Gordon, K. (2010). A focus on purpose: Using a genre approach in an EFL writing class. *ELT Journal*, 64(3), 283-292. doi:<https://doi.org/10.1093/elt/ccp057>
- Nesselhauf, N. (2005). Corpus linguistics: A practical introduction. Retrieved from <http://www.as.uni.heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf>
- Osman, H. (2013). Pedagogic genre analysis: A module for copy-writing. *English for Specific Purposes World*, 13(37).
- Pendar, N., & Chapelle, C. (2008). Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, 25(2), 189-206.
- Romer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research*, 2(2), 99-127. Retrieved from [https://www.researchgate.net/profile/Ute\\_Roemer/publication/45534846\\_Applying\\_corpus\\_methods\\_to\\_written\\_academic\\_texts\\_Explorations\\_of\\_MICUSP/link/s/54d8d2a70cf24647581c47a7.pdf](https://www.researchgate.net/profile/Ute_Roemer/publication/45534846_Applying_corpus_methods_to_written_academic_texts_Explorations_of_MICUSP/link/s/54d8d2a70cf24647581c47a7.pdf)
- Rice, R. E., & Bair, J. H. (1984). New Organizational Media and Productivity. In R. E. Rice (Ed.), *The new media: Communication, research and technology* (pp. 185- 215). California, USA: Sage Publications. Retrieved from <http://www.comm.ucsb.edu/faculty/rice/c1.htm>
- Sadjirin, R., Aziz, R. A., Nordin, M. N., Ismail, R. M., & Baharum, M. D. (2018). The Development of Malaysian Corpus of Financial English (MaCFE). *GEMA Online Journal of Language Studies*, 18(3), pp. 73-100, <http://doi.org/10.17576/gema-2018-1803-05>
- Safayenji, F., Lee, E., & Macgregor, J. (1988). An empirical investigation of two electronic mail systems. *Behaviour & Information Technology*, 7(4), 361-372. doi:<https://doi.org/10.1080/01449298808901883>
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Singh, M. K. S. (2014). *A corpus-based genre analysis of quality, health, safety and environment work procedures in Malaysian petroleum industry*. Unpublished PhD Thesis, Universiti Teknologi Malaysia, Faculty of Education. Retrieved from <http://eprints.utm.my/id/eprint/78163/1/ManvenderKaurSarjitPFP2014.pdf>
- Singh, M. K. S., Shamsudin, S., & Zaid, Y. H. (2012a). Revisiting genre analysis: Applying Vijay Bhatia's approach. *Procedia - Social and Behavioral*, 66, 370-379. doi:<https://doi.org/10.1016/j.sbspro.2012.11.280>
- Singh, M. K. S., Zaid, H. Y., & Shamsudin, S. (2012b). A Computer-Assisted Corpus Analysis (CACA) of professional discourse. *Sino-US English Teaching*, 9(6), 1236-1245.
- Subtirelu, N. C., & Baker, P. (2017). Corpus-based approaches. In J. Flowerdew, & J. E. Richardson (Eds.), *The Routledge Handbook of Critical Discourse Studies* (pp. 107-120). London, UK: Routledge.
- Swales, J. M. (1990). The concept of discourse community. In J. M. Swales, C. A. Chapelle, & S. Hunston (Eds.), *Genre analysis: English in academic and research settings* (pp. 21-32). New York, USA: Cambridge University Press.
- Tono, Y. (2003). Learner corpora : Design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 800-809). Lancaster, UK: University Centre for Computer Corpus Research on Language.
- Tostlebe, E. (2011). *Teaching writing in the textbook series profiles: A genre-based approach*. [Unpublished Thesis], University of Tampere, School of Language, Translation and Literary Studies.
- Upton, T. A., & Cohen, M. A. (2009). An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies*, 11(5), 585-605. doi:<https://doi.org/10.1177/1461445609341006>
- Vicic, S. G. (2008). Potentials and challenges of ESP learner corpora: The case of modal auxiliaries in Slovene ESP learners' written interlanguage. In Brkan et. al. (Eds), *Inter alia 1* (pp. 15-30). Ljubljana: Slovenian Association of LSP Teachers.