GEMA Online® Journal of Language Studies
Volume 21(4), November 2021 http://doi.org/10.17576/gema-2021-2104-05

The Diachronic Malaysian English Corpus (DMEC): Design, Development and Challenges

Hajar Abdul Rahim [a]
hajar@usm.my
School of Humanities,
Universiti Sains Malaysia, Malaysia

Raihana Abu Hasan
raihana_20001813@utp.edu.my
Management and Humanities Department,
Universiti Teknologi PETRONAS, Malaysia

Ang Leng Hong [b]
lenghong@usm.my
School of Humanities,
Universiti Sains Malaysia, Malaysia

Siti Aeisha Joharry
aeisha@uitm.edu.my
Akademi Pengajian Bahasa,
Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia

ABSTRACT

Malaysian English (ME) has received much research attention in terms of its linguistic system, and use. In the last two decades, with the development of language corpora and corpus methods, research in ME as an Outer Circle and postcolonial variety increased significantly. These corpus-based studies are important and indeed interesting but they are limited to descriptions and discussions on contemporary ME, as they are based on synchronic ME corpora. Research in diachronic changes in ME, to date, is rare to say the least. To address this gap and facilitate systematic examination of changes in ME necessitates the development of diachronic ME corpora. This article reports on the design and development of the first diachronic Malaysian English corpus (DMEC), comprising Malaysian English written texts from the 1960s until the 2010s. The six decades represent three phases of English in Malaysia, end of colonial era, postcolonial and contemporary ME. The corpus is designed to facilitate research in ME changes in terms of its linguistic system, identity and trajectory. Besides the design and development of the DMEC, three sample analyses based on the existing corpus are discussed to demonstrate the value of the corpus as a resource for diachronic research in ME. The current article, overall, contributes to existing knowledge on the development of language corpora in Malaysia and foregrounds the potential of diachronic research in ME.

Keywords: Malaysian English; diachronic corpus; diachronic changes; corpus development; corpus design

[a] *Main author*
[b] *Corresponding author*

eISSN: 2550-2131
ISSN: 1675-8021

INTRODUCTION

Corpus research in English in Malaysia began in the 1990s but only garnered attention among local language researchers in the early 2000s. Two crucial areas of research that spurred the rise in corpus research in English in Malaysia are language pedagogy and Malaysian English as a variety of New Englishes. The interest in corpora and language pedagogy encouraged the development of Malaysian learner corpora (e.g. EMAS, The English of Malaysian School Students corpus; CALES, Corpus Archive of Learner English in Sabah/Sarawak; MACLE, Malaysian Corpus of Learner English) for research in learner language and second language issues, while corpus-based research in the field of New Englishes significantly contributed to the development of several ME corpora, and the rise in corpus-based research in Malaysian English (henceforth ME). Corpus-based studies on ME structure, lexis and phonology (e.g. Hajar, 2008, 2014; Hajar & Harshita, 2003; Low, 2021; Newbrook, 2006; Pillai et al., 2010; Tan, 2009, 2013) contributed considerably to the body of knowledge on ME as a new variety of English. In recent years, more synchronic ME corpora of different genre and size, such as the Malaysian Online English Sports News Corpus (MOSNEC) (Tan, 2015), the Malaysian Corpus of Financial English or MaCFE (Roslan et al., 2018) and Mesolectal Malaysian English Corpus (Ong, 2019) have been and are being developed. Existing English language corpora in Malaysia are generally synchronic corpora. They are essential resources for studying and describing contemporary ME features and use, and have facilitated much research in ME as a local variety and as a postcolonial variety. Research in ME from a postcolonial lens (e.g. Azirah, 2002, 2007; Hajar, 2008, 2014; Nurul Farhana, 2014; Shakila, 2014) involves analysing linguistic features in literary and non-literary texts to foreground issues of representation, identity and voice of the users of the language.

While there is a steady increase in the number of ME corpora and a surge of research based on them, to date, there has been no diachronic study on ME as a local variety. This is despite its rich literature, dating back to the 1970s with seminal works such as Tongue's (1974) descriptive framework of Malaysian and Singaporean English, Platt and Weber's (1980) systematic description of the speech variety used in Singapore and Malaysia to more contemporary corpus-based studies.

Diachronic changes in new varieties of English, particularly postcolonial varieties such as ME are important as the changes in the status and role of the English language throughout colonial and postcolonial contexts factor significantly in the way the language has evolved from the original native form into the contemporary localised variety. The changes that postcolonial varieties of English experienced at different periods, from the end of the colonial era through the postcolonial phase until the present day necessitate research that traces the changes in varieties of English language such as ME.

Bybee (2012) explains that the conventional areas of study in language change are sound change, analogy, morphosyntactic change and semantic change. Beyond this, the evidence from diachrony is crucial in showing lexicon-grammar interconnections and "the nature of the cognitive representation of phonological and grammatical form" (Bybee, 2012, p.980). These are significant in detailing changes in a language "such as gradual phonological reduction, new inferential meanings, or new contexts of use" (Bybee, 2012, p.980). Examining changes and developments in English from different periods necessitates the development of diachronic or historical corpora, which are "textual resources that represent comparable types of language use over sequential periods of time" (Hilpert & Gries, 2016, p.36). Given this and the gap in the literature on diachronic changes in ME, a diachronic corpus comprising data on English used in Malaysia and by Malaysians over a period of six decades was designed and developed. The six decades, the 1960s to the 2010s, represent three different phases of English in Malaysia, that is, end of the colonial era, postcolonial and present day ME. This paper

discusses the development of the first diachronic corpus in Malaysia, named the Diachronic Malaysian English Corpus (henceforth DMEC). As diachronic corpora typically contain written language, the DMEC represents written Malaysian English from the 1960s until the 2010s. It is designed to facilitate analysis in ME diachronic changes and evolution from a colonial variety to the present-day localised variety.

MALAYSIAN ENGLISH

The history of English in Malaysia dates back to over 150 years ago when the British first came to Malaya. English was the language of the administrators during the colonial period, but the country's independence in 1957 saw the status of English as the official language replaced by Malay. Within the World Englishes framework, ME is an Outer Circle variety, given its history as a colonial language and its status as the official second language of the country post-independence. English is a familiar language in Malaysia and a second language for many Malaysians. It is used widely in the workplace, schools, printed and non-printed media, internet, entertainment, and in the private sector. English is an important lingua franca alongside the Malay language in the country with an extended functional range in various socio-cultural, economic, political and creative domains (Asmah, 1996; Newbrook, 2006; Rajadurai, 2004). As a new variety, ME is described in terms of an acrolect-mesolect-basilect lectal continuum or cline. ME acrolect is the standard form of Malaysian English, a "prescribed pedagogical norm necessary for international communication" while ME mesolect is the unofficial Malaysian English used for intranational communication, and ME basilect, also sometimes known as "Manglish" or broken Malaysian English, is "regarded as almost unintelligible outside of the speech communities in which it is developed" (Gill, 2002, p.52).

As a vibrant language in a multi-lingual and multi-ethnic society, English in Malaysia has received much research attention, most of which falls into two general categories: the standard of English in Malaysia and ME as a new variety of English. Research in the former generally revolves around educationists' complaints about the quality of English in the country, focusing mostly on Malaysian students' proficiency, English language teaching and learning problems, English language curriculum and materials, teacher training and other related issues. Research in these areas dates back to the 1970s, when educationists began to raise concerns regarding the change from English to Malay as the medium of instruction in Malaysian national schools. Research in the decline of the standard of English among students and efforts to improve it forms the bulk of the literature on English in the country. In the last two decades however, the literature has seen an increase in research in the language itself, as a new variety of English.

Research in ME from the New Englishes perspective is different as its interest is in the evolution of ME as an English language variety used by the speakers. New varieties of English are offshoots of the native form which scholars argue experience stages of evolution due to various socio-cultural and political reasons. Thus research in new Englishes involves 'evolutionary approaches to language change' (Mukherjee, 2007, p.172) such as Schneider's dynamic model of New Englishes which traces the change of English over a five-stage trajectory known as the foundation stage, exonormative stabilisation stage, nativisation stage, the endormative stabilisation and differentiation (Schneider, 2007). Research in ME in this regard (based on spoken and written registers, standard and non-standard forms), suggests a strong inclination towards nativisation at all linguistic levels (see example Baskaran, 1994, 2005; Hajar, 2008; Hajar & Harshita, 2003; Lim, 2001; Lowenberg, 1991, 1992; Morais, 2001; Newbrook, 2006; Schneider, 2007; Low, 2021). Scholars therefore suggest that ME has reached the third stage of evolution, i.e., the nativisation stage.

The dynamic nativisation of ME is evident from past studies on the standard form (e.g. the use of lexical borrowings, code switching and mixing among educated Malaysian elites (Lowenberg, 1991, 1992), use of local lexical forms in Standard ME (Hajar & Harshita, 2003) the phonological, lexical and syntactic characteristics of ME (Baskaran, 2005), the acrolect form in ME in print and in careful speech (Newbrook, 2006), variety used by educated Malaysians in their speeches (Kirkpatrick, 2007), and different types of borrowing in ME newspaper (Tan, 2009). The inclination to nativise is further supported by research in ME in the creative domain where nativisation is distinct "through the use of syntactical variation, rhetorical devices and figurative language by creative writers" who use the language in their work as "vehicle to transmit local cultural heterogeneity, values, mores, and Malaysia's varied sociolinguistic realities" (Hajar & Shakila, 2014, p.27). Mukherjee (2007) considers these creative processes as "progressive forces" that encourage nativisation.

Diachronic changes are important indications of how a language has evolved. Hence, the development of a diachronic corpus containing written texts representing standard ME from the 1960s to the 2010s discussed in this paper. The decision to create a corpus of one variety of ME is in keeping with Hilpert and Gries' (2016, p.36) view that, "it is a design feature of most diachronic corpora to hold the type of text constant, so that diachronic language change within a given text type may be studied with as few confounding factors as possible". The corpus is designed to facilitate research in the changes in ME at the lexical and structural levels. It also provides opportunity to analyse changes in the language due to the different language policies instituted nationally (e.g. the 1961 National Education Policy, the 1967 National Language Act), as well as to investigate creative changes that emerged in ME representing identity negotiations and reconstruction. And importantly, the development of a diachronic ME corpus is an important resource for studying its trajectory as a new variety of English.

The following section discusses the development of the diachronic corpus, beginning with a discussion of the key issues taken into consideration in the design of the corpus, followed by a description of the corpus building and challenges that emerged at different stages of the corpus development. The final section presents samples of analysis based on the current data available in the corpus.

BUILDING THE DIACHRONIC MALAYSIAN ENGLISH CORPUS: DESIGN, DEVELOPMENT AND CHALLENGES

THE DESIGN OF THE CORPUS

Diachronic corpora are textual resources that represent comparable types of language use over sequential time periods, i.e. at least two periods, as in the Diachronic Corpus of Present-Day Spoken English (DCPSE, Wallis et al., 2006), while others such as Davies' (2010) Corpus of Historical American English (COHA) comprise written texts sampled from a sequence of decades (Hilpert & Gries, 2016). And importantly, as stated earlier, the text type of a diachronic corpus should be constant (Hilpert & Gries, 2016).

Besides text type, representativeness is also a crucial factor in corpus design. A corpus is a collection of samples used in real life contexts by language users from the same or different backgrounds, therefore it should be representative of the language and its speakers. Information on the samples of language users and the type of language represented by the corpus relate to the ability to translate and generalise results from the corpus to a context outside the corpus (McEnery & Hardie, 2011). While representativeness is central in the design of a corpus, scholars generally agree that there is no absolute representativeness in designing and developing a corpus, because it is a sample taken from the population in the real world (Gablasova, Brezina & McEnery, 2019). Essentially "there are no generally agreed objective

criteria that can be applied to this task: at best, corpus designers strive for a reasonable representation of the full repertoire of available text types" (Kilgariff et al., 2006, p.129). So in striving for a reasonable corpus representativeness, it is recommended that researchers "document corpus design criteria explicitly and make the documentation available to corpus users so that the latter may make appropriate claims on the basis of such corpora" (McEnery, Xiao & Tono 2006, p.18). In relation to this, Mackey and Gass (2015) suggest that to achieve corpus representativeness, there should be information on language samples and methods of collection to allow the reader to determine the degree to which the findings of the study are indeed generalisable in a new context.

Guided by Weisser's (2016b) discussion on corpus design, the creation of the DMEC takes into consideration the following: it is essentially a diachronic, as opposed to synchronic corpus; a written, as opposed to spoken or mixed, corpus; a general, as opposed to specific corpus. While the decision of a dynamic versus static corpus was not made at the design stage of the corpus, it must be noted that the DMEC has the potential of being maintained as a dynamic corpus, given the relative ease in updating the corpus with texts from the most recent decades from online resources.

The basic idea that guided the process of selecting materials for the corpus is that ME is English produced by Malaysians. Crucially therefore, all selected texts must have been written by Malaysians and, as much as possible, published locally. Secondly, the written texts considered for the DMEC comprise fiction and non-fiction texts. These two categories which correspond to the BNC's imaginative and informative categories, respectively, were decided on to fulfil the requirement of a general corpus. The fiction category includes novels and short stories, while the non-fiction category comprises newspapers, magazines, biographies and (written) speeches. The details of the two main categories of texts (by target number of words) for every decade are shown in Table 1 below.

TABLE 1. The DMEC text categories and sub-categories

Fiction	**No. of words**
Novels	200,000
Short Stories	150,000
Non-Fiction	
Newspapers	200,000
Magazines	50,000
Biographies	30,000
Speeches	70,000

The target total number of words for each category, fiction and non-fiction, is 350,000 words. Each decade, from the 1960s to the 2010s, therefore, is represented by 700,000 words of written texts. In total, the DMEC, will have approximately 4.2 million words.

BUILDING THE CORPUS

The corpus building stage involves three different processes, namely data sourcing, cleaning and documentation. Data sourcing should have been completed before carrying on with the next process, which is cleaning; however, the challenges in obtaining certain types of texts from the 1960s and 70s required adjustments to be made in the work process. Due to this, data sourcing and cleaning for these two decades had to be carried out simultaneously.

SOURCING THE DATA

Materials obtainable online and publications accessible through libraries were opted to facilitate this stage. However, texts from the 1960s and 1970s were almost exclusively in print form and were sourced from the National Archives in Kuala Lumpur as well as the Malaysiana and archives section of Universiti Sains Malaysia library. Texts in the fiction category were selected based on consultations with specialists in Malaysian English literary studies, and through literary bibliographies and online databases searches. Only texts for the 2000s and 2010s were available in readily digitised format and in the form of web data. Texts sourced from newspapers such as mainstream newspaper *The Star* were available in digitised format from 2003 onwards only. Similarly, *Malaysiakini*, an alternative news website, had articles dating back to the year 2000. Other texts that were sourced online include short stories accessible from a publisher's blog (i.e. Silverfish Book) and a writer's personal blog (i.e. Fadzlishah Johanabas bin Rosli).

PREPARATION OF DATA

The preparation of the data relied heavily on manual work. Materials in print form were scanned using a flatbed scanner. Converting physical materials to digital form was achieved as follows: (1) scan the printed pages and convert them to PDF, (2) perform OCR on the PDF – different OCR software has different ranges of capabilities, and in our case we used Adobe Acrobat's text recognition tool as the default and FreeOCR as our alternative, (3) copy-paste the relevant texts directly from the PDF into a plain-text editor and save as a TXT file. Each TXT file contains one text written by one author, or co-authors, on one content. Excerpting from an entire text has been a common practice in various corpora (e.g. International Corpus of English, ICE) but for the DMEC, this method applied only to books where several complete chapters were selected, while for non-book materials, the whole text was included.

There was a trade-off between the time spent on scanning and the time spent on cleaning. Getting high-quality scans, which typically involved scanning a single page multiple times to get the best quality, had the benefit of reducing the time needed to clean the resulting text, and vice versa. This is because the OCR process can be very exacting, and blurred texts do not necessarily have the same readability for the OCR program even if they appear readable to our eyes. So texts not readable by any software were digitised manually using the Microsoft Word program. As manual typing is prone to mistakes, and given the AutoCorrect facility in Microsoft Word, manually typing had its drawbacks. To address this problem, typed texts were afterwards checked against the original texts.

Once a text was in the plain-text file, regardless of whether it was typed or pasted from the OCR results or both, it went through a cleaning process. This was manually done as there was no special programs or algorithms to automate the cleaning process. Although the process was tedious, the whole process was done meticulously as the quality of the cleaning can directly affect the quality of the computational analysis and the accuracy of the results and conclusions (Weisser, 2016b). The cleaning process went through at least two levels of checking. The first involved placing the original source and the plain-text file containing the resulting text side by side for line-by-line cross-examination to check for characters that had not been detected accurately. The letter 'O', for instance, might have been changed to the number '0', or 'h' might have been read as a 'b'. The next level was to paste the text into Microsoft Word to let its spelling checker identify errors that had been missed and to correct them in the plain-text file.

The principle followed in cleaning the texts was to be as faithful to the original material as possible while ensuring that there is continuity to the text. Spelling errors in the original text were not corrected. Page numbers and page headers, however, were deleted because they only

serve as navigational aids characteristic to published materials, and are not part of the actual texts. The cleaning process also involved removing image captions, reference lists, footnotes and endnotes and their superscript numbers. Single words broken up and hyphenated (including compound words) appearing in two lines were merged to facilitate the analysis later. There was more leniency where the mechanics of the text were concerned in that the order and presence/absence of punctuations still followed the original source but the whitespace between punctuations and their preceding word were removed, even if the original source appeared to have it. This involved, for instance, the space between the final word and the full stop or question mark or exclamation mark. Types of dashes (e.g. m- and n-dashes) and the spaces before and after them also did not strictly follow the original source.

DOCUMENTATION OF DATA

The last step is the documentation of the metadata. Again, this process did not make use of any automatic information retriever but involved keying in the data manually in a spreadsheet file. Information that was deemed important and useful for users of this corpus are:

1. Title of the text
2. Year of publication
3. Author
4. Category
5. Sub-category (or text type)
6. Publication information
7. Word tokens
8. Source
9. Remarks

The number of word tokens in a text file was determined using Microsoft Word's word count function, although it should be noted that different software that can perform word counts may have different ways of counting. In the case of Microsoft Word, any character or group of characters is considered as one word as long as they are not separated by a space.

CHALLENGES IN DEVELOPING THE CORPUS

This section discusses the main issues and dilemmas in building the corpus to offer some insights into the challenges of developing a regional-variety corpus. The drawbacks that stemmed from the problems will be helpful for researchers when using the corpus and in making inferences and conclusions based on their analysis.

MATERIAL AVAILABILITY

The first problem is the availability of materials for the corpus. Weisser (2016b) refers to insufficiency of materials as a natural limitation and it is likely a recurring problem for most corpora developments. Taking a more practical approach, online resources aside, we focused on printed materials that the university library could offer. As expected, the pool of resources shrank from materials that were supposed to exist, to what have been preserved and are available, to what are accessible, and further reduced due to missing books, fussy librarians and prolonged downtime in the library and archive systems. With a diachronic corpus, this problem spilled over to the maintenance of a balanced distribution of the corpus composition throughout the sub-periods that the corpus covers. Kohnen (2007) highlights this in discussing problems in corpus design and development which Helsinki Corpus, one of the most

established English diachronic corpora, tried to remedy. In the case of the current corpus development, the 1960s' decade emerged as the "prototype" to determine the target number of words for every text type in a sub-period. In other words, what the 1960s had to offer determined the proportion of data according to text types compiled for each decade. Despite this blanket decision, throughout the data collection and compilation of texts for the other decades, circumstantial decisions and negotiations still had to be made. For instance, it became apparent that, after the 1960s' data compilation was completed, there was a dip in the availability of short stories for the 1970s. The assumption is that the English language fiction-publishing scene at the time was affected following the change in the national language policy in 1967 which saw English being replaced by the Malay language as the national language. As a result, to meet the target number of words for short stories for the 1970s, a large portion of the data tested the boundaries of the criteria set for what constitutes a short story, and this made the data more heterogeneous than originally intended.

Kohnen (2007) also reports that a corollary of trying to maintain a balanced composition of the data across sub-periods is the inadequate representation of certain text types. This emerged in the current corpus from the pattern of publication of non-fiction books across the five decades. In the decade immediately following the country's independence (i.e. 1960s), it was very difficult to find non-fiction books written by local writers. Among the few that were available, most of them were too specialised. Even novels were not as short in supply as non-fiction books. To address this, only 8.6 per cent (30,000 words) of non-fiction data were considered for this text type. In the successive decades, however, non-fiction writing seemed to be more in fashion, hence the inadequacy in the representation of non-fiction books. Another text type inadequately represented is novels. It is interesting to note that while both the novel and newspaper categories each comprise 250,000 words, but given the nature of the two genres, the data for the former comes from only six texts while for the latter, the data comes from 518 texts. The novel category therefore exemplifies a long-and-thin distribution while the newspaper category a thick-and-short distribution.

Challenges in sourcing data influenced to some extent the distribution of texts but this was inevitable. And it seemed that the actual sampling and selection processes were, for us, the most challenging in compiling the corpus. In working with what is accessible, the kind of access that was afforded was generally of two kinds. On the one hand, text types such as novels and short stories were hard to come by and so there was no question of which novels or short stories were to be included in the corpus. On the other hand, text types like speeches, magazines and newspapers have a larger reserve to sample from, so there had to be a systematic way of selecting them. It was necessary to balance the number of texts in a year to limit repetitive topics or writers. With regard to newspaper data, random sampling was possible when sourcing the online archive of Singapore's newspapers, which contains the whole "population" of newspaper issues. However, random selection was not possible anymore when sourcing online newspaper data from the *Bernama* database. *Bernama* is Malaysia's national news agency which has a research portal, namely the *BERNAMA Library and Infolink Service* (BLIS). The portal provides paid-subscription access to archived news and speeches dating back to the 1960s. Unlike Singapore's newspaper database, *Bernama*'s archived news collected individual articles, older ones as scanned paper cuttings and later ones as digital texts. The paper cuttings were limited to certain topics such as statements by Malaysian prime ministers and reports on the *Bernama* news agency.

eISSN: 2550-2131
ISSN: 1675-8021

AUTHOR

One of the most fundamental problems in developing the corpus is the difficult task of determining the delimitations of a concept and observing them scrupulously when data deficiency in a category arose. This was alluded to earlier in highlighting the inadequacy of representation of text types but is expanded in this section in relation to the metadata spreadsheet. The issues highlighted here are probably unique to the Malaysian context.

The primary concept that needed defining is what constitutes ME. In ensuring that the corpus comprises texts written by Malaysians, it was necessary to understand the history of the formation of Malaysia. In the early days of Malaya's independence, citizenship status was ambiguous while the government sorted out who belonged to Malaysia. In addition, Malaysia only came into existence in September 1963 with the merging of Singapore, Sarawak and Sabah with the Federation of Malaya. Sabah and Sarawak were therefore latecomers and Singapore's membership as a state lasted for only two years. The process of untangling Malaysia from Singapore took time. Even until now, in literary research practices especially, works from both countries are studied together but linguistic research prefers to view them as separate entities (Azirah & Tan, 2012; Baskaran, 2008; Low & Tan, 2016).

Ideally, texts for the DMEC are those written by Malaysian nationals who have stayed in Malaysia for a significant period of their lives, preferably during their schooling years. Given this, works by diasporic writers were considered only if they maintained some connection with Malaysia as their homeland. Malaysia as the birthplace is not a prerequisite, and this refers to writers born in Malaya but claimed Singapore citizenship. Guided by this, native-English expatriate writers were differentiated from Malaysians who have anglicised names, and Singaporean writers were differentiated from Malaysian writers. It was much trickier for the latter as Malaysians and Singaporeans share similar name conventions. As a result, for the 1960s decade when Singapore was still a part of Malaysia, texts written by Singaporean writers during that period were included in the corpus. This group, however, forms only a small portion of the data.

Confirming the nationality of the authors involved meticulous web searches as well as books and documents where their work is published. It entailed reading through the author/writer's profile description accompanying the actual text, as well as narrations about the authors in research literature discussing the identity of Malaysian writers (e.g. Fernando, 1966). In cases where the profile was absent or not accessible through those sources, a general Google search and Google Books search of the author's name was performed. In instances where the search did not yield the required information, an educated inference of the writer's nationality was made based on his or her attitude in the text. With newspapers, not all articles were credited to an author. When there was no mention of the author, only those reporting on events that happened in Malaysia were included.

Another problem with Malaysian authors' names is those with honorific titles. As titles may change in a matter of years, it was necessary to standardise name writing during the documentation of the metadata, and the most efficient way was to dispense with all titles except for prime ministers' and the royalty's.

GENRE AND CATEGORY

The classification of genres and text types may seem straightforward but in reality it is not. The terms "newspaper" and "magazine" as sub-categories of non-fiction, for instance, are quite generic. Newspapers comprise news reports, articles, editorials, letters to the editor, opinion pieces, advertisements, cartoons, and other genres. Magazines too contain a mixture of genre types. To be consistent, only news reports were collected for the corpus to represent newspaper data, and only magazine articles were selected for the corpus category of magazine. Non-fiction

books category was initially labelled as "Biographies, etc." because originally only biographies were intended for this category. As the data sourcing process went on, it became apparent that biographies written by Malaysians were non-existent in the 60s. Thus, the category had to be broadened to include other non-fiction writings such as autobiographies, memoirs, commentaries, anecdotes and serial writings published as books.

Decisions on texts that do not appear in their conventional forms also had to be made. Texts, including novels, that are non-fiction were subsequently categorised as biographies, etc.; short stories published in magazines were considered under the short stories category; speeches by Malaysian public figures published in newspapers and magazines in the 1960s were considered under newspapers and magazines sub-categories, respectively, for that particular decade; anecdotal columns originally published in a newspaper and later compiled into a book was placed under biographies, etc.; and finally, novels with less than 20,000 words were categorised as novels, and short stories of more than 20,000 words and published in anthologies were categorised as short stories.

With texts that cross genres, it was necessary to check whether they had been edited to suit the genre they were published in. For instance, in dealing with short stories that bordered on novels (i.e. a short novel, also called novella, or a long short story), it was decided that a story in a stand-alone book was considered a novel, and a story compiled in an anthology was considered a short story unless specifically identified as a "novel" by the publisher or author. In short, justification for the categorisation of these irregular texts was necessary.

SOURCE

It is customary for published texts, whether they appear as books or in newspapers and magazines and even public speeches, to go through an editing process. While external interference to the original writing is inevitable, it was necessary that texts that were selected for the corpus did not include those published externally. To this end, interference from non-ME writers was minimised by screening the publishers or editorial teams. This is because foreign publishers, in editing and correcting certain linguistic idiosyncrasies that do not fit their writing preference or conventions, may remove the essence of ME. Thus, in selecting texts from the 1970s onwards, it was necessary to omit foreign-published books including those published in Singapore and other Asian countries. The 1960s was a special case because for this sub-period there had been a serious shortage of books published locally.

It is worth mentioning that there has not been an exhaustive initiative to digitise newspapers in Malaysia, nor has there been a coordinated effort to synchronise a database that collects all titles of local English publications that are available. This became evident while scouring the internet for potential data and made more obvious when considering the digitisation of newspapers published in Singapore that is maintained by its National Library Board.

A large percentage of newspaper data for the DMEC comes from the *New Straits Times* (*NST*), Malaysia's longest surviving English newspaper. The background of the newspaper's company bears some significance to explain why some texts had been sourced from Singapore's newspaper database and others from *Bernama*'s database. Coincidentally, it also foregrounds an exercise undertaken by Singapore and Malaysia in untangling themselves from each other. *NST* is an offshoot of Singapore's *The Straits Time* that was established in 1845. Even prior to the independence of Malaya, *The Straits Time* maintained separate editorial and production offices in Singapore and Kuala Lumpur (Lent, 2001). However, both offices still shared the same content with some minor difference in the technical aspects and an occasional change in words and phrases of the texts. The newspaper circulated freely across the states until the Malaysian government started to restrict Singaporean newspapers that entered Malaysia

after the separation of Singapore in 1965. After almost a decade, the Malaysian office finally rebranded the newspaper's name to New Straits Times in 1974. To stay faithful to the aim of using English texts produced only by Malaysians, no materials were sourced from the Singapore's database for data beyond 1963.

With regard to the inclusion of translated[i] works in the DMEC, the decision on the selection of these texts was to not include them. Exceptions were made only for texts translated by the author of the original work, and this was limited to selected fictional works only.

YEAR OF PUBLICATION

The year of publication in the metadata spreadsheet refers to the year the text was first published. The task of assigning the year of publication was made difficult when a text was published in subsequent editions. Very rarely did the later publications retain the text in its most original form. There were also changes in the conventions of punctuation, replacement of words, and restructuring of sentences. So the question that ensued was which year was to be assigned to the text if the text that had been included in the corpus was from a revised edition. This is because revisions might have been taken up to suit the latest language usages. The practice, therefore, was to refer to the earliest possible editions as the source, and if that was not possible, the year the book was copyrighted to the author was considered. If the later edition still retained its copyright notice the year it was originally published, then it is safe to say that the text had not been revised in a substantial way.

As for speeches, the year of publication would be the year the actual speech was delivered. Speeches from the 1960s particularly, were published in a form of a book and copyrighted. In the early stage of data compiling, the year of publication of the speech texts had been assigned to the year of publication of the book of speeches. Later, this was rectified when it became evident that the dates of the speeches were different.

Another important issue to consider in compiling a corpus is what and how much to include in a text file. With regard to newspaper texts, for instance, headlines, leads and author names were included according to the actual source. In the case of longer news reports in printed newspapers broken up into different pages, the signposting as found in the actual source (e.g. "See back page") was included, and the part of the sentence that had been broken up from the former page would be brought up to complete the sentence. The navigational aids were included even though they do not contribute to the meaning of the text because newspapers create a secondary headline for the separated text. Without the navigational aid, the corpus user would not understand the presence of that additional headline. So when analysing a text, the navigational aid will be included in the word count. It is crucial therefore to document the process so that when texts like these are to be processed, they will have a uniform format.

The final issue involves legality and copyright. As there is no intention to make the DMEC available for wide distribution, the corpus currently adopts the fair-use policy (see Crews, 1993, for an enlightening discussion).

The data for the fifth and sixth decades, the 2000s and 2010s, in the corpus comprise mainly digitally available texts. As they are available online, a special technique was used to export HTML links into TXT format, namely the ICEWeb, a small and simple utility for compiling, downloading and analysing web corpora (Weisser, 2016a). According to Weisser (2016a, p.1), "the name was chosen because the original intention was to create corpora that are similar in nature to the International Corpus of English (ICE) data". Although ICEWeb allows for typical corpus analytical methods such as concordancing and extracting n-grams, for compiling the 2000s dataset, only the URL retrieval and data conversion feature of the tool was used. The first step in using ICEWeb involved copying URL links of selected, relevant websites to a folder created with ICEWeb[ii]. Upon having a list of URL links on ICEWeb, the

tool then retrieved them automatically from the 'Retrieval' tab. This means that URL links were now captured/crawled/copied and saved as HTMLs. Since corpus analytical tools such as WordSmith Tools are mostly compatible with data in TXT format, ICEWeb provides a straight-forward mechanism where HTMLs can be converted automatically to TXT format. Once all HTMLs were converted, the files were inspected to ensure that proper conversion had taken place and that other non-relevant information (or noise) were removed. These files were saved together with the rest of the files in DMEC for documentation.

Overall, data sourcing and collecting for the last two decades were easier because of the availability of digitised texts. News, opinion pieces and speeches are available online compared to short stories, novels, magazines and biographies. One explanation for this could be the transition of printed news to online and social media, which have become predominant sources of news for Malaysian news users (Zaharom Nain, 2019). According to Alivi, Ghazali, Tamam and Osman (2018), this is a result of the emergence of new media that has seen an impact on the Malaysian political scene for instance in 2008. Opinion pieces and speeches have also followed suit where major newspaper agencies like *The Star* and *NST* have developed their online platforms to achieve wider readability. Even though short stories and novels have also seen more transition to the Internet, challenges in collecting these texts stemmed from restricted accessibility to the full texts. As a result, most of the texts for stories and novels were sourced online, converted to TXT format, and saved for compilation.

ANALYSING DIACHRONIC CHANGES IN ME

In their discussion on approaches in English historical linguistics, Hilpert and Gries (2016) highlight the significance of corpora and corpus-based methods, citing established resources with rich and long-standing tradition of corpus-based work (see the survey in Rissanen, 2008), for instance the Helsinki Corpus, the Brown family of corpora and ARCHER which have garnered much research in language change both lexical and grammatical. Citing other important researchers' work (namely Biber & Gray, 2011; Hundt & Mair, 1999; Tagliamonte, 2006), they also stress the importance of corpus resources in informing research in diachronic variation in genres, registers and varieties (Hilpert & Gries, 2016). In relation to this, Hilpert and Gries (2016, p.37) discuss how the use of quantitative corpus methods can be employed in analysing various linguistic features in diachronic corpora to address "When and how does a given change happen? Can a process of change be broken down into separate phases? Do formal and functional characteristics of a linguistic form change in lock-step or independently from one another? What are the factors that drive a change, what is their relative importance, and how do they change over time? How do cases of language variation in the past compare to variation in the present?"

This section presents sample analyses based on the DMEC to demonstrate the potential applicability of the corpus in addressing diachronic changes in ME.

SAMPLE DIACHRONIC ANALYSIS 1: MODALITY IN ME

Using *AntConc* 3.5.9 linguistic software (Anthony, 2020), six modals, namely, *may, might, must, ought to, shall and should* were selected to analyse diachronic change in modality in ME. These modals have been undergoing changes in contemporary American and British English (AmE and BrE) (Collins et al., 2014; Hansen, 2017; Leech et al., 2009). Diachronic investigation of English modals in contemporary AmE and BrE by Leech et al. (2009) shows that English modals have decreased in popularity, and a similar trend was observed in the studies of Asian Englishes such as PhilE (Collins et al., 2014) and Hong Kong English (HKE)

(Hansen, 2017). These findings are the main reason for the analysis of the six modals in the DMEC, particularly in the 1960s and 2000s subcorpora.

Table 2 shows the frequencies for the modals in ME and the percentage rise-and-falls in their frequencies between the 1960s and the 2000s. Table 3 presents the frequencies for the same modals in BrE, AmE and PhilE and the percentage rise-and-falls in their frequencies between the 1960s and the 1990s. Interestingly, most of the six modals declined in use, except for "may" which shows a noticeable overall increase (+14.1%). This is similar to findings on PhilE by Collins et al. (2014) which increased 12.9% between the 1960s and the 1990s. However, "may" decreased in use sharply in AmE (-31.8%), followed by BrE (-17.4%) (Leech et al., 2009). The ME and PhilE divergence from the native varieties of English (AmE and BrE) could be seen as evidence of grammatical change in the two non-native varieties of English.

TABLE 2. Frequencies (per 10000 words) of Modals in ME

	1960s	2000s	Change (%)
may	4.98	5.68	+14.1
might	3.09	1.99	-35.6
must	11.37	11.28	-0.8
ought to	0.15	0.11	-26.7
shall	2.78	0.54	-80.1
should	10.57	10.43	-1.3
Total	32.94	30.03	-8.8

TABLE 3. Frequencies (per 10000 words) of Modals in BrE, AmE, and PhilE (Collins et al., 2014)

	BrE			AmE			PhilE		
	1960s	1990s	Change (%)	1960s	1990s	Change (%)	1960s	1990s	Change (%)
may	13.33	11	-17.4	12.98	8.78	-31.8	12.06	13.61	+12.9
might	7.79	6.4	-17.7	6.65	6.35	-3.7	4.18	2.54	-39.3
must	11.47	8.14	-29.0	10.18	6.68	-33.8	10.28	7.77	-24.4
ought to	1.03	0.58	-43.6	0.69	0.49	-28.4	0.58	0.13	-78.1
shall	3.55	2.0	-43.6	2.67	1.50	-43.3	6.56	2.70	-58.9
should	13.01	11.48	-11.7	9.10	7.87	-12.8	11.67	11.55	-1.0
Total	50.18	39.60	-21.1	42.27	31.67	-25.1	45.33	38.30	-15.5

The modal, which declined most in frequency in ME, is "shall" (-80.1%). In the three other varieties, however, the decline is not as drastic ranging from -43.3% to -58.9%. Despite the moderate decline of "shall" in PhilE (-58.9%), its frequency remains superior to that in AmE, BrE and ME. Another interesting phenomenon involves the popularity of "might" in the native varieties (AmE and BrE) and non-native varieties of English. The modal has shown a declining trend in all varieties of English between the decades. Nonetheless, its decline in AmE and BrE has been mild (-3.7% and -17.7%, respectively), but its decline in both ME and PhilE stronger (-35.6% and -39.3%, respectively). The obsolescent modal "ought to" has shown a mild decline in ME (-26.7%), compared to its sharp decline in PhilE (-78.1%) and moderate decline in BrE (-43.6%) and AmE (-28.4%). The drop in the use of "ought to" in all varieties may be due to its competition with "should", also used as a moderate deontic expression. The modal "should" has undergone a very mild decline in all four varieties, ranging from -1.0% to -11.8%. The modal "must", with a rate of decline of only -0.8%, emerges as the most conservative among all modals in ME. Its decline rate is subtle, with its overall frequency (11.28) the highest compared to other varieties after several decades (BrE=8.14; AmE=6.68; PhilE=7.77).

The analysis suggests that modals in ME are decreasing in frequency, consistent with past research findings on modals in other varieties of English (e.g. Collins et al., 2014; Hansen,

2017). Future research could explore quasi-modals from the diachronic perspective to provide more insights into the trajectory of this grammatical category. The analysis, seemingly simple, nonetheless, demonstrates the potential of the DMEC as research resource.

SAMPLE DIACHRONIC ANALYSIS 2: DISTRIBUTION OF PREPOSITIONS 'ON' AND 'OF' IN ME

This is an analysis of the distribution of prepositions 'on' and 'of' based on the DMEC 1960s and 2000s subcorpora. The collocates of keywords in the two sub-corpora were compared in terms of their proximity within the collocation network visualisations. Following Sagi et al. (2012), second-order collocates, i.e., collocates of collocates were examined to overcome data sparsity. The distribution of 'on' in the two subcorpora (Figure 1) suggests that it is used more frequently in the 2000s (82.54 per 10k tokens) compared to the 1960s (60.03 per 10k tokens). The opposite is observed in the distribution of preposition 'of' (Figure 2) which occurs 362.43 per 10k tokens in the 1960s and 206.62 per 10 tokens in the 2000s, respectively.

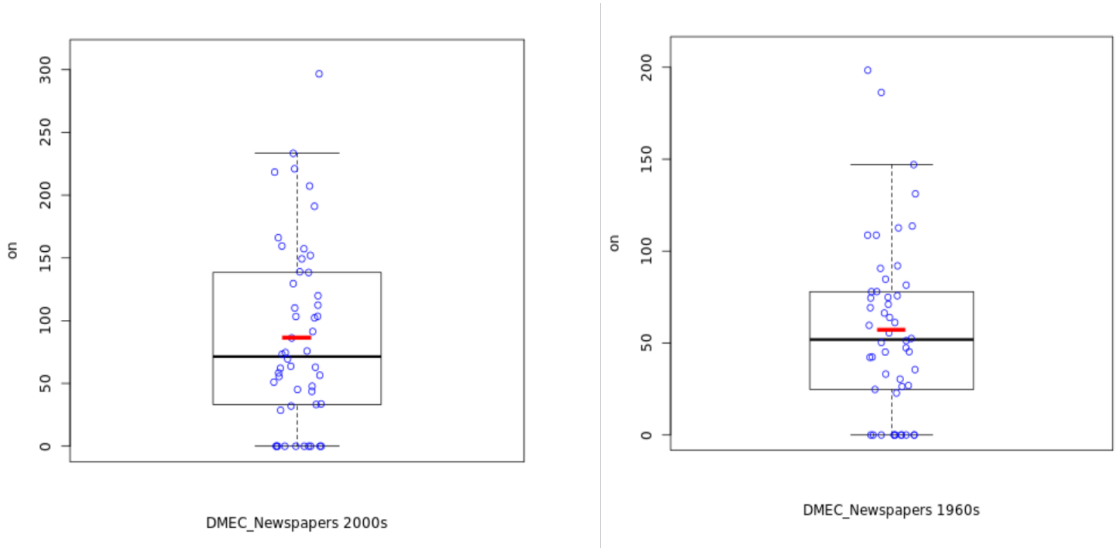

FIGURE 1. Distribution of 'on' in the DMEC newspapers 1960s and 2000s subcorpora

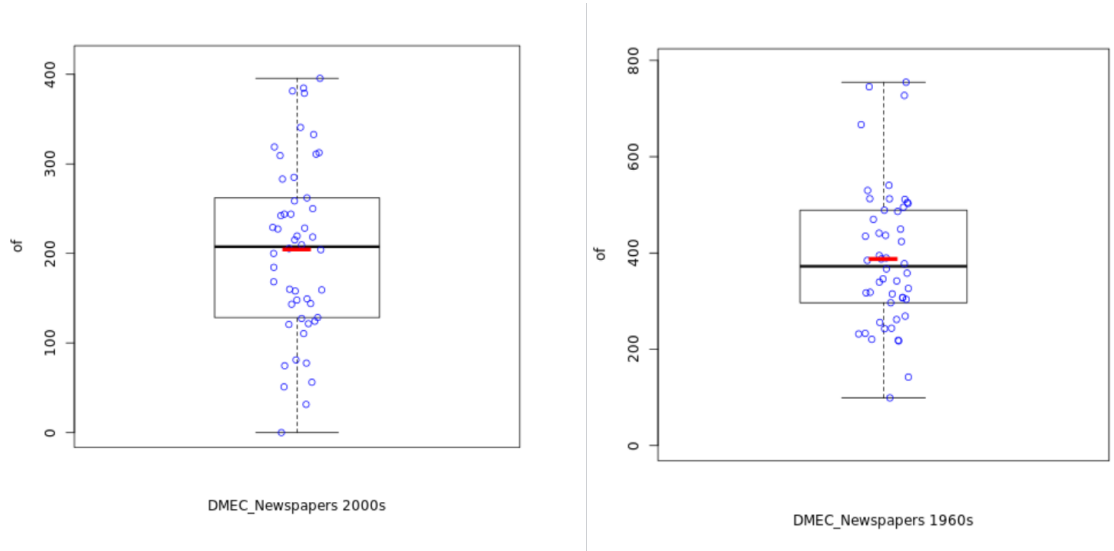

FIGURE 2. Distribution of 'of' in the DMEC newspapers 1960s and 2000s subcorpora

eISSN: 2550-2131
ISSN: 1675-8021

The GraphColl package of LancsBox (Brezina et al., 2020) was employed to examine the collocates (words that occur five or more time within the span of five words to the left and right). MI-value was used as the statistical measure with a threshold of five or higher (to narrow the number of statistically significant collocates). The analysis identified first-order collocates of the prepositions and the network of collocates (collocation networks). Figures 3 and 4 are visualisations of the collocational networks of 'on' and 'of' respectively.

FIGURE 3. A comparison of the collocation networks of preposition 'on' in the DMEC 2000s and 1960s newspaper subcorpora

Interestingly, 'on' occurs more frequently in 'based on' and 'focus on' in the 2000s compared to 'speaking on' and 'talks on' in the 1960s. News reporting and hedging preference (e.g. ***Speaking on*** *the Supply Bill, the Alliance Member of Kuala Trengganu Utara…*; *but as far as Malaysia is concerned,* ***talks on*** *the Sabah claim is closed*) in the 1960s were different from the more direct style in the 2000s (e.g. *He said building a sustainable economy* ***based on*** *innovation and quality…*; *The campaign will* ***focus on*** *educating young people on protecting themselves against…*). One explanation for this is Partington's (2010, p.11) view that "conversationalisation and informalisation [are examples of] changes in societal developments regarding face (Brown & Levinson, 1987) and social space (increased democratisation, or the decline of respect, depending on one's point of view)".

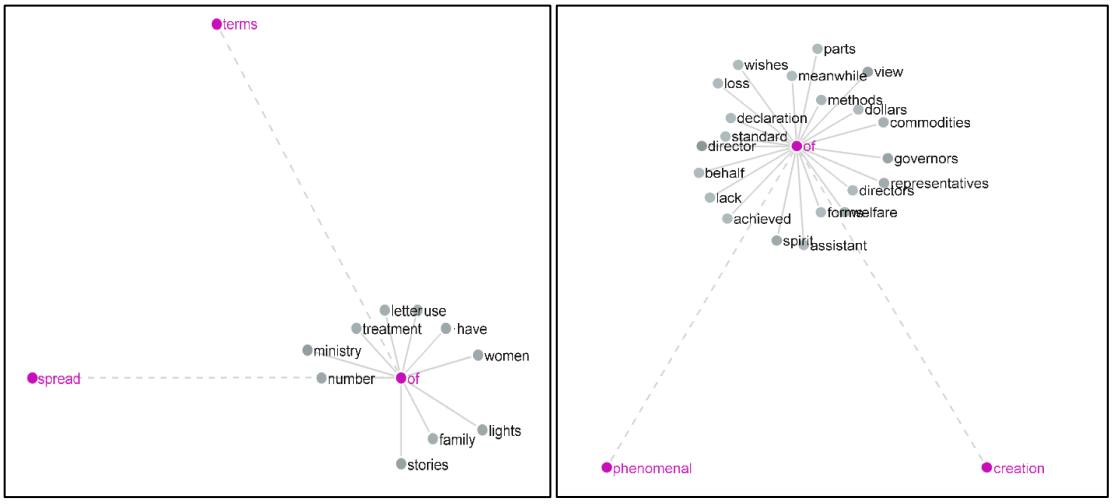

FIGURE 4. Collocation networks for 'of' in the 2000s followed by 1960s

eISSN: 2550-2131
ISSN: 1675-8021

As regards 'of', the analysis suggests that its collocations are more topic-related. For instance, in the top two most significant collocates in the 2000s (*spread* and *terms*) suggest that the phrase 'spread of' collocates frequently with HIV (see examples below), while 'creation of' co-occurs with descriptive progress reported in the 60s (e.g. *the* ***creation of*** *a Ministry of Rural Development*; *the* ***creation of*** *vast capital works of water control*).

1	NF_NE_174_2009.txt	is, will testing help curb the **spread**	of	the disease? Here we go again. The
2	NF_NE_174_2009.txt	latest government initiative to curb the **spread**	of	HIV. But chronology and technicalities aside, the
3	NF_NE_174_2009.txt	have limited impact on controlling the **spread**	of	HIV infection. A 1993 study of mandatory
4	NF_NE_174_2009.txt	premarital testing was "useless in the control	of	the **spread of** HIV, as refusal of
5	NF_NE_174_2009.txt	"useless in the control of the **spread**	of	HIV, as refusal of a license to
6	NF_NE_174_2009.txt	of the **spread of** HIV, as refusal	of	a license to marry does not prevent
7	NF_NE_174_2009.txt	to its effectiveness in preventing the **spread**	of	HIV as women are exposed to the

FIGURE 5. Concordance lines for samples of preposition 'of' with 'spread'

One reason for the higher use of preposition 'of' in the 1960s may be the preferred way of nominalising verb phrases (in that these instances take adjectival modification and realise participants periphrastically by means of an *of*-phrase) to achieve a more formal style of writing. A similar process of nominalising the gerund is evident in the use of 'phenomenal' in the description of the word 'expansion' as illustrated in the examples below extracted from the generated data. The nominalisation 'expansion' adds to the particularised style of employing the preposition 'of' which was not as frequent in the 2000s.

8	NF_NE_060_1960.txt	achieved **phenomenal expansion** of education	of	which the country was proud.
9	NF_NE_060_1960.txt	achieved the **phenomenal expansion** of education	of	which Malaya was so proud today.
10	NF_NE_060_1960.txt	achieved **phenomenal expansion**	of	education of which the country was proud.
11	NF_NE_060_1960.txt	achieved the **phenomenal expansion**	of	education of which Malaya was so proud

FIGURE 6. Concordance lines for samples of nominalisation with the preposition 'of'

Another interesting difference in the use of preposition 'of' is its collocation with 'terms' in the 2000s subcorpus (e.g. *in terms of*; *better terms of*; *new terms of*; *the terms of*), found in the 1960s subcorpus. This may indicate the growing use of the preposition with 'terms' to describe conditions or stipulations (as in *better terms of*; *new terms of*; *the terms of*) and a basis of expression or thought (as in *in terms of*).

This analysis shows how corpus tools, such as LancsBox, can facilitate corpus-based diachronic analysis using sophisticated techniques which enable language patterns and differences to be more visible than otherwise possible (Hilpert & Gries, 2016).

SAMPLE DIACHRONIC ANALYSIS 3: CHANGE IN LEXICAL USE IN ME

A present-day speaker's intuition of a particular word, structure or linguistic form that is not in fashion in contemporary use is enough to establish language change (Hilpert & Gries, 2016). However, such subjectivity is insufficient when the question goes "beyond the mere detection of a change and into the internal dynamics of that change". Quantification or language change by numbers becomes necessary. In this analysis, three words, namely 'colony', 'Malaya', 'Malaysia' and their lemma, generated from the wordlists of the 1960s and 2000s non-fiction (newspapers and speeches) DMEC subcorpora were examined in terms of change in their use. The two decades represent two different phases of ME, i.e. the 1960s represent a time when the country's colonial past still lingered while the 2000s, at the turn of the century, colonial ties are only in the history lessons.

The data for analysis is the frequency of the words and their lemma generated using *AntConc* 3.5.9 linguistic tools software (Anthony, 2020). The total number of words in the 1960s and 2000s subcorpora are 271610 and 229557 respectively. The frequency analysis reveals that the lemmas for 'colony', 'Malaya' and 'Malaysia' are more varied in the 1960s subcorpus compared to the 2000s subcorpus, as shown in Table 4.

TABLE 4. Frequency of 'colony', 'Malaya', 'Malaysia' and their lemmas in two subcorpora of the DMEC

	60s	**2000s**		**60s**	**2000s**		**60s**	**2000s**
Colony	5	4	Malaya	516	27	Malaysia	779	753
Colonial	42	8	Malayan	249	10	Malaysian	194	298
Colonialism	19		Malayanised	1	0	Malaysianisation	7	0
Colonies	1		Malayans	43	2	Malaysians	60	104
Colonisation	1					Malaysianised	7	0
Colonise	3					Malaysianism	9	0
Colonised	4					Malaysia's	0	2
Total	75	12		809	39		1056	1157

The lemmas for 'colony' are more varied in the 1960s compared to the 2000s, and their frequencies are higher. This may be attributed to the socio-political scenarios of each decade. The higher frequency of 'colony' and its lemmas in the 1960s, for instance, may be because at the time, Malaysia had only achieved independence, and issues surrounding the governing of British colonies such a Malaya and colonial matters were still familiar in everyday language. The same trend in the case of 'Malaya' (over 500 times in the 1960s compared to only 27 times in the 2000s) and its lemmas 'Malayan' and 'Malayans', may be due to the country's transition from a colonised country known as Malaya to the independent nation of Malaysia. In the case of 'Malaysia', the 1960s data shows that its lemmas are more varied but its total number of occurrence and that of its lemmas are higher in the 2000s. Interestingly, its lemmas 'Malaysianisation', 'Malaysianised' and 'Malaysianism' in the 1960s are not available in the 2000s. These lemmas are revealing of the socio-political climate at the time as the country changes from Malaya to Malaysia.

A collocational analysis of the words under study can provide insights into changes in lexical use due to different time and socio-cultural/political situation. As Hilpert and Gries (2016, p.38) argue, "more rigorous quantification of diachronic data becomes necessary when

research questions go beyond the mere detection of a change". This is achievable using corpora in tandem with sophisticated corpus analysis techniques.

CONCLUSION

A speaker's intuition of a particular linguistic element or use in contemporary use, according to Hilpert and Gries (2016), is sufficient to establish language change. However, quantification or language change by numbers is crucial in understanding the "internal dynamics" of change in a linguistic element. This is where diachronic corpora and corpus methods become invaluable. The DMEC, as the first diachronic corpus of Malaysian English, is therefore necessary in addressing the dearth of research in diachronic changes in ME. As the sample diachronic analyses show, the DMEC offers great potential in examining change in the ME variety such as the trend of modal verb use, preference for certain prepositional collocations as well as lexical change across time. The corpus can also facilitate stylistic, genre and other linguistic differences, diachronically. The diachronic dimension of the corpus is extremely important as it presents much opportunity to explore how ME linguistic forms changed over time, examine socio-political factors that caused the changes, compare variation in ME in the past and the present, as well as investigate its trajectory as a new variety of English. The systematic analyses of diachronic changes based on diachronic corpora can answer questions regarding the development of a language not just in terms of its linguistic profile but also the adjustments and negotiations it experienced in relation to temporal, socio-cultural and political factors. With regard to ME therefore, the development of the DMEC is timely and necessary as a resource for much needed diachronic research of a unique and rich variety of English. The development of the DMEC is not just a novel attempt in compiling a diachronic corpus to facilitate diachronic research in ME but an important contribution to the field of corpora development and corpus research in Malaysia. In line with this view, the DMEC, once completed will be made available to researchers who are interested in studying ME diachronic changes.

ACKNOWLEDGEMENTS

This work was supported by the Universiti Sains Malaysia Research University (RUI) grant (1001/ PHUMANITI/8016023).

END NOTES

[i] The inclusion of translated work in a monolingual corpus has reasonably been justified (see Zanettin, 2011).
[ii] Given the specific design criteria of the DMEC, ICEWeb's function for an automatic seed term search was unnecessary.

REFERENCES

Alivi, M. A., Ghazali, A. H. A., Tamam, E., & Osman, M. N. (2018). A Review of new media in Malaysia: Issues affecting society. *International Journal of Academic Research in Business and Social Sciences*, *8*(2), 12–29. Retrieved Sept 28, 2020 from https://www.researchgate.net/publication/323547457_A_Review_of_New_Media_in_Malaysia_Issues_Affecting_Society

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available at https://www.laurenceanthony.net/software

Asmah Haji Omar. (1996). Post-imperial English in Malaysia. In J. A. Fishman, A. W. Conrad & A. Rubal-Lopez (Eds.), *Post-Imperial English: Status Change in Former British and American Colonies, 1940-1990* (pp. 513–533). New York: Mouton de Gruyter.

Azirah Hashim. (2002). Culture and Identity in the English Discourses of Malaysians. In A. Kirkpatrick (ed.), *Englishes in Asia: Community, Identity, Power and Education* (pp. 75–94). Melbourne: Language Australia Ltd. Retrieved March 26, 2021 from https://www.academia.edu/26244861/Englishes_in_Asia_Communication_Identity_Power_and_Education

Azirah Hashim. (2007). The use of Malaysian English in creative writing. *Asian Englishes*, *10*(2), 30–43.

Azirah Hashim & Tan, R. S. K. (2012). Malaysian English. In Low E. L. & Azirah Hashim (Eds.), *English in Southeast Asia: Features, policy and language in use* (pp. 55–74). Amsterdam: John Benjamins.

Baskaran, L. M. (1994). The Malaysian English mosaic. *English Today*, *10*(1), 27–32.

Baskaran, L. M. (2005). *A Malaysian English primer: Aspects of Malaysian English features*. Kuala Lumpur: University of Malaya Press.

Baskaran, L. (2008). Malaysian English: Phonology. In R. Mesthrie (Ed.), *Varieties of English 4: Africa, South and Southeast Asia* (pp. 278–291). Berlin: Mouton de Gruyter.

Biber, D., & Gray, B. (2011). Grammar emerging in the noun phrase: the influence of written language use. *English Language and Linguistics*, *15*, 223–250.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). *LancsBox* (version 5.0). [computer software]. Available at http://corpora.lancs.ac.uk/lancsbox.

Brown, P., & Levinson, S. C. (1987). *Studies in interactional sociolinguistics, 4. Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Bybee, J. L. (2012). Diachronic Linguistics In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 945–987). Oxford: Oxford University Press.

Collins, P., Borlongan, A. M., & Yao, X. (2014). Modality in Philippine English: A Diachronic study. *Journal of English Linguistics*, *42*(1), 68–88.

Crews, K. D. (1993). *Copyright, fair use, and the challenge for universities: Promoting the progress of higher education*. London: The University of Chicago Press.

Davies, M. (2010). *Corpus of Historical American English (COHA)*. United States of America: Brigham Young University. Available at https://www.english-corpora.org/coha/

Fernando, L. (1966). Malaysia. *The Journal of Commonwealth Literature*, *1*(1), 55–59.

Gablasova, D, Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–158.

Gill, S. K. (2002). *International communication: English language challenges for Malaysia*. Selangor: Universiti Putra Malaysia Press.

Hajar, Abdul Rahim. (2008). The evolution of Malaysian English: Influences from within. In Shakila, Abdul Manan & L. Sinha (Eds.), *Exploring Space: Trends in Literature, Linguistics and Translation* (pp. 1–19). Newcastle: Cambridge Scholars Publishing.

Hajar, Abdul Rahim. (2014). *Malaysian English lexis: postcolonial and beyond*. In Hajar Abdul Rahim & Shakila Abdul Manan (Eds.), *English in Malaysia: Postcolonial and Beyond* (pp. 35–54). Frankfurt: Peter Lang.

Hajar, Abdul Rahim & Harshita Aini Haroon. (2003). The use of native lexical items in English texts as a codeswitching strategy. In S. Granger & S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges* (pp. 159–175). Amsterdam: Rodopi.

Hajar, Abdul Rahim & Shakila, Abdul Manan. (2014). Postcolonial Malaysian English: Realities and Prospects. In Hajar Abdul Rahim & Shakila Abdul Manan (Eds.), *English in Malaysia: Postcolonial and Beyond* (pp. 9–34). Frankfurt: Peter Lang.

Hansen, B. (2017). The ICE metadata and the study of Hong Kong English. *World Englishes*, *36*(3), 471–486.

Hilpert, M., & Gries, S. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 36–53). Cambridge: Cambridge University Press.

Hundt, M., & Mair, C. (1999). "Agile" and "uptight" genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, *4*(2), 221–42.

Kilgariff, A., Rundell, M., & Uí Dhonnchadha, E. (2006). Efficient Corpus Development for Lexicography: Building the New Corpus for Ireland. *Language Resources and Evaluation*, *40*, 127–52.

Kirkpatrick, A. (2007). *World Englishes: Implications for international communication and English language teaching*. Cambridge: Cambridge University Press.

Kohnen, T. (2007). From Helsinki through the centuries: The design and development of English diachronic corpora. In P. Pahta, I. Taavitsainen, T. Nevalainen, & J. Tyrkkö (Eds.), *Towards Multimedia in Corpus Studies*. Helsinki: Research Unit for Variation, Contacts and Change in English. Retrieved March, 5 2021 from http://www.helsinki.fi/varieng/series/volumes/02/kohnen/

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical stud*y. Cambridge: Cambridge University Press.

Lent, J. A. (2001). New Straits Times. In D. Jones (Ed.), *Censorship: A world encyclopedia* (pp. 1701–1702). New York: Routledge.

Lim, G. (2001). Till divorce do us part: The case of Singaporean and Malaysian English. In V. B. Y. Ooi (Ed.), *Evolving identities: The English language in Singapore and Malaysia* (pp. 125–139). Singapore: Times Academic Press.

Low, E. L. (2021). English in Singapore and Malaysia: Differences and similarities. In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes* (pp. 298–318). London: Routledge.

Low, E. L., & Tan, R. S. K. (2016). Convergence and divergence of English in Malaysia and Singapore. In G. Leitner, A. Hashim, & H-G. Wolf (Eds.), *Communicating with Asia: The future of English as a global language* (pp. 43–54). Cambridge: Cambridge University Press.

Lowenberg, P. H. (1991). Variation in Malaysian EnglishL. The pragmatics of language in contact. In J. Cheshire (Ed.), *English around the world: Sociolinguistics perspectives* (pp. 364–375). Cambridge: Cambridge University Press.

Lowenberg, P. H. (1992). The marking of ethnicity in Malaysian English literature: Nativization and its functions. *World Englishes*, *11*(2/3), 251–258.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. New York: Routledge.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Taylor & Francis.

Morais, E. (2001). Lectal varieties of Malaysian English. In V. B. Y. Ooi (Ed.), *Evolving identities: The English language in Singapore and Malaysia* (pp. 33–52). Singapore: Times Academic Press.

Mukherjee, J. (2007). Steady States in the Evolution of New Englishes: Present-Day Indian English as an Equilibrium. *Journal of English Linguistics*, *35*(2), 157–187.

Newbrook, M. (2006). Malaysian English: Status, Norms, Some Grammatical and Lexical Features. In K. Bolton & B. B. Kachru (Eds.), *World Englishes: Critical concepts in Linguistics, Volume II* (pp. 390–417). London: Routledge.

Nurul Farhana, Low Abdullah. (2014). Malaysian English in postcolonial adaptations of Shakespeare in Malaysia. In Hajar Abdul Rahim & Shakila Abdul Manan (eds.), *English in Malaysia: Postcolonial and Beyond* (185–220). Bern: Peter Lang.

Ong, S. B. Christina. (2019). *Mesolectal Malaysian English Corpus*. On-going project. Penang: Universiti Sains Malaysia.

Partington, A. (2010). *Modern diachronic corpus-assisted discourse studies on UK newspapers*. Edinburgh: Edinburgh University Press.

Pillai, S., Zuraidah Mohd. Don, Knowles, G., & Tang, J. (2010). Malaysian English: an instrumental analysis of vowel contrasts. *World Englishes*, *29*(2), 159–172.

Platt, J., & Weber, H. (1980). *English in Singapore and Malaysia: Status, features, functions*. Kuala Lumpur: Oxford University Press.

Rajadurai, J. (2004). The faces and facets of English in Malaysia. *English Today*, *20*(4), 54–58.

Rissanen, M. (2008). *Helsinki Corpus of English Texts*. Retrieved May, 7 2021 from www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus

Roslan Sadjirin, Roslina Abdul Aziz, Noli Maishara Nordin, Mohd Rozaidi Ismail & Norzie Diana Baharum. (2018). The Development of Malaysian Corpus of Financial English (MaCFE). *GEMA Online® Journal of Language Studies*, *18*(3), 73–100.

Sagi, E., Kaufmann, S., & Clark, B. (2012). Tracing semantic change with latent semantic analysis'. In K. Allan & Justyna A. Robinson (Eds.), *Current methods in historical semantics* (pp. 161–183). New York: Mouton de Gruyter.

Schneider, E. (2007). *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

Shakila, Abdul Manan. (2014). The linguistics of creativity: Nativising Malaysian postcolonial creative writings in English. In Hajar Abdul Rahim & Shakila Abdul Manan (Eds.), *English in Malaysia: Postcolonial and Beyond* (pp. 221–252). Bern: Peter Lang.

Tagliamonte, Sali A. (2006). Historical change in synchronic perspective: the legacy of British Dialects. In A. V. Kemenade & B. Los (Eds.), *The Handbook of the History of English* (pp. 477–506). Oxford: Blackwell.

Tan, K. H. (2015). *Malaysian Online English Sports News Corpus (MOSNEC)*. Selangor: Universiti Kebangsaan Malaysia. Available at https://tankimhua.com/mosnec/

Tan, S. I. (2009). Lexical borrowing in Malaysian English: Influences of Malay. *Lexis*, *3*, 11–62.

Tan, S. I. (2013). *Malaysian English: Language contact and change*. Frankfurt: Peter Lang.

Tongue, Ray K. (1974). *The English of Singapore and Malaysia*. Singapore: Eastern University Press.

Wallis S., Aarts, B., Gabriel, O., & Kavalova, Y. (2016). *The Diachronic Corpus of Present-Day Spoken English (DCPSE)*. London: Survey of English Usage. Retrieved April, 22 2020 from https://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm

Weisser, M. (2016a). *ICEWeb*. Available at http://martinweisser.org/ling_soft.html#iceweb

Weisser, M. (2016b). *Practical corpus linguistics: An introduction to corpus-based language analysis*. Oxford: Wiley-Blackwell.

Zaharom Nain. (2019). *Reuters Institute Digital News Report 2019*. Retrieved Sept, 28 2020 from http://www.digitalnewsreport.org/survey/2018/malaysia-2018/

Zanettin, F. (2011). Translation and corpus design. *SYNAPS – A Journal of Professional Communication*, *26*, 14–23.

ABOUT THE AUTHORS

Hajar Abdul Rahim is a professor of linguistics at the School of Humanities, Universiti Sains Malaysia, Penang. Her current research areas include Malaysian English and culture in ELT. She is currently leading a project on the trajectory of Malaysian English and development of a diachronic Malaysian English corpus.

Raihana Abu Hasan is currently a Ph.D student at Universiti Teknologi PETRONAS. She has been involved in a number of research projects related to language education and linguistics. Her main areas of interest are morphosyntax and reading.

Ang Leng Hong (Ph.D) is a senior lecturer at the School of Humanities, Universiti Sains Malaysia, Penang. Her main research interests include corpus linguistics, phraseology and vocabulary studies. She is currently working on a project on the phraseology in academic writing.

Siti Aeisha Joharry (Ph.D) is a senior lecturer at the Akademi Pengajian Bahasa, UiTM Shah Alam. Her research interests include corpus linguistics, media discourse and English for professional communication, and actively works with the Malaysian Corpus Research (MCRN) that hosts monthly webinars on corpus research in Malaysia (https://malaysiancorpusnetwork.wordpress.com/webinars/).

eISSN: 2550-2131
ISSN: 1675-8021