Kajian Malaysia, Vol. 32, Supp. 1, 2014, 17–43

CORPUS RESEARCH IN MALAYSIA: A BIBLIOGRAPHIC ANALYSIS

Siti Aeisha Joharry[1] and Hajar Abdul Rahim[2*]

[1]Department of Linguistics, Faculty of Arts and Social Sciences, University of Sydney, Australia
[2]School of Humanities, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia
[*]Corresponding author: hajar@usm.my

The literature on corpus-related research in Malaysia suggests that studies that have been done include a spectrum of issues on the Malay language as well as those on the use of the English language. However, to date, there has been no attempt to analyse the development of corpus research in Malaysia. To fill this gap and to better understand the state of corpus research in Malaysia, a bibliographic analysis of corpus-related studies in Malaysia was carried out. In an attempt to use published research as data for the bibliographic analysis, the Google Scholar search system was used to source for corpus-related studies in Malaysia. The bibliographic analysis, based on published studies between 1996 and 2012, suggests that there is an upward trend in the use of the corpus method in language research in Malaysia. And a thematic analysis of the studies shows that corpus research in Malaysia, thus far, have focused on five main areas of research : (1) English language use in Malaysia, (2) Malaysian English learner language, (3) Malaysian textbook content, (4) Malay language description and lexicography, and (5) corpora development. Based on research issues and number of studies, corpus research in English and in Malay differs to some extent in terms of research concerns. As well, studies on English outnumber those on Malay despite the fact that corpus research on the latter began much earlier.

Keywords: corpus research in Malaysia, bibliographic analysis, Malaysian English, Malay, corpora development

INTRODUCTION

Corpus research has gained much importance and revolutionised research in various areas of linguistics in the last three decades. The empirical nature of corpus research has also enriched related fields of knowledge through numerous interdisciplinary and cross-linguistic studies. Realising the contribution that corpus research can make to language description, pedagogy and planning, more and more local linguists, researchers and institutions have developed and are developing corpora (English language and Malay language) for research. The availability of these language databases has facilitated local and international researchers' work on English as well as on Malay language use in Malaysia and contributed to a considerable increase in the use of corpora and corpus-based research in Malaysia, particularly in the last decade.

Despite the rise in the number of language corpora and corpus-based research in Malaysia, thus far, no attempt has been made to understand the state of corpus research in Malaysia. This paper addresses the gap in the literature based on a bibliographic analysis of corpus-related studies in Malaysia. The findings of the study, it is hoped, will provide new knowledge on corpus research in Malaysia with regard to research focus and issues, as well as, its development as a research area.

METHODOLOGY

The study involved two main phases, that is, the data collection and data analysis stages. The former involves a survey of published corpus-related studies in Malaysia while the latter involves a qualitative analysis of the published research works. To obtain the data needed for the analysis, a survey of published corpus-related works using Google Scholar was carried out over a period of three months, from June to August 2012. The Google Scholar search generated 100 pages of references on corpus-related research which formed the primary source of data on corpus-related works in Malaysia. The process of ploughing through the 100-page Google Scholar citations found only 42 corpus-related studies on either Malay or English in Malaysia. The 42 published studies formed the data for the bibliographic analysis that was carried out.

Analysis

The bibliographic analysis essentially involved reading the references, identifying the research focus and summarising the content. The data then had to be systematically categorised to facilitate the discussion on the state of corpus research in Malaysia. While a chronological analysis of the published research is a possible way of structuring the data, a thematic analysis of the bibliography was opted for instead. This is because the aim of the study is to better understand the state of corpus research in Malaysia, and not just to review the area. A thematic analysis "is a method for identifying, analysing and reporting patterns (themes) within data" (Braun and Clarke, 2006: 79). According to Braun and Clarke (2006) there are two primary ways to analyse themes, namely inductive or deductive. The former approach "means the themes identified are strongly linked to the data themselves" (Braun and Clarke, 2006: 83) while the latter is driven by the researcher's hypothesis or theoretical interest. Where the present study is concerned, an inductive approach was employed to carry out an analysis that is solely based on the data. In other words, an inductive approach was taken so that the themes or patterns that emerge from the analysis are data-driven.

ANALYSIS AND DISCUSSION

Corpus Research in Malaysia

The thematic analysis of the corpus-related studies first and foremost confirms that corpus research in Malaysia in general involves studies that are related to the Malay language as well as the English language. In terms of language focus, corpus-related studies on English outnumber those on the Malay language. Of the 42 studies, only approximately 20% relate to Malay language corpus development and research. This is quite interesting given that corpus-related research projects on Malay began much earlier than work on English in Malaysia. Apart from the above observations, the thematic analysis also suggests that corpus studies in Malaysia can be grouped loosely according to five major themes or categories of research. They are:

1. English language use in Malaysia
2. Malaysian English learner language
3. Malaysian textbook content
4. Malay language and lexicography
5. Corpora development

The above categories are defined according to the themes or issues that emerged from the analysis of the collection of studies which were found to be of similar values. These categories of research do not only provide some insight into Malaysian corpus research focus areas but also where there may be areas for potential research. In the following sections, the five categories of research form the basis for the discussion on corpus research trends to better understand the state and development of corpus research in Malaysia.

Corpus Research in English Language Use in Malaysia

Studies of language can be divided into two main areas: studies of structure and studies of use (Biber, Conrad and Reppen, 1998). The latter involves investigating how particular linguistic structures occur in differing contexts by a group of speakers/writers. The studies reported below are those that emerged in the search that implemented the corpus-based approach in investigating English language use in Malaysia.

One of the earliest corpus-based studies on English in Malaysia (Imran, 1996) compared the use of "by" between English as a native language variety and

English as a second language variety. The study utilised data from the fiction section of the Wellington Corpus of Written New Zealand English (NZE) and a corpus of Malaysian English (ME) short stories. The study concluded that (a) "by" is higher in frequency in ME than in NZE, (b) "by" is most frequently used to indicate the agentive, and (c) in NZE spatial usage of "by" ranks higher than temporal usage and vice versa in ME. The study suggests Malaysian English language users tend to construct more passive sentences than native speakers, in this case New Zealanders.

Another corpus-based study that looked at English language use in Malaysia is Hajar and Harshita's (2003) study on the use of local lexis (i.e., Malay words) in standard written Malaysian English. Based on an analysis of newspaper reports and editorials, the study found that local words are not just used to fill linguistic gap (e.g., names of local dishes, flora and fauna, artefacts etc.) but also sometimes preferred over English words with equivalent meaning. The reason for this preference, the analysis suggests, is due to the cultural and local connotations that are conveyed in the local words. The study concludes that the pattern of local lexis use in standard Malaysian English to some extent goes against the linguistic convention of borrowing.

Other corpus-based studies that looked at standard Malaysian English use include Banafsheh's (2005) study that compared the use of local lexis in ME and standard Singapore English (SE) and Tan's (2009) study of ME from the perspective of language contact. Banafsheh's study of the frequency and distribution of local lexis in the two varieties contributes to a better understanding of ME in comparison to SE while Tan's (2009) analysis of Chinese lexical borrowings in ME based on the Malaysian English Newspaper Corpus (MEN Corpus) found that among the Chinese varieties, Hokkien, Cantonese and Mandarin appear to be dominant in Malaysian English.

Other studies on English language use in Malaysia include Yee and Bahiyah's (2008) investigation of the personal attributes used by Malaysian female and male adolescents (of Malay, Chinese and Indian ethnicities) in constructing their gender and ethnic identities based on a corpus of personal advertisements in *Galaxie* magazine. Malaysian females and males, according to the study, use neutral traits such as "simple", "open-minded", "friendly", "crazy", "happy go lucky" and "funny" to describe themselves. The study concludes that Malaysian youths are mature and expressive when it comes to disclosing their identity and that these youths are "[aware] of language use for strategic purposes" (Yee and Bahiyah, 2008: 27).

Corpus Research in Malaysian English Learner Language

One of the areas of language research that has benefited the most from the corpus method is learner language, particularly with regard to the English language. In Malaysia, as in many other ESL/EFL contexts, the increase in learner language research is partly due to the development of learner corpora. A learner corpus is a representation of learner language and can be a very useful tool to investigate language learning processes and to improve language learning strategies (Granger, 1998).

One of the earliest corpus-based studies on Malaysian learner language was Arshad's (2004) study on Malaysian school learners' English language development. The study utilised the English of Malaysian School Students (henceforth EMAS) corpus, a collection of untagged and unedited learner corpus containing essays written by approximately 800 students from three age groups (11 year olds: Year 5 of primary school education, 13 year olds: Year 1 of secondary education, and 16 year olds: Year 4 of secondary education).[1] To study the students' language development, the researcher compared the performance of the three age levels in terms of their vocabulary use and language productivity (indicated by the number of sentences per essay and the words per sentence). The results show an increasing rate of sentences per essay as well as words per sentence for the three age levels. The study essentially shows that corpus data can be utilised in investigating non-native language development for language educators.

Others who have used the EMAS corpus to research on Malaysian English learner language include Vethamani, Umi Kalthom and Omid (2008a; 2010); Ang et al. (2011). In their study, Vethamani, Umi Kalthom and Omid (2008a) examined the use of modals in the corpus. The study found that Malaysian students prefer to use the modals *can, will* and *could* and that modals of probability/possibility were much less used in the writings. The study calls for more attention to be given to the learning and teaching of modal verbs, and the need for textbook writers to adhere to the English language syllabus so that students are exposed to other modals like *would* and *shall*. The researchers also found that students tend to repeat the same modals. Their later work, involving descriptive statistics derived from a concordance analysis, reveals Malaysian learners' tendency to use simplification strategies when faced with modals in written work (Vethamani, Umi Kalthom and Omid, 2010).

Ang et al. (2011) used the EMAS subcorpus of 130 essays written by 16 year old Malay students to investigate the types and sources of verb-noun collocational errors in their essays. The analysis reveals that prepositional errors were rated most problematic (41.72%), followed by verb (16.56%) and noun errors (14.24%). With regard to errors in the use of prepositions, it was found that students prefer to include prepositions rather than avoid them (*go *for* fishing) and made most errors with the prepositions *in*, *to* and *into*. The study also found that most students paid little attention to rule restrictions (59.60%). Other collocational errors include approximation (21.19%), L1 transliteration (9.60%), false concept hypothesised (5.63%), and language switch (2.32%). The least influential factor is overgeneralisation (1.66%). Given these findings, Ang et al.

(2011: 42) suggest that the language of Malaysian learners is an Interlanguage (IL) "as it possesses the IL characteristics suggested by Adjemian (1976)."

Another Malaysian learner corpus that has facilitated research on various L2 writing issues is the Corpus Archive of Learner English in Sabah-Sarawak (CALES). The corpus is modelled after the International Corpus of Learner English (henceforth ICLE) (Granger, 1998; Granger, Hung and Petch-Tyson, 2002) which contains argumentative essays written by higher intermediate to advanced learners of English developed in different countries around the world. In keeping with the methodological and design principles of ICLE (Granger, 1998; Granger, Hung and Petch-Tyson, 2002), CALES contains argumentative essays of between 200 and 800 words by Diploma and Degree level students taking English proficiency courses at three institutions of higher learning in the states of Sarawak and Sabah in East Malaysia. One study (Botley and Dillah, 2007) used CALES to explore Malaysian undergraduate spelling errors. The study found up to 1,018 spelling errors in the Degree-level sample and 867 errors in the Diploma-level sample. An analysis of the errors based on James' (1998) framework reveals that the errors include doubling, omission, addition, misordering, misuse of punctuation, replacement, L1 influence, US spellings, mispronunciations, word coinage and direct borrowing. In another study Botley (2010) compared the use of idiomatic expressions in Malaysian, British and American students' essays to investigate the nature of the interlanguage of Malaysian learners. The analysis of idiom use and distribution in essays by Malaysians (based on CALES) when compared against those by their native counterparts (based on essays written by British and American undergraduates in the LOCNESS)[2] suggests that Malaysians are highly influenced by their L1 which is the Malay language. Idiomatic forms such "as we can see" (*seperti yang kita lihat*), "as we know" (*seperti yang kita tahu*) and "in our life" (*dalam hidup kita*) frequently show up in the CALES sample but are hardly found in the LOCNESS sample. This, Botley argues, indicates "widescale natural language interference" in Malaysian students' use of idiomatic expressions (Botley, 2010: 141).

Besides the EMAS and CALES corpora, there are others that have been used to study Malaysian English learner language. Kaur and Sarimah (2011) for instance used the Business and Management English Language Learner Corpus (BMELC) to investigate the use and patterns of nouns by Malaysian undergraduates in business and management courses in two different settings in Malaysia (public vs. private higher learning institution). Based on a corpus of different written genres, namely essays, journal writing and a media invitation, the study found that the students were capable of using a variety of word choices according to the different forms of noun such as neutral, singular, plural and proper nouns. There also appears to be a significant number of noun types according to different written genres. Simpler words appear more in journal entries compared to essays and the media invitation writings. The findings offer a better understanding of "the structural patterns of a particular genre and further provide useful insights on structural-based linguistics investigations such as syntactic and morphology analyses" (Kaur and Sarimah, 2011: 24).

Other genre-specific corpus-based studies of learner language include the study by Kamariah and Su'ad (2011) which investigated the collocational competence of undergraduate law students at Universiti Sultan Zainal Abidin, (UniSZA). The study examined the inaccurate use of prepositions in the production of colligations in essays by law undergraduates. The study found that the students do not only produce a small number of prepositional sequences, but also inaccurate collocations of prepositions (*discuss *about*; *sued him *on* the promise). While the errors suggest the existence of L1 interference and overgeneralisation strategies, the study also brought to light the overall attempt of the students to construct possible collocations. The researchers thus stress the importance of teaching prepositions explicitly as "collocations of prepositions are the most essential aspects of language in legal discourse and are worthy of being seriously taught" (Kamariah and Su'ad, 2011: 191).

While the above studies are based on corpora of written English, the corpus method has also been used to research on Malaysian English learner oral communication. A study by Sharifah Zakiah et al. (2009) for instance used a small corpus of group discussion to identify and classify grammatical errors and other performance factors in students' oral communication. The study found that the participating students displayed frequent lexical, grammatical and formal errors. The findings accord Botley, Haykal and Monaliza's (2005) study of a large corpus of university students' essays which found that noun and verb tense errors were most frequent. The study also reveals that the students appeared to have hesitations, filled pauses, repeats and reformulations in their oral discussion.

Corpus-based learner language research in Malaysia focuses on, but is not limited to school and undergraduate students' language. The bibliographic analysis shows that the corpus method has also facilitated research on postgraduates' language. A study by Yunisrina (2009) on the colligations "to" and "for" for instance is based on a corpus of essays written by postgraduates in the Faculty of Engineering at the University of Malaya. The study found that postgraduates, despite being academically advanced, still have problems with the use of the prepositions "to" and "for".

The studies reported in this section suggest that the use of corpus method to analyse learner language began and increased in tandem with the development of learner corpora by local researchers in the last ten years. Besides large corpora such as the EMAS and CALES, there are smaller and genre-specific learner corpora that have been developed by researchers for their own research. The increase in the number and types of learner corpora does not only signal local researchers' interest but also their confidence in the corpus method to investigate learner language.

Corpus Research in Malaysian Textbook Content

The textbook is the main material used by teachers and learners in the classroom. In Malaysia, textbooks for schools are developed by Malaysians (usually teachers and educators) according to the guidelines set by the Ministry of Education. Corpus studies in the area of textbook content and development in Malaysia to a large extent have concentrated on language in textbooks used in Malaysian schools.

The studies include an investigation on common word classes among keywords (Mukundan and Sujatha, 2006) in Science, Mathematics and English textbooks used in Form 1 (first year secondary school) classrooms. The study was among the first corpus-based studies of Malaysian textbooks. More recent corpus-based research focuses on issues of grammar and structure. One study (Mukundan and Khojasteh, 2011) compared modal auxiliary verbs in prescribed Malaysian textbooks with those in the British National Corpus (BNC). Interestingly, the study reveals that modal auxiliary verbs in the Malaysian prescribed textbooks were not completely similar to those used by native speakers. Khojasteh and Kafipour (2012a; 2012b), in their studies extended the research on modal auxiliaries and modal verb phrase structures use in Malaysian English language textbooks. Other studies on grammar and structure in Malaysian English language textbooks include Zarifi and Mukundan (2012); Mukundan, Leong and Nimehchisalem (2012). The former investigated the distribution of phrasal verb combinations in Malaysian English textbooks with regard to the justification of the selection and presentation of these combinations. And the latter examined the distribution of articles in Malaysian secondary school English language textbooks. The study found that all three articles (*a, an, the*) appear in all five textbooks with an increasing frequency from the lowest level (secondary 1) to the highest (secondary 5). However, colligational patterns of the articles showed inconsistencies as some patterns were more emphasised than others. Research on language patterns and structure for the most part is intended to reveal language issues that need attention by textbook developers. Others, such as Sarimah and colleagues' study on the language patterns in Science textbooks, aim to provide insight into how corpus research can inform the development of teaching and learning materials (Sarimah et al., 2008).

Besides the above concerns, textbook research has also delved into cultural matters. Azlina (2004) for instance carried out a study on the representation of "gratitude" in the lexis of Malay school textbooks (year 1–6). Among others, the study found that the expression *terima kasih* (thank you) was most common, and interestingly, children are more often represented as the expresser rather than the receiver of thanks in the school textbooks. Another cultural issue that gained interest among textbook researchers is gender representation. This issue gained researchers' attention because of the global move for "inclusivity" or "non-sexist approach" (Gray, 2002: 157) with regard to course book content. In Malaysia, one of the earliest corpus-based treatments of this issue was by Yuen et al. (2003) who analysed linguistic sexism and sex role stereotyping in Malaysian English language. More recently, Mohd Faeiz et al. (2011) reported findings on gender inequality based on an analysis of action verbs by males and females in secondary textbooks.

The studies show that corpus-based investigation of textbooks is still a fairly recent phenomenon in Malaysia. Nonetheless, the increasing number of studies at the end of the last decade signals a preference for the corpus approach in textbook research. This is possibly because it provides local researchers with a systematic approach to textbook content and development research that allows them to make recommendations for the improvement of national textbooks. Indeed, the employment of the corpus approach has impacted textbook research in Malaysia in a positive way.

Research in Malay Language and Lexicography

Corpus-related research in the area of Malay language in Malaysia has for the most part revolved around issues concerning the development of Malay corpora and the description of the Malay language. The corpus approach made its way into the area of Malay language research with the development of the Malay language corpus by the Dewan Bahasa dan Pustaka (henceforth DBP)[3] in the 1980s. The corpus was developed to facilitate systematic and objective analyses of the Malay language to generate new knowledge on the Malay language that would contribute to the improvement of Malay dictionaries, grammar books and others. The corpus, to date, is the largest Malay corpus in Malaysia (see the section on corpora development for details).

In recent years, more Malay corpora have been developed to facilitate research on Malay language description and translation. Some of the corpora, such as the one developed by the Practical Grammar of Malay Project, use data extracted from the large collections of electronic texts in the DBP corpus. The corpus, known as the Malay Practical Grammar Corpus (MPGC) is an on-going Malay corpus project that involves analysis of written texts in Malay from several major genres (mainly newspapers, magazines and books). The MPGC corpus is not only developed to examine aspects and grammar of Malay but also to acquaint teachers and students with corpus analysis and its application in the classroom (Imran et al., 2004).

Another research project (Zuraidah, 2010) called MALEX (MaLay LEXicon), uses data from the DBP corpus in conjunction with other Malay texts. The MALEX project is an attempt to produce a working system for the automatic processing of Malay texts. MALEX "is an annotated lexicon designed as a relational database" (Zuraidah, 2010: 90). According to Zuraidah (2010), the data for the project is a compilation of texts from novels (approximately 800,000 words), newspaper corpus (approximately 5 million words), academic text

(20,000 words), as well as speeches by the former Malaysian Prime Minister, Dr Mahathir Mohamad (1.3 million words). The researcher regards the compilation of the different genres of texts as an "archive" rather than a corpus. The integrated design of MALEX comprises a spelling normaliser, a tag set, a list of lemmas, morphological derivations, and a pronouncing dictionary.

More recently, Lee and Low (2011) report on the development of an online Malay language lexical database based on a corpus of Malay textbooks for primary schools. Due to the dearth of corpora on children's reading materials, the corpus is developed to facilitate research and the teaching and learning of Malay to primary level students. The corpus comprises data from "the Malay language textbooks used in bilingual National schools and trilingual National-type schools in Malaysia" (Lee and Low, 2011: 97).

Besides developing corpora for the automatic processing of Malay texts and for the teaching and learning of Malay, local researchers have also developed parallel corpora for machine translation research[4]. Suhaimi and Normaziah (2004) for instance report on the development of a 250,000 word English-Malay bilingual parallel corpus in the domain of agriculture and health. Another research project worth mentioning is the Translation Memory of legal texts and a Glossary of legal terminology (Tengku Sepora et al., 2009). This team of researchers has now developed a legal English-Malay parallel corpus of about 210,000 words (at the time of study) and a Termbase Glossary of legal terminology. More recently, Norsimah, Intan Safinaz and Imran (2011) investigated the linguistic features of the Malay preposition *untuk* (equivalent to "for" in English) in a translation corpus. The study found various patterns of use for *untuk* between the Malay translation and its original text. Among the recurrent patterns of *untuk* in the corpus is *untuk*+verb ("*untuk mengatasi masalah*" or "to overcome the problem") and *untuk*+noun ("*untuk anak-anak*" or "for the children"). The findings of this study support the "translation universals hypothesis which claims that translation language resembles the normative standard language of the original language" (Norsimah, Intan Safinaz and Imran, 2011: 642).

The availability of corpora and the familiarity of the corpus approach have also increased work in Malay language description. The DBP corpus for instance has made lexical and grammatical studies on Malay more dynamic and empirical with the use of attested data and computational methods. Hajar's (2005) study for instance compared the connotations of the words *perempuan* (female) and *wanita* (women). Based on the DBP newspaper subcorpus, a collocational analysis was carried out to reveal the semantic prosody of the two words. The findings reveal that *wanita* has a better reputation than *perempuan* due to the cultural connotations that arise from their collocations. The study also brought to light the important sense of *perempuan* as a gender marker in Malay, which surprisingly was not represented in the entry for the word in Kamus Dewan Bahasa, an authoritative Malay dictionary published by the DBP.

Chung (2010; 2011) examined the Malay numerical classifier *buah* and the uses of the morpheme *ter-* using the corpus approach. In his study of *buah* (numeral classifier for things), Chung observes that *buah* not only serves as a classifier but also has a cultural role. The study examined a list of noun collocates classified by *buah* based on 5,009 newspaper articles randomly collected via the search archive system of *Utusan Malaysia*, a mainstream Malay daily. In his study of the prefix *ter-* (prefix that implies something happening by accident) in Malay, he examines the complexities of the usage of *ter-* in both modern and historical texts. Findings from this study reveal that "the active construction of *ter-* appeared to decrease in modern texts while the passive construction is increasing" (Chung, 2011: 808).

The development of various Malay corpora indicates researchers' recognition of the significance of the method in Malay language research. Yet, the limited number of studies that emerged in the Google Scholar survey suggests that the corpus method has yet to be fully exploited by Malay language researchers.

Language Corpora Development in Malaysia

One of the first corpora to be developed in Malaysia is the Malay language corpus that is being developed by the DBP. The corpus as it stands contains approximately 128 million words from various Malay written and spoken texts. The development of the corpus dates back to the 1980s and presently, this corpus may be considered the largest corpus in Malaysia. With the increased interest in the corpus approach in Malay language research and lexicography, more corpora have been developed. Many of them, such as MPGC, the DBP-UKM corpus, and the MALEX corpus are developed for specific linguistic or language teaching purposes. Despite these developments in Malay language research, Malaysian English corpora still outnumber Malay language corpora. This is evident from the preceding discussion and reports on various corpus-based studies in the Malaysian contexts. In the past two decades there have been a number of important English corpora that have been developed, especially by language researchers in public universities in Malaysia. Some include those mentioned earlier such as the EMAS and CALES. Others include MACLE, COMEL, ICE and ELC, which are discussed below.

As mentioned earlier, the EMAS corpus is a learner corpus. It consists of written essays by Malaysian learners of English from selected primary and secondary schools in three states in Malaysia (Arshad et al., 2002). The CALES which stands for Corpus Archive of Learner English Sabah-Sarawak (Botley, Haykal and Monaliza, 2005) is also a learner corpus but which has not been made public. MACLE is the Malaysian Corpus of Learner English, a collection of students' essays (Knowles and Zuraidah, 2003) based at the University of Malaya. Another corpus that is based there is the Corpus of

Malaysian English (COMEL), a spoken corpus project which is still in the process of development.

Malaysia also participates in the International Corpus of English (ICE) project. The ICE project began in 1990 and aimed at collecting material for comparative studies of English used in L1 and L2 contexts around the world. Each completed ICE corpus contains one million words of spoken and written texts. Each country that is involved in this project collects both written and spoken data according to guidelines and categorisation based on the ICE framework. The Malaysian component of ICE, known as ICE Malaysia, is based at Universiti Sains Malaysia. The corpus is still being developed and currently stands at approximately 250,000 words. The corpus has been used to facilitate a number of studies on Malaysian English (e.g., Hajar and Harshita, 2003; Banafsheh, 2005; Hajar, 2012).

Another important Malaysian English corpus is the Engineering Lecture Corpus (ELC) which is a small corpus of academic lectures, modelled on the BASE (British Academic Spoken English) corpus. The ELC is a growing collection of transcripts of English-medium engineering lectures from around the world. The project aims to gain insights into engineering discourse in English which can be put to immediate use in various student and staff development programmes by the participating institutions. As well, the corpus serves as a pilot for the development of a full scale Malaysian Academic Spoken English (MASE) corpus in the future.

CONCLUSION

In the present study, observations on the state of corpus research in Malaysia are based on the analysis of 42 studies generated from the Google Scholar search that was carried out. It is important to note that these were the only studies that were found to be relevant despite the effort to survey as many studies as possible. There are perhaps other studies that have been carried out, but which inadvertently have been left out in the analysis of the present study because they are not cited by Google Scholar or have not been published. Publications in Malay particularly may not have been picked-up by Google Scholar, hence the limited data on corpus research in the Malay language.

Notwithstanding the limitations stated above, it can be deduced from the bibliographic analysis of corpus-related studies that there has been an upward trend in corpus research in Malaysia. The analysis of data generated from the Google Scholar search shows that the earliest publication was in 1996 (Imran, 1996), and this is also the only publication for the 1990s that was generated from the Google Scholar search. The trend changed dramatically in the following decade, because as evident from the bibliographic analysis, 25 studies were published between 2001 and 2010. And crucially, the Google Scholar search found 13 studies published after 2010. These studies, published within the span of two years (2011–2012), are half as many as those published over the whole of the last decade. This suggests that the corpus approach, especially in the last decade, has gained popularity as a language research methodology in Malaysia.

The rise in corpora development and corpus-related research in Malaysia, as shown in the analysis of the present study involves research in Malay and English. This commonality between the two languages however cannot be extended to their corpus research focus areas. The thematic analysis that was carried out suggests that corpus research in the Malay language has centred mostly on the description, lexicography and the translation of the language. This owes much to the publicly accessible DBP Malay corpus. The corpus, as shown in the discussion above, is not only available for research but has also been used by researchers as a resource for the development of other Malay corpora (e.g., DBP-UKM, MALEX). The availability of the DBP corpus and other Malay corpora has enabled Malay language researchers to move away from purely theoretical analysis to data-driven empirical studies on Malay[5]. Malay linguistics and grammar research has been empowered with the corpus method and it would not be wrong to suggest that this is one of the strengths of Malay corpus linguistics in Malaysia. With regard to English in Malaysia, the development of Malaysian English corpora (e.g., ICE, MEN), English learner corpora (e.g., EMAS, CALES, MACLE) and textbook corpora has motivated many corpus-based studies on Malaysian English use, learner language and English teaching materials. Based on the number and research concerns of the studies that were analysed, corpus research in English in Malaysia largely focuses on the use of English as a local variety and on issues pertaining to its teaching and learning as a second language.

Based on the research trend, it is very likely that the corpus method will continue to appeal more local language researchers. This phenomenon will not only increase the number of studies but also widen the spectrum of research issues in Malay and English in Malaysia. With the development of other types of corpora, such as Malay language learner and pedagogic corpora, research in Malay may extend beyond language description, translation and lexicography to include Malay language learning and teaching issues. As regards English in Malaysia, emerging issues surrounding its use as a second language and as a new variety of English will certainly continue to encourage the development of more Malaysian English corpora, learner corpora as well as genre-specific corpora. These developments, among many others, will undoubtedly ensure the continued employment of corpora and the corpus approach in language research in Malaysia.

ACKNOWLEDGEMENT

The study reported in this paper was funded by the Universiti Sains Malaysia Short Term Research Grant entitled "The Development of the International Corpus of Malaysian English–Phase 3" (2012–2014). Grant number: 304/PHUMANITI/6312007.

NOTES

1. The essay by the youngest group was based on a picture series. Students were given an hour to write an essay describing events that occur during a fishing trip. The essay by the 13 year olds is entitled "The happiest day of my life" which teachers of selected schools administered to the respondents. The collection of essays by the 16 year olds was selected by teachers from essays that respondents had completed as part of their regular school work.
2. The Louvain Corpus of Native English Essays (LOCNESS) is a corpus of native English essays comprising British A-level and university students' essays as well as American university students' essays.
3. The Dewan Bahasa dan Pustaka or the DBP is the Malaysian national language and literature agency.
4. Work in machine translation in Malaysia began more than two decades ago at the translation unit at the School of Computer Sciences, Universiti Sains Malaysia (see Chuah and Zaharin, 2002 for details).
5. Many of these studies are written in Malay and published in local journals (e.g., *Jurnal Bahasa* published by Dewan Bahasa dan Pustaka). Unfortunately, many of these studies are not cited in Google Scholar.

AN ANNOTATED BIBLIOGRAPHY

Ang, L. H., Hajar Abdul Rahim, K. H. Tan and Khazriyati Salehuddin. 2011. Collocations in Malaysian English learners' writing: A corpus-based error analysis. *3L: The Southeast Asian Journal of English Language Studies* 17(special issue): 31–44.

This study investigates types and sources of verb-noun collocational errors in a subcorpus of EMAS. The study found seven types of collocational errors, the most frequent being those related to prepositions.

Arshad Abdul Samad. 2004. Beyond concordance lines: Using concordances to investigating language development. *Internet Journal of e-Language Learning & Teaching* 1(1): 43–51.

This article reports on a study that examined the language development of Malaysian learners of English based on data in the EMAS corpus.

Azlina Murad Sani. 2004. Gratitude as a moral value: A corpus-based analysis of its representation in the lexis of Malay primary school textbooks. *Malaysian Journal of Learning & Instruction* 1(1): 73–83.

This is a study on the representation of "gratitude" in the lexis of Malay school textbooks (year 1–6) used in primary schools in Malaysia. A corpus of textbooks was analysed to investigate the portrayal of gratitude via the three corpus techniques, namely frequency, phraseology and collocational analysis. It was found that the expression *terima kasih* (thank you) was most common. The analysis reveals that children are more often represented as an expresser of thanks, but less as a receiver. There are also fewer occurrences of gratitude expressions between children and their peers in the texts.

Banafsheh, B. 2005. A corpus-based comparative study of the use of native lexical items in Malaysian and Singaporean English. MA diss., School of Humanities, Universiti Sains Malaysia.

This study investigated the pattern of use of local words, namely Malay, Chinese and Tamil in the use of Standard English in Malaysia and Singapore. Newspaper texts from the subcorpus of ICE Malaysia and ICE Singapore were used to generate data for the comparative analysis.

Botley, S., Haykal Hammaad Zin and Monaliza Sarbini. 2005. Lexical and grammatical transfer by Malaysian student writers. Paper presented at the Proceedings of the 10th International Conference on Translation, Universiti Malaysia Sabah, Kota Kinabalu. 2–4 August.

Botley, S. and D. Dillah. 2007. Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research* 3: 74–93.

Botley, S. 2010. A corpus-based comparison of idiom use by Malaysian, British and American students. In *Proceeding of International Conference on Science and Social Research*, 139–144. Kuala Lumpur: Institute of Electrical and Electronics Engineers (IEEE).

This study exploits the CALES corpus to identify the most frequently-occurring features of IL + compare to sample of essays produced by native speaker students. The most frequent idiomatic expressions or n-grams were identified in the sub-corpus and they seemed to be influenced by the Malay language. Comparisons were then made with native speakers' corpus (LOCNESS corpus).

Chuah, C-K. and Zaharin Yusoff. 2002. Computational linguistics at Universiti Sains Malaysia. Paper presented at the International Conference on Language Resources and Evaluation (LREC'02), 1838–1842. The University of Las Palmes de Gran Canaria, Canary Island-Spain. 29–31 May.

The paper provides a brief history of the translation unit in the School of Computer Sciences, Universiti Sains Malaysia, known as UTMK (Unit Terjemahan Melalui

Komputer) and its projects and research collaborations. At the time of publication, the unit's focus was to collaborate on computing linguistic resources on Malay. The paper reports on a three-way collaborative effort between Malaysia, Indonesia and Brunei to set up a Malay language portal that would serve as an online linguistic resource for the Malay language.

Chung, S-F. 2010. Numeral classifier *buah* in Malay: A corpus-based study. *Language and Linguistics* 11(3): 553–577.

This is a corpus-based analysis of the classifier *buah* in Standard Malay. Concordancers (AntConc 3.2.1 and Wordsmith Tools ver5) were used to extract and examine the list of collocates for the classifier *buah*. Based on a corpus of newspaper articles from the *Utusan Malaysia* local daily (online version) published between 2005 and 2010 collected, the study found that *buah* does not seem to function as a general classifier. The researcher therefore argues for the elimination of *buah* as a general classifier.

Chung, S-F. 2011. Uses of *ter-* in Malay: A corpus-based study. *Journal of Pragmatics* 43(3): 799–813.

This study focused on the use of the prefix *ter-* in Malay. The analysis is based on the use of words with the suffix *ter-* in comparison to the prefix *di-* and the adversative passive *kena*. Based on an analysis of a modern corpus and a historical corpus, the study found that *ter-* appears less frequently than the prefix *di-* butmore frequently than *kena*. The study provides a useful methodology for the analysis of Malay morphology based on corpus data.

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

This is the first published study that is based on ICE Malaysia. The focus of the study is the use of local words, namely Malay in standard written Malaysian English (ME). The source of the data is the newspaper subcorpus of ICE Malaysia. Based on a semantic analysis of words generated from the corpus, the study found that that local words found in standard ME include those that have English equivalents. The pattern of use of the local words suggests ME defies the linguistic convention of borrowing as they are used in some places as a codeswitching strategy to construct meaning.

Hajar Abdul Rahim. 2012. Corpora in ESL/EFL teaching. In *English in multicultural Malaysia: Pedagogy and applied research*, ed. Zuraidah Mohd Don. Kuala Lumpur: University of Malaya Press.

This paper discusses the issue of corpora in language teaching with a focus on issues surrounding the use of corpora in teaching in non-native contexts such as Malaysia.

Imran Ho-Abdullah. 1996. By ESL writers vs by native writers: A corpus analysis of native and non-native speakers' written English. *Deep South* 2(3). http://www.otago.ac.nz/deepsouth/vol2no3/imran.html/

This paper presents the results of a preliminary 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). The aim of the study is to contrast prepositional usage between English as a native language variety and English as a second language variety.

Imran Ho-Abdullah, Zaharani Ahmad, Rusdi Abdul Ghani, Nor Hashimah and Idris Aman. 2004. A practical grammar of Malay – A corpus-based approach to the description of Malay: Extending the possibilities for endless and lifelong language learning. Paper presented at the First COLLA Regional Workshop, Putrajaya, Malaysia. 28–29 June.

The paper discusses the Malay Practical Grammar Corpus (MPGC) and highlights the potential of the corpus in language teaching and learning. They recommended that a conceptual framework for the use of corpus linguistics in language teaching and learning must take into consideration access to technology and minimal IT skills.

Kamariah Yunus and Su'ad Awab. 2011. Collocational competence among Malaysian undergraduate law students. *Malaysian Journal of ELT Research* 7(1): 151–202.

Kaur, M. and Sarimah Shamsudin. 2011. Extracting noun forms: A lesson learnt. *International Journal of Language Studies* 5(4): 19–32.

This is an analysis of NOUN usage in business and management texts from two higher learning institutions in Malaysia. The study looked at frequencies of neutral, singular, plural and proper nouns in written texts. The analysis reveals that there is an over-use of the singular noun form in the corpus.

Khojasteh, L. and R. Kafipour. 2012a. Are modal auxiliaries in Malaysian English language textbooks in line with their usage in real language? *English Language Teaching* 5(2): 68–77.

______. 2012b. Have the modal verb phrase structures been well presented in Malaysian English language textbooks? *English Language and Literature Studies* 2(1): 35–41.

______. 2012c. Non-empirically based teaching materials can be positively misleading: A case of modal auxiliary verbs in Malaysian English language textbooks. *English Language Teaching* 5(3): 62–72.

Khojasteh and Kafipour looked at modal auxiliaries in Malaysian English textbooks from Form 1–5 by adopting a corpus-based study. The authors discovered that a

discrepancy exists between the frequencies of modal auxiliaries in the Malaysian textbook corpus and native English corpus.

Knowles, G. and Zuraidah Mohd Don. 2003. Tagging a corpus of Malay texts, and coping with "syntactic drift". In *Proceedings of the corpus linguistics 2003 conference*, 422–428. UK: University of Lancaster, Centre for Computer Corpus Research on Language.

This study initiates the tagging of 120,000 words containing literary texts of four modern Malay novels.

Lee, L. W. and M. L. Hui. 2011. Developing an online Malay language word corpus for primary schools. *International Journal of Education and Development Using ICT* 7(3): 96–101.

This study highlights common words that occur in the Malay language textbooks based on a corpus of Malaysian primary school Malay textbooks.

Mohd Faeiz Ikram Mohd Jasmani, Mohamad Subakir Mohd Yasin, Bahiyah Abdul Hamid, Yuen Chee Keong, Zarina Othman and Azhar Jaludin. 2011. Verbs and gender: The hidden agenda of a multicultural society. *3L: The Southeast Asian Journal of English Language Studies* 17(special issue): 61–73.

This study reports the findings from an analysis of gender inequality in textbooks. An analysis of action verbs in English textbooks used in secondary schools was carried out to unravel male and female occurrences in the text and gender roles. The corpus-based study brings to light interesting facts about gender stereotyping and inequality in the textbooks that were studied.

Mukundan, J., A. C. H. Leong and V. Nimehchisalem. 2012. Distribution of articles in Malaysian secondary school English language textbooks. *English Language and Literature Studies* 2(2): 62–70.

Mukundan, J. and L. Khojasteh. 2011. Modal auxiliary verbs in prescribed Malaysian English textbooks. *English Language Teaching* 4(1): 79–89.

Mukundan, J. and Anealka Aziz. 2009. Loading and distribution of the 2000 high frequency words in Malaysian English language textbooks for Form 1 to Form 5. *Pertanika Journal of Social Sciences and Humanities* 17(2): 141–152.

Mukundan, J. and M. Sujatha. 2006. Lexical similarities and differences in the mathematics, science and English language textbooks. *K@ta* 9(2): 91–111.

Studies by Mukundan et al. (2006‒ 2012) include analyses of common word classes in Science, Mathematics and English language secondary textbooks (Form 1), the patterns and distribution of words in textbooks in comparison to the 2000 High

Frequency word list, and English prepositions in Forms 1, 2 and 3 textbooks. The researchers also compared Malaysian textbook corpus with the BNC in looking at modal auxiliaries. In a recent study on articles used in a textbook corpus (Forms 1–5), the researchers found that the frequency of occurrences of articles increase in trend, but the colligation patterns of the articles were inconsistent.

Norsimah Mat Awal, Intan Safinaz Zainuddin and Imran Ho-Abdullah. 2011. Use of comparable corpus in teaching translation. *Procedia – Social and Behavioral Sciences* 18: 638–642.

This study looked at the linguistic nature of Malay preposition *untuk* in a comparable corpus (Translation Corpus with DBP-UKM corpus of Malay text).

Sarimah Shamsudin, Zaidah Zainal, Yasmin Hanafi Zaid and Salbiah Seliman. 2008. Corpus analysis of primary one science textbooks for designing ELT materials. In *Research in language teaching*, eds. Noor Abidah Mohd Omar and Zaidah Zainal, 135–155. Johor: Penerbit Universiti Teknologi Malaysia.

The study examined the language patterns in science authentic texts based on a corpus of texts from Science textbooks for Primary 1 students in Malaysia. The aim is to provide insight into how corpus analysis can provide Science teachers with new ways to design teaching and learning materials. Methods of using the frequency list and corpus of Science texts for teaching are also suggested in the paper.

Sharifah Zakiah Wan Hassan, Simon Faizal Hakim, Mahdalela Rahim, J. F. Noyem, Sueb Ibrahim, Johnny Ahmad and Kamaruzaman Jusoff. 2009. The communicative ability of Universiti Teknologi MARA Sarawak's graduates. *English Language Teaching* 2(2): 84–91.

This is a study on the oral proficiency in English of Malaysian students based on a small corpus of group discussions. The analysis found that there are frequent lexical, grammatical and formal errors in the subjects' oral communication. The study also reveals that the subjects' communicative ability is affected by communication problems such as hesitation, repetition, incomplete structure and redundancy.

Suhaimi Ab Rahman and Normaziah Abd Aziz. 2004. Improving word alignment in an English–Malay parallel corpus for machine translation. Paper presented at the Language Resources and Evaluation (LREC) workshop on the amazing utility of parallel and comparable corpora, Lisbon, Portugal. 25 May.

The focus of this study is a compilation of an English-Malay bilingual parallel corpus (250,000 words) in the domain of agriculture and health. The goal of the study is to address the language barrier issue in narrowing the digital divide problems in Malaysia.

Tan, S. I. 2009. Lexical borrowing from Chinese languages in Malaysian English. *World Englishes* 28(4): 451–484.

As the title suggests, this is an analysis of the use of Chinese lexis in Malaysian English (ME). Based on an analysis of the Malaysian English Newspaper (MEN)

Corpus, the study found that the vast majority of ME features of Chinese origin are borrowed from Hokkien and Cantonese. Chinese is identified as an important source of novel lexical features in ME.

Tengku Sepora Tengku Mahadi, Helia Vaezian, Mahmoud Akbari, Nor Aini Ali and C. S. Cheng. 2009. Building a legal TM and glossary from an English-Malay parallel corpus. In *The sustainability of the translation field: The 12th International Conference on Translation*, eds. Hasuria Che Omar, Haslina Haroon and Aniswal Abd. Ghani, 362–377. Kuala Lumpur: Malaysian Translators Association.

This paper reports on a project which began in January 2009 to develop a Translation Memory of legal texts and a glossary of legal terminology based on an English-Malay Parallel Corpus. At the time of publication, the corpus had reached 210,000 words.

Vethamani, M. E., Umi Kalthom Abd Manaf and Omid Akbari. 2010. Students' use of modals in their written work: Compensation strategies and simplification features. *Studies in Languages and Language* Teaching 14(2): 13–26.

The focus of this study is on ESL learners' use of modals in 2 written tasks (EMAS corpus) and their strategies/simplification features to compensate this. Descriptive statistics were derived from a concordance and findings suggest that Malaysian learners use simplification more than compensation strategies to overcome limitations.

______. 2008a. Students' use of modals in narrative compositions: Forms and functions. *English Language Teaching* 1(1): 61–74.

This is an investigation of the distribution and functions of modals used in Malaysian ESL learners writing. The study examined learners' use of modals in their written tasks.

______. 2008b. ESL learners use of English modals in narrative compositions: Syntactic and semantic accuracy. *TEFLIN Journal* 19(2): 141–159.

This is a study on the use of modals in the EMAS corpus (two written tasks by Form 4 Malaysian secondary school ESL learners).

Yee, C. L. and Bahiyah Dato Hj Abd Hamid. 2008. "I'm a simple, easy going Chinese girl…": Gender and ethnic identity constructions of Malaysian adolescents in the personal advertisements. http://dspace.uniten.edu.my/xmlui/bitstream/handle/123456789/616/chinese.pdf?sequence=1

This is an investigation of Malaysians' construction of their gender and ethnic identities based on an analysis of a corpus of personal ads in *Galaxie* magazine.

Yuen, C. K., Mohd Subakir Mohd Yasin, Kesumawati Abu Bakar, Azhar Jaludin and Bahiyah Abdul Hamid. 2003. Unraveling linguistic sexism & sex role stereotyping in Malaysian English Language textbooks: The Wordsmith Tools way. *Jurnal Pengajian Umum* 8: 101–111.

This study is an investigation of linguistic sexism and sex role stereotyping in Malaysian KBSR (primary) and KBSM (secondary) English language school textbooks. The study utilised a combination of quantitative method (Wordsmith Tools) and qualitative analysis (CDA) in its approach. The researchers conclude that the adoption of the quantitative and qualitative framework proved to be a reliable method in analysing large volumes of language data.

Yunisrina Qismullah Yusuf. 2009. A corpus-based linguistics analysis on written corpus: Colligation of "TO" and "FOR". *Journal of Language and Linguistic Studies* 5(2): 104–122.

The focus of this study is on the colligations of the word "TO" and "FOR" in a written corpus. The findings of the study show that there are errors in the use of "TO" and "FOR".

Yunisrina Qismullah Yusuf. 2010. A corpus-based linguistic analysis on spoken corpus: Semantic prosodies on "Robots". *Journal of Language and Linguistics Studies* 6(1): 49–64.

The focus of this study is on the semantic prosody of the word "robot" based on an analysis of words that it collocates with in spoken data. The study shows that words that collocated the most with "robot" are service (8), machines (20), surgical system (15), intelligence (13).

Zaharin Yusuff. 1995. Towards a language information centre for Malay. Paper presented at MT Summit V Proceedings, Luxembourg. 10− 13 July.

This is a joint effort between Computer-Aided Translation Unit at Universiti Sains Malaysia (USM) with Dewan Bahasa dan Pustaka (DBP) to set up a kind of language information centre for Malay. Placed on the Internet and accessible to all nationally and internationally, the system includes: (1) a corpus system, (2) dictionary systems and (3) a general lexical database with as much information as possible about words in Malay. It also has various language processing tools plus translation systems. The corpus system also includes: (1) corpus study for grammar writing in machine translation research, (2) concordance for lexicographical work, (3) archive of examples of Malay literature of the 20th century.

Zarifi, A. and J. Mukundan. 2012. Phrasal verbs in Malaysian ESL textbooks. *English Language Teaching* 5(5): 9–18.

This study compares Malaysian ESL textbooks and empirical corpus findings with regard to the inclusion of phrasal verb combinations. The study compared the frequency of phrasal verb combinations in the Malaysian textbook corpus against the BNC which was used as reference corpus.

Zuraidah Mohd Don. 2010. Processing natural Malay texts: A data-driven approach. *Trames: Journal of the Humanities and Social Sciences* 14(1): 90–103.

The study focuses on MALEX, an integrated relational lexical database (MALay LEXicon – an annotated lexicon designed as a relational database).

Zuraidah Mohd Don, G. Knowles and C. K. Fatt. 2010. Nationhood and Malaysian identity: A corpus-based approach. *Text & Talk* 30(3): 267–287.

Based on a corpus of several million words of Tun Mahathir's published speeches in Malay and English, the researchers carried out a lexical analysis using standard corpus linguistic techniques, i.e., key words and collocations to investigate how nationhood and identity is represented in the speeches.

REFERENCES

Adjemian, C. 1976. On the nature of interlanguage systems. *Language Learning* 26(2): 297–320.

Ahmad Zaki Abu Bakar. 1993. Utilization of machine translation in Malaysia machine-aided book translation. Paper presented at the MT Summit IV, Kobe, Japan. 20–22 July.

Ang, L. H., Hajar Abdul Rahim, K. H. Tan and Khazriyati Salehuddin. 2011. Collocations in Malaysian English learners' writing: A corpus-based error analysis. *3 L: The Southeast Asian Journal of English Language Studies* 17 (special issue): 31–44.

Arshad Abdul Samad, Fauziah Hassan, J. Mukundan, Ghazali Kamarudin, Sharifah Zainab Syd Abd. Rahman, Juridah Md. Rashid and Malachi Edwin Vethamani. 2002. *The English of Malaysian school students (EMAS) corpus.* Serdang, Selangor: Universiti Putra Malaysia.

Arshad Abdul Samad. 2004. Beyond concordance lines: Using concordances to investigating language development. *Internet Journal of e-Language Learning & Teaching* 1(1): 43–51.

Azlina Murad Sani. 2004. Gratitude as a moral value: A corpus-based analysis of its representation in the lexis of Malay primary school textbooks. *Malaysian Journal of Learning & Instruction* 1(1): 73–83.

Banafsheh, B. 2005. A corpus-based comparative study of the use of native lexical items in Malaysian and Singaporean English. MA diss., School of Humanities, Universiti Sains Malaysia.

Biber, D., S. Conrad and R. Reppen. 1998. *Corpus linguistics*. Cambridge, UK: Cambridge University Press.

Bolton, K., G. Nelson and J. Hung. 2003. A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong ICE-HK. *International Journal of Corpus Linguistics* 7(2): 165–182.

Botley, S. 2010. A corpus-based comparison of idiom use by Malaysian, British and American students. In *Proceeding of International Conference on Science and Social Research*, 139–144. Kuala Lumpur: Institute of Electrical and Electronics Engineers (IEEE).

Botley, S. and D. Dillah. 2007. Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research* 3: 74–93.

Botley, S., Haykal Hammaad Zin and Monaliza Sarbini. 2005. Lexical and grammatical transfer by Malaysian student writers. Paper presented at the Proceedings of the 10th International Conference on Translation, Universiti Malaysia Sabah, Kota Kinabalu. 2–4 August.

Braun, V. and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3: 77–101.

Chuah, C-K. and Zaharin Yusoff. 2002. Computational linguistics at Universiti Sains Malaysia. Paper presented at the International Conference on Language Resources and Evaluation, The University of Las Palmes de Gran Canaria, Canary Islands – Spain. 29− 31 May.

Chung, S-F. 2011. Uses of ter- in Malay: A corpus-based study. *Journal of Pragmatics* 43(3): 799–813.

______. 2010. Numeral classifier buah in Malay: A corpus-based study. *Language and Linguistics* 11(3): 553–577.

Granger, S. 1998. *Learner English on computer*. London; New York: Longman.

Granger, S., J. Hung and S. Petch-Tyson. 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Gray, J. 2002. The global course book in English language teaching. In *Globalization and language teaching*, eds. D. Block and D. Cameron, 151–67. London and New York: Routledge.

Hajar Abdul Rahim. 2012. Corpora in ESL/EFL teaching. In *English in multicultural Malaysia: Pedagogy and applied research*, ed. Zuraidah Mohd Don. Kuala Lumpur: University of Malaya Press.

______. 2005. Impak konotasi budaya terhadap leksis: Satu kajian semantik berasaskan korpus ke atas perkataan "Perempuan" dan "Wanita". *Jurnal Bahasa* 5(1): 83–111.

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

Hoey, M., M. Mahlberg, M. Stubbs and T. Wolfgang, eds. 2007. *Text, discourse and corpora: Theory and analysis*. London: Continuum.

Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2): 249–268.

______. 2002. Pattern grammar, language teaching and linguistic variation: Applications of a corpus-driven grammar. In *Using corpora to explore* *linguistic variation*, eds. R. Reppen, S. M. Fitzmaurice and D. Biber, 167–186. Amsterdam: John Benjamins.

Imran Ho-Abdullah. 1996. By ESL writers vs. by native writers: A corpus analysis of native and non-native speakers' written English. *Deep South* 2(3). http://www.otago.ac.nz/deepsouth/vol2no3/imran.html/

Imran Ho-Abdullah, Zaharani Ahmad, Rusdi Abdul Ghani, Nor Hashimah and Idris Aman. 2004. A practical grammar of Malay – A corpus-based approach to the description of Malay. Paper presented at the First COLLA Regional Workshop, Putrajaya, Malaysia. 28–29 June.

James, C. 1998. *Errors in language learning and use: Exploring error analysis*. Haslow, Essex: Addison-Wesley Longman.

Kamariah Yunus and Su'ad Awab. 2011. Collocational competence among Malaysian undergraduate law students. *Malaysian Journal of ELT Research* 7(1): 151–202.

Kaur, M. and Sarimah Shamsudin. 2011. Extracting noun forms: A lesson learnt. *International Journal of Language Studies* 5(4): 19–32.

Khojasteh, L. and R. Kafipour. 2012a. Are modal auxiliaries in Malaysian English language textbooks in line with their usage in real language? *English Language Teaching* 5(2): 68–77.

______. 2012b. Have the modal verb phrase structures been well presented in Malaysian English language textbooks? *English Language and Literature Studies* 2(1): 35–41.

______. 2012c. Non-empirically based teaching materials can be positively misleading: A case of modal auxiliary verbs in Malaysian English language textbooks. *English Language Teaching* 5(3): 62–72.

Knowles, G. and Zuraidah Mohd Don. 2003. Tagging a corpus of Malay texts, and coping with "syntactic drift". In *Proceedings of the corpus linguistics 2003 conference*, 422–428. UK: University of Lancaster, Centre for Computer Corpus Research on Language.

Lee, L. W. and H. M. Low. 2011. Developing an online Malay language word corpus for primary schools. *International Journal of Education and Development Using ICT* 7(3): 96–101.

Louw, B. 1993. Irony in the text or insincerity in the writer. In *Text and technology: In honour of John Sinclair*, eds. M. Baker, G. Francis and E. Tognini-Bonelli, 157–176. Amsterdam: John Benjamins.

McEnery, T. and A. Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mohd Faeiz Ikram Mohd Jasmani, Mohamad Subakir Mohd Yasin, Bahiyah Abdul Hamid, Yuen Chee Keong, Zarina Othman and Azhar Jaludin. 2011. Verbs and gender: The hidden agenda of a multicultural society. *3L: The Southeast Asian Journal of English Language Studies* 17(special issue): 61–73.

Mukundan, J. and Anealka Aziz. 2009. Loading and distribution of the 2000 high frequency words in Malaysian English language textbooks for Form 1 to Form 5. *Pertanika Journal of Social Sciences and Humanities* 17(2): 141–152.

Mukundan, J. and L. Khojasteh. 2011. Modal auxiliary verbs in prescribed Malaysian English textbooks. *English Language Teaching* 4(1): 79–89.

Mukundan, J. and M. Sujatha. 2006. Lexical similarities and differences in the mathematics, science and English language textbooks. *K@ta: A biannual publication on the study of language and literature* 9(2): 91–111.

Mukundan, J., A. C. H. Leong and V. Nimehchisalem. 2012. Distribution of articles in Malaysian secondary school English language textbooks. *English Language and Literature Studies* 2(2): 62–70. http://www.ccsenet.org/journal/index.php/ells/article/view/17558/

Nair, R. and Rosli Talif. 2010. Lexical choices and the construction of gender in Malaysian children's literature. *Kajian Malaysia* 28(2): 137–159.

Norsimah Mat Awal, Intan Safinaz Zainuddin and Imran Ho-Abdullah. 2011. Use of comparable corpus in teaching translation. *Procedia – Social and Behavioral Sciences* 18: 638–642. http://linkinghub.elsevier.com/retrieve/pii/S1877042811012079/

Norwati Roslim and J. Mukundan. 2011. An overview of corpus linguistics studies on prepositions. *English Language Teaching* 4(2): 125–131.

Ooi, V. B. Y. 2001. Upholding standards or passively observing language?: Corpus evidence and the concentric circles model. In *Evolving identities: The English language in Singapore and Malaysia*, ed. V. B. Y. Ooi. Singapore: Times Academic Press.

______. 2000. Asian or Western realities? Collocations in Singaporean-Malaysian English. *Language and Computers* 30: 73–92.

______. 1997. Analysing the Singapore ICE corpus for lexicographic evidence. *Language and Computers* 20: 245–260.

Partington, A. 2004. Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1): 131–156.

______. 1993. Corpus evidence of language change: The case of the intensifier. In *Text and technology: In honour of John Sinclair*, eds. M. Baker, G. Francis and E. Tognini-Bonelli. Philadelphia: John Benjamins.

Ridwan Wahid. 2011. The use of corpus-based techniques in literary analysis: Exploring learners' perceptions. *Asiatic* 5(1): 104–128.

Sarimah Shamsudin, Zaidah Zainal, Yasmin Hanafi Zaid and Salbiah Seliman. 2008. Corpus analysis of primary one science textbooks for designing ELT materials. In *Research in language teaching*, eds. Noor Abidah Mohd Omar and Zaidah Zainal, 135–155. Johor: Penerbit Universiti Teknologi Malaysia.

Sharifah Zakiah Wan Hassan, Simon Faizal Hakim, Mahdalela Rahim, J. F. Noyem, Sueb Ibrahim, Johnny Ahmad and Kamaruzaman Jusoff. 2009. The communicative ability of Universiti Teknologi MARA Sarawak's graduates. *English Language Teaching* 2(2): 84–91.

Sinclair, J. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.

Stubbs, M. 2007. On texts, corpora and models of language. In *Text, discourse and corpora: Theory and analysis*, eds. M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, 187–214. London: Continuum.

______. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

______. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23–55.

Suhaimi Ab Rahman and Normaziah Abdul Aziz. 2004. Improving word alignment in an English – Malay parallel corpus for machine translation. Paper presented at the Language Resources and Evaluation (LREC) 2004 Workshop on the amazing utility of parallel and comparable corpora, Lisbon, Portugal. 25 May.

Tan, S. I. 2009. Lexical borrowing from Chinese languages in Malaysian English. *World Englishes* 28(4): 451–484.

Tengku Sepora Tengku Mahadi, Helia Vaezian, Mahmoud Akbari, Nor Aini Ali and Chew Saw Cheng. 2009. Building a legal TM and glossary from an English-Malay parallel corpus. In *The sustainability of the translation field: The 12th International Conference on Translation*, eds. Hasuria Che Omar, Haslina Haroon and Aniswal Abd. Ghani, 362–369. Kuala Lumpur: Malaysian Translators Association.

Teubert, W., M. Hoey, M. Mahlberg and M. Stubbs. 2007. Parole-linguistics and the diachronic dimension of the discourse. In *Text, discourse and corpora: Theory and analysis*, eds. M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, 57–88. London: Continuum.

Vethamani, M. E., Umi Kalthom Abd Manaf and Omid Akbari. 2010. Students' use of modals in their written work: Compensation strategies and simplification features. *Studies in Languages and Language* Teaching 14(2): 13–26.

______. 2008a. Students' use of modals in narrative compositions: Forms and functions. *English Language Teaching* 1(1): 61–74.

______. 2008b. ESL learners use of English modals in narrative compositions: Syntactic and semantic accuracy. *TEFLIN Journal* 19(2): 141–159.

Yee, C. L. and Bahiyah Dato Hj Abd Hamid. 2008. "I'm a simple, easy going Chinese girl…": Gender and ethnic identity constructions of Malaysian adolescents in the personal advertisements. http://dspace.uniten.edu.my/xmlui/bitstream/handle/123456789/616/chinese.pdf?sequence=1

Yuen, C. K., Mohd Subakir Mohd Yasin, Kesumawati Abu Bakar, Azhar Jaludin and Bahiyah Abdul Hamid. 2003. Unraveling linguistic sexism & sex role stereotyping in Malaysian English Language textbooks: The Wordsmith Tools way. *Jurnal Pengajian Umum* 8: 101–111.

Yunisrina Qismullah Yusuf. 2010. A corpus-based linguistic analysis on spoken corpus: Semantic prosodies on "Robots". *Journal of Language and Linguistics Studies* 6(1): 49–64.

______. 2009. A corpus-based linguistics analysis on written corpus: Colligation of "TO" and "FOR". *Journal of Language and Linguistic Studies* 5(2): 104–122.

Zaharin Yusuff. 1995. Towards a language information centre for Malay. Paper presented at MT Summit V Proceedings, Luxembourg. 10− 13 July.

Zarifi, A. and J. Mukundan. 2012. Phrasal verbs in Malaysian ESL textbooks. *English Language Teaching* 5(5): 9–18.

Zuraidah Mohd Don, G. Knowles and C. K. Fatt. 2010. Nationhood and Malaysian identity: A corpus-based approach. *Text & Talk* 30(3): 267–287.

Zuraidah Mohd Don. 2010. Processing natural Malay texts: A data-driven approach. *Trames: Journal of the Humanities and Social Sciences* 14(1): 90–103.