Kajian Malaysia, Vol. 32, Supp. 1, 2014, 1–16

CORPORA IN LANGUAGE RESEARCH IN MALAYSIA

Hajar Abdul Rahim

School of Humanities, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia
Email: hajar@usm.my

Since the creation of the first generation of computer-readable corpora, pioneered by the Brown Corpus in the 1970s, corpora have been extensively used in virtually all branches of linguistics. This article discusses corpus-based research and its development in the Malaysian linguistic scenario over the last three decades in considering its status and prospect as a research method. The discussion shows that research in Malay and English in the country has benefited from the use of corpora. In the last decade especially, there has been a surge of interest in corpora as a research tool in Malay and English, evidenced by the rise in research and publications on Malay linguistic description, English learner language, and Malaysian English as a local variety. Despite the encouraging development, there is still much to be gained from the use of corpora and corpus analysis in the linguistic description of Malay dialects, variation and pedagogy, English language pedagogy, Malaysian English and other language varieties including endangered and indigenous languages in the country.

Keywords: language corpora, Malay language, Malaysian English, corpus linguistics, corpora and pedagogy

INTRODUCTION

Since the early days of pre-electronic corpora dating back to the 1700s, corpus research has evolved into a sophisticated and rigorous research methodology, impacting the breadth and depth of language and linguistics research. According to Wilson, Archer and Rayson (2006), not too long ago, the focus of most corpus linguistic research was on the grammar and vocabulary of standard English, particularly British and American English. However, the computer and the arrival of digitised corpora, particularly in the 1980s and 1990s (McCarthy and O'Keeffe, 2010), have enabled access to a tremendous amount of texts and brought about a "theoretical shift…from the focus on a notion of central grammar, core lexicon and general rules to a more decentralised notion of contextual appropriacy, geographical and social variety, restricted language, idiolect and style" (Tognini-Bonelli, 2001: 6). Thus by the turn of the century,

areas of language research such as linguistic description, language variation, lexicography, computational linguistics, language education, translation, forensic linguistics, are among many that had benefited from the corpus methodology (Tognini-Bonelli, 2001; Kachru, 2008). In the last decade, it has become apparent that the scope of corpus-based language research has widened even more considerably as "the range of languages, research questions, and indeed, methodologies which are addressed by corpus linguists" diversify and research questions extend beyond the "traditional" concerns to include learner language and other world languages (Wilson, Archer and Rayson, 2006). And in recent years, language corpora have begun to be employed in other fields as new research that is emerging around the world extends beyond language and linguistic inquiry into exciting "areas such as cross-cultural rhetoric and social psychology…and even economic forecast" (Wilson, Archer and Rayson, 2006: preface). These developments clearly indicate that corpus linguistics "has much to offer other areas by providing a better means of doing things" (McCarthy and O'Keeffe, 2010: 7).

Despite the widening application of corpora in various languages and fields, the fact remains that "the majority of corpora are of the English language" (Lee, 2010: 108). To a large extent, this has been due to the significant interest in corpora and language pedagogy in the 1990s, during which time also saw a marked increase in conferences and publications on corpora and pedagogy (McEnery and Xiao, 2011)[1]. Leech (1997) in his oft-cited discussion on corpora in language learning suggests three ways in which corpora and pedagogy converge, (1) the indirect use of corpora in teaching, (2) the direct use of corpora in teaching, and (3) the development of teaching-oriented corpora. Where Malaysia is concerned, the literature suggests that the convergence between corpora and pedagogy, particularly with regard to English language, is most noticeable in terms of the development of learner and pedagogic corpora. Besides pedagogy, corpus-based studies on English in Malaysia have also centred on issues surrounding Malaysian English (ME) as a local English variety. In the last decade especially, the theoretical developments in World Englishes (WEs) as well as the increase in corpus-based studies on a number of WEs have prompted the creation of ME corpora for research in the description and use of English as a new local variety.

While corpora development and research in English have been dominant, "corpora for a huge variety of languages have sprung up all over the world" (Lee, 2010: 107), including the Malay language. In Malaysia, the earliest corpus development project was not on English but was in fact on the Malay language. The literature shows that corpus research in English in Malaysia began in the 1990s but the first attempt to create a Malay language corpus dates back to the early 1980s. This was a project to develop a Malay corpus by Dewan Bahasa dan Pustaka (henceforth DBP),[2] a government institution responsible for monitoring and planning Malay language and literature in the country. As the

first Malay corpus development endeavour in the country, the project marked an important milestone in the Malaysian linguistic research scenery and the advent of corpora in the country as a tool for the linguist.

Despite the thirty-year existence of corpora as a research methodology in studies on Malay and English in the country, there has been hardly any serious discussion on the development and progress of corpus research in Malaysia. The dearth of literature on this issue was the motivation for the *Malaysia Corpus Research Colloquium*[3] held on 5 April 2012 at the School of Humanities, Universiti Sains Malaysia. With the theme "Corpus Research in Malaysia: Progress and Prospects," the colloquium participants, comprising corpus researchers from various local universities, presented and discussed corpora development, corpus-based research issues, opportunities and limitations as well as new directions in corpus research in Malaysia. The colloquium discussions suggest that corpora and the corpus-based approach have gained the interest of an increasing number of researchers who work in various fields of Malay and English language research, which means that the scope of corpus-based research in the country has expanded. This observation is the point of departure for the current discussion which aims to take stock of corpus research in Malaysia in the last three decades. The discussion traces the developments in corpus creation and studies particularly on Malay and English, highlights the challenges and progress in the field, and considers the future directions of corpora development and corpus-based research in Malaysia.

CORPORA DEVELOPMENT AND RESEARCH IN MALAYSIA: PROGRESS

Corpora and Malay Language Research

In Malaysia, Malay is the main lingua franca covering the urban and rural areas and is also the official language of the country (Asmah, 1996). Therefore it is not surprising that the first Malay corpus development project in Malaysia was spearheaded by the very agency that is responsible for the use and development of the language in the country. Inspired at the time by the Brown Corpus, the project began in 1983 and involved the compilation of texts for language analysis to develop a database of two million Malay words (Rusli, Norhafizah and Chin, 2006). Following the Brown Corpus sampling technique, the project set out to collect and compile 2,000-word samples of Malay texts from various genres. However, as reported by Rusli, Norhafizah and Chin (2006), before the corpus reached half a million, a shift in DBPs Malay language research focus required a change in the sampling technique. In place of samples of texts, complete texts were collected and compiled.

The change in the sampling technique was motivated at the time by the need to provide Malay language researchers with resources for lexicographic, grammatical and linguistic analyses of the Malay language that are objective and authentic in nature, based on real contexts and discourse, and that can generate information on the typical behaviour of Malay expressions (words and phrases) (Rusli, Norhafizah and Chin, 2006). To this end, the corpus designers took into consideration the inclusion of complete old Malay texts as well as modern texts in developing the corpus. The former essentially include texts on Malay legends and scriptures while the latter are represented by texts that were sourced from books, newspapers and magazines. The corpus currently does not contain spoken data, although there are plans to include them in order to create a representative corpus of the Malay language.

Since its inception, the corpus, known as Pangkalan Korpus DBP (DBP Corpus Database) has been built in stages, and to date is the largest corpus on the Malay language in the country. According to Rusli, Norhafizah and Chin (2006), in its early stages, the DBP corpus was designed as an archive of texts with a system for processing selected texts to generate concordances as well as statistical information on word frequencies and total number.[4] The reason for such a design was the need to compile digitised texts in a short amount of time for lexicographical work undertaken by DBP. Thus, as underlined by Rusli, Norhafizah and Chin (2006), the data collection approach taken in the beginning stages of the corpus development was opportunistic in nature in that texts published by DBP that were available in digital format were included along with digital texts by other publishers. A corpus system works in tandem with the corpus for analysing texts, and information that is derived from the analysis can be on word forms, derivatives and phrases that are presented in concordance sets. The current DBP corpus database is web-based to allow the public access to the database for text analysis.[5]

The DBP corpus database currently comprises 128 million words[6] which are compiled in 10 subcorpora representing different genres of texts as follows: books (novels, academic books, general reading, textbooks), magazines (general and covers various fields), newspapers (daily, tabloids, Sunday editions), translations (academic books and general readings), ephemerals (flyers, brochures, advertisements), drama (bound form), poems (bound form), material cards, and traditional texts (primary and secondary school textbooks). It is necessary to note at this juncture that the 128 million words available in the corpus are based on texts collected and compiled until 2008 only. However, this does not mean that the development of the corpus has stopped. As explained by DBP representatives in a recent interview, since 2009, data have continued to be collected but have yet to be uploaded. Once the collected data have been cleaned, the corpus will be updated with 25 million more words of Malay texts by 2015.

As the first major publically-accessible Malay language corpus, the DBP corpus has been employed in a range of studies on Malay language and has

informed lexicography and translation projects (particularly Malay-English/Malay-English).[7] Between 1995 and 2014, the corpus has been used to inform work on Malay dictionaries published by DBP, including the much referred to *Kamus Dewan* which has reached its fourth edition. As stated in the foreword of the most recent edition of *Kamus Dewan* (2005), computerised and manual corpora were consulted for explanations of new words and updates on meanings. The computerised corpus comprised 85 million words that were taken from the DBP corpus database. Informed by corpus data, and published 10 years after the previous edition, the *Kamus Dewan* 4th edition has almost 6,000 more entries from various fields including science and technology and linguistics.

Besides facilitating the analysis of Malay language for the above-mentioned studies and publications, the DBP corpus database is also a resource for researchers who wish to develop their own corpus for research in Malay. The Practical Grammar of Malay Project (Imran et al., 2004) for instance uses texts taken from the DBP corpus to develop a Malay Practical Grammar Corpus (MPGC) for corpus-based investigations of Malay linguistic and grammatical features. The corpus, known as the UKM-DBP corpus, contains 5 million words taken from the text archive in the DBP corpus database. Another project that also used texts from the DBP corpus database to create a corpus for Malay language analysis is the MALEX (MALay LEXicon) project (Knowles and Zuraidah, 2006). The project essentially created a small corpus of novels extracted from the DBP corpus database in analysing grammatical class in Malay. The success of the projects in using texts from the DBP database to create a corpus for specific analysis, validates DBP corpus designers' aim of creating an archive of texts instead of a corpus per se. As argued by Rusli, Norhafizah and Chin (2006), the design of the DBP corpus allows researchers access to the texts to create their own corpus for Malay linguistic research. Indeed, researchers from within and outside DBP have benefitted from this facility and the web-based system of the database, as evidenced by various publications that have been generated.

While studies based on the DBP corpus database have generally focused on the linguistic aspects and features of Malay for lexicographical, descriptive and translation studies, there are other studies that go beyond these concerns. One such project is the development of an online Malay lexical database from a corpus of texts from Malay textbooks used in Malaysian primary schools (Lee and Low, 2011). As reported by the researchers of the project, the database is a repository of Malay words that are normally encountered by school-going children in Malaysia and is designed to facilitate evidence-based teaching and research practices pertaining to Malay language literacy. Besides studies that are based on general and pedagogic corpora, there are corpus-based studies that rely on online and internet-based Malay texts such as online newspapers, articles and novels. Indeed, the availability of digitised or online Malay texts of various genres has made it possible for researchers to create their own corpus for Malay linguistic analysis (e.g., Chung, 2010; 2011) and provides the opportunity for

researchers to design their own corpora in researching more challenging and topical issues such as Malay as a second language, and Malay language change and variation.

Corpora and English Language Research

Research in English language has benefited much from the corpus approach. In Malaysia, in the last two decades especially, there has been a steady increase in the development of Malaysian English corpora. The availability of corpora, not just in terms of number but also types, has encouraged research with regard to the use of English as a learner language, as a second language as well as a language variety.

Where English language and pedagogy is concerned, research based on learner corpora and pedagogic corpora has increased significantly over the last ten years. This phenomenon owes much to the proliferation of English language learner corpora such as the EMAS (English of Malaysian School Students) corpus (Arshad et al., 2002), MACLE (Malaysian Corpus of Learner English) (Knowles and Zuraidah, 2005), CALES (Corpus Archive of Learner English Sabah-Sarawak) (Botley et al., 2005), as well as genre-specific learner corpora (e.g., the Engineering Lecture Corpus (ELC), the Business and Management English Language Learner Corpus (BMELC). These corpora, some of which are publically-accessible for research, have been employed in a number of studies (e.g., Arshad (2004); Botley and Dillah (2007); Vethamani, Umi Kalthom and Omid (2010); Kamariah and Su'ad (2011) among many others) which are steadily adding to the repository of corpus-based studies on Malaysian English learner language. Besides learner corpora, local educators and researchers' concern over language learning and teaching issues also saw the development of English pedagogic corpora. The literature shows that there was a sudden increase in textbook research (e.g., Mukundan and Aziz (2009); Mukundan and Khojasteh (2011)), particularly English school textbooks, between 2006 and 2012 (see the article *Corpus Research in Malaysia: A Bibliographic Analysis* in this volume for more studies).

Beside English pedagogy, corpus-based studies have also fascinated researchers in the field of World Englishes (WEs). In discussing language variation within the contexts of WEs, Kachru argues that corpus-based research produces "revealing results in the areas of variation in the use of grammatical and lexical devices…in coming to grips with dialect and variety differentiation and will deepen our understanding of register and genre variation" (Kachru, 2008: 5−6). These theoretical and methodological contributions of the corpus-based approach have motivated the development of corpora of varieties of English. In Malaysia, the last decade especially saw the development of Malaysian English (ME) corpora, (e.g., The International Corpus of English (ICE) Malaysia (Hajar and Su'ad, 2012); The Malaysian English Newspaper

(MEN) Corpus (Tan, 2013); Corpus of Spoken Malaysian English (COSME)) and an increase in corpus-based studies as well as published literature on ME as a local variety. ICE Malaysia, while still being developed, has been used by researchers from within and outside Malaysia to carry out various linguistic analyses. The MEN corpus and COSME, despite being restricted in their use, have also been employed in researching ME features. Studies based on these and other corpora have generated publications on ME structure, lexis and phonology (e.g., Collins, 2014; Hajar, 2008; 2014; Pillai et al., 2010; Pillai, Zuraidah and Knowles, 2012; Tan, 2009; 2013; Newbrook, 2006) which contribute to the body of knowledge on the development of ME as a local variety.

CORPORA DEVELOPMENT AND RESEARCH IN MALAYSIA: PROSPECTS

The attention corpus research has gained from Malay and English language researchers, is an indication of its importance as a language research methodology in Malaysia. In the last 30 years, the number of corpus-based studies has progressively increased in tandem with the development of language corpora (Malay and English) and also the accessibility of attested language materials in digitised form via online and internet sources. Where Malay is concerned, the corpus-based approach has been successful in enhancing research in language description, lexicography and translation. Also importantly, the availability and accessibility of the corpus database housed in DBP have begun to affect change in the analysis of Malay linguistics in Malaysia. Corpora have essentially initiated a shift in the dominant intuitive-based analysis of Malay that formed the basis of important publications on Malay linguistics such as Asmah's (1993) *Nahu Melayu Mutakhir* and the commonly accepted Malay reference grammar known as *Tatabahasa Dewan* (Nik Safiah, Farid and Hashim, 2008). As Zaharani and Nor Hashimah (2012: 18) explain, "the analysis of this grammar, particularly the syntactic aspects are based on transformational generative grammar propounded by Chomsky". In recent years, corpus-based studies on Malay linguistic features (e.g., Noraini, Maslida and Karim, 2013; Zaharani, Shakira and Nor Hashimah, 2012; Zaharani and Nor Hashimah, 2012; Lam, 2011; Chung, 2010; 2011; Zuraidah, 2010; Knowles and Zuraidah, 2006; Hajar, 2005) have brought to light findings that are interesting and in some cases, challenge conventionally accepted descriptions of Malay.

Notwithstanding the encouraging progress, there are still many Malay language issues that are under-researched which can benefit from corpora as an empirical linguistic research tool. Knowles and Zuraidah (2006: 3) argue that despite being a major language in the region, the Malay language is "one of the least known to contemporary linguists in the western world" and perhaps "the least regulated." In this connection, DBP as the epicentre of Malay corpus

research in Malaysia has much to offer. With its expanding Malay corpus and its role as the Malay language resource centre, DBP in collaboration with corpus linguists from within and outside the country can step up corpus-based research, not only for a systematic analysis and description of Malay linguistics but also Malay language pedagogy. The DBP corpus database can lead to new descriptions of Malay language, which can in turn influence content, material development and syllabus design for Malay language teaching and learning in the country. An immediate outcome of the DBP corpus database that has important pedagogical implications is the publication of a dictionary that is completely based on the corpus. One such dictionary titled *Kamus Besar Bahasa Melayu Dewan*, is reportedly[8] in the making. This potentially major corpus-based Malay dictionary is projected to house approximately 100,000 entries including new Malay derivatives and terms.

The materialisation of a major Malay dictionary is no doubt an important milestone in corpus-based Malay lexicography in Malaysia. However, in terms of pedagogy, dictionaries are but one of the many ways in which corpora can be of relevance. Johansson's (2009) diagrammatic depiction of the uses of corpora in language teaching and learning in Figure 1 puts into perspective how corpora can fully contribute to language pedagogy.

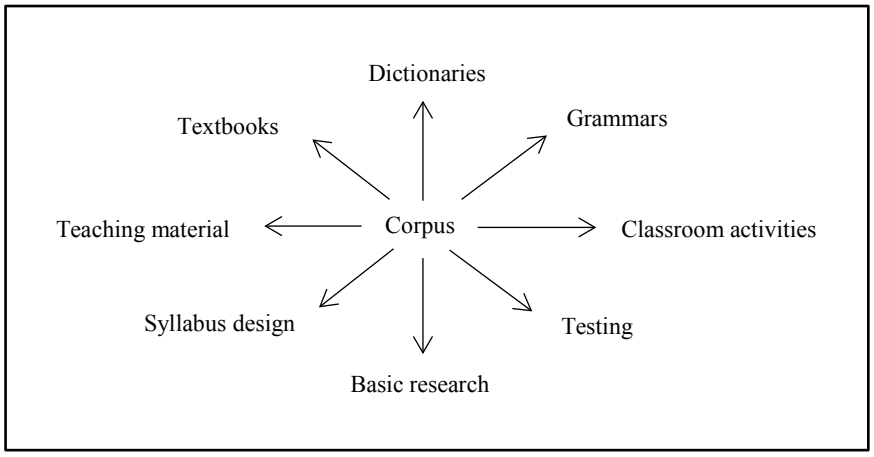

Figure 1: Uses of corpora in language teaching and learning
Source: Johansson (2009: 40)

Going by Johansson's suggestion, there is still much to be gained from the use of corpora in Malay language pedagogy in the country. For instance, Malay language pedagogy must also be informed by learner and pedagogic corpora. To date, there are no known major Malay learner corpora. However, pedagogic corpora such as the one created for the online Malay lexical database project (Lee and Low, 2011) are available. Corpus-informed teaching and

learning activities in Malay, at least at the primary level is already possible. However, the extent to which the database is effective or indeed used by teachers has yet to be established (see also the article *The Development and Application of an Online Malay Language Corpus-based Lexical Database* in this volume).

Where corpora and English language pedagogy are concerned, as reported earlier, research has been rather dynamic. Indeed English language corpora development and corpus-based studies outnumber those on the Malay language in the country. The continuous and increasing concerns among researchers and educators, over the learning and teaching of English may be the reason for this. Learner corpora have been employed to investigate a spectrum of issues concerning learners of English in Malaysia, ranging from L2 learners' spelling errors (e.g., Botley and Dillah, 2007) to the development of their collocational competence (e.g., Kamariah and Su'ad, 2011; Ang et al., 2012). Pedagogic corpora although not as many, have generated a number of studies and publications on English textbook content and approach. These studies contribute new knowledge to the areas of Malaysian learners' interlanguage and language teaching material development respectively (more discussion on this in the article *Corpus Research in Malaysia: A Bibliographic Analysis* in this volume).

Corpus researchers' keen interest in Malaysian English learner language is also evident in the collection of articles in the current volume. Three of the articles are concerned with English learner language and teaching. The article *Argument Structure in Learner Writing: A Corpus-based Analysis using Argument Mapping* is an analysis of the argument structure in Malaysian learners written English based on the CALES. While most studies on learner writing focus on L2 grammar and vocabulary issues, this study goes beyond to focus on learners' critical thinking and argumentation abilities in English based on their argument strategies. Another article in this collection *The Impact of Data-driven Learning Instruction on Malaysian Law Undergraduates* colligational competence' discusses data driven learning in non-native English language learning context. The article essentially tested the effect of data-driven learning (DDL) against conventional learning based on Malaysian law undergraduates' production of colligations of prepositions. The third article, compares the use of adjectives of evaluation by native and non-native speakers, based on spoken native (UK) and non-native (Malaysian) lecture corpora. The study *"...so it is very important...": Evaluative Adjectives in Engineering Lectures* is an attempt to reveal native and non-native lecturers' awareness of students' needs in processing lecture content. These studies suggest that English learner language research is expanding to include critical issues, beyond proficiency and learner errors, that have theoretical implications and contribute new knowledge to the field.

Beyond research in Malay and English, corpora have much potential in studies on other languages and dialects in the country. Lee (2010), in a survey of

corpora that are available worldwide points out that although English language corpora are the most common, there are now more corpora of other languages, including endangered languages. In Malaysia, endangered and marginalised language research is an important agenda. However, the main focus of most of the studies has been on sociolinguistic issues such as language change and maintenance and also linguistic description of the languages. To the best of the researcher's knowledge, corpus development of these languages has never been attempted. Yet, the research potential of these languages in digitised corpus format is boundless. As well, digitising the language data of these languages is important because, as Lee correctly argues, corpora of endangered languages may "even be the last repositories of knowledge about those language in the future" (Lee, 2010: 108). Indeed, in digitised form, the data on these languages can be analysed from various linguistic and non-linguistic aspects, which may in turn have important implications for the speakers of the language who are essentially indigenous groups who live in Peninsular Malaysia and northern Borneo. The development of endangered language corpora will also serve as a resource for researchers, locally and internationally, to research on matters that are beyond linguistics, into cross-disciplinary concerns such the environment, culture, local knowledge, etc. The availability of such corpora may encourage exciting comparative research in indigenous and marginalised languages from various geographical localities.

CONCLUSION

Emerging research issues in Malay and English will very likely motivate the development of more corpora, not just in terms of numbers but also types. These developments will help elevate the quantity and quality of language research in Malaysia and also perhaps inspire new research issues. Despite these encouraging developments, it is not wrong to suggest that there is still much to be gained from the use of corpora and corpus methods in the Malaysian linguistic scenario. To fully utilise corpora in the Malaysian context, there are a number of issues that need to be addressed. These may be considerations for future corpus research in Malaysia.

Firstly, with regard to Malay and English language corpora in Malaysia, there are more written corpora than there are spoken ones in both languages. The DBP, in its plans to improve the corpus database, as explained earlier will increase the number of words in the corpus by 25 million by 2015 (besides making the system more user friendly, upgrading the analytical capability of the system, and including a dictionary).[9] Yet, the development of a spoken subcorpus does not seem to be in the horizon. With regard to English, spoken corpora are available but not as many as written corpora. The MEN corpus for instance comprises written data only. ICE Malaysia, which is still being

developed, will ultimately have 1 million words comprising 40% written and 60% spoken data. To date only approximately 10% of the spoken data has been collected. The development of major spoken corpora in Malay and English in country is necessary for a balanced representation of both languages.

Secondly, research in the Malay language that DBP and many others are involved in currently centre on the linguistic features of Malay as a first language. Yet, in Malaysia, Malay is also an important second language to many speakers. So, future directions in Malay language research should therefore involve the development of Malay learner corpora. As regards English, the earlier discussion shows that corpus-based research in English in Malaysia has centred on English language pedagogy and Malaysian English as a new variety of English. In relation to corpora and pedagogy, it is clear from the above that translating research findings into practice whether directly or indirectly is a challenge. Although corpora have gained familiarity in research, teachers in general are still not aware about corpora in pedagogy. Thus, despite the increasing number of studies, findings and new knowledge, "it does not seem as if learner corpora (or any corpus for that matter) have a role in the Malaysian ESL classroom" directly or indirectly (Hajar, 2012: 98). There are a variety of reasons for this including those that are universally true such as the problem faced by teachers in integrating corpora in the lesson plan and reconciling "minute details of the phraseology of particular words…with the 'big themes' of language teaching, such as 'tenses' or 'articles'" (Hunston, 2002: 184). Where Malaysia is concerned, limited number of hours per week allotted for language classes as well as "the shortage of hardware, lack of software and programs, and inadequate lab space are major constraints in majority of schools" (Hajar, 2012: 98). Despite these challenges, making corpora part of school students' learning experience is possible and should probably begin with the teachers. In this connection, introducing the corpus approach as a methodology in pedagogy in teacher training programmes in the country may be the first step towards making corpora and pedagogy converge in a meaningful way in the Malaysian classroom.

There is also much to be gained from a new breed of corpus known as multimodal corpora. Lee points out that a "growing number of corpora are now fully multimedia in the sense of having transcripts that are aligned or synchronised with the original audio or video recordings" (Lee, 2010: 114). In the Malaysian context, creating such corpora is a feat that may still be years away. Nonetheless, a multimodal corpus, even a small one, can be useful in various situations. Lee cites Braun (2005) in suggesting that in a language classroom, integrating text and other materials can help learners "authenticate decontextualized corpus materials and thus get the most out of them" (Lee, 2010: 115). A multimodal corpus can also be very useful in documenting endangered languages so that beyond spoken words, the language users' gestures, use of prosody, and other elements of the discourse situation can be captured for

analysis (Lee, 2010). The realisation of this type of corpus will inspire new and exciting inter-disciplinary and cross-cultural studies based on corpora.

In conclusion, since the early days of pre-electronic corpora dating back to the 1700s, corpora have evolved and their impact on language and linguistics study in terms of the breadth and depth of research is unprecedented. In Malaysia, corpus linguistics and the use of corpora for language research have gained familiarity especially in the English language and Malay language research scenery. After 30 years, the impact of corpus-based research is felt, evident from the increasing number of researchers who employ corpora and corpus analysis in their research, the widening scope of corpus-based studies by local researchers and the introduction of corpus linguistic courses in various universities in the country. Nonetheless, there is still much to be gained from the use of corpora in Malaysia, not just in language and linguistic research but also in other fields of enquiry.

ACKNOWLEDGEMENT

The work reported in this article is supported by a short-term research grant from Universiti Sains Malaysia (304/PHUMANITI/6312007). I wish to thank Rusli Abd. Ghani and Saidah Kamin at Dewan Bahasa dan Pustaka (DBP) in Kuala Lumpur for providing information on the DBP corpus status and development.

NOTES

1. See McEnery and Xiao (2011: 367) for a list of references on this.
2. Dewan Bahasa dan Pustaka, commonly known as DBP, is the national planning body for language and literature in Malay in the country.
3. The one-day colloquium was intended to be the first of its kind and was organised to convene local corpus linguists and researchers to engage in discussions on the current status and future prospects of corpus research in Malaysia. In all, 11 papers on corpora development and application, corpus-based translation, corpus-based English learner language studies and Malay linguistics were presented.
4. The archive system (instead of a corpus per se), according to Rusli et al. (2006) gives researchers the opportunity to define what they need for their own analysis without being bound by DBP's criteria for a corpus.
5. The web-based system replaces the original UNIX system that was developed in collaboration with researchers from the computer-aided translation unit of the School of Computer Sciences, Universiti Sains Malaysia. The original system was also equipped with text analysis capabilities known as MATA (Malay Text Analysis) which allow for statistics on a text to be generated.

6. In an interview with Rusli Abd. Ghani and Saidah Kamin at DBP, it became clear that by 2008, the corpus data had reached 135 million words. Since then, the process of cleaning up the corpus has been in progress which explains the 128 million words available for use currently.
7. The DBP corpus was consulted for sample sentences.
8. This information was obtained from an interview carried out with Rusli Abd. Ghani and Saidah Kamin at DBP.
9. Based on the interview with Rusli Abd. Ghani and Saidah Kamin at DBP in January 2014.

REFERENCES

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.

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

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

Asmah Haji Omar. 1996. Post-imperial English in Malaysia. In *Post-imperial English: Status change in former British and American colonies, 1940−1990*, eds. J. A. Fishman, A. W. Conrad and A. Rubal-Lopez, 513−531. Berlin, New York: Mouton de Gruyter.

———. 1993. *Nahu Melayu mutakhir*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Baker, P., A. Hardie and T. McEnery. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

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

Botley, S. P., C. De Alwis, L. Metom and Isma Izza. 2005. CALES: A corpus-based archive of learner English in Sarawak. Final project report, Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Braun, S. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17(1): 47−64.

Chapelle, C. A. 2001. *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

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.

Collins, P. 2014. Modal expressions in Malaysia English. In *English in Malaysia: Postcolonial and Beyond*, eds. Hajar Abdul Rahim and Shakila Abdul Manan, 127-160. Bern: Peter Lang.

Hajar Abdul Rahim. 2014. Malaysian English lexis: Postcolonial and beyond. In *English in Malaysia: Postcolonial and Beyond*, eds. Hajar Abdul Rahim and Shakila Abdul Manan, 35-53. Bern: Peter Lang.

———. 2012. Corpora in ESL/EFL teaching. In *English in multicultural Malaysia: Pedagogy and applied research*, ed. Zuraidah Mohd Don, 85-103. Kuala Lumpur: University of Malaya Press.

———. 2008. The evolution of Malaysian English: Influences from within. In *Linguistics, Literature and Translation*, eds. Lalitha Sinha and Shakila Abdul Manan, 1−19. Newcastle: Cambridge Scholars Publishing.

———. 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 Su'ad Awab. 2012. ICE Malaysia: The first decade. Paper presented at the Malaysia Corpus Research Colloquium, Universiti Sains Malaysia, Penang. 5 April.

Imran Ho-Abdullah, Zaharani Ahmad, Rusdi Abdul Ghani, Nor Hashimah Jalaludin 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.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Johansson, S. 2009. Some thoughts on corpora and second-language acquisition. In *Corpora and language teaching*, ed. K. Aijmer, 33−44. Amsterdam: John Benjamins.

Kachru, Y. 2008. Language variation and corpus linguistics. *World Englishes* 27(1): 1−8.

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

Kamus Dewan. 2005.Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kennedy, G. 1998. *An introduction to corpus linguistics*. Harlow: Addison Wesley Longman.

Knowles, G. and Zuraidah Mohd Don. 2006. *Word class in Malay: A corpus-based approach.* Kuala Lumpur: Dewan Bahasa dan Pustaka.

———. 2005. Introducing MACLE: The Malaysian Corpus of Learner English. In *The first national symposium on corpus linguistics: Selected papers*.

Wang Longyin & He Anping, Guong Zhou: North East Normal University Press.

Lam, K. C. 2011. Penggunaan di mana dan yang mana sebagai Kata Hubung: Analisis linguistik korpus. *Jurnal Bahasa* 11(2): 297− 230.

Lee, D. Y. W. 2010. What corpora are available? In *The Routledge handbook of corpus linguistics*, eds. Anne O'Keeffe and Michael McCarthy, 107−121. Oxford: Routledge.

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

Leech, G. 1997. Teaching and language corpora: A convergence. In *Teaching and language corpora*, eds. A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles, 1−23. London: Longman.

McCarthy, M. and A. O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In *The Routledge handbook of corpus linguistics*, eds. Anne O'Keeffe and Michael McCarthy, 3−13. Oxford: Routledge.

McEnery, T. and R. Xiao. 2011. What corpora can offer in language teaching and learning. In *Handbook of research in second language teaching and learning (Volume 11)*, eds. Eli Hinkel, 364−380. New York: Routledge.

McEnery, T., R. Xiao and Y. Tono. 2006. *Corpus-based language studies*. Oxford: Routledge.

Mukundan, J. and A. 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.

Newbrook, M. 2006. Malaysian English: Status, norms, some grammatical and lexical features. In *World Englishes: Critical concepts in linguistics, Volume II*, eds. K. Bolton and B. B. Kachru, 390−417. London: Routledge.

Nik Safiah Karim, Farid M. Onn and Hashim Haji Musa. 2008. *Tatabahasa dewan edisi ketiga*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Noraini Salleh, Maslida Yusof and Karim Harun. 2013. Klasifikasi *aktionsart* predikat keadaan Bahasa Melayu. *Jurnal Bahasa* 13(2): 192–216.

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

Pillai, S., Zuraidah Mohd Don, G. Knowles. 2012. Towards building a model of standard Malaysian English pronunciation. In *English in multicultural Malaysia: Pedagogy and applied research*, ed. Zuraidah Mohd Don, 167–181. Kuala Lumpur: University of Malaya Press.

Rusli Abdul Ghani, Norhafizah Mohamed Husin and L. Y. Chin. 2006. Pangkalan data korpus DBP: Perancangan, pembinaan dan pemanfaatan. In *Aspek nahu praktis Bahasa Melayu*, ed. Zaharani Ahmad, 21–25. Bangi: Universiti Kebangsaan Malaysia Press.

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

———. 2009. Lexical borrowing in Malaysian English: Influences of Malay. *Lexis* 3: 11–62.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins Publishing Company.

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.

Wilson, A., D. Archer and P. Rayson. 2006. *Corpus linguistics around the world*. Amsterdam: Rodopi.

Zaharani Ahmad and Nor Hashimah Jalaluddin. 2012. Incorporating structural diversity in the Malay Grammar. *GEMA Online™ Journal of Language Studies* 12(1): 17–34.

Zaharani Ahmad, Shakira Khairudin and Nor Hashimah Jalaluddin Ahmad. 2012. Perilaku morfologi awalan *ber-* dalam Bahasa Melayu klasik dan Bahasa Melayu moden: Satu kajian perbandingan. *Jurnal Bahasa* 12(2): 181–203.

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