LSP International Journal, Vol. 9, Issue 2, 2022, 85–96
© Universiti Teknologi Malaysia
E-ISSN 2601–002X
DOI: https://doi.org/10.11113/lspi.v9.19469

The Potential Contribution of General and Specialized Corpora to Research on Malay and Malaysian English

Zuraidah Mohd Don*
Language Academy, Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia

Gerry Knowles
Independent Scholar

Submitted: 6/11/2022. Revised edition:6/12/2022. Accepted: 6/12/2022. Published online: 26/12/2022

ABSTRACT

Today's linguists are increasingly concerned with high-level properties of texts, and tend to work top-down in some branch of discourse analysis, while corpus linguists are concerned with low-level properties such as grammatical class, syntactic constructions and different kinds of text annotation, and tend to work bottom-up. This paper seeks to close the gap, using a general corpus and a specialised corpus. The point of departure is the assumption that a corpus is compiled to study the language of texts in some language for some special purpose beyond the existence of the corpus itself. The particular languages in mind are Malay and Malaysian English. The introduction deals with matters that have to be considered when a corpus project is planned, and with the problems that can arise, some of which have been reported. The methodology section concentrates on the groundwork that has to be done for just about any corpus-based project, and starts with a project undertaken long before computers were invented, and describes the role of computational expertise in modern corpus-based projects. The results section reports some preliminary work on a specialised corpus containing the speeches of Tun Mahathir Mohamed, which attempts to go beyond the groundwork to ascertain objectively what the speeches are about. The paper ends with a combined discussion and conclusion that summarises the content of the paper.

Keywords: Malay, Malaysian English, Corpora, Specialised Corpora, Empirical Methodology, Frequency Word Lists

1.0 INTRODUCTION

The general idea of a corpus is that it is a compilation of texts for some special purpose of interest to linguists. The word *corpus* itself (plural *corpora*) is the Latin word for 'body', which has also found its way into English in the form *corpse*. There is an older usage connected with coins, as in *Corpus Nummorum Romanorum*[1] 'Corpus of Roman Coins', which is concerned with coins not so much as physical objects, but rather as entities in coinage systems, often represented by images; and a corpus of this kind has the clear purpose of bringing items together for close observation and study as a system. However, in contemporary usage, corpora are generally assumed by default to contain large collections of texts. Text corpora are likewise unconcerned with texts as physical objects, but as entities brought

**Correspondence to:* Zuraidah Mohd Don (email: zuraidah.mohddon@utm.my)

[1] https://www.corpus-nummorum.eu

together for systematic study by means of linguistic procedures to reveal something new of interest to linguists. Because corpora are compiled for some special interest or purpose, corpus linguistics is the academic next-door neighbour of LSP.

The key criterion is that corpora are compiled for some special purpose beyond their own existence. The LSP researcher might reasonably expect the groundwork to have been completed by corpus-based methods, enabling the analysis to start at a higher level, and avoid the problem of handpicked examples by using examples motivated by the preliminary analysis. For example, a corpus dealing with the Malay press coverage of Covid-19 could use information about words in the actual text to discuss the question of English loan words in Malay, and whether (and if so how) the pandemic could be covered using only traditional Malay words. A corpus of English texts on the Malaysian palm oil industry is likely to raise political questions concerning bias on the one hand, and monoculture and the extinction of flora and fauna on the other. In this case, the groundwork can be expected to provide evidence to detect political bias. It has to be said, however, that reasonable expectations of this kind are more of an aspiration than a present-day reality. Although there is a growing literature concerning the extension of corpus-based analysis to higher levels, especially in the case of discourse analysis (Baker, 2006, 2010; Lee, 2008), what is possible in practice depends on the current state of the art. The overall aim of this paper is to indicate how the use of a specialized corpus (in this case the Mahathir corpus) in addition to a general newspaper corpus make it possible to get from the current state of the art to a situation in which corpus linguistics is indeed able to provide the groundwork for the higher-level analysis of texts.

2.0 CORPUS-BASED PROJECTS IN MALAYSIA

It is essential for the purpose of a corpus project to be in focus during the planning stage (for a good example see Ummul & Chai (2014)), because otherwise it is easy to get carried away with the collecting as an end in itself, and end up with a corpus without knowing what to do with it. Real examples are alas not hard to find. In some cases, the purpose is ill defined; for example, Ong (2019) sets out to produce a mega-corpus without explaining what it is for, or what information will be created that could not be obtained from a first generation million word corpus. Although in some special cases, such as the British National Corpus (BNC)[2], a corpus may be designed for a wide range of possible purposes, including some unknown at the time of planning, corpora need in general to be designed with some specific research output in mind.

A further problem is that a large corpus-based project has to be planned for the longer term, and is not something that can be scaled down for a short-term funded project. Some related problem of this kind has beset the intended Malaysian contribution to the International Corpus of English (ICE). The concept of this corpus was put forward by Sydney Greenbaum in the late 1980s, and was to contain separate corpora of about a million words each from different countries around the world. As recollected by both present authors, the proposal for a Malaysian contribution dates to about 2002. By 2014, the ICE Malaysia project had collected just a quarter of the target million words (Siti Aeisha Joharry & Hajar Abdul Rahim, 2014,

[2] https://www.english-corpora.org › bnc

p. 28), and although the project is still listed on the internet for 2017[3], and some unspecified deliverable was expected for 2020 (Kirk & Nelson, 2018), the Malaysian contribution is not included in the list of available corpora on the ICE website[4].

Nevertheless, the general problem does not seem to be to design an interesting and worthwhile project, but rather some uncertainty about what to do when the corpus has been compiled, and perhaps more importantly, how to set about doing it. A survey of corpus-based research undertaken in Malaysia (Siti Aeisha Joharry & Hajar Abdul Rahim, 2014) includes reports on several interesting projects which set out with a clear purpose, but end up with disappointing findings. For example, the Emas corpus (Arshad Abdul Samad, 2004) of essays written by Malaysian schoolchildren is reported (p. 21) to have found that as learners progress, they write more words per sentence and more sentences per essay. The Emas corpus project was a valuable pioneering venture, and had the potential to produce less expected and more significant results. The 2014 survey reports equally disappointing findings from other projects. It also refers (p. 25) to some of the work undertaken by the present authors, but does not include two publications (Knowles & Zuraidah Mohd Don, 2006, 2008), which address this problem directly.

Between the planning and the research which fulfills the purpose of the corpus project is a significant amount of preliminary work which is here called the groundwork. In the first place, material has to be taken from its source and stored in a form ready for processing which also facilitates the analysis. For example, texts from possibly different sources have to be formatted consistently and stored as text files. This is easy enough for text material, but for some corpora, text alone will be insufficient. A corpus of promotional material for retail selling or a general election, or material related to the Covid-19 pandemic, will probably present problems for storage, because images will have to be scanned and saved in a consistent format, and stored separately from text. Preparations for storage need to be made before collecting begins, because corpus material tends to arrive in large amounts that have to be kept somewhere.

The groundwork will typically concern the structure of some language, and in general, corpus linguists can expect to undertake some automatic or automated processing of corpus texts in preparation for the intended main analysis. A corpus will often be compiled for purposes requiring grammatical tagging or syntactic analysis, but there are many other possible kinds of annotation to be undertaken, for example the annotation of errors in a learner corpus, or to highlight anaphora or given and new information, or even the location of metaphors (Charteris-Black, 2004). The annotation may all be undertaken manually, but if huge amounts of data are to be annotated, it is essential in practice to identify what parts of the work can be done by machine, and to make arrangements for automatic processing. Both data storage and text annotation can in principle be undertaken manually, but there is a problem of scale, and where manual processing would in practice take far too long, computational support is essential.

In this way, the groundwork brings together linguistic expertise and expertise in computer science. It is the responsibility of the corpus linguist to identify the linguistic task to be undertaken, and the responsibility of the computer scientist to find a way of carrying it out. A modern corpus-based project needs therefore to be undertaken by a team that has not only the necessary linguistic expertise but also sufficient practical knowledge of computer science to scale up and process large amounts of corpus data efficiently and within a reasonable time. This is a contemporary problem, and it is addressed in this paper.

[3] https://www.ice-corpora.uzh.ch/en/joinice/Teams/icemal.html

[4] http://ice-corpora.net (last accessed 23 October 2022)

The next section deals with methodology, starting with a project from long before computer science, and leading into the design of the present research project in section 4, followed by some preliminary results in section 5. The paper ends with a combined discussion and conclusion.

3.0 THE METHODOLOGY OF CORPUS LINGUISTICS

Because the groundwork combines linguistic and computational expertise, the two can sometimes be confused with each other and with the research design. This section deals with the linguistic part of the groundwork, starting with the first recorded project of its kind.

The First Recorded Corpus-Based Research Project

The first recorded research corpus was used to develop the Oxford English Dictionary (OED) in the nineteenth century, long before computers were invented. The purpose was motivated by dissatisfaction with existing dictionaries, and the desire to create a more complete one, taking into account the contemporary interest in the history of words. To collect data for the dictionary, volunteer readers selected suitable examples, in the course of their reading, of words in contexts which illustrated their meaning, and wrote the examples down on slips of paper. The slips were then sent for sorting in a specially-constructed building in the editor’s back garden (Murray, 1977). First the slips were sorted according to words, and then the slips for the same words were sorted according to their different meanings. Since many words change their meanings over time, the meanings had to be analysed and sorted according to their first recorded use. Finally, the words and meanings were sorted and placed in alphabetical order to form the content of the dictionary. Sorting was such a huge task that although planning for the dictionary began in 1857, publication did not begin until 1884, and the first edition was completed only in 1928. The main editor of the OED was James Murray, who joined the project in 1879 and died in 1915, before publication was complete.

The success of the project was due in the first place to a clear purpose, which in 1857 was rather vaguely to make a better dictionary. The notion of making good the shortcomings of existing dictionaries is acceptable as a purpose in everyday conversation, as is the later idea of making a more comprehensive dictionary; but for a major research project, the purpose needed to be expressed in terms of explicit objectives (or perhaps research questions). The original purpose was eventually replaced by a more explicit and ambitious objective, which was to record all the words used in English since about the eleventh century, and which motivated the groundwork and processing. The groundwork involved the collection of huge amounts of relevant data stored in a consistent format for manual processing, and the processing itself consisted of manual sorting. As the groundwork was completed, it was possible to get on with the main task of producing a dictionary. Understanding the problems involved in reaching the objective required the application of linguistic expertise, and in accordance with the prevailing historical approach to linguistics, the dictionary was designed on historical principles. Understanding the linguistic problems was the key factor that made it possible to design a sequence of procedures to turn masses of data into an organized dictionary.

Managing the Data

Today’s researchers do not need a special building for data storage, because almost unlimited amounts of data can be held on a laptop and an external storage device if necessary. Complex sorts can be carried out by the computer in minutes or even seconds. Nevertheless, we still face the same fundamental research problems. The starting point has to be a research objective. For example, this paper draws on the MaLex project, which sets out to make a description of contemporary Malay that simulates the intuitive linguistic knowledge of a native speaker of the language. The corpus used for this project was then planned and compiled to reach the objective. In this case, there were some corpora already available, including a corpus of Malay novels and a corpus of speeches delivered by Tun Mahathir Mohamed as prime minister. The problem was that both of these were likely to have bias in the selection of lexical items (see below) and possibly also in the grammar. The decision was then taken to compile a new corpus of newspaper articles, which is added to every day by the inclusion of new articles, so that it remains up to date and now even records the language used to refer to Covid-19. Note that the corpus provides data to solve the research problem, and not the other way round. It might be possible in some cases to invent a problem to be solved by the data in an existing corpus compiled for some other purpose, but this does not represent a logical approach to research.

Before work starts on compiling a corpus, there has to be a clear idea of the kind of corpus to compile. If the decision is to compile a general newspaper corpus, the relevant data is in practice to be found in on-line newspapers. The simplest way of making an on-line article available as a corpus text is to select it for copying using CTRL-C and then pasting it into Notepad (or preferably Notepad++). From there it can be saved into a folder in the normal way. This way, just the text is copied, and images are ignored. (Typing CTRL-A is not a good idea as it includes irrelevant unwanted matter.) This is fine for a small corpus, and for learning how to get an article into computer storage. However, as work proceeds it becomes more important to get more information about the text, and if thousands of articles are to be downloaded, to make the process less labour-intensive.

An alternative method is to right click somewhere in the text of the article, i.e. not in an image, and to select “View page source” from the menu that appears. Behind the visible page is a source page written in a language called HTML (Hypertext Markup Language), and consisting of a sea of markup that includes instructions exactly how to present the text on the printed page, most of which is irrelevant for the corpus linguist. However, depending on how the page is designed, typing CTRL-F for Find, and then “<p” (without the double quotes of course), brings the cursor to the first paragraph of the text. The end of the paragraph is indicated by “</p>”, where “/” means ‘end of’. The headline can sometimes be found by searching for “<h1”, and other headings by typing “<h”. The text is hidden away there among the markup, and fortunately it is possible to find it. MaLex downloads the source page for selected articles from the internet, extracts the text, and stores it in a folder as a text file, all without any human intervention.

Books are more useful for specialised corpora on account of the information they contain, but are otherwise unsuitable because of the difficulty of obtaining a useable text, and of creating consistently formatted text files. Unless a soft copy of the text is available, the pages of the book have to be scanned, and the images saved on computer. The images then have to be processed using an OCR (optical character reader) to turn the images into text, and unless the OCR works perfectly, it will be necessary to go through the text making corrections. It is possible to make a word frequency list, and find the errors by checking the list rather than the whole text, because most errors are unlikely to have occurred more than once or twice, and so will be at the bottom of the list. Having found the errors, it is still necessary to use Find and

Replace to make the corrections. Unless there is a strong reason to include them, it is probably better to avoid books when compiling a corpus.

If the corpus is intended for a lexical study, care is required to ensure that it is reasonably balanced (Lee, 2001), and for example not overloaded with words for such things as jungle flora and fauna. But if the object of study is the grammar, it is difficult to see how texts can be unbalanced, for example by having too many passive constructions or insufficient relative clauses. If there are in fact significant differences in the frequency of grammatical constructions, this is something that can be expected to emerge in the course of research, not something that can be known in advance. The experience of the MaLex project is that texts certainly differ in the frequency of grammatical constructions, but that no pattern of variation has so far emerged from the study of newspaper articles selected arbitrarily. Unless and until some indication of bias emerges, it seems safe to compile a newspaper corpus without restriction, as long as care is exercised in drawing inferences about the frequency of lexical items.

Doing the Groundwork

As the corpus is being compiled, the corpus linguist has to prepare for the groundwork required in order to reach the objective. Structural linguists of a century ago associated with Bloomfield (1933) would probably start with phonetics, but this is inappropriate for a written corpus. Corpus linguists have in practice from the beginning started with grammatical tagging. If the OED had been available for corpus-based research, it could perhaps have been used for initial tagging, but in the circumstances, linguists had to invent their own approach. The development of the CLAWS tagger (Garside, 1987; Leech *et al*., 1994)[5] illustrates the importance of collaboration between linguists and computer scientists. Linguists cannot ask computer scientists to get on with the task of analysing corpus texts, because understanding linguistic structure is the special province of linguists, and cannot be expected of computer scientists. In the case of CLAWS, the linguistic expertise was the responsibility of the linguist Geoffrey Leech, and the task of creating a device to meet the linguistic requirements was the responsibility of the computer scientist Roger Garside. In the early days of corpus linguistics, chips were slow by modern standards, and storage space was severely limited, and developers had to do what was possible in the circumstances. Some common words could be included in a tagged list, and many words could be identified by a process later known as stemming. For example, *eggs* consists of the stem *egg* and the ending -*s*, and could be tagged as a plural noun. The problem for early taggers came from words with several possible tags, e.g. *round* is a noun in *a round of toast*, an adjective in *a round face* and a preposition in *round the corner*. The correct tag is chosen by placing the word in an appropriate syntactic context; but forty years ago, tagging and syntactic analysis were undertaken by different research teams, and tagging had to be completed in its own context. Garside's solution was to simulate the syntax by using the probabilities of occurrence of individual tags to calculate the least improbable sequence of tags, e.g. *round* is probably an adjective following a determiner and preceding a noun. In this way, Garside was able to devise a brilliant tagging system that is still in use. Malaysian corpus linguists still use the free tagging facility for English texts available on the UCREL website[6].

[5] https://ucrel.lancs.ac.uk › papers › coling

[6] http://ucrel-api.lancaster.ac.uk/claws/free.html

Empirical Procedures

An essential characteristic of the methodology of corpus linguistics is that it is empirical (Boulton, 2012). The empirical approach uses observable data as the source of knowledge, and corpus linguists likewise base their findings on observable corpus data. This approach is fundamental to scientific method, and leads to the formulation and testing of hypotheses. Valid hypotheses must be testable and therefore falsifiable, but how they are tested varies from one discipline to another. Experiment is the normal method in the natural sciences, and tests for statistical significance are widely used in the social sciences. In corpus linguistics, the hypotheses tend to be informal and even implied, but the testing is done using large amounts of naturally produced authentic data. In this way, corpus linguistics is at the scientific end of linguistics. It also contrasts with the theoretical mainstream in linguistics, for example head-driven phrase structure grammar (Müller *et al.*, 2021), which is still largely concerned with the well-formedness of sentences (Abeillé & Borsley, 2021). Well-formedness is certainly an important consideration, and plays a crucial role in computer programming, for the code written by a programmer is tested for well-formedness by a compiler, and the code fails if it is not well formed. In the case of the compiler, code is tested against the formal rules of the computer language, so that a decision for or against well-formedness is objective. There are no objective rules to test the sentences of a natural language, and so there is an element of subjectivity in judgements of well-formedness. Natural languages are not like computer languages in this respect. Since the development of sociolinguistics in the 1960s, it is not even clear what is meant by “the language”. Although corpus linguistics might be thought of as theoretically lightweight in comparison with the theoretical mainstream, closer inspection of the methodology groups corpus linguistics more closely with the theoretical mainstream in wider academic research.

Many corpus data findings constitute testable hypotheses, but these are probably called something else, such as grammatical classes, syntactic rules, or even affixes. Although it might superficially seem that corpus linguists need huge amounts of data in order to set up hypotheses, no one consciously examines thousands of corpus examples in order to form a hypothesis. Recent brain research confirms that this is unlikely to be the case. A special characteristic of human learning as opposed to present-day machine learning is that we can form hypotheses having examined very small amounts of data, and it is by testing hypotheses that we learn (Dehaene, 2014, 2021, pp. 29–30). If we see a group of swans all of which are white, we already begin to form the idea that all swans are white. This is reasoning by induction, which does not necessarily lead to true conclusions; for example, the next swan we see could be an Australian black swan. We start with a simple hypothesis, and when we encounter contradictory evidence, we replace it with a better one, in this case that all swans other than Australian black swans are white. In this way, we begin to increase our knowledge of swans.

Linguistic hypotheses can be set up using very small amounts of evidence. Noticing that some Malay words beginning *ber-* are intransitive verbs, we can intuitively set up the hypothesis that *ber-* is a prefix marking intransitive verbs. Suppose then that we encounter *berlian* ‘diamond’ which is clearly not a verb, and requires the hypothesis to be amended. The original hypothesis explains a large number of cases, and so it is not completely false but rather insufficient, and taking *berlian* into account is evidence of learning. This is how the initial processing of texts leads to knowledge. For the researcher, it might seem like failure when a black swan is found in the data, so that the hypothesis is falsified and needs to be replaced.

On the contrary, formulating the new hypothesis is an important step forward. The corpus linguist does not need lots of data to set up hypotheses but to test them, and to look for black swans. The description of a language such as Malay begins with a large number of simple hypotheses that are gradually improved to form the rules and declarative knowledge that make up linguistic expertise.

4.0 RESEARCH DESIGN

The research reported in this paper is supported by two digital language models, MaLex for Malay and EngLex for English. MaLex has been produced by both present authors, originally working on a corpus of Malay novels amounting to about 1.8 million words, and more recently working on a general newspaper corpus, which is added to daily and now contains about 1.5 million words. EngLex is the work of the second author, and has been developed over several years through involvement in different projects. These models are intended as a contribution to the Digital Humanities (Zuraidah Mohd Don & Knowles, 2022) in that they use computational means to address traditional problems in the humanities, in this case language description. The objective in both cases is to simulate the intuitive knowledge of the language possessed by its native speakers, and this knowledge is presented as declarative knowledge which can in principle be made available to interested groups such as researchers and language teachers unconnected with corpus linguistics. The methodology implemented in these language models follows the general outline presented in section 3, and uses computer code to imitate what linguists do when making a language description, concentrating on the groundwork which is likely to be essential for any further analysis.

Much corpus-based research is concerned with the description of some language system (Conrad, 2010; Moon, 2010), and for this purpose a general corpus is appropriate, for which a newspaper corpus as used in this research can be considered the prototype. Although the words of each article will reflect its particular topic, a newspaper corpus includes so many different articles that their individual biases cancel each other out, and the corpus as a whole will be lexically diverse and provide suitable data for the intended purpose. However, research that goes beyond the groundwork for some higher purpose requires not a general corpus but a specialised corpus containing texts related to the intended analysis.

The results presented below relate to a specialised corpus of 1,728 political speeches delivered by Tun Mahathir Mohamed during his first term of office as Malaysian prime minister from 1981 to 2003. It contains 1,758,133 English words, 1,222,145 Malay words, and a residue of 172,331 miscellaneous items such as numerals and names. This corpus is currently being analysed using MaLex for Malay and EngLex for English, although the preliminary results presented in section 5 below are concerned exclusively with EngLex and English.

5.0 SOME PRELIMINARY RESULTS

The methodology section was concerned with the groundwork of corpus linguistics, and the kind of research outcome normally expected at that stage. This discussion section raises the question how corpus linguistics can contribute to linguistics as a discipline, and more specifically to LSP. Some research findings are reported here, but they belong to early-stage research that is best regarded as in progress.

The texts in the Mahathir corpus raise several topics of interest at the level of discourse analysis, including Mahathir's use of metaphor and his criticism of unrestrained global capitalism (Aliakbar Imani *et al.*, 2021). To cross the enormous gulf between the requirements of discourse analysis and the text annotation characteristic of conventional corpus linguistics, work has already begun on the Mahathir corpus to find out how corpus-based methods can be used to give an objective account of the content of a text. Some initial findings are available for presentation here.

A raw word list, i.e. one containing all the words in the corpus, is unlikely to be of much interest, as it contains function words at the top end. The most frequent English words (with their frequencies of occurrence in brackets) in the Mahathir corpus are *the* (128,914); *of* (68,497); *and* (65,724); *to* (62,104); *in* (41,303); and *is* (26,847). These function words include the trio *the*, *of* and *and* which is routinely found at the top of the frequency list of any English corpus, and they tell us nothing about how Mahathir used these or any other English function words. The raw list needs to be pruned of function words, which can be kept in a list known as a stop list, so that they can be automatically removed.

Nouns are a more promising category. The twelve most frequent nouns (with very much lower frequencies) are *countries* (8,866); *world* (5,927); *government* (4,893); *people* (4,251); *trade* (3,741); *development* (3,504); *country* (3,351); *market* (2,658); *growth* (2,629); *time* (2,413); *business* (2,355); and *economy* (2,304). These are the kind of words one might expect in a prime minister's speech. Nevertheless, the list contains some anomalies. It includes both *country* and *countries*, which raises the question whether these should be merged as a single item known as the lemma; this is a complex question to which no simple answer is yet available. Secondly, *time* is a general word, which does not really belong in what is otherwise a more specialized list. What seems to be the case is that the frequency of relevant words is boosted in the context, so that they overtake words that are usually more frequent in the language as a whole. Even so, some words are sufficiently frequent to retain their position at the top of the frequency list. The same phenomenon is observed in the frequency list of Malay nouns.

Adjectives are another promising category. The top ten in the frequency list are *economic* (5,305); *other* (4,254); *new* (3,572); *international* (2,621); *foreign* (2,203); *good* (2,015); *private* (1,625); *able* (1,492); *important* (1,480); and *free* (1,456). Again, this list includes both words such as *other*, *new* and *good*, which are frequent anyway, and words of relevance to the speaker's topic, the frequency of which has been boosted in the context. This particular list led to an interesting insight into a difference between English and Malay. Where English uses a derived adjective in a phrase such as *economic problems*, Malay tends to use a noun, as in *masalah ekonomi*. Many examples of Malay kata nama + kata nama corresponding to English adjective + noun have since been found elsewhere (in the general newspaper corpus). Malay sometimes has a kata sifat 'adjective' available, as in the case of *antarabangsa* 'international', but large numbers of English adjectives are being borrowed into Malay to create a kata nama + kata sifat construction parallel to (and perhaps as a translation of) the English adjective + noun. This is an interesting problem that merits further consideration in due course.

Frequency lists of nouns and adjectives can be made by associating each word of the corpus with its grammatical class. The next step in the research involved syntactic information, and counting English bigrams consisting of an adjective followed by a noun. The top ten (with another sharp drop in frequency) are *private sector* (1,032); *economic growth* (583); *economic development* (375); *other countries* (360); *free trade* (349); *same time* (325); *human rights* (301); *free market* (278); *poor countries* (270); and *other hand* (267). These are indeed easily recognized as recurrent themes in Mahathir's speeches. That this is not a statistical accident is confirmed by *economic cooperation* (247), *foreign investors* (212), and *international trade* (197) in the next three positions in the list.

As before, although most of these bigrams relate to familiar themes, there are two exceptions, namely *same time* and *other hand*, which probably get into the list as part of the frequent expressions *at the same time* and *on the other hand*. *Private sector* is perhaps a surprising item at the top of the list. *Private* (1,625) gets into the list of most frequent adjectives, but the more frequent *sector* (2,065) falls below *economy* (2,304) in the nouns list above. No fewer than 1,032 of the 1,625 occurrences of *private* are followed by *sector*, and likewise no fewer than 1,032 of the 2,065 occurrences of *sector* are preceded by *private*. *Private sector* is a good example of a collocation, which is a sequence of words which occurs more frequently than might be predicted from the separate frequencies of the individual words.

6.0 DISCUSSION AND CONCLUSION

Although computer science has made dramatic changes to the research environment since 1857, the work on the Mahathir corpus shows that it is still necessary to follow the same basic steps as those pioneered by the editor of the OED:

1. Identify the purpose and objectives of the research project;
2. Find a source of relevant data;
3. Organise a means of storing the data systematically;
4. Undertake the groundwork;
5. Undertake the high-level analysis required to reach the objectives.

Before work begins, the researchers have to have clear objectives in mind to motivate and guide the collection of data. At least, that is true in principle. In practice, researchers may start off with little idea of what they are trying to do. PhD students may take a year or more before they understand what their thesis is going to be about. Academic books and articles tend to have a life of their own, so that the finished product is not much like the original plan. This is normal, because we cannot know in advance what we are going to learn in the future. What is important is that the project is updated as it develops, and reorganized according to the most recently formulated objectives. The OED started off with vague ideas and even started data collection, but the success that made it possibly the greatest publishing venture of the nineteenth century was due to the later objective to make exhaustive coverage of all English words for the previous 800 years. The objectives have to be or to become sufficiently precise, for otherwise the project makes a collection of words with no obvious reason for existing.

Finding a source of relevant data is much easier than it was just a few decades ago, because huge amounts of data can now be downloaded from the internet. Storing the data systematically can still be a problem, and for some projects this is the stage at which advice is required from computer science. This needs to be foreseen and planned for, because some projects seem to encounter an insurmountable obstruction at the point at which computational support is required; and if computer support is needed and not available, the project cannot proceed.

The groundwork will probably involve some kind of text annotation. Deciding on the nature of the annotation is the linguist's task, as is the provision of resources such as a grammatical tagset or a phonetic alphabet for work on spoken language; and is for the computer scientist to work out how to apply the annotation process to the text. It may be necessary in this part of the work to re-think how to undertake old tasks. For example, in early corpus linguistics, a grammatical tagger would create a new text with tags associated with each word, e.g. in a tagged Malay text, *kereta* 'car' will be tagged as a nama am 'common noun', thus "kereta_NA", and the tagged text would have to be stored as a separate document in addition to the original text. Each new annotation would require the storage of a new set of files. A neater alternative is to make a corpus word list, i.e. the list referred to above as the "raw" word list, probably accompanied by frequency of occurrence, so that words can be sorted into descending rank order. This list can be linked to many different kinds of information, including pronunciation, morphological derivation and grammatical class. If the tags are included in the word list, the text can be tagged "on the fly", so that wherever *kereta* occurs in a text, it is automatically linked to the tag NA. The tagging and other annotations can be carried out whenever required, and it is not necessary to save a separate annotated text. The use of a word list in this way creates a more complex infrastructure for the annotation, but once it is in place, it simplifies the annotation process itself. In this connection, the linguist may, for laudable reasons, have to develop new knowledge and skills, and learn to look at linguistic data in new ways, and develop an understanding of language structure at a deeper level. A charge that corpus linguists sometimes have to face is that their discipline is theoretically lightweight. As will be clear from this paper, such a claim is not true at all. In fact, although the structure of English is now well known, the structure of Malay is not. Zaharin Yusuff (1995) made a proposal for a language centre for Malay, and had it been set up, it would have dealt with these problems long ago.

When the groundwork has been completed, the researcher can get on with the task which is the main purpose of the project. Since each project has its own task, it is impossible to list here the different tasks that might be involved. The preliminary results reported above take a small step beyond annotation and upward in the hope and expectation that this will lead to new techniques that will eventually meet up with the work of discourse analysts and others working down from the level of the text.

REFERENCES

Abeillé, A., & Borsley, R. D. (2021). *Basic properties and elements* (S. Müller, Abeillé, R. D. Borsley, & J.-P. Koenig, Eds.; pp. 3-45). Language Science Press.

Aliakbar Imani, Hadina Habil, & Zuraidah Mohd Don. (2021). Metaphor in Mahathir's political speeches in the context of economic crisis. *South East Asia Research, 29*(4), 434-449.

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

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (Ed.). *Research Methods in Linguistics* (pp. 93–113). Continuum.

Bloomfield, L. (1933). *Language*. Holt.

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (Eds.). *Corpus-Informed Research and Learning in ESP : Issues and Applications* (pp. 261-291). John Benjamins.

Charteris-Black, J. (2004). *Corpus Approaches to Critical Discourse Analysis*. Palgrave Macmillan.

Conrad, S. (2010). What can a corpus tell us about grammar? In A. O'Keefe & M. McCarthy (Eds.). *The Routledge Handbook of Corpus Linguistics.* (pp. 227-240). Routledge.

Dehaene, S. (2014). *Consciousness and the Brain.* Viking.

Dehaene, S. (2021). *How we learn*. Penguin.

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. N. Leech, & G. Sampson (Eds.). *The Computational Analysis of English: A Corpus-based Approach*. Longman.

Kirk, J., & Nelson, G. (2018). The International Corpus of English project: A progress report. *World Englishes*, *37*(4), 697-716.

Knowles, G., & Zuraidah Mohd Don. (2006). *Word Class in Malay: A corpus-based approach*. Dewan Bahasa dan Pustaka.

Knowles, G., & Zuraidah Mohd Don. (2008). *Natural Data in Linguistic Description: The case of adverbs and adverbials in Malay*. Dewan Bahasa dan Pustaka.

Lee, D. Y. W. (2001). Defining Core Vocabulary and Tracking Its Distribution across Spoken and Written Genres: Evidence of agradience of variation from the British National Corpus. *Journal of English Linguistics*, *29*(3), 250-278.

Lee, D. Y. W. (2008). Corpora and discourse analysis: New ways of doing old things. In V. K. Bhatia, J. Flowerdew, & R. H. Jones (Eds.), *Advances in Discourse Studies* (pp. 86-99). Routledge.

Leech, G. N., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. *COLING 94*, 622-628.

Moon, R. (2010). What can a corpus tell us about lexis? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics.* (pp. 197-211). Routledge.

Müller, S., Abeillé, A., Borsley, R. D., & Koenig, J.-P. (Eds.). (2021). *Head-Driven Phrase Structure Grammar: The handbook*. Language Science Press.

Murray, K. M. E. (1977). *Caught in the Web of Words: James Murray and the Oxford English Dictionary*. Yale University Press.

Ong, C. (2019). Development of a Mesolectal Malaysian English Corpus. *ICLLIC 2019*, 123-126.

Siti Aeisha Joharry, & Hajar Abdul Rahim. (2014). Corpus Research in Malaysia: A bibliographic analysis. *Kajian Malaysia, Universiti Sains Malaysia*, *32*(1), 17-43.

Ummul K. Ahmad, & Chai, H. C. (2014). Language of promotion in Malaysian banking brochures. *Southeast Asian Journal of English Language Studies*, *20*(3), 135-146.

Zaharin Yusuff. (1995). Towards a language information centre for Malay. *MT Summit V Proceedings, Luxembourg*.

Zuraidah Mohd Don, & Knowles, G. (2022). The digital humanities and re-imagined language description: A linguistic model of Malay with potential for other languages. *Digital Scholarship in the Humanities*, *37*(4), 1084-1096.