LSP International Journal, Vol. 8, Issue 2, 2021, 79–91

E-ISSN 2601–002X
DOI: https://doi.org/10.11113/lspi.v8.17972

Language and Disciplinary Concepts in Corpus Linguistics: Investigating Corpus Data

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

Gerry Knowles
Independent Scholar

Submitted: 13/11/2021. Revised edition: 4/12/2021. Accepted: 4/12/2021. Published online: 15/12/2021

ABSTRACT

This paper is intended for researchers involved in or contemplating research in corpus linguistics, and is concerned in particular with the language of corpus linguistics. It introduces and explains technical terms in the context in which they are normally used. Technical terms lead on to the concepts to which they refer, and the concepts are related to the procedures, including tagging and parsing, by which they are implemented. English and Malay are used as the languages of illustration, and for the benefit of readers who do not know Malay, Malay examples are translated into English. The paper has a historical dimension, and the language of corpus linguistics is traced to traditional usage in the language classroom, and in particular to the study of Latin in Europe. The inheritance from the past is evident in the design of MaLex, which is a working device that does empirical Malay corpus linguistics, and is presented here as a contribution to the digital humanities.

Keywords: Corpus Linguistics, Malay, Empirical, MaLex, Digital Humanities

1.0 INTRODUCTION

Corpus linguistics has become increasingly popular in recent years as a field of research, but what is involved in doing corpus linguistics has perhaps not always been fully understood or taken into account. Part of the problem has to do with language, because the less experienced researcher has to become familiar not only with new procedures but also with the language used for the specific purpose of doing corpus linguistics. This paper sets out to explain what research in corpus linguistics involves, paying particular attention to the language in this specific context of use.

The researcher who comes to corpus linguistics from mainstream linguistics, and who is accustomed to invent sentences as data ad hoc, faces the unfamiliar task of handling huge amounts of naturally produced authentic data, and has to learn new procedures with their own terminology. Even discourse analysts, who pay close attention to real texts but in small amounts, have to become familiar with the initial stages of text analysis, which they probably take for granted when they

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

begin their work at a higher level. The argument put forward here is that the language of corpus linguistics is not new, but was inherited from the language and procedures traditionally used in the context of language teaching and learning. While it has much in common with the language of general linguistics, it also has its own specific technical terms and senses.

Like any academic discipline, corpus linguistics has an inherited disciplinary language which has been adapted for its own specific purposes. New words are sometimes invented, and some existing words are given new meanings. The technical terms label concepts, which have to be understood in order to use the terms appropriately. The concepts are implemented in procedures, or explain the content of data sets, and these provide the context in which the concepts can be understood. Learning the specific language involves learning the discipline. For example, phonetic transcription involves several data sets, including the text to be transcribed, the output transcription, the phonetic alphabet used, and a table of phonemes and perhaps allophones. Doing phonetic transcription involves a procedure that associates speech sounds with symbols, and when transcribing a written text, first associating spellings with speech sounds. Understanding phonetic transcription as a practice requires an understanding of the procedures and data sets, and of the concepts involved, and familiarity with the technical terms used to label the concepts. Doing corpus linguistics likewise involves practices that include data sets, such as texts, a tagset and sets of rules, and also procedures such as tagging and parsing. Developing expertise in corpus linguistics requires knowledge of research practices, and also experience of using data sets and procedures to extract linguistic information from a corpus.

Perhaps it is appropriate to include in this introductory section a comment on the word *corpus* itself. It is a Latin word meaning 'body', which came into medieval English in the form *corpse*. French *corps* came to be used for an organized body of people, originally soldiers, and the form is now used more generally in English, e.g. *press corps*. *Corpus* /ˈkɔːpəs/ is used in linguistics to refer to an organized body of texts, typically stored in electronic form on computer. Like *body*, *corpus* is a count noun, the plural being *corpora* /ˈkɔːpərə/, although the regular English plural *corpuses* is in occasional use. *Corpus* is not a mass noun, and so expressions such a *lots of corpus* or *more corpus* are inappropriate. A corpus provides data for linguistic research, and appropriate expressions include *lots of corpus data*, and *more corpus data*.

The next section is unavoidably long, and includes the historical background, methods and results. This structure is adopted because in this case it would not be possible to combine coherence with separate methods and results sections. The paper ends with a discussion and a conclusion.

2.0 DOING CORPUS LINGUISTICS

The Origins of Corpus Linguistics in Language Teaching and Learning

Corpus linguistics derives from a long tradition of language teaching and learning, and is closely connected with grammar, which was one of the three Arts of the medieval course of study known as the trivium. Today's ELT grew out of the European language tradition, which itself was modelled on

the study of Latin, which began in Scotland in the sixth century, and more or less came to an end in the 1960s.

English grammar schools were established to teach Latin, and admitted children at about the age of eleven, aiming to enable them to read Latin texts, especially works of literature. (This is the origin of the notion that linguistics is specially connected to the study of literature.) Beginners would be taught to analyse simple sentences such as *Brutus Caesarem occidit* 'Brutus killed Caesar', and in order to do this they would have to assemble a wide range of relevant linguistic information. Understanding the linguistic information would require familiarity with an extensive terminology developed by teachers for the specific purpose of analysing Latin sentences. *Brutus* is a proper noun and the name of a person, and *-us* indicates that the person is male and the subject of a verb. *Caesarem* is also a proper noun and the name of a person, and *-em* indicates that the person is probably male and the object of a transitive verb. *Occidit* is a verb meaning 'kill', and *-idit* indicates an event carried out in the past by some person or entity. The first stage in using this information is to group the object with the verb, thus *Caesarem occidit*, and the next stage is to group verb and object with the subject, thus *Brutus Caesarem occidit*. Having successfully analysed the sentence and put the parts back together again, the learner would then have access to the meaning 'Brutus killed Caesar'. This procedure is traditionally known as construing a sentence, where *construe* is an archaic variant of *construct*.

For someone who does not know Latin, the information in this last paragraph must come across as a jumble of word endings, meanings, gender, time and grammatical rules. In fact, the learner would have even more terminology to master, for *Brutus* is in the nominative case and *Caesarem* in the accusative, while *occidit* is the third person singular form of the perfect tense of the verb of which the infinitive is *occidere*, and which is also referred to informally as *occido* 'I kill', which is the first person singular form of the present tense. In addition, nouns are said to belong to declensions, while verbs belong to conjugations. The language developed by Latin teachers was extended for other languages, and laid the foundations for the terminology of corpus linguistics and modern linguistics as a whole. The terminology would be accompanied by a procedure which today would be called an algorithm, or a formal series of steps leading to the solution of a problem, in this case understanding a sentence.

Of course, the teacher would also provide learners with a language description (for the exemplar, see Kennedy 1888) including a systematic explanation of its structures and systems. The problem is that a systematic description can also be confusing. Someone who does not know Latin cannot be expected to understand the information that the first declension contains mainly feminine nouns but also words such as *poeta* 'poet', *agricola* 'farmer' and *nauta* 'sailor'. The solution is for the learner to approach the theoretical description and the practical ability to read texts in tandem, and learn the terminology to tackle texts of increasing difficulty. In the last century, the theoretical knowledge and practical ability were taken out of the Latin class and developed into the notions of langue and parole (de Saussure, 1916), and later competence and performance (Chomsky, 1965).

Written and Spoken Language

Latin learners would not be required to say anything in Latin to anybody except their teachers, and the same applied to many of the school students who took obligatory French. In these circumstances,

the focus was on the written language, and not much attention was paid to the spoken form. In the first half of the last century, linguists began to study unwritten languages, and developed phoneme theory (Jones, 1950; Pike, 1947), which dominated the emerging spoken component. However, linguists typically use orthographic transcription to study spoken language, and even phoneticians use phonetic transcription, which likewise turns spoken data into written data for study. The outcome was that although spoken corpora began to appear (Knowles & Wichmann, 1996; Knowles, Williams, & Taylor, 1996), and spoken data constituted a significant proportion of major corpus projects such as the *Survey of English Usage* carried out at University College, London (Svartvik & Quirk, 1980), and also the collaborative project called the *British National Corpus* (http://www.natcorp.ox.ac.uk/), speech continued to be studied through writing.

After 1945, the need increased for people to use a foreign language in communication, and by the 1970s a new approach to language teaching had emerged, called communicative language teaching. Whenever we speak or write, we make use of grammatical knowledge; and when we learn a new language, we do not keep it in a separate compartment in our brains, but connect it to our existing linguistic knowledge, for otherwise we would not be able to translate. It is unfortunate that the communicative approach was associated with the rejection of traditional language teaching, now deprecated as the grammar-translation method. The result is that many language teachers came to lack the disciplinary knowledge to teach a language effectively, while the disciplinary knowledge itself was inherited by corpus linguists, who subsequently developed it in the context of their own discipline. Like traditional language teachers, most corpus linguists confine their activities to the study of written materials.

Early Corpus Linguistics

The first machine-readable corpus, containing a million words of American English, was produced at Brown University (Rhode Island) by Henry Kučera and Nelson Francis (1967), and is accordingly known as the Brown Corpus. This was followed by a matching million words of British English, produced in a joint venture by the universities of Lancaster, Oslo and Bergen, and generally known as the LOB /lɒb/ corpus. In the early days, text had to be fed into the computer on punched cards, and getting a million words into memory was a massive undertaking. Until then, the notion of having huge amounts of data to analyse belonged to science fiction, and the first generation of corpus linguists faced the task of working out exactly what to do with a million words, now that it was available. Because different research groups would probably have access to only one corpus, the same data would have to be used for many different purposes.

Compiling a Corpus

In the present century, it has become a relatively trivial task to download several million words from the internet, and several corpora can fit together into the memory of a laptop. Before beginning to compile a corpus, it is now essential to plan in advance what to do with it. There has been a tendency in the past to collect data as though for its own sake, so that the finished corpus seems to be an exhibit in itself, like a stamp collection or a butterfly collection, and many corpora must have been

compiled and never used for anything. Although by default, corpora are used for tasks corresponding to those undertaken by language learners in the days before computers, it is important to think of and exploit the new possibilities offered by computers.

Word Lists and Concordances

Although corpus linguistics is associated with text processing, some interesting and useful tasks can be undertaken on the raw text. One of the simplest computational tasks is to construct a word frequency list. Any English frequency list will identify *the*, *and* and *of* as the most frequent words, but the interesting words are not function words but content words. For example, *xerotic* and *diastole* are likely to be more frequent in a medical corpus than in the language at large, while *finial* and *architrave* are likely to be particularly frequent in an architectural corpus. Conversely, in a corpus of stories for young children, words such as *prostitute* and expletives are likely to be less frequent than usual to the extent of not occurring at all. The expression *key words* is used in corpus linguistics in the special sense of words that are significantly more or less frequent than usual. For researchers in LSP, word frequency lists are a very basic tool.

The corpus text can also be searched by a concordance program for examples of words in use. A KWIC ("key word in context") concordance retrieves the word searched for together with words to the left and the right. (The reader might observe that the term "key word" is used differently in making word lists and making concordances.) Although the concordance program used here retrieves just one word left and one word right, it can be modified to retrieve a wider context. A set of texts searched for the word *sekali* 'one occasion, once' yielded 298 occurrences, of which 129 were followed by *gus*. In fact, *sekali gus* 'simultaneously' can also be written solid as a single word *sekaligus*. There were 59 cases of *sekali lagi* 'once again', and a surprising 9 cases of *sekali guna* 'use once'. All 9 cases were preceded by *plastik* in the expression *plastik sekali guna* 'single use plastic'. This program has access to grammatical class, and in 48 instances, *sekali* is preceded by a kata sifat 'adjective', in which case *sekali* is an intensifier, roughly equivalent to *indeed*, e.g. *jauh sekali* 'far indeed'. However, in 21 of these cases, the kata sifat is *sama* 'same', which followed by *sekali* can be translated 'the very same'.

A language description presents a language as an orderly and organized system. A concordance can reveal what a text is like before it is brought to order; different kinds of pattern may come to light, and have to be handled in some way as the text is processed. One of the uses of a concordance is to identify collocations, groups of words that occur together more frequently than would be expected by chance. The status of *sekali gus* as a collocation is reflected in the solid *sekaligus*. Although *sekali* can follow several kata sifat, it seems specially attracted to *sama*, which indicates that *sama sekali* is another collocation. The interesting collocation is *sekali guna*, which is itself collocated with *plastik* to form the phrase *plastik sekali guna*, and which is presumably a relatively recent formation in connection with climate change. Although the units of syntax are generally assumed to be words, there is also a case to be made for collocations patterning as though they were single words; for example *sekali guna* patterns like a kata sifat.

Tagging

The first stage in processing corpus texts is grammatical tagging. Many researchers on English texts go to the website http://ucrel.lancs.ac.uk/claws/ entitled "CLAWS part-of-speech tagger for English" and the related website http://ucrel.lancs.ac.uk/claws1tags.html entitled "UCREL CLAWS1 (LOB) Tagset", and take advantage of the free tagging service which enables them to get their English texts tagged. UCREL is the University Centre for Computer Corpus Research on Language at Lancaster, and CLAWS is the Constituent Likelihood Automatic Word-tagging System developed at Lancaster in the 1980s (Garside 1987). CLAWS1 is the earliest of several tagsets, and is the one made generally available, and consists of over 130 tags and corresponding explanatory keys, including the tag PP$ which corresponds to the key "prenominal possessive personal pronoun (her, your, my, our ...)". LOB is the name of the corpus used to develop the tagset.

Understanding the tagging process begins with the keys. It is essential to know what a personal pronoun is, and what *possessive* and *prenominal* mean in this context. Before using the tagger, it may be advisable to read up on the relevant grammatical terminology. The tag itself is just a short identifier equivalent to the key; for example, "PP" corresponds to *personal pronoun*, and "$" corresponds to *possessive*. Each tag must be unique and is preferably mnemonic, so that humans can remember it. A computer program would work perfectly well if personal pronouns were tagged "DX" or consistently given any arbitrary tag, but humans are more comfortable with "PP". The keys contain the information essential to process the text, and in this way correspond to the linguistic information assembled by the Latin learner when tackling a sentence such as *Brutus Caesarem occidit*. Whereas the Latin learner uses the information immediately in order to understand a sentence, the program processes the whole text at once and associates a tag with each word; for example, wherever "my" occurs in the text it may be re-written "my_PP$", which can be read aloud as "*my* is a possessive personal pronoun".

As computer science has advanced, new ways have been found to make information available. For example, since the 1980s it has been possible to present sets of related data in the form of tables, and the English pronoun system can be presented in this way. This means that *my* could be tagged simply as a pronoun, thus "my_P", and quite separately "my_P" could be looked up in a pronoun table to recover the full tag "PP$". The tagset would be simpler and easier to understand, and for humans a table of pronouns is also easy to understand. Now suppose that someone researching English pronouns came up with a better way of representing the pronoun system: just the table would have to be revised, and when *my* was looked up it would be given a new and more appropriate tag. The researcher would not be constrained by the way the tagset developer understood the pronoun system. This is unlikely to happen in the case of English, because the language has been studied intensively for centuries. In the case of a language that has been studied little or not at all, the researcher may only know that certain words are pronouns, and have no deeper understanding of the pronoun system at all. The pronoun table could in this case be developed in the course of research, leaving the tagging system unchanged. Significant advances can be made by separating different language systems, in this case the tagging system and the pronoun system. The fact that information can be presented in different ways explains why there is no single tagging system, even for a single language. Different tagsets are developed for different purposes. However, there is some information

common to different tagsets, including tagsets for different languages, and this is connected with the so-called parts of speech.

The study of the parts of speech began in Ancient Greece, and must be the oldest area of study in western linguistics. The main types are noun, pronoun, verb, adjective, adverb, preposition and conjunction, while article and interjection are minor types. The list is not exhaustive, for numerals do not really fit in, and ordinal numbers as in *the third day* seem to be intermediate between numerals and adjectives. A modern tagset has to provide tags for every word in a corpus and also include different kinds of linguistic information for processing, and this is why a typical tagset includes more than 100 tags. The tags are said to represent grammatical classes rather than parts of speech, although the terms part-of-speech tags and even postags are also used. There is some general correspondence between different languages, but this is far from exact. For example, although the English adjective *big* corresponds to the Malay kata sifat *besar*, and both have superlative forms (*biggest* and *terbesar*), the adjective needs a verb to form a clause (e.g. *that house is big*), whereas the kata sifat does not (cf. *rumah itu besar*, also 'that house is big'). For this reason, when work began on tagging Malay texts (Knowles & Zuraidah Mohd Don 2006, 2008), Malay tags were used as a matter of principle for Malay grammatical classes, because the use of English tags was bound to lead to confusion. As research progressed, it was found that some Malay classes did indeed correspond exactly to English classes, but this was certainly not known at the beginning.

When developing a tagset for a new language, it is tempting to copy the tags from a language that already has a tagset. If English *house* is a noun, then Malay *rumah* must also be a noun. But this is false reasoning. English *sleep* is a verb, but it does not follow that Malay *tidur* is likewise a verb. *Ali tidur* can be translated 'Ali is sleeping', but it can also be translated 'Ali is asleep'. *Tidur* is one of many Malay words that belong to a class that corresponds to the border area between English verbs and adjectives. Copying English tags unthinkingly can lead to absurdity. The following example is based on a published article available on the internet, using a different example. English *round* can be an adjective, as in *a round shape*, a noun as in *a round of golf*, a preposition as in *round the corner*, or an adverb as in *she came round*. If *round* is translated into Malay as 'bulat', it does not follow that *bulat* can be an adjective, a noun, a preposition or an adverb. Confusion is avoided if classes are based on the language being investigated, not on the classes of some other language.

Inappropriate classification can lead to other forms of confusion. The word *sebuah* often corresponds to the English indefinite article, as in *sebuah rumah* 'a house', while *seorang* 'one person' is used for people, as in *seorang guru* 'a teacher'. There are other words, including *tersebut* 'mentioned' and *terbabit* 'involved', which are sometimes equivalent to English definite articles, as in *rumah tersebut* / *terbabit* '[the] house mentioned / involved'. Such examples can lead to the mistaken belief that Malay has definite and indefinite articles much like English. In the wider context of Malay grammar, these examples illustrate syntactic structures that have in principle nothing to do with articles at all. In centuries gone by, scholars sought to base English grammar on the grammar of Latin, but eventually realised it was not a sensible thing to do. In our own time, it is likewise not sensible to base the grammar of Malay on the grammar of English.

Corpus Linguistics and Empirical Linguistics

The general point made in the last section is that tagsets have to be based on evidence from the language under investigation, not from some other language or source. This is consistent with the general principles of empirical research, and empirical linguistic research in particular. The word *empirical* derives from a Greek word meaning 'experience', and incidentally, it has nothing to do with *empire*, which is etymologically related to Latin *imperium* 'power'. The empirical approach was originally associated with medical doctors guided by experience, but has since become the scientific approach *par excellence*. Empirical research typically involves establishing hypotheses on the basis of evidence, and then testing them in order to support or falsify them. Hypotheses are tested by experiment in the natural sciences, and typically by inferential statistics in the social sciences. In empirical linguistics, hypotheses are most effectively tested by evidence from corpora. When a hypothesis is falsified, it has to be replaced by a better one. When an initial tagset developed for a new language is tested against corpus data, it will almost certainly be found inadequate in some way. It is then replaced by an improved tagset. Although researchers naturally want to defend their hypotheses and a falsified hypothesis might be regarded in the short term as a failure, the development of improved hypotheses is what makes scientific progress.

Not all linguistic research is empirical in nature, at least on the surface. The dominant research paradigm in the second half of the last century began with a theoretical position which was then supported by evidence often invented by the researchers themselves and so not independent. For example, phrase structure grammar begins with the theoretical claim that *S = NP + VP* 'a sentence consists of a noun phrase followed by a verb phrase'. Noun phrase and verb phrase are then successively analysed down to individual words, and the structure is presented in the form of a tree diagram. What is not usually explained is where *S = NP + VP* comes from. This structure has actually been well known to language teachers for hundreds of years, and probably emerged intuitively at some unknown time from attempts to explain language structure to learners. Understanding the nature of language depends, directly or indirectly, on the empirical study of written or spoken naturally produced data.

A tagset is tested by being used as input to a parser, which takes the steps necessary to group words together to form sentences. Take for example the well-known information that whereas English adjectives come before nouns, Malay kata sifat follow kata nama. Given the sequence *big_J house_N* (where "J" is the tag for adJectives), an English parser will group the words together as the phrase *big house*, and likewise a Malay parser will group *rumah_NA besar_SA* as the phrase *rumah besar*. If by some mischance one of the words has been incorrectly tagged, the rule will not operate, and the error will have to be traced to its source. In other cases, rules may not have been identified and so not included in the parser. For example, the Malay prefix *ber-* is usually thought of as a verbal prefix as in *bergerak* 'move', but it can also be added to nouns, as in *berteknologi* 'with technology'. Although now part of another word, this noun can still be followed by a kata sifat in *berteknologi tinggi* 'with high technology'. Although not a kata sifat, this kind of phrase can be used to modify a preceding noun, as in the expression *keretapi berkelajuan tinggi* 'high speed train'. Although it may seem that English and Malay grammar are quite different in this area, the word order in each case is a mirror image of the other.

Introducing MaLex

MaLex is a digital device containing a collection of data sets and practical procedures designed to process Malay texts. It began with a commission from Dewan Bahasa dan Pustaka to develop a tagger for Malay using a corpus of novels amounting to about 1.8 million words. The first stage was to process basic words such as *rumah* and *besar*. It is impossible to identify the grammatical class of basic words, and so processing has to begin with sufficient linguistic expertise and intuitive native speaker knowledge to choose suitable names for grammatical classes. The obvious source for Malay grammatical classes and appropriate names is the work of Asmah Hj. Omar (1993). Suitable tags have to be devised to represent the classes. *Rumah* is thus kata nama am 'common noun' and tagged *rumah_NA*, while *besar* is a kata sifat am 'ordinary adjective' and tagged *besar_SA*.

As work progressed, it soon became clear that Malay has many complex words that are related by *imbuhan* 'affixation' to the basic words, e.g. *terbesar* 'biggest' (on Malay morphology, see Abdullah Hassan, 1974). Another morphological pattern is reduplication, e.g. *rumah-rumah* 'houses'. Although the aim was to produce a tagger, this could not be done without at the same time investigating the morphology. A typical Malay dictionary lists derived words under the headword, so that for example *terbesar* will be listed under the headword *besar* as "ter~", where the swung dash "~" represents the stem *besar*. The corresponding concept in corpus linguistics is the *lemma*. The lemma BESAR contains many words, but so far we have just a set of two items, which can be represented BESAR = {besar, terbesar}. Note that the name of the lemma is presented in small capitals, and that the form "besar" occurs both as the name of the lemma and as a member of the lemma. The lemma is a set, and like any set, it is logically distinct from its members. The form *besar*, which has no imbuhan or reduplication, is here called the "simplex" form in contrast to the complex members of the lemma.

The investigation of morphology soon encounters another complication. There are several prefixes which seem to be shortened versions of the complete prefix *meng-*: *meng-* itself occurs in *menghantar* 'send', *men-* in *mendengar* 'hear', *mem-* in *membaca* 'read' and just *me-* in *melintas* 'cross (the road)'. There must be some explanation for these variant forms, but it cannot be found just by looking at the spellings; and it is clearly necessary to go behind the spellings and investigate the phonology. The source form is /məŋ/, and /ŋ/ assimilates to /n/ before /d/, and to /m/ before /b/. A widespread rule in Malay phonology simplifies a sequence of two identical consonants to a single consonant, and this is the presumed explanation for *melintas*, in which a former assimilated form /məllintas/ has been simplified to /məlintas/. Linguists traditionally prefer to divide language into distinct layers, and investigate a layer at a time. This is sometimes neither desirable nor possible in the processing of corpus data.

The first few hundred words can be processed manually a word at a time, but this is unrealistic if there are tens or hundreds of thousands of words to be processed. The initial investigation will have suggested some ideas of how words are related to each other, and these ideas can be put together in the form of a program. A stemmer is a program that automatically identifies the stem of complex words and identifies the lemma to which they belong. For example, given the word *membesarkan* 'make big', the MaLex stemmer identifies *mem-* as a prefix and *-kan* as a suffix, and removes them leaving the stem *besar*, and in the process *membesarkan* is identified as a member of the lemma BESAR. In the processing of corpus data, new words are constantly encountered for the first time, and a practical tagger has to be able to handle them in one way or another. The CLAWS1 tagger referred to

above just works out the most likely grammatical class, but MaLex goes further, and identifies the lemma, and works out the meaning of the new word as far as this is predictable. For example, the meaning 'make big' is predictable for *membesarkan*, given that *besar* is a kata sifat.

The stemmer makes a huge contribution to the tagging of a corpus, but of course it can only process complex words, and new simplex words continue of necessity to be processed manually one by one. Simplex words include names, which tend to occur in large numbers in Malay texts, in different languages, and formed according to different syntactic rules. CLAWS1 identifies words with an initial capital letter, but a parser is required to take the analysis further: *May*, for example, could be the name of a woman, a family or a month, or the modal verb *may* in sentence initial position. The problem of names is in practice so great that MaLex treats their analysis as a separate problem.

When words have been successfully tagged, the tags are used as input to the parser. The procedures operate in much the same way as those used by language learners, including Latin learners, to understand a text. The direction is bottom-up, in contrast to the top-down approach of phrase structure grammar. Each syntactic rule corresponds to a hypothesis which can be falsified. The rule that a kata nama can be grouped with a following kata sifat works for simple cases such as *rumah besar*, but sometimes fails in the case of complex noun phrases. The solution to this problem cannot be given here, because it is currently under investigation.

The parser builds constituents of increasing complexity. Well-formed constituents are of interest because they are the appropriate units for translation. As complex constituents are formed, MaLex puts together the translations of their parts to generate a translation at the higher level. At the time of writing, the translation falls far short of literary quality, but is sufficient for someone who does not know Malay to work out what the text is about.

3.0 DISCUSSION

Corpus linguistics as presented here is clearly connected with the traditional approach to linguistic analysis in language learning, to the extent that it can be considered a continuation of the traditional humanities approach. The important difference is that computational methods are employed. Since about the 1960s, new approaches to traditional research problems in the humanities have been developed using the computer as a research tool, and these approaches to research are known collectively as the digital humanities. Corpus linguistics clearly belongs to the digital humanities, and can also be described as kind of digital linguistics.

Corpus linguistics (together with digital linguistics) is part of a much broader discipline known as computational linguistics, which includes any activity that involves both computers and language. For example, the development of speech and language technologies such as speech synthesis and recognition necessarily make use of input from computational linguistics. Herein is a problem understood by all corpus linguists. English has been studied intensively for so long that when the development of these technologies began, the necessary linguistic information was already available. In the case of less studied languages, including Malay, that information was not available, so that engineers and computer scientists had to make do with whatever information seemed to be relevant,

and in many cases that meant borrowing solutions developed for English. However, English is not necessarily a good model, especially for non-Indo-European languages.

Corpus linguistics as pioneered by MaLex has the potential to provide linguistic information for the development of technologies and many other purposes based on the empirical analysis of Malay texts. The first corpus-based dictionary was published by Collins in 1987 based on the work of COBUILD at the University of Birmingham (https://collins.co.uk/pages/elt-cobuild-reference), and marked a major step forward in lexicography. With a lexicon approaching 40,000 words and with access to several kinds of lexical information, MaLex as it stands could be used to produce a corpus-based dictionary of contemporary Malay. It can also provide data for traditional humanities research on Malay lexis, morphology, phonology, and syntax. Long experience in English corpus linguistics led to the publication in 1985 of a massive corpus-based English grammar by the illustrious "gang of four" (Quirk *et al*., 1985). MaLex has built on the essential groundwork already covered (Abdullah Hassan, 1974; Asmah Hj. Omar, 1993) to take the next step towards a corpus-based grammar of contemporary Malay. MaLex can even provide Malay-English translators with a rough draft to work on, thus potentially saving huge amounts of time. The real question is not what MaLex can do, but given a research culture focusing on the empirical study of Malay texts, what can be made available to expedite research beyond the current state of the art.

Recent decades have seen the development of several approaches to the study of language above the level of the sentence under the general heading of discourse analysis. In some cases, such as doctor-patient talk, permission may be given to use only a small amount of data, but in others the researcher is free to collect large amounts of data, and in this way discourse analysis overlaps with corpus linguistics. Since discourse analysis depends explicitly or implicitly on the previous formal analysis of texts, an important question is what preliminary processing can be carried out automatically and made available to the discourse analyst, so that the research can concentrate on higher levels of analysis. Two emerging overlap areas are discussed here, namely critical corpus linguistics and forensic corpus linguistics.

Many researchers have used the methodology of corpus linguistics to study language use from a critical perspective. Critical Discourse Analysis or CDA (Fairclough, 1995) begins with a social issue and investigates the connection between the use of language and ideology and power. Contributions to CDA are inevitably critical of society and the exercise of power, but the purpose of CDA is not just to criticise some person or persons. Since there is a wide range of social issues to investigate, different contributions to CDA do not necessarily have much in common beyond the critical aims themselves. To illustrate the potential contribution of corpus linguistics, and taking as an example a current social issue, it would be possible to make a start on a set of texts in which unemployed graduates are deemed unemployable and blamed for their own unemployability by extracting samples containing the strings "graduat" and "employ". Simulating the work of a researcher with expertise in CDA would also require access to related words and collocations, an ontology, and corresponding frequencies in a comparator corpus such as the 100M-word British National Corpus. Critical corpus linguistics would in practice need to be undertaken by a research group with track records in both corpus linguistics and CDA, and with expertise in inferential statistics and artificial intelligence.

Forensic research is associated with solving crimes, typically for presenting a case to a court of law. Forensic corpus linguistics requires more than a corpus containing texts dealing with crime and crimes. It has an important antecedent in the traditional humanities in the determination of

authorship, and requires the study of the idiolect, which is frequently mentioned in linguistic publications but not developed to the point of practical application. The longest-running case is related to the question who wrote the plays attributed to Shakespeare, which has rumbled on for over 150 years. Given a megacorpus of Renaissance literature and other relevant writings, it would be necessary to identify objective criterial linguistic features and use them to measure the credibility of proposed true authors, taking into account the probability that the true author is unknown and so not in the list. Similar unsurmountable problems would be faced by researchers currently working on any forensic corpus. Real forensic questions, such as identifying the maker of an obscene telephone call, or ascertaining whether or not a suicide note is genuine, require high levels of expertise in linguistics and phonetics, but do not necessarily involve corpora. Nevertheless, it has long been recognised that corpus linguistics opens up the possibility of identifying the linguistic usage of individuals, and forensic corpus linguistics undoubtedly has a promising future in research.

4.0 CONCLUSION

This paper began with the problems of the language used for the specific purpose of research in corpus linguistics. As in the case of other disciplines, the understanding of technical terms typically requires the understanding of the concepts they are used to label, and understanding the concepts typically requires an understanding of the procedures in which they are implemented. The paper is not intended to cover the implementation itself, and stops short of explaining to the new researcher how to set about tagging a corpus or parsing one. These things belong in separate papers, and are in any case covered in general introductions to corpus linguistics.

Linguistic knowledge, like other forms of knowledge, develops over time and can take surprising twists and turns. Although it might not be self-evident that the language and practices of language teaching and learning are the source of the language and practices of corpus linguistics, the link is of course the empirical route to knowledge. The empirical approach is also what connects corpus linguistics to the general culture of science. Corpus linguists may not set up and test explicit hypotheses like natural and social scientists, but tagging and parsing imply hypotheses that have to be rejected when found wanting and replaced by better hypotheses. Linguistic descriptions based on the analysis of corpus data are the best and most reliable linguistic descriptions available at the present stage of linguistic knowledge and research. The new researcher in corpus linguistics has much to learn, but has the potential to build on what has been achieved so far, and like the first corpus linguists, take language description into a new turn that was formerly unforeseeable and unimaginable.

REFERENCES

Abdullah Hassan. 1974. *The Morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
Asmah Hj. Omar. 1993. *Nahu Melayu Mutakhir*. 4th ed. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge. MA: MIT Press.

Fairclough, N. 1995. *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.

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

Jones, D. 1950. *The Phoneme: Its Nature and Use*. Cambridge: Cambridge University Press.

Kennedy, B. H. 1888. *Revised Latin Primer*. London: Longman, Green & Co.

Knowles, G., & Wichmann, A. 1996. *The Lancaster/IBM Spoken English Corpus*. London: Longman.

Knowles, G., Williams, B., & Taylor, L. eds. 1996. *A Corpus of Formal British English Speech: The Lancaster / IBM Spoken English Corpus*. London: Longman.

Knowles, G., & Zuraidah Mohd Don. 2006. *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Knowles, G., & Zuraidah Mohd Don. 2008. *Natural Data in Linguistic Description: The Case of Adverbs and Adverbials in Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kučera, H., & Francis, W. N. 1967. *Computational Analysis of Present-day American English*. Providence, R.I.: Brown University Press.

Pike, K. L. 1947. *Phonemics: A Technique for Reducing Languages to Writing*. Ann Arbor: University of Michigan Press.

Quirk, R., Greenbaum, S., Leech, G. N., & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

de Saussure, F. 1916. *Cours de Linguistique Générale*. Lausanne: Payot.

Svartvik, J., & Quirk, R. eds. 1980. *A Corpus of English Conversation*. Lund: CWK Gleerup.