LANGUAGES AND LINGUISTICS IN 2003: THE POTENTIAL CONTRIBUTION OF CORPUS LINGUISTICS

Gerry Knowles
Department of Linguistics, Lancaster University, UK
& Faculty of Languages and Linguistics, University of Malaya

Abstract

The nature of linguistic research and even the goals of research are changing as a result of information technology This paper discusses what counts as legitimate linguistic data, and the new standards of data collection, organisation and analysis associated with the methodology of corpus linguistics. Two of the more familiar kinds of text annotation are described, namely tagging and parsing, and attention is drawn to the problems of working on Asian languages, including the pitfalls of applying European categories. Two corpus-based projects currently underway in Malaysia are described, one on English and the other on Malay. The paper ends with a look forward to the possible contribution of corpus linguistics to language-based research in Malaysia.

Introduction

The development of information technology has had a profound effect on the way linguists examine language and undertake linguistic research. Language data has always been plentiful, for libraries are stacked with written texts, and speech is all around us, but relatively little of it can be used effectively in research using conventional methods. The situation has changed radically in the last twenty years or so, and it is now possible to store huge text databases (or "corpora") on computer for detailed investigation. Corpus linguists have developed techniques for researching these databases, and corpus linguistics has emerged as a new methodology which takes advantage of the opportunities offered by information technology, and which is rapidly becoming the accepted paradigm for language research.

In this paper I will begin by explaining what corpus linguistics is, and why it is important. I shall then describe some current initiatives which are currently underway and which are relevant to Malaysia. Finally, I will outline some possible ways forward for corpus linguistics in Malaysia

1. Corpus Linguistics

Data in language research

People normally use language for meaning, to convey meaning to others and to interpret what other people mean. In this normal communicative use of language, linguistic form is at most of marginal interest, and probably of no interest at all In order to make a formal study of language, on the other hand, we have to focus on the form of the text itself, and linguistic forms found in texts constitute the data for empirical linguistic research. The problem is that focussing on linguistic form is something human beings are not good at. We process written texts very slowly, and it would take a long time, for example, to collect 100 examples of something like relative clauses, and then we would find that they vary so much that we would need far more examples to make any useful generalisations. It took the compilers of the Oxford dictionary over 70 years to collect and sort the amount of data required for a comprehensive dictionary, and all write the entries.

In view of the difficulties of processing language data manually, a convention has arisen of inventing data to illustrate linguistic points. The problem is that invented data can be misleading. For example, *the man kicked the ball*, or *saya membaca buku itu* might superficially seem to be good simple examples of sentences in English and Malay respectively If they are closely modelled on real data, such examples of structures stripped down to their minimum essentials can be used to provide useful insights into the workings

of the language. But if they are used instead of real data, they give a false impression of what language is like and of what it is used for in the first place. It is actually quite difficult to think up a convincing social situation in which either of the above invented sentences would occur naturally What real people say and write is nothing like the kind of sentence invented by linguists.

Spoken data is even more difficult to handle. Phoneticians have developed methods of representing the sounds of speech in some detail, and phonologists have investigated sound patterning, but there remains much that cannot adequately be handled by traditional impressionistic methods. Techniques of phonetic transcription were developed in the late nineteenth century and the first half of the twentieth, and in order to learn to transcribe reliably and accurately, phoneticians underwent long and rigorous 'ear training' Phonetic transcriptions made by linguists without that training – and that includes the vast majority of linguists – are of the same scientific value as musical scores written by people without training in musical notation. Since it has not been practical for several decades to provide full phonetic training, there is now no way forward at all using these conventional techniques. What most linguists do in practice is to write a spoken text down using conventional orthography, and then to treat it as though it were a written text. This was not too serious a problem for much of the twentieth century, when most linguists concentrated on language systems. But orthography throws away all the phonetic information in a text that is not captured by the orthographic system, and as the focus of attention shifts to discourse, this becomes an increasingly serious problem, as linguists have a totally inadequate model of the spoken language to work on. There are too many areas of normal language use that we are inadequately prepared to deal with using conventional techniques.

The development of large corpora

In order to study language effectively we need natural data in large quantities. For some purposes, such as morphology, we can make do with 100,000 words or even less. The Lancaster / IBM Spoken English Corpus, compiled at Lancaster University in collaboration with IBM UK (Knowles *et al*, 1996), contains just over 50,000 words and yet is enough for the study of phonological patterns, and to a limited extent for prosody. So-called 'first generation' corpora of one million words – including the Brown Corpus of American English and the corresponding LOB ("Lancaster-Oslo-Bergen") Corpus of British English – proved sufficient for the study of most areas of the syntax. Already by the 1980s (Sinclair, 1987) it was clear that for the study of individual words, and particularly collocations and other combinations of words, much bigger corpora were required containing tens or hundreds of millions of words.

Datasets of this size really need to be handled by modern computers. There were certainly attempts in the pre-electronic age to compile large corpora, the most important being the million-word Survey of English Usage. This began in 1959 at University College, London, but by the time it was completed, computers had long been in use. In the early years of computing, getting the data into computer storage was a time-consuming and laborious matter, and this was a major factor in keeping early electronic corpora down to one million words. Keyboarding and the use of optical character readers speeded up the process and made bigger corpora possible. Corpora began to appear not only for English, but for other languages, and for different varieties of English. The Kolhapur Corpus, for example, modelled on Brown and LOB, is a corpus of Indian English, which is of particular interest to linguists in East Asia. By the 1990s, the availability of huge amounts of data on the internet meant that in many cases finding data was no longer a serious problem. For example, the Malaysian government website includes the text of the prime minister's speeches in English and Malay for the last twenty years and more. It is thus now possible to obtain easily a million words spoken by one man alone, and with a choice of language.

A data-driven methodology

The compilation of huge databases of this kind has inevitable consequences for methodology, and for what linguists think they are doing when they analyse a language. Corpus linguistics has crossed a watershed, and the linguist who has gained access to large quantities of natural data will never again be satisfied with phrase structure trees of imaginary sentences.

It must be emphasised that developing a new methodology is not the same thing as developing a new theory. Corpus based methods can be used in conjunction with different theories of language, although the methodology will promote certain theoretical approaches and downgrade others. For example, corpus linguists are unlikely to set as their theoretical goal the definition of all and only the well formed grammatical sentences of the language. The more one examines real data, the less convinced one is of one's own ability in the first place to pronounce on what is grammatical and what is not. And the more one examines individual words in context, the more one's doubts grow about traditional concepts of grammaticality. On the other hand, as one becomes more familiar with real data, it seems increasingly perverse to examine sentences out of context, and so the study of corpus texts naturally leads on to an interest in discourse analysis. While corpus data does not lead directly to a theory, it certainly provides constraints what can be regarded as sensible or realistic theories.

One important principle on which corpus-based research is based and one surely central to any kind of scientific investigation – is the logical independence of data, analysis, and conclusions. By contrast, there is a logical circularity in the use of invented data. For example, it would be easy to invent hundreds or even thousands of Malay sentences in an attempt to prove that Malay syntax is really like English. But because the conclusion is built in as an assumption to the invention of the data, the procedure is logically invalid. A cursory study of real data immediately reveals the great differences between the two languages. A related principle is that research is data-driven. Claims made about the language are based on what is found in the data, and can be substantiated by reference back to the data. This contrasts with the approach that adduces data which allegedly supports some preconceived theoretical explanation.

Large scale text processing

To process millions of words we need computer programs, either to carry out tasks automatically, or to automate the repetitive tasks that can be done by a computer, leaving the researcher to deal with higher order problems that require human intelligence. The basic tool is a concordance program, such as WordSmith, which provides a KWIC ("Key Word In Context") concordance for a key word. All the examples of the chosen word are displayed with (by default) five words to the left and five words to the right, so that the user can get an idea of how the word is used in context. In an informal corpus of nearly a million words of Malay texts taken from a collection of novels, for example, the word *hati* is found to occur 2076 times. One might expect *sakit hati* to be a common collocation, but in fact it occurs only 12 times. *Sakit lelaki* occurs once, but *sakit perumpuan* not at all, *sakit suami(nya)* occurs 10 times, and *sakit isteri(nya)* only 6 times. In a novel about Anisah and Seman, *hati Seman* occurs 34 times, and *hati Anisah* only 23 times. This may be a trivial example, but it illustrates an important point. We do not have good intuitions about what is frequent or normal in language. Conventional expressions, such as *sakit hati*, which we might expect to be common may on the contrary prove to be relatively rare. Our expectations about word associations may also be wide of the mark. For example, one might expect *hati* to be associated with females rather than with males. But unless the corpus happens by chance to be misleadingly unrepresentative, this would appear not to be the case. The use of a concordance program has a salutary effect on the linguist's assumptions about what is normally to be found in texts.

Text annotation

A concordance program can be used directly with conventional texts, without any modification or processing, and can produce much enlightening information. However, to go beyond superficial processing, we need to make use of linguistic information in the text, and for most purposes that means adding expert annotations to the words of the text.

We can annotate anything that requires expert knowledge. We can mark the part of speech of words, we can identify metaphors, or even words of Latin origin. Another possibility is anaphora: we can mark the antecedent for pronouns such as *he* and *she*, i.e. the previous item in the text that these words point back to. Annotations can be checked by someone else, so that if we mark *orang* as a preposition, or link *dia* back to the wrong person, the errors can be pointed out and corrected. In this way annotations can be validated independently, and any conclusions drawn from the study of the annotations are logically independent of the annotation process itself. This is of fundamental importance, since in much 'theoretical' discussion, there is a logical circularity linking the annotation or categorisation on the one hand, and the conclusions drawn on the other.

In conventional linguistics, using a small amount of data, it is possible for the linguist to process the data manually. Using corpora of millions of words, this is of course unrealistic, and in practice we have to adopt or devise automatic or automated procedures. Computers are good for certain routine tasks – such as counting that human researchers would not have the patience to do accurately, or indeed do at all. In the corpus of 120,000 words taken from novels, *hati* and *rumah* are the commonest nouns; in prime ministerial speeches, among the high frequency words are *kerajaan*, *rakyat* and *kita*. Such facts may be unsurprising in themselves, but knowing what is frequent and normal changes the way we look at texts, and prompts us to ask interesting new questions.

Most corpora contain written language, for the obvious reason that written texts are easier to obtain. To handle speech, the traditional procedure is to make a phonetic transcription. But if the transcription is on a piece of paper and the original recording is on an audio cassette, it is impossible to check the transcription in detail, even by listening back to the recording. Without checking, there is no reason to believe an impressionistic phonetic transcription. To work on speech we have to start by recognising the nature of the primary data, which consists not of something written down, but of speech waveforms. Speech data has first to be digitised and stored on disk, and then accessed using speech analysis software. The waveform can be annotated at phoneme level, and claims that a certain part of the waveform represents an instance of a particular phoneme can be checked. Similarly if we claim that a

certain syllable is "stressed", the waveform can be checked for the appropriate attributes of a stressed syllable. Because of the huge amount of expert labour required, it is still only possible to process relatively small amounts of speech data. The London-Lund Corpus (Svartvik, 1990), which was originally the spoken part of the Survey of English Usage (discussed above) appeared with a detailed prosodic transcription, but alas this cannot be checked. For the Spoken English Corpus, it was only possible to mark outline prosody, although this can be checked. The British National Corpus, with much larger samples of speech, has only an enriched orthographic transcription.

Tagging and parsing

In order to go beyond the the kind of searching of the raw data that can be undertaken by a concordance program, we need in practice to gain access to the grammatical structure of texts. This is done using two basic kinds of annotation known as tagging and parsing. To illustrate the need for grammatical information, consider the case of homonyms, which are a frequent nuisance in concordances made on raw text. For example, a concordance-based study of *can* and *may* in English will throw up examples of *can* in the sense 'tin can', and a study of words with the *–kan* ending in Malay will inevitably throw up *akan* and *makan*. To study the data effectively we need to know the grammatical class of words, e.g. whether they are nouns or verbs, and we have to identify grammatically relevant strings within words, and distinguish them from arbitrary strings of letters.

Identifying the grammatical class of the words of a text is known as *grammatical tagging*, or simply as "tagging" The point of departure for tagging is the traditional set of "parts of speech", and a "tagger" identifies each word as a noun, a verb or an adjective, etc. The tag can be associated with the word in different ways, but perhaps the simplest is the use of the underline character, e.g. *The*_article *man*_noun *kicked*_verb *the*_article *ball*_noun. When we tag a text, we have to tag it exhaustively and not leave any words untagged, and we need more detail than is given by traditional parts of speech. Whereas there are traditionally only eight parts of speech, a typical modern tagset will include at least a hundred different categories. There are also conventional mnemonic labels, e.g. VBD 'verb in the past tense', or NNS 'plural common noun'

A "parser" uses the tags to group words into phrases, and phrases into sentences. For example an article combines with a noun to form a noun phrase, and looking at the tags rather than the individual words in the invented sentence above, we can form the noun phrases *the man* and *the ball*. A (transitive) verb combines with a noun phrase (object) to form a predicate. thus *kicked the ball*, and a noun phrase (subject) combines with a predicate to form a complete grammatical sentence, thus *the man kicked the ball*. In this way the tags

provide the essential information needed by the parser to complete the grammatical analysis of the sentence. This is of course the analysis of a sentence invented specially for ease of analysis, and while it may be useful for illustration, it can be misleading, for real sentences are rarely as simple as this to analyse.

Processing Asian languages

The system of "parts of speech" was originally developed for the teaching of Latin, and it was then applied to modern European languages, with differing degrees of success. English, for example, is not at all like Latin in its grammatical structure, and many English words do not fit easily into Latin categories. To some extent the problem has been eased by the development of large tagsets, but other problems remain. For example, English has for hundreds of years had words like *telephone*, which can pattern like a noun or a verb, and this is handled by claiming that English has two words of the form *telephone*, one a noun and the other a verb. Much of the effort of tagging goes into resolving so-called "ambiguities" of this kind.

The differences between English and Latin are relatively minor compared to Asian languages, with their very different grammatical structures (Knowles and Zuraidah, under review). This problem has not been solved, for example, in current corpus work on Chinese and Korean. What is happening in practice (and in the absence of a better approach, by default) is that English categories are being imposed on Asian languages, notwithstanding the fact that they manifestly do not fit. The source of the difficulty is that English is in such a dominant position that most linguistic analysis is carried out on English. It can be very difficult for scholars working on other languages to tell which properties of English are peculiar to English, and which are shared by other languages. In the short term, it is easier and actually less controversial to analyse Asian languages as though they had the same categories as English. If we use invented examples such as *saya membaca buku itu*, we can even persuade ourselves that Asian languages are really like English after all, leaving aside such things as the order of nouns, adjectives and determiners. It is not so simple, however, dealing with real corpus texts.

One of the interesting properties of Malay is that it presents this problem in such an extreme form that in empirical work it just cannot be ignored. The problem is disguised in traditional linguistics by the logically circular methodology that allows the researchers to invent the data that they are going to analyse. Malay linguists learn this approach in the US, and then apply it to Malay. It works, and sentences appear to have nouns and verbs and conjunctions just like English, but only if we start off in the first place with sentences that are structurally similar to English. The logic is again circular. Real natural

texts of Malay have structures which are nothing like English at all. Adjectives have the habit of turning up as adverbs, and when they are predicators they function as verbs, verbs turn up as adjectives or adverbs, and even as nouns. For linguists trained on "parts of speech", real Malay texts can be very confusing. There are many Malay words – including *lepas* and *lalu* – which defy inclusion in any of the English 'parts of speech'

2. Current initiatives: MALEX and MACLE

Two projects are outlined here, one on Malay and the other on English. These are the MALay LEXicon ("MALEX"), and the Malaysian Corpus of Learner English ("MACLE").

MALEX

The problem of Malay grammatical class is being tackled on the MALEX project. The pilot study for this project was undertaken in September 2001 with the support of Dewan Bahasa dan Pustaka, and work proper began in 2002. We have so far processed 150,000 words of Malay text taken from the corpus held at Dewan Bahasa dan Pustaka (Knowles and Zuraidah, in preparation). For this we have set up an annotated lexicon (or 'computerised dictionary') of about 15,000 words, which we then use to tag the texts using Malay categories that we have actually discovered in the texts. A prototype parser (again based on the empirical study of the texts, and implementing grammatical rules some of which are quite unlike anything to be found in English) checks the tags, to see if they are correctly assigned. We plan to increase the coverage to 1 million words in 2003, and automate the process, so that we can predict the grammatical class of most words the first time they are encountered in a text.

Grammatical tags may not sound very exciting in themselves, but they provide the key to any intelligent text processing. Any access to the meaning of a texts or the words in it will require the kind of information stored in our lexicon. To get a computer to read the text aloud, or for automatic speech recognition, we need to know about grammatical tags, tag sequences, and the formation of phrases and sentences. But apart from computer-based applications, the annotated corpus gives a wealth of information about what is frequent and normal in the Malay language (Knowles and Zuraidah, in press). Traditional linguistics concentrates on what it is theoretically possible to say in a language. For many purposes, such as language teaching or speech and language therapy, it is essential to know what speakers normally say

MACLE

The work on MALEX is complemented by the MACLE project, which began as a pilot project at the University of Malaya in October 2002, and which deals with learner English. Corpora of conventional texts have for some time been used for teaching purposes (Wichmann *et al*, 1997), and attention is now being paid to texts produced by learners themselves. The traditional approach to learner English involves so-called error analysis, which concentrates on things learners fail to do. Any teacher will be familiar with such learner problems. In real life we want to know what people can do in English, and this is difficult to assess positively using traditional methods. Our initial aim is to collect 200,000 words of English written by Malaysian learners of English, which will form the Malaysian contribution to an international initiative based in Belgium and dealing with learner English. Where the learner writes grammatically well formed English, we can expect a tagger and parser to process the text reasonably accurately; and where the grammar is faulty, we can expect the automatic processing to break down. In other words, we expect to collect accurate information on what learners can and cannot do in English.

Why are these projects important?

From the point of view of research, these projects reflect a major step forward taken by corpus linguistics. On the surface, a Malay lexicon may seem to have nothing whatsoever to do with learner English; but viewed from a research perspective, they are instances of what is basically the same class of problem. In this respect the methodology marks a major step forward in linguistics research. By making the data logically independent of the analysis, and the analysis logically independent of the problem to which it provides a solution, corpus linguistics is developing a methodology of general application. Annotations and conclusions can be contested if necessary, and falsified when they are in error. In an area where researchers constantly have to address the question whether their work counts as scientific or not, this methodology gives greater confidence.

To begin with we can be precise about our primary data: waveforms for speech, and wordstrings for written language. Our annotations are precise, and can be checked; and indeed checking and cross-checking are standard procedures. To automate our procedures we have to write formal algorithms. Subjective judgements of grammaticality are replaced by probabilities and frequencies per million words of text. Our corpora are samples of a language, and we take for granted that independent samples come up with similar results. Our conclusions are not personal beliefs but precise and falsifiable claims. There is still room for argument over whether our work should count as

"scientific", but at least we have come a long way since linguists earnestly debated unrealistic sentences such as *The man kicked the ball*, or *Saya membaca buku itu*.

In addition, linguistics is able to make a greater contribution to the solution of language problems in the real world. Dictionary makers need information on the frequency of words, and on the range of meanings of words. This is precisely the kind of information most easily obtained from a corpus and concordance. Grammar writers, too, need access to the same kind of information. At the present time, it is impossible to write an adequate grammar of Malay for use in schools, and that is because there is no agreement among scholars on word class in Malay. Without word class, attempts to describe Malay syntax are futile. These problems are being addressed directly by the MALEX project. Automatic methods of retrieving information from large sets of documents, e.g. from websites, requires a stemmer, a program that strips words down to their basic form, or lemma, so that, e.g. *penulisannya* is associated with the lemma *tulis*. Our lexicon incorporates a stemmer, and texts can be lemmatised automatically

The development of a knowledge base for English Language Teaching surely need no defence. Belief and guesswork are replaced by hard information, which will in turn enable materials developers to design taught courses in response to known needs. The corpus itself is an important resource for developing a new culture of autonomous language learning. Apart from teaching, our corpus will have an important role in clarifying the relationship of Malaysian English to global English. By comparing Malaysian data with corresponding data collected in other parts of the world, we will have a clearer view of what to expect in 2020.

3. The way ahead

The scale of corpus-based research

Corpus linguistics requires a research operation larger in scale than is usual in the Arts or Humanities. To begin with, research has to be interdisciplinary, and that requires cooperation across departments and faculties. Corpus linguistics has from the beginning had a close relationship with computer science. In a sense corpus linguistics forms the linguistic end of the new discipline of computational linguistics. Some corpus linguists only work on linguistic data, using ready-made tools, while others are actively involved in writing software. (Working on Malay, incidentally, means that one has to write special software.) This means that collaborative research across faculties is possible and desirable, and beyond the early stages of research absolutely essential. Com-

puter scientists do not in general know enough about language, and linguists do not know enough about computer science.

But linguistics is not only linked to computer science. The study of language is so broad that it is necessarily inter-disciplinary, and involves sociology and psychology, educational studies, history and geography, literature and music, anatomy and physiology, physics and paleography, anthropology and even in some cases genetics. Our current plans for MALEX and MACLE will require us, sooner or later, to collaborate with colleagues in most of these disciplines. The twentieth century view that the study of language was the special preserve of scholars called linguists is already beginning to look rather dated.

Future prospects

This new linguistics is timely for a country like Malaysia. While Malaysia has produced linguistic scholars of the first rank, in this new situation it has the advantage of not being tied down by a particular linguistic tradition. Corpus linguistics offers a new direction, building on the best empirical work of the past, and a means of introducing or upgrading research in many different areas. It is actually difficult to think of any area of linguistic enquiry which would not benefit significantly from corpus-based techniques. There is no virtue in tolerating inaccuracy in the identification and annotation of data, in sets of rules containing logical discontinuities or circularities, or in generalising from inadequate amounts of data. Nor, now that empirical methods have been developed, is there a case for armchair linguistics, which largely consists of sitting and thinking beautiful thoughts about language.

The first initiatives are beginning in Malaysia at several universities. In this paper I have briefly outlined two with which I am involved, and I am aware of other initiatives elsewhere. The size of modern research projects puts them beyond the capacity of a single institution, even with collaboration across faculties. To make a success of corpus linguistics, we therefore have to start by recognising that any one university is unlikely to have sufficient expertise and available research time, and that collaboration across institutions is essential. Competition is fine if the competitors are able to compete; but if they all have inadequate resources, the result is that everybody fails. The ideal would be a Malaysia-wide centre for corpus linguistics research, bringing together scholars from different Malaysian universities. The aim of such a centre should be to integrate research into real-life language problems using the latest technology. For example, the Malaysian government's policy on English raises a number of research questions, and linguists working together could make a significant contribution to its effective implementation.

Beyond Malaysia, corpus linguistics is already organised on a global scale. Interest in corpus linguistics is growing in East Asia, in China and Hong Kong, in Korea and Japan, and in Singapore and Brunei. Initiatives in Malaysia need to be linked up to what is going on elsewhere. In the coming years, success in research will depend on the development of centres of excellence with international partners. There is no reason at all why Malaysia should not be able to build up a centre of excellence in the field of corpus linguistics to play a leading role at a regional level and a major role at a global level. A country that can start a car industry at the end of the twentieth century instead of closing one down, build the twin towers and the Penang bridge, and face the future with Wawasan2020 can surely make a success of corpus linguistics too. What is needed is the vision and determination, the active contributions of high calibre scholars, and the confidence that Malaysia boleh.

References

Knowles, G., B.J Williams & L.Taylor, (1996) eds, *The Lancaster/IBM Spoken English Corpus. a corpus of formal British English Speech.* London. Longman.

Knowles, G. and Zuraidah Mohd Don (under review) 'The notion of a "lemma": headwords, roots and lexical sets' submitted for publication in the *International Journal of Corpus Linguistics*

Knowles, G. and Zuraidah Mohd Don (in press) 'In pursuit of normal Malay speech. designing a spoken corpus of Malay' to appear in *Proceedings of the International Symposium of Linguistics and Speech and Hearing Sciences.*

Knowles, G. and Zuraidah Mohd Don (in preparation) *Word Class in Malay. a corpus-based approach.* To be published by DBP

Sinclair, J.M. (1987) ed, *Looking Up. an account of the COBUILD project in lexical computing* London & Glasgow: Collins ELT

Svartvik, J (1990) *The London-Lund Corpus of Spoken English.* Lund. University Press.

Wichmann, A., S. Fligelstone, T McEnery and G. Knowles (1997) *Teaching and Language Corpora.* London. Longman.