TRAMES, 2010, **14**(**64**/**59**), 1, 90–103

PROCESSING NATURAL MALAY TEXTS: A DATA-DRIVEN APPROACH

Zuraidah Mohd Don

University of Malaya

Abstract. This research represents the first attempt to produce a working system for the automatic processing of texts of Bahasa Melayu 'Malay'. At the heart of the system is an integrated relational lexical database called MALEX, which draws on the experience of working on English and other languages, but which is specifically tailored to the conditions of Malay. The development of the database is from the beginning entirely data driven, and is based on the analysis of a corpus of naturally produced Malay texts. In designing procedures which access the database, properties of the text are consistently and rigorously distinguished from properties of the lexicon and of the grammar. The system is currently used to provide information for a range of applications, for grammatical tagging, stemming and lemmatisation, parsing, and for generating phonological representations. It is hoped and intended that the design features of MALEX will be transferable, and provide a model for the development of working systems for other Asian languages.

Keywords: corpus, lexicon, text, part of speech, Malay

DOI: 10.3176/tr.2010.1.06

1. Introduction

MALEX (MALay LEXicon) is an annotated lexicon designed as a relational database. It brings together in a systematic and logically consistent manner several different kinds of linguistic information required to process Malay texts automatically. The intention is to create a linguistic resource that is useful, and make linguistic information available in a form in which it can be exploited by the wider research community. What we are trying to do is to make available the kind of expert information we possess as linguists.

MALEX is designed according to the logical relationships among different kinds of linguistic information, and can generate suitable output for a range of computer-based applications. For example, using grammatical and phonological

information, we can output a detailed phonological representation of a text, which has useful applications in speech science. Although MALEX is not intrinsically designed as part of an information retrieval system, we believe that it has potential applications in this field. We routinely retrieve data on a small scale using complex Boolean descriptor combinations, e.g. verbs formed with the prefix *ber-* and derived from nouns formed with the suffix *-an*; and if necessary we could retrieve sentences containing words related to *merah* 'red', or all the variant spellings of polysyllabic words borrowed from English. As the number of documents in our corpus approaches 2000, we are considering how best to classify our documents automatically, so that we can retrieve different subsets for analysis. Since our database contains a wide variety of interlinked information, it could presumably be used for data retrieval or document retrieval on a larger scale.

Although Malay is one of the most widely spoken languages on earth, perhaps ranked fourth after Spanish if Bahasa Malaysia and Bahasa Indonesia are grouped together as a single language, it is one of the least studied and known about, to the extent that it is even left out of rank orders of the world's major languages. Researchers working on other major languages have rich linguistic traditions to draw upon, and while there is an extensive literature on the Malay language, there is little systematic information for the researcher beyond a few ill-defined labels for grammatical word classes (e.g. *kata nama* 'noun' and *kata sifat* 'adjective') and informal descriptions of derivational processes in the morphology. This lack of scholarship can be regarded either as a major obstacle, or as an interesting opportunity. In the first place, it is possible to develop a language system that is entirely data driven. That is, theoretical positions are based on generalisations to be made about large amounts of naturally occurring language data. This contrasts with the usual approach in linguistics, which imposes theory on to data, and even invents data to support a predetermined theoretical position. Secondly, instead of having to reconcile partial descriptions inspired by a range of (possibly competing) theoretical positions, it is possible from the beginning to design an integrated database underpinned by a consistent theoretical framework.

2. The data

Work on the database began with the analysis of the texts of a set of novels, some 800,000 words, provided by Dewan Bahasa dan Pustaka (DBP) in Kuala Lumpur. DBP has also provided us with a newspaper corpus of about 5M words, of which we have so far processed only about 1%. In addition, we use (with permission) a corpus of 1.3M words of speeches of the former Malaysian Prime Minister, Dr Mahathir Mohamad. Finally, we have included an academic text of some 20,000 words. All the data collected so far consists of written texts. The longer-term intention is to include spoken data (which will necessitate an extension to the database to handle prosody), and to do this the analysis of written data is the necessary prerequisite. Such a collection of data is not what is normally

thought of as a 'corpus'. Ours is a collection of computer-readable texts that happened to be available, and is perhaps better thought of as an archive. For compiling lists of words, for investigating the morphology and the syntax, and for designing the database generally, this archive has provided large amounts of suitable data in the most economical manner. What we cannot do is to generalise from our data and make frequency claims about the language as a whole.

3. The integrated design

The main table contains the lexicon, the list of over 30,000 different orthographic words found in our data. Related tables include a spelling normaliser, a tag set, a list of lemmas, morphological derivations, and a pronouncing dictionary. Setting up these tables has involved the writing of software, including a stemmer and a pronunciation program implementing a spelling-to-phoneme algorithm. What in linguistics is normally thought of as "theory" takes the form of formal procedures and logical relationships.

3.1. Normalising text

Words can appear in several different orthographic forms in texts, but the aim is to have a single entry in the lexicon for each lexical word. The first stage in normalising the orthographic form is to strip off punctuation. The entity bounded by spaces (or the beginning or end of the line) can in principle be divided into pre-punctuation (e.g. opening quotes and brackets), the lexical item itself, and punctuation (including closing quotes and brackets). Each of these items matches our intuitive understanding of how the writing system works, and separates the item that belongs in the lexicon from non-lexical features of text. The approach taken here contrasts with the one taken in CLAWS tagging (Garside 1987), in which punctuation marks are treated as separate 'words'. Since punctuation constitutes an important environment for the 'disambiguation' of tags, the CLAWS approach has the advantage of simplicity, at the cost of obscuring the relationships among the 'words' that make up the text.

Text normalisation ensures that each lexical item occurs only once in the lexicon. This can be difficult to achieve in the case of compounds. As in English, compounds can be written solid, hyphenated or as separate words, e.g. *kakitangan*, *kaki-tangan* or *kaki tangan* 'employees'. For compounds written as one word, we only have to decide on the solid or hyphenated form. For example, if we decide that *kaki-tangan* is normally written with a hyphen, we can correct *kakitangan* at the normaliser stage and so link *kakitangan* in the text to *kaki-tangan* in the lexicon. However, if we are linking text to lexicon at the level of the orthographic word, we encounter problems with compounds, e.g. kaki *tangan*, written as two words. We have no option but to link *kaki* in the text to *kaki* 'foot' in the lexicon, and *tangan* to *tangan* 'hand'. The two words have to be joined up against at a later stage, when the text is annotated for compounds as a prerequisite to syntactic

parsing. In this way, even text normalisation is not an isolated procedure, and has to be linked to parsing.

3.2. Stemming and lemmatising

A stemmer strips the words of a text down to their basic forms so that English walking is stripped down to walk. More formally, WALK is a set of morphologically related lexical items, or a 'lemma'. In corpus linguistics it is more common to refer to lemmatising a text that is, associating each word in the text with its lemma. It is important not to confuse simplex forms and lemmas: simplex forms are found in the text, and lemmas are found in the lexicon.

A well designed lexicon will identify the lemma to which any lexical item belongs, and thus do the job of a stemmer as a matter of course. If we encounter the word *membacakannya* in a Malay text, we first need to have included *membacakan*, and that in turn presupposes the inclusion of *membaca* and *baca*. If we have the lexicon set up in such a way that we know the stem of each word, we can look up *membacakannya* to find its stem, and then look up the stems recursively until we find the simplex form *baca*. In this way, the lexicon contains a complete morphological analysis of each word.

3.3. Tagging and parsing

For English, tagging and parsing are in practice treated as complementary but independent processes. A tagger associates with each word of the text a label indicating its grammatical class, so for example women will be tagged as a noun and old as an adjective. The parser groups words of certain classes to form phrases and higher level structures up to the sentence, so for example the adjective old and the noun women will be grouped together to form the noun phrase old women.

For Malay, these two processes have to be treated together. In the absence of any general agreement on the nature of the grammatical classes of Malay, we have produced a tag set based on our corpus (Knowles and Zuraidah Mohd Don 2006), and as far as we know, this is the first time a tag set has been used to tag Malay texts. The problem is that anybody can make up a tag set and claim that it represents the grammatical classes of Malay, and in these circumstances a tag set has to be validated in a way that is unnecessary for English. This is because the English parts of speech are very familiar, and a tag set which refines and subdivides the parts of speech can be assumed to be valid. The syntactic rules of English are also well known, and any linguist with expertise in English knows how to use English tags in the analysis of the syntax. In the case of Malay, precisely because these things are not known, a tag set can only be accepted as valid to the extent that the grammatical information it provides can successfully be used by a parser to analyse syntax. We have a parser currently under development, and it makes use of the information in our grammatical tags.

3.4. Pronunciation

Pronunciation fields have been included in English dictionaries for some 250 years, and so it would seem natural to link pronunciations directly to the entries in the lexicon. This could be done for Malay words, and for the vast majority, the results would be unproblematic. However, entries in the lexicon are set up as unique entities with respect to the grammar, and there is no guarantee that they will always be phonological entities and therefore pronounced the same. The normalisation table treats *tetapi* and *tapi*, or *tidak* and *tak*, as variants of the same lexical item, and this is quite reasonable as far as the grammar is concerned. Since they are pronounced differently, this has to be accounted for in the design of the normalisation table, and the spellings *tidak* and *tak* have to be linked to different entries in the pronouncing dictionary.

Entries in the pronunciation dictionary are generated by a spelling-to-phoneme algorithm which has access to morphological derivations to map spellings on to phoneme strings; it divides words into syllables, and adds some phonetic detail according to the position of the sound in the syllable.

4. Discussion: text, lexicon and grammar

Processing techniques for English texts have been developed over a long period using to a large extent a ‘common sense’ approach to the structure of English. This cannot be done for Malay, partly because the structures of Malay remains largely uninvestigate, and partly because it is strikingly different from English anyway. The development of the database incorporates two important design principles, namely the logical organisation of data, and the separation of properties of text, lexicon and grammar.

By organising the data from the beginning, and never actually losing the links between related pieces of data, we start off in a very different way than is possible for a well-researched language like English. Grammatical tagging, stemming, morphological analysis and lemmatising are all ancient and closely related practices, and they were formally brought together in the Latin classroom; but the connections are not always clear in modern computational practice. Parsing derives from a different tradition of school grammar, formalised in generative-transformational grammar and other modern approaches to grammatical structure. English dictionaries were first devised to explain hard words, and had nothing to do with texts or with grammar. Modern wordlists may be grammatically tagged, and might have an incidental connection with a corpus, but they are unlikely to be formally linked to a grammar. English phonetics started with the study of spelling, and went on to study speech sounds; and while pronouncing dictionaries have been around for over 200 years, phonetics is not essentially seen as connected to the dictionary. The same is true of modern spelling-to-phoneme algorithms. The study of prosody began as an extension of phonetics, and was unconnected to grammar or other aspects of the text. The study of natural texts, or ‘discourse

analysis', has been approached from several angles, mainly by social scientists unfamiliar with the detail of linguistic structure.

In this way, formal linguistics is divided into a number of unconnected components, none of which is designed to fit any of the others. When we examine an individual component, we may find its very manifestation surprisingly elusive. For example, we might say that the word information belongs to the lexicon of English. But where and in what form does this lexicon exist? Is it the Oxford Dictionary or some other dictionary, or is it a list of words in some English-speakers head? Where do conventional rules of grammar exist, other than in the head of the linguist who studies them? Many textbooks contain lists of English phonemes, but phonemes are assumed to exist not in textbooks but in the pronunciation of lexical items. In the case of MALEX, such entities are explicitly located. The lexicon is a table containing the set of different lexical items found in our corpus; while phonemes are members of the set of sound segment types used to pronounce the words in the lexicon.

Computer scientists inevitably inherited this fragmented approach to the analysis of texts, and the connections which should be explicit may not be apparent at all. For example, a tagger based on n-grams may not appear to have any connection with a stemmer which just chops up individual words. This cannot by any stretch of the imagination constitute the best of all possible approaches to the study of a language, and may be a contributory factor to the generally pessimistic assessment of the contribution of NLP to information retrieval (e.g. Strzalkowski 1999). The computational study of English has long since developed a momentum of its own, and it would be unrealistic to expect any substantial change in the foreseeable future. Nevertheless we shall here draw attention to some of the problems created by a fragmented approach.

In the first place, lurking behind the 'NLP approach' (e.g. Sparck Jones 1999:2–6) is a particular set of models of transformational-generative syntax that happened to be dominant in the last few decades. While these models have the great attraction of lending themselves to implementation on computer, and may be linked in some way to semantics and phonology, they are not in general linked to text, and have nothing much to do with real data at all. It is unrealistic to expect models designed for simple or invented sentences such as "Germany invades Poland" or "The man hit the ball", to lead to the meaning of unrestricted natural texts.

Secondly, to take the specific case of stemming, it is not always clear how the results of stemming are expected to relate to text meaning. For example, Sparck Jones (1999:6) reduces words to substrings of their spellings, such as centr* or redevelop*; but it is not clear why such substrings should correspond to units of meaning or behave consistently in texts. Grouping historically related forms, e.g. magnesia and magnet under magnes (reported by Jacquemin and Tzoukermann 1999:26) is almost bound to lead to confusion in view of semantic change, as indeed the authors point out (p30). Stemming surely needs to be conducted in accordance with morphological principles either known in advance (as is usually the case for English) or discovered in the data (as in the case of MALEX). The

value returned by the stemmer needs to have some status with respect to the lexicon if it is to link up effectively with meaning.

In constructing our database alongside the corpus (or 'archive'), we have from the beginning not only preserved relevant links, but also maintained a rigorous three-way distinction between properties of the text, properties of the lexicon, and properties of the grammar. These are properties at three levels of increasing generality, and can be illustrated by reference to the word. Punctuation marks are a good example of what we mean by properties of the text. An individual word token occurring in a text may be followed by a comma, but this comma would not normally be regarded as part of the word type. At the middle level, the properties of word types include their morphological derivation; for example, *bernilai* is divided into the prefix *ber-* and the stem *nilai* 'value', and this analysis applies equally to all its tokens occurring in texts. At a higher level, *bernilai* is one of a set of words formed from nouns by the addition of *ber-*; and these words can be followed by an adjective, e.g. *bernilai tinggi* 'of high value'. When we generalise about sets of words in this way we are dealing with properties of the grammar. Properties of the text are appropriately represented by annotations inserted into the text itself, while properties of word types are represented in a lexicon linked to the text, and properties of the grammar constitute the rules of a morphological stemmer or syntactic parser.

The rigorous separation of text, lexicon and grammar enables us to design closely related tools which draw upon the same expert knowledge of the language. The morphological analysis performed by the stemmer is used to predict entries for the pronouncing dictionary. The same tag set is used to label words at different stages of their morphological analysis and to label words in texts. This contrasts with what seems to be the normal practice in computational linguistics, in which stand-alone tools are designed to operate on text.

A stemmer strips off affixes and perhaps undoes other phonological processes (such as reduplication in Malay) in order to get to the basic form of the words of the text. This represents a 'common sense' approach to text processing. However, since the outcome of stemming is always the same for a given word type, the most efficient method is to stem the word types once in the lexicon, and then look up the words of a text to find the analysis. Stemmers written by computer scientists, including a stemmer written for Malay (Ahmad et al. 1996), tend to operate on individual word tokens in the text. Jacquemin and Tzoukermann (1999:39) contrast the 'list' and 'dynamic' approaches to morphological analysis (i.e. using a lexicon and operating directly on text respectively), expressing a preference for the latter on the assumption that the list is closed and based on existing dictionaries. The only reason for operating on word tokens is that a comprehensive list of word types is not available, and restricting a search to the coverage of a conventional dictionary (which is bound to be out of date with respect to a contemporary text) is an idea with little to recommend it. Our stemmer draws on known solutions for the vast majority of words which are already in the lexicon, and operates in a dynamic fashion when new word types are encountered in corpus texts.

There are other reasons for using a lexicon. Stemmers often use lists of stop words which can be excluded from the search, but an annotated lexicon can provide a rational basis for constructing the stop list in the first place, especially if different lists are used for different kinds of text. For example, a pronoun like we is a good candidate for a stop list, but we have found *kita* '(inclusive) we' to be a key word in political speeches, alongside *kerajaan* 'government' and *rakyat* 'people'.

A second reason is to improve the stemmer itself. As we have collected more and more successful stemming in the lexicon, we have been able to refine the rules to deal with rarer cases. For example, the final string *-nya* is a frequently a clitic pronoun 'his, her, their', as in *bukunya* 'her book'. Nouns can be reduplicated to indicate an indefinite plural, e.g. *buku-buku* 'books'. By removing the *-nya* from *buku-bukunya* and undoing the reduplication we get back to *buku*. This is the normal case. However, in *tanya-tanya* 'questions', *-nya* is an arbitrary string. In order to avoid generating **tanya-tan*, we therefore have to make a preliminary check for reduplication before removing *-nya*. Some stems are modified when affixes are added, and this causes difficulties for the stemmer. For example, *tulis* 'write' loses its /*t*/ when the agentive prefix *pen-* is added, so that *pen+tulis* becomes *penulis* 'writer'. Unless one knows Malay, it is not at all obvious whether *penulis* derives from **nulis* or *tulis*. However, since we have a large number of examples of this type, we can at least formulate rules which will correctly identify the simplex form of new word types more often than not.

A typical modern Malay text contains a significant number of English words, and these are looked up in an English lexicon. If by mistake English words are sent to the Malay stemmer, the results tend to be nonsensical, e.g. an initial string *di-* will be identified as the passive prefix *di-*, and so *division* will be analysed as the passive form of *vision*. There always remains a hard rump of new words which we cannot process automatically. These include members of lemmas which we have not previously encountered, and the current bottleneck in the development of the lexicon is in handling new lemmas. Part of the problem of identifying new lemmas is that so many of them are borrowed from English, sometimes using new patterns of word formation, e.g. English words ending *-tion* are given the new ending *-asi*, thus *globalisasi* 'globalisation'. English has provided a number of new prefixes, including *pro-* and *anti-*, while *pasca-* is modelled on English *post-*. The set of affixes we have identified in contemporary texts is rather different from those given in conventional descriptions, such as Abdullah (1974) and Sneddon (1996), excellent though these are for their own purposes.

Although our work so far has concentrated on the exhaustive analysis of corpus texts, our model could be modified to operate on unrestricted text at a high level of accuracy. When we process new text, the vast majority of words have been encountered before and are already entered into the lexicon, and for these our analysis is almost 100% accurate. The stemmer handles most of the remaining words by locating the stem from which a new word is derived and for these the accuracy is not far short of 100%. In dealing with new lemmas we have the means to obtain more accurate results than a stand-alone stemmer could achieve. Our

main lexicon table contains over 30,000 entries each marked for grammatical class, the name of the lemma to which it belongs (Knowles and Zuraidah Mohd Don 2006), and the stem from which the word is immediately derived. Since any word can be looked up successively in the table until the simplex form is reached, the table contains thousands of complete derivational histories. Since we know the class of each word, we have information for thousands of words on how grammatical class is affected by morphological derivation, which means that the grammatical class of new derived words can be predicted with a high degree of accuracy. MALEX can thus do what a stemmer does, and a lot more besides, and lemmatisation, morphological analysis and stemming all turn out to be different aspects of exactly the same lexical problem.

A tagger is another stand-alone tool designed to operate directly on text. Modern tagsets are based on the traditional parts of speech, which were used to label headwords in dictionaries; but the original classes have been extended to mark detailed properties of the text to the extent that grammatical tags now have to be regarded as properties of the text rather than of the lexicon. A parser operates on a previously tagged text, and is based on equally traditional notions of how words fit together to form phrases and how phrases fit together to form sentences. The essential relationship between tagging and parsing is that the tagger provides the grammatical information needed by the parser to analyse sentences; but beyond that there is no necessary connection between any particular tagger and any particular parser.

A tagger has to have access to a lexicon or some equivalent look-up procedure to identify the as an article and of as a preposition. The problem is that some words can belong to more than one class, e.g. book can occur as a noun (e.g. my book) or as a verb (e.g. book a room). This is reflected in the practice in grammatical tagging, e.g. the CLAWS tagger, whereby all possible tags are entered in the lexicon with their respective frequencies, so that book occurs x% as a verb and y% as a noun. An important task for the tagger is then to 'disambiguate' words like book and decide their part of speech in texts. The emphasis here is on getting the right tag in the text. The lexical entry is of no interest in itself, but merely serves as an ad hoc device to hold information on the ambiguities. But this is surely the wrong way round. In the case of Malay, this 'ambiguity' is central to the way large numbers of words are used, and this fact has to be recognised in the design of the tag set and in the mode of operation of the parser.

To 'disambiguate a tag' one has to examine the context: after my, book must be a noun, and after will it is more likely to be a verb. But this examination of the context is also what the parser does. At a later stage of processing, the parser examines the noun book after the possessive my and groups the two words together as a noun phrase. In this way tagger and parser overlap in examining book in its context. But they are doing the same job. When we encounter book in the lexicon, what we can say about it is that it will function as a noun or a verb in the text, and in this respect it patterns like many, many other English words. That is to say, 'noun-or-verb' is an important grammatical class in English. This may not correspond to a conventional

part of speech, but we can argue that the failure to recognise such classes is a serious shortcoming of the conventional grammatical classification of words. There is, for example, another class of words including before, after, and since which can occur as prepositions, adverbs or subordinating conjunctions. These facts are surely properties of the grammar of English. Although such examples are relatively rare in English, comparable examples are found in abundance in Malay. In fact, one of the salient characteristics of Malay grammar is what we have called elsewhere (Knowles and Zuraidah Mohd Don 2003) 'syntactic drift', the ability of words to drift from one syntactic context to another.

If we examine conventional parts of speech closely, we actually find that they implicitly incorporate notions of syntactic drift. For example, we might think of an English adjective as a noun modifier, e.g. the old man, but a typical adjective can also follow BE as a complement, as in the man is old. The corresponding class in Malay can function as a noun modifier, thus *orang tua itu*, or as a predicator, i.e. without a copula corresponding to BE, in the form *orang itu tua*. But a typical Malay 'adjective', in addition to modifying a noun can also modify a verb, and thus function as what in English grammar is called an 'adverb'. Although we might translate adjective into Malay as '*kata sifat*', it is clear that adjectives and *kata sifat* do not occur in the same syntactic contexts. If the familiar class of adjectives is to be associated with a set of syntactic contexts, then there is no reason not to set up a super class such as noun-or-verb which is found in the union of noun contexts and verb contexts. If we think of a grammatical tag not as a label for a single class but as the identifier of a set of syntactic contexts, then instead of having several tags for a word in the lexicon, we need just a single tag. A word like book can be classed as noun-or-verb in the lexicon, and when the parser encounters my book in a text, it can simultaneously identify book as a noun and my book as a noun phrase. This is surely close to what a human being does when reading a text: we do not first mentally tag a text and then parse it.

It has to be recognised that English taggers have achieved a remarkably high success rate, and as far as processing English texts is concerned, it might be argued that there is simply no point in worrying about the niceties of text and lexicon, or the overlap between tagging and parsing. Nevertheless, when dealing with a language like Malay, when we do not have an established tradition to draw upon, these are precisely the things with regard to which logical rigour is essential. To analyse the syntax of Malay texts we need to find out from the lexicon what syntactic environments a word can occur in, and from the parser which environment is involved in any particular case. The concept of a text-level grammatical tag is redundant, and is a candidate for Occam's razor.

Another area in which the 'common sense' approach does not always handle data at the appropriate level of generality is the area of meaning. According to 'common sense', meaning is an attribute of words, and therefore something that belongs in the lexicon. Dictionaries of course contain word meanings, and an early approach to semantics, often referred to as 'componential analysis' (Katz and Fodor 1963) attempted to identify the components of meaning that make up the

meanings of words. Taking this approach, we can add a semantic tag to our lexicon entries (making use of work on semantic tagging currently underway for English at Lancaster University), and expect to gain thereby some useful information about meaning in Malay texts.

Behind the grouping of words under headwords in the dictionary is the assumption that meaning relations are to some extent a property of the grammar. Someone who knows what an egg is can infer the meaning of eggs, and similarly the meaning of walked is inferable from that of walk. This is of course the commonsense assumption on which stemming is based. How reliable the grammar actually is depends on the type of language, and here western practice follows the precedent of Latin, which is a language (at least as it is taught) in which meaning relations can be inferred with total confidence. In the case of languages with large complex lemmas (or sets of morphologically related words) stems have to be reasonably consistent in meaning and the effect of morphological processes has to be predictable if confusion is to be avoided. Arabic is a good example, for a general area of meaning is associated with a triconsonantal root shared by large numbers of words, e.g. dozens of words including *kutub* 'books', *maktab* 'office' and *kataba* 'he wrote' all share the root *KTB*, which is associated with writing. Maučec et al. (2004) lemmatise the words of Slovenian texts in order to make an 'automatic separation of grammatical and semantic information encoded in text'. This may work for a highly inflected language like Slovenian, but it is a matter of good fortune rather than of theoretical vision, and not something to be taken for granted in advance.

By contrast, a language with small lemmas possibly has more freedom for individual words to develop idiosyncratic meanings. For example, in Malay, which has on average about 3 words per lemma, meaning relationships may be unpredictable. *Mata-mata* 'policeman' is stemmed to *mata* 'eye', while *talian*, which has to do with 'connection', is stemmed to *tali* 'rope'. Paice (1996) draws attention to the dangers of under- and over-stemming, but when the grammar does not help, it is difficult to know how far the stemmer should go. Here the lexicon is surely indispensible. In the normal case, a reduplicated form such as *buku-buku* can be stemmed to *buku* 'book' and the meaning inferred using the grammar, but *mata-mata* has to be linked directly to meaning and the inference from *mata* blocked.

A stemmer might already have gone too far when the text is divided into orthographic words. Word combinations as a whole can link to meaning. For example, *sakit* 'sick' combines with *hati* 'liver' to mean not 'suffering from cirrhosis' but 'jealous'. Sequences of words can combine to form opaque compounds, such as *kaki* 'foot' + *tangan* 'hand' = *kakitangan* 'employee', or *kereta* 'cart' + *api* 'fire' = *keretapi* 'train'. As shown in these examples, compounds can be written solid in modern orthography, but they are often found in texts written as separate words. In a language like Malay (and for that matter English) there is no convenient 1:1 relationship between the units identified by a stemmer and units of meaning in the text. When the stemming has been carried out appropriately, we still need to know at what level words are linked to meanings in texts.

A second problem in handling meaning is the relationship between meaning in the lexicon and in the text. Semantic networks set up at the lexical level can look very convincing until we try to link them up to text. For example, the definition of bitch as 'female dog' is unproblematical until we look for it in texts, where it is hardly ever used in that sense except perhaps in the context of dog breeding. The problems of relating lexical meaning to text are parallel to those of lexical grammatical class and syntax. A conventional dictionary will list different possible meanings of a word, and it might seem that all we need to do is to design a disambiguation procedure so that we know which meaning is appropriate in any given case in the text. However, the concept of discrete meanings at text level is logically equivalent to grammatical tags at text level, and redundant for the same reason. The problem is related to parsing but far more complex, because instead of dealing with combinations of word classes as a whole, we have co-occurrences of individual words. It is worthy of note that when we tag or parse my book, we also decide on the meaning of book. Again there are many examples of this kind in the Malay data; for example *nanti* as a main verb means 'wait', but before the main clause it is interpreted as an adverbial 'later'. Tagging, parsing and semantic disambiguation are not as logically discrete as might at first appear. At the present stage, we have not yet included semantics in our model, but the signs indicate that the traditional separation of semantics and syntax would cause problems in tackling the structure of Malay. It is also clear that a common-sense thesaurus modelled on Roget will not take us very far, and that we will have to extract a thesaurus from the corpus.

To summarise, we are increasingly finding that conventional 'common sense' linguistic practices in general, and procedures designed for English in particular, are inappropriate for the task of processing our Malay corpus texts. We have had to ask some fundamental questions, why we want to tag and parse a text in the first place, or how a stemmer is related to lemmatising and morphological analysis. By addressing such questions, we have done the groundwork essential for an organised approach to accessing meaning in our texts.

5. Conclusion

We have argued in this paper for an integrated approach to the linguistic processing of texts, using closely related tools. The advantages are so obvious that it is difficult to understand how anyone could really prefer a stand-alone tool with no linguistic backup. The answer, of course, is that computer scientists do not have the linguistic expertise to set up an integrated language system, and linguists do not in general see it as their job to create one. For English it would be impossible in practice anyway, because there are so many competing fragments of systems which could never be made to work together. The state of the art in Malay linguistics is such that the best way to proceed is to take an integrated approach from the start.

It is important to question the pessimism expressed in the IR literature about the value of linguistic expertise. If the linguistic ideas employed are imposed from without on to the data, then it is hardly surprising if they do not throw much light on it. The linguistic approach must be data driven, and linguistic ideas which are applied must be relevant to the task. There is no point in adding a tagger or n-gram extractor in the hope that they will somehow improve the performance of a stemmer, for these tools have to be designed to work together to perform a task. If the task is to access the meaning of a text, we have to start with the right set of tools.

Many of the kinds of processing we have discussed in this paper belong to a kind of popular 'common-sense' linguistics. Any literate person knows how to look words up in the dictionary to find the appropriate sense, and anyone who knows English will know about parts of speech and will look up walking under the headword walk. To some extent, therefore, modern stemming, lemmatising and tagging (and to some extent parsing) involves doing with a computer what has long been done in other ways. In these circumstances, the emphasis has been on carrying out the task more effectively, e.g. to automate it rather than do it manually, or make a tagger run faster or increase its success rate, rather than on asking whether the right tasks are being done in the most appropriate way.

This development of computer-based technology has coincided with the rise of English as an international and indeed global language. The result, perhaps understandably, is that procedures developed for English have been imposed on other languages, whether they fit or not. In the case of Asian languages, this can lead to serious distortion. Malay is not like English at all. Of course the two languages share many features by virtue of the fact that they are both natural human languages, and that means that English procedures can be transferred to some limited extent (and a much greater extent if one is prepared to adopt Procrustean methods). In working on Malay, we have in practice to be aware of and sensitive to the important differences. This has led us to raise fundamental questions about what we are trying to do when we process texts. Answering these questions has in turn led to the design of the database, and the development of procedures which relate text to lexicon and grammar in the most appropriate and efficient way.

We would venture to suggest that the approach we have taken to Malay – strictly data-driven and sensitive to the structure of the language – could serve as an appropriate model for work on other Asian languages. Problems of grammatical class recur in Chinese: Mandarin hao translates as 'good' but some scholars call it a 'verb' rather than an 'adjective'. Arabic has a superb grammatical tradition which from a computational point of view offers an interesting alternative to the conventional Western approach. The problems of processing different languages are of course different in detail, but the point can hardly be made too strongly that non-Indo-European languages are not variants of global English: procedures should follow the language, and the language should not be forced to look like English.

Address:
Zuraidah Mohd Don
Faculty of Languages and Linguistics
University of Malaya
Lembah Pantai
50603 Kuala Lumpur
Malaysia
Tel.: +603 7967 3177
E-mail: zuraida@um.edu.my

References

Abdullah Hasan (1974) *The morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ahmad, F., M. Yusoff, and T. M. T. Sembok (1996) "Experiments with a stemming algorithm for Malay Words". *Journal of the American Society of Information Science* 47, 12, 909–18.

Blair, D. C. (1990) *Language and representation in information retrieval*. Amsterdam: Elsevier.

Garside, R. (1987) "The CLAWS word tagging system". In *The computational analysis of English: a corpus-based approach*, 30–41. R. Garside, G. Leech, and G. Sampson, eds. London: Longman.

Jacquemin, C. and E. Tzoukermann (1999) "NLP for term variant extraction: synergy between morphology, lexicon and syntax". In *Natural language information retrieval*, 25–74. T. Strzalkowski, ed. Dordrecht: Kluwer.

Katz, J.J. and J.A. Fodor (1963) "The structure of a semantic theory". *Language* 39, 170–210.

Knowles, G. and Zuraidah Mohd Don (2003) "Tagging a corpus of Malay texts, and coping with 'syntactic drift'". In *Proceedings of the corpus linguistics 2003 conference*. (UCREL Technical Paper, 16.) D. Archer, P. Rayson, A. Wilson, and T. McEnery, eds. Lancaster University: UCREL.

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

Knowles, G. and Zuraidah Mohd Don (2004) "The notion of a 'lemma': headwords, roots and lexical sets". *International Journal of Corpus Linguistics* 9, 1, 69–82.

Maučec, M. S., Z. Kačič, and B. Horvat (2004) "Modelling highly inflected languages". *Information Sciences—Informatics and Computer Science* 166, 1–4, 249–269.

Sneddon, J. (1996) *Indonesian: a comprehensive grammar*. London: Routledge.

Sparck Jones, K. (1999) "What is the role of NLP in text retrieval?". In *Natural language information retrieval*, 1–24. T. Strzalkowski, ed. Dordrecht: Kluwer.

Strzalkowski, T., ed. (1999) *Natural language information retrieval*. Dordrecht: Kluwer.