Tagging a corpus of Malay texts, and coping with 'syntactic drift'

Gerry Knowles, Lancaster University, UK
Zuraidah Mohd Don, University of Malaya.

The structure of Malay presents the corpus linguist with an extremely interesting problem. At high syntactic levels, the language is familiar enough, and one can talk of direct objects in transitive constructions, and even of agentless passives. The dominant sentence order is SVO. Parsing at this level is therefore relatively straightforward. The problem is at lower levels, where Malay patterns quite differently from Indo-European languages. If the linguist tries to process Malay using categories and techniques designed for Indo-European, then it comes across as at best confusing and at worst in a state of chaos. Malay is neither confusing nor in chaos; but it does need to be analysed using techniques which are sensitive to its own patterns.

Conventional tagging is based on the assumption that grammatical class is static, allowing for some 'ambiguities' such as *telephone* as a noun or as a verb. In Malay, grammatical class is dynamic: adjectives can occur as verbs or adverbs and some verbs as adjectives; some verbs can pattern as nouns and others can take on the role of function words. In order to cope with this, we have to make a rigorous distinction between lexical class and the syntactic slots which words fill. Words are given a single class label in the lexicon, and the parser then has to identify cases in which words have drifted away from their default slot.

Tagging and parsing can be carried out for English as separate if independent processes. , for Malay they have to be treated as at least complementary.

1. Introduction

This paper arose out of research supported by Dewan Bahasa dan Pustaka in Kuala Lumpur into the automatic grammatical tagging of a corpus of Malay texts. It became clear from the beginning and for several reasons that tagsets of the kind developed for Indo-European languages would not be suitable for Malay. In the first place, there is no simple way of matching Indo-European grammatical classes with those of Malay. There are words in Malay that seem to correspond to "verbs", but they are also likely to appear as "nouns"; and similarly "adjectives" appear in the guise of "verbs" or even "adverbs". "Nouns" can even appear as "prepositions". In this way the simplest text of Malay gives linguists ample opportunity to disagree over the class of the words. Until we have ascertained the major classes, there is no possibility of adding the kind of refinement required for a modern tagset, which essentially adds more detail to the traditional 'parts-of-speech'.

Tagging a European language also has the advantage that the categories behind the tags are familiar and immediately comprehensible to other linguists. Another researcher can use the tags independently as input to a parser, so that tagging and parsing can in practice be tackled as independent operations. The word classes of Malay, by contrast, are unfamiliar both in their membership and in their mode of operation in texts; as a result they require extensive explanation, and there is no reason for another linguist to accept their validity without proof. In practice, therefore, a tagset has to be justified by a working parser that correctly identifies high-level syntactic structures using the information given in the tags. For Malay, tagging and parsing have to be treated as complementary operations, and they have to be formally integrated into a syntactic model of Malay. Our task in this paper is to show how the slippery parts-of-speech of Malay can be linked to categories which can be recognised by a parser.

2. The Malay corpus

The texts we have studied are taken from the DBP corpus, a collection of Malay texts amounting to some 80 million words, and held by the Dewan Bahasa dan Pustaka in Kuala Lumpur. We have tagged an arbitrary sample of 120,000 words containing literary texts, actually four modern novels. We have made no attempt to take a representative sample, because at this early stage any sample is as good as any other.

To tag the texts, we have constructed an annotated lexicon of some 15,000 words. This contains all the words found in the texts, and also the simpler forms of complex words. The equivalent for English would be to encounter *writings* in a text, but also include *writing* and *write* in the lexicon. In this way, the lexicon includes a full morphological analysis of each word of the text, and provides the information needed for a future detailed morphological study of Malay. For each word the lexicon contains just one grammatical tag. Given the nature of syntactic drift in Malay, corresponding to what in English tagging is known as "ambiguity", it might seem perverse to give just one tag per word. On the other hand, it enforces an integrated approach to the design of the tagger and the parser, and an approach to parsing that obviates the need for multiple tags.

We are developing a rudimentary parser to check the tagging system. Although at this stage it only deals with straight-forward grammatical constructions, it seems to be working not just well, but suspiciously well. In general, Malay uses much less formal grammatical coding than is expected in western languages, and correspondingly relies more heavily on inferencing at the pragmatic level. Our parser makes use of formal grammatical coding, and its success may point to an interesting bias in our literary texts. Malay has long been in contact with languages such as Sanskrit, Arabic and English, and this may have led to a re-modelling of the syntax of literary texts on western lines.

3. 'Parts of Speech'

European 'parts-of-speech' are the accepted point of departure for considering grammatical class in Malay (see Asmah, 1993; Sneddon, 1996). In the initial stages of research, these lead to quick results, but in the longer term they inevitably generate difficulties. It would be easy to jump to the conclusion that the Malay language is in a state of chaos, and that anything goes. In order to identify valid grammatical classes for Malay, it is essential to take a rigorous data-driven approach and identify classes based on the direct evidence of Malay texts. But an exclusively Malay tagset would create its own problems, for it would present Malay as a language *sui generis*, when in reality a comparison of Malay syntax with that of English reveals a number of remarkable similarities. It is important to avoid imposing English categories on to Malay, but it is also important to recognise the similarities that really exist.

The 'parts-of-speech' approach is not the only way of dealing with grammatical class. Arab grammarians devised a three way classification of *ism* 'noun', *fi9l* 'verb' and *harf* 'particle', backed up by a morphology based on *jithr* 'root' and *wazn* 'frame, template'. These make an efficient analysis of a language with a structure very different from that of English. Abdullah (1974) also has three classes in the morphology, which is sufficient for his morphological analysis, but this approach needs to be complemented by a more detailed model to handle the syntax of Malay.

All these classifications need far more detail to handle syntactic relations, and tagsets are much larger than traditional classification systems. Our tagset currently contains 119 tags in 19 different major classes. Some of our class labels look like traditional parts of speech, but the underlying definitions are entirely different.

4. Ambiguity

In designing a tagset for English and other western languages, the parts of speech are used as a primary classification system. However, many words do not fit neatly into just one category, and may belong to more than one part of speech. This lack of fit is handled by an extension of the notion of homonymy. A good example of homonymy is found in the case of the two unrelated words of the form *can*, one a

modal verb and the other a noun. By extension, given a word like *telephone*, we can say that there are two homonyms, i.e. a noun *telephone* and a verb *telephone*. An alternative is to say that *telephone* is used sometimes as a noun and sometimes as a verb. In either case, the homonymy is treated as a source of "ambiguity" which has to be resolved by the grammatical tagger.

In order to resolve the ambiguity, we have to examine the syntax. After *the*, *telephone* is likely to be a noun, and after *will* it is more likely to be a verb. In other words we use parsing to ascertain the tag of a word in context. At a later stage, a parser uses the tag to make parse. At the very least, there is some overlap here between tagging and parsing. An alternative approach is to make a clear distinction between properties of the lexicon and properties of the text. When we see a word like *telephone* out of context, we know it usually fits contexts where either a noun or a verb is expected, and which one it is is the property of a particular text. This is not an isolated case, and English has had large numbers of similar examples for hundreds of years. Another set contains words like *after*, which we might think of as essentially prepositions, but which can also be used as adverbs or subordinating conjunctions.

This discussion of English is relevant because Malay operates in a very similar way, only much more so. The language simply does not work in the manner presupposed by the part-of-speech classification. It is preferable to study words as entities in their own right, independently of preconceived grammatical classes, and to allow the natural grammatical classes of the language to emerge.

5. Syntactic drift in Bahasa Melayu

The class of many words in European languages is made unambiguous by their morphology, and to a lesser extent this is true of Malay. However, in view of the lack of any inflectional morphology, Malay has a large number of simplex forms which belong to no clearly defined class, and appear to 'drift' from one class into another. For example, *masuk* is the normal word for 'enter', which makes it a kind of verb; but it is used in such a way on buildings and in carparks that it could also be taken to be a noun 'entrance'. This indeterminacy is also found in words in syntactic contexts, and has to be recognised as an important characteristic of the language. What we here call syntactic drift is so widespread and indeed systematic in Malay that the homonymy explanation is unrealistic. We need a grammatical model that allows us to trace words as they cross from one class to another. In the case of *telephone*, there are many similar nouns that also occur as verbs, to the extent that English can be said to have a class of "noun-or-verb", and the tagger and parser together have to decide on a more precise class in any given case. Consider now the behaviour of what in Malay might be called an "adjective". In *buah keras itu* 'that hard fruit', *keras* behaves as an attributive "adjective". In *buah itu sudah keras sekarang*, 'that fruit has become hard now', where it follows the time marker *sudah*, it is a predicator and behaves more like a "verb". In *kami bekerja keras* 'we work hard', it is more like an "adverb".

We have invented these examples to illustrate different uses of (what we now consider) the same word. For automatic processing we enter a single form in the lexicon with a single tag, in this case kata sifat "adjective". The parser then has to recognise "adjectives" in different syntactic positions, and identify their differing syntactic roles. For example, in *bulan samar* 'dim moon', the "adjective" *samar* behaves as expected and follows the noun as a modifier. In *Seman terlalu gembira* 'Seman was extremely happy', the "adjective" *gembira* follows the intensifier *terlalu*. The English translation makes it still look like an adjective, but the structure is one of a large set relating to the verbal group, and our parser treats *gembira* as a kind of "verb". In *ibu bapanya membangkang keras* 'his mother and father disagree strongly' the parser treats the "adjective" *keras* as an "adverb" after the "verb" *membangkang* .

To a large extent, we can identify sets of words that drift in the same way, and to that extent drift is a property of the grammar. In the parser, rules normally refer to tags, and words with the same tag function in the same way in different syntactic contexts. There remains a residue of words that drift in an entirely idiosyncratic manner. This situation can be illustrated from examples in English. The drift of *telephone* from noun to verb applies equally to *e-mail* and even *table* and *chair*. This is part of the grammar of English. Idiosyncratic drift is illustrated by a work like *provided*, which is a past participle that can be used as a subordinating conjunction. This is a property specific to the word *provided*, and not a general property of past participles. A Malay example is *menarik* which is tagged in the lexicon as a "verb" and which can occur in texts as an "adjective". (It is part of the grammar of Malay that "adjectives" can drift to "verbs", but the reverse is not generally true.) There are a number of important words referring to past time – including *sudah*, *lepas* and *lalu* – which are important for an

understanding of Malay syntax, and which drift in highly idiosyncratic ways. To handle words like this in the parser, we need rules that operate not on tags, but on individual word forms.

When individual words are referred to by syntactic rules, it is of course irrelevant what tag we give them initially. For example, *o'clock* is only found in expressions of the time of day, and no other word patterns anything like it. However we choose to tag *o'clock*, it is only found after a numeral, as in *three o'clock*. This is a very useful fact, because it means we can sometimes tag a word according to its normal use, and disregard a specialised use. For example, the Malay expression corresponding to *o'clock* is a "verb" *pukul* normally meaning 'to hit, strike (e.g. a gong)'. To handle an expression such as *pukul tiga* 'three o'clock', the parser has to test for the specific word *pukul* before a numeral, and so the fact that *pukul* is tagged in the lexicon as a "verb" causes no problem at all.

Behind the conventional tagging of European languages is the assumption that words belong in principle to one class. In practice they do not, and "ambiguity" is seen as a problem of exceptions that we have to solve. In view of syntactic drift in Malay, this is not a good starting point. Grammatical class in Malay is not a static property, as we assume for European languages, but a property that is by nature variable.

6. Lexical tags and syntactic slots

We have so far used scare quotes to refer to classes we might want to think of as "adjectives" or "verbs". This usage masks a logical inconsistency that is commonly found in linguistic discussion. When we say that *big* is an "adjective", we are referring to its lexical class, and we can predict morphologically related forms such as *bigger* and *biggest*. We expect to find it labelled as an "adjective" in the dictionary. When, on the other hand, we say that "adjectives" come before "nouns" in English, we are referring to positions in syntactic structure. It so happens that Indo-European languages are in general organised in such a way that there is a highly predictable relationship between lexical classes and syntactic positions, with the result that ambiguity in the use of terms such as "adjective" (and indeed labelling words as adjectives in the dictionary) does not lead to confusion.

For Malay, these distinctions are essential and must be maintained rigorously and consistently. First we can separate out the lemma, the concept behind the dictionary headword. It is usually defined as a set of words related by inflectional morphology and consequently of the same "part of speech", so that dictionary headwords can be given a single part-of-speech field. For example, {sing, sings, sang, sung, singing} are all grouped under the headword SING and classed as a verb. This is relatively unproblematical for languages with an inflectional morphology, as long as one does not ask about the status of words such as *song* and *songs*.

For a language like Malay, with only derivational morphology, this is not a sensible thing to do at all. For example, under the headword *besar* 'big' in a dictionary we might find not only *terbesar* 'biggest' but also *membesarkan* 'enlarge' and *kebesaran* 'size'. Lemmatising is an important step in processing a Malay text, but it does not and cannot have a fixed link to parts of speech. The lemma BACA 'read', for instance, has to be defined in such a way as to include "nouns" such as *pembaca* 'reader' and *bacaan* 'reading'. Since members of Malay lemmas belong to different grammatical classes, it is meaningless to ask the part of speech of a Malay lemma. Malay lemmas can be classified, but on a quite different basis. Malay dictionary entries do not have a part-of-speech field, and the relationship between words in dictionaries and words in texts is different in Malay and Indo-European languages.

The distinction between adjective as a syntactic position and a lexical class is essentially that between a container and its contents. In algebra we distinguish the name of a variable from its value, in programming we distinguish a memory location from the information stored in it, and in social life we distinguish a house from its inhabitants. A distinction is sometimes made in linguistics between "slots" and "fillers", but in grammatical tagging it is not always made clear whether the tags are slots or fillers. In fact we logically need two sets of terms, one for the slots and the other for the fillers. In our work on Malay, we use the term "tag" to label a lexical class, and "slot" to refer to a position in syntactic structure. We maintain the distinction consistently by giving lexical classes Malay labels, and syntactic slots and constructions English labels.

A given slot can be filled by words of different lexical classes. A head noun slot, for example, can be followed by a modifier slot, and the modifier can be a kata sifat 'adjective', e.g. *pintu hijau* 'green door', or a kata nama 'noun', e.g. *pintu rumah* 'house door'. More surprisingly, it can be filled by the simplex form of a kata kerja 'verb', e.g. *pintu masuk* 'entrance door'. By distinguishing lexical classes and slots, we arrive at a straight-forward statement of the distribution of simplex kata kerja forms, without having to resort to a metaphysical process by which a "verb" transforms itself into a "noun" or "adjective". Nor is there anything remotely ambiguous in a phrase such as *pintu masuk*. (In a surreal world, *pintu masuk* could conceivably be taken to mean 'a door enters'; but such bizarre possibilities can safely be ignored in the practical processing of normal texts.)

Illustration: kata sifat

Kata sifat or "adjectives" are a relatively straight-forward class by the standards of Malay, but nevertheless appear in unexpected places, and drift into "verbs" and "adverbs". They can even follow certain verbs as "adjectives" without becoming "adverbs". In the phrase *berbaju merah* 'wear a red shirt', the "verb" *berbaju* 'wear a shirt' is formed from the "noun" *baju* 'shirt', and this noun can still be followed by an "adjective" such as *merah* 'red'. This situation is actually an artefact created by the Malay writing system in writing *ber-* solid with *baju*: the structure is not strictly ((berbaju)(merah)) but (ber(baju merah)).

Of the 120,000 words of text which we have tagged so far, we have made a more detailed study of one corpus text of 21635 words, using a prototype parser. Of these words, 1610 (7.4%) were tagged as kata sifat. These appeared in a number of different positions:

1	immediately after a noun	541
2	as a predicator	505
3	after the verb	215
4	berbaju merah type	25
5	after dengan	9
6	residue	315

Table 1: Distribution of kata sifat

The most frequent position is after a noun, where the kata sifat plays the same role as an attributive adjective in English. The second most frequent group corresponds to English predicative adjectives, but in Malay they are much more like "verbs" and this is one of several situations in which it is difficult to make a clear distinction between "adjective" and "verb". Our parser analyses them successfully by treating them as "verbs", the main difference being that they do not take the full set of verbal markers.

A surprisingly large proportion of kata sifat immediately follow the verb. From an English point of view the "adjective" functions like an "adverb", but one can equally argue that the "verbs" are really like "nouns" in being followed by "adjectives". There is unlimited scope here for confused argument over the presence or absence of "adverbs" in Malay. Whatever we want to call them, there is a slot immediately after the verb which is commonly filled by a kata sifat. The point is that both head "nouns" and "verbs" can be followed by modifers, and kata sifat can fill the modifier slot. Originally our figures for this structure included formations of the *berbaju merah* type, but these have now been separated out in Table 1 above. There are just 9 cases in which a kata sifat follows *dengan* to form an adverbial expression. There remains a residue of cases which our parser is unable to process satisfactorily at this stage, or for which we are unable to extract meaningful information from our database.

We have so far related a lexical class to syntax. We can also look at these patterns from the point of view of the syntactic slot. The most common nominal modifier is actually another kata nama or "noun", as in *pintu rumah* 'house door', and these are more common than kata sifat. It is often claimed in grammars of Malay that adjectives are for preference linked to nouns not directly, but by using the relative particle *yang*, so that *pintu yang hijau* 'door that is green' is more natural than the simple *pintu*

hijau ‘green door’. This is the reverse of what we found in our admittedly very small sample, as seen in Table 2.

kata nama + kata sifat	388
yang + kata sifat	114

Table 2: “Adjectival” expressions

Nor does *yang* plus kata sifat seem to be a particularly common use of the yang-clause. Of the 655 *yang*-clauses counted, only 114 had *yang* immediately followed by a kata sifat, and only another seven were formed with the negative particle *tidak* ‘not’, thus *yang*+*tidak*+kata sifat. Note that after *yang*, the kata sifat is actually in the predicator slot. If it were true that kata sifat are normally used with *yang*, the most frequent use of kata sifat overall would be as predicators rather than as nominal modifiers.

While there is much debate over whether Malay has adverbs or not, it is quite clear that, like any other language, it has adverbial expressions. Descriptions of Malay point to an unusual construction in which *dengan* ‘with’ is followed by a kata sifat, thus *dengan betul* ‘with correct, correctly’, and to noun phrases using the noun *secara* ‘manner’, thus *secara betul* ‘in a correct manner, correctly’. These two types proved to be rare in our data, occurring only 9 and 4 times respectively, as shown in Table 3.

(verb +) kata sifat only	215
(verb+) *dengan* + kata sifat	9
(verb +) *secara* + kata sifat	4

Table 3: “Adverbial” expressions

The dominant type of manner adverbial was formed by using a kata sifat as a verbal modifier.

It would not be unreasonable to regard the most frequent use, namely as a nominal modifier, as the prototype, so that *pintu hijau* represents the prototypical use of a kata sifat. The nominal modifier slot can thus be regarded as the ‘home slot’ for kata sifat. But kata sifat drift from the home slot into the predicator slot, the verbal modifier slot, and can follow *dengan* to form an adverbial expression. As this example shows, drift is not random, but follows clearly defined directions. Kata sifat contrast in the direction of drift, for example, with kata kerja, or “verbs”. A word such as *berikut* ‘follow’ is a good example of a “verb”, and its home slot is the predicator slot. However, in an expression such as *tren berikut* ‘following train’ it drifts into the nominal modifier slot.

7. Conclusion

Malay is of considerable interest to the corpus linguist because it defies analysis using assumptions taken for granted in work on Indo-European languages. Conventional notions of “parts of speech”, in particular, really belong to Indo-European and can be highly confusing when imposed on Asian languages such as Malay. There are two basic approaches. The first, and alas probably the more common, is to proceed in a Procrustean manner, forcing the words of a Malay text into categories designed for European languages, and ignoring the obvious fact that they do not fit. The alternative is a genuine data-driven approach, identifying the categories that actually appear in the texts. This second approach recognises that grammatical class is organised on different bases in Malay and Indo-European languages, and that grammatical class sequences found in texts are consequently very different. To some extent, Malay grammatical classes are a language-specific property of Malay. Ironically, this point was well known to grammarians earlier in the twentieth century. Lewis (1947: xvii) was quite explicit: ‘Malay words change their function according to context. Be prepared for this, and do not attempt to force the language into a set mould. It will escape’. On this see also the valuable review by Azhar (1988:46-53). Perhaps there is an important insight here that has been lost, and which needs to be rediscovered.

The slots which grammatical classes fill are by contrast not generally specific to Malay at all, and are in many cases recognisably similar to those of Indo-European and other languages, e.g. noun + adjective,

negative particle + verb, number + noun. As we move to higher syntactic levels, Malay structures increasingly correspond to those found in many other languages. General linguistic terminology can safely be used, e.g. 'noun phrase', 'prepositional phrase', 'agentless passive', or 'direct object of transitive verb'. If we analyse invented sentences such as *saya membaca buku itu* 'I read that book', the high level syntax and the lack of low-level detail give the false impression that the structure of Malay is remarkably similar to English. The syntax of real corpus texts, with their wealth of low-level syntactic detail, offers a real challenge to the corpus linguist, a challenge that any linguist who tries to impose low-level English categories is bound to fail.

We are still at a very early stage in this research, and our target for 2003 is to complete the processing of our first million words. Our understanding of the basic categories is still growing, and our tagset is still evolving. However, it is already clear that the foundation of a well-defined set of categories for Malay is the clear logical distinction between lemmas, lexical classes and syntactic slots, and the integration of tagger and parser. Setting up word classes and labelling them with tags is complementary to using those tags in a working parser.

References

Abdullah Hasan 1974 *The Morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka

Asmah Haji Omar 1993 *Nahu Melayu Mutakhir*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Azhar M. Simin 1988 *Discourse-Syntax of "Yang" in Malay (Bahasa Malaysia)*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Lewis, M.B. 1947 *Teach Yourself Malay*. London: English Universities Press.

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