**Computational analysis of affixed words in Malay language**

Bali RANAIVO-MALANÇON
School of Computer Sciences
Universiti Sains Malaysia
ranaivo@cs.usm.my

A full-system Malay morphological analyser must contain modules that can analyse affixed words, compound words, and reduplicated words. In this paper, we propose an algorithm to analyse automatically Malay affixed words as they appear in a written text by applying successively segmentation, morphographemic, morphotactic, and interpretation rules.

To get an accurate analyser, we have classified Malay affixed words into two groups. The first group contains affixed words that cannot be analysed automatically. They are lexicalised affixed words (e.g. *berapa*, *mengapa*), idiosyncrasy (e.g. *bekerja*, *penglipur*), and infixed words because infixation is not productive any more in Malay language. Affixed words that belong to the first group will be filtered, treated and take off before the segmentation phase. The second group contain affixed words that can be segmented according to segmentation and morphographemic rules.

Affixes are classified following their origin (native vs. borrowed), their position in relation to the base (prefix, suffix, circumfix), and the spelling variations they may create when they are added to a base. At the moment, the analyser can process affixed words containing only native prefixes, suffixes and circumfixes. In Malay language, only five native prefixes (*ber-*, *per-*, *ter-*, *me-*, *pe-*) may create some modification (deletion, insertion, assimilation) at the point of contact with the base. Thus the main problem in segmenting prefixed words in Malay is to determine the end of the prefix and the beginning of the base. To solve the problem, we state that each affix in Malay has only one form (i.e. no allomorph), and if there is any morphographemic alteration, it belongs to the base. Following this idea, we built segmentation rules. They are applied iteratively from the left-side of the word to search prefixes and from the right-side to find suffixes. The iterative search avoids creating specific rules for combined affixes like '*memper-*' and '*ke-…-an*'.

To generate prefixed words in Malay, only one morphographemic rule can be applied. But if one wants to recognise the structure of a given prefixed word, different morphographemic rules may be applied. The segmentation of the word *mengecam* gives 'me+ngecam'. Three morphographemic rules can be applied and give at the end four possible forms of the base: 'me+NGECAM', 'me+KECAM', 'me+ECAM', and 'me+CAM'. To look up in a list of Malay roots may help to choose between the four segmentations. But in this case, the ambiguity still remains because the two root-words *kecam* and *cam* exist, and both can generate the form *mengecam*.

After being divided into its components and rebuilt into the form of the base, the results are checked by the morphotactic rules. Impossible sequence**s** of prefixes, suffixes and impossible combination**s** of prefix and suffix are eliminated. The possible circumfixes (prefix-suffix) are grouped. At this level, the affixed word is ready to be interpreted.

In Malay, affixes are not attached to any specific category as in English where the suffix '*-ation*' is attached to verbs and produced nouns. The morphological interpretation of affixed words in Malay is no trivial matter. At the moment, the analyser can calculate the part-of-speech, subcategorisation of verbs (transitive/intransitive; active/passive) from the morphological structure, and this without using any linguistic information about the base.