# THE MORPHOLOGICAL ANALYSIS OF BAHASA MALAYSIA

Chang May See

School of Mathematical Sciences
Universiti Sains Malaysia
Penang
Malaysia

## Abstract

This paper describes a model for the automated morphological analysis of Bahasa Malaysia (the Malay language) via the ATEF system, a component of the mechanical translation system known as ARIANE, which was developed by G.E.T.A. at Grenoble. This model serves two purposes, that is, to test the capability of handling Bahasa Malaysia morphological analysis using ATEF and also to provide a first working model.

This grammar covers the three main morphological processes in Bahasa Malaysia analysis, that is, affixation, reduplication and compounding.

Reduplication is the process whereby a base or some part of the base is repeated. There are three types of reduplication - proper, rhyming and chiming. Reduplication of nouns generally gives a semantic category of heterogeneity or indefinite plural while reduplication of verbals result in one of the following semantical features: repetition, continuity, habituality, intensity, extensiveness and resemblance.

Compound forms are constructions that have two or three free forms as their constituents and each of the constituent forms may either be a rootform or derived form.

Affixation is a morphological process whereby a base may be extended by one or more affixes. Affixes may be classified as prefixes, suffixes, infixes and circumfixes. Multiple affixation is also not uncommon in Bahasa Malaysia though no construction exceeds three layers of affixation. Several features are obtained through affixation. On affixation, morphographemic changes may occur depending on the initial segment of the rootform, word classes of the derived words are set and also semantical features are set. The setting of semantical features may be further complicated as a result of multiple affixation.

Unlike affixation, the handling of reduplication and compounding do not present much of a problem for ATEF and is quite straightforward. Affixation is a more complicated process but also a more important process.

A simple finite state diagram is used to depict the basic overall structure for the handling of multiple affixation as a more detailed finite state diagram is not justifiable.

Morphographemic changes occur mainly with the prefixes pe_N and me_N in which different allomorphs are used depending on the initial segments of the derived word. On deletion of these allomorphs to obtain the resultant form, segments may have to be added to the resultant form, the form remains unchanged or substituted. This means that rules have to be provided for the treatment of each of these allomorphs individually.

The main word classes in Bahasa Malaysia are nominals, verbals, auxiliaries, adverbals and particles and these again can be subcategorised. The word class of a derived word is dependent on its affix and on affix deletion to obtain the rootform, its word class is set. For multiple affixation, the outermost prefix (if any) determines the word class.

Affixation also results in modifications or additional semantical features. Each affix carries a set of possible semantical features. For example, the prefix pe_N may cause the wordform to be agentive, instrumental, the object of action, etc. Which is the correct 'role' depends on the base on which the prefix was attached. In this model, no decision is made as to which semantical feature is the correct one. Instead, the whole set of features are set when the affix has been detected.

Although it may be possible to subcategorise the word classes into groups with common semantical features, this model so far only considers grouping according to word classes and does not consider subgroupings for semantical features. Work is now in progress to include such subgrouping to provide a more complete morphological analysis of Bahasa Malaysia. This model not only handles these three main morphological processes, but also handles idiomatic expressions as well. On completion of morphological analysis, all the information gathered is submitted to the next stage of the ARIANE system, that is, the syntactical analysis stage in order to build up a more complete 'picture' of Bahasa Malaysia.

## Introduction

The Malay language has been the national language of Malaysia since 1955 and with the formation of Malaysia in 1963, it has been known as Bahasa Malaysia (B.M.). B.M. belongs to the Western Group of the Austronesian Family and is spoken by people through Malaysia, Singapore, Indonesia and Brunei. There are a number of 'varieties' of B.M. - the regional type, pidginised B.M. as well as standard B.M., that is, that 'variety' used formally and official in government establishments, formal institutions as well as in mass communication. The morphological analysis described in this paper is based on this particular variety, that is, standard B.M.

B.M. uses both the Romanised and Arabic scripts for its writing system. For this purpose, the Roman script proposed in 'Pedoman Umum Ejaan Bahasa Malaysia'[1] will be used. This system was an attempt to standardise the spelling system of both B.M. and the Indonesian language.

This paper suggests a morphological model for B.M. using the ATEF system which was developed by the GETA group at Universite Scientifique et Medicale Grenoble.[2] The ATEF system is part of an interactive system known as ARIANE-78. ARIANE-78 is a software tool for machine-aided translation to which linguistic data (grammars, dictionaries, heuristic) formalised in some external artificial language is given. It includes the following components:

1. ATEF - a non-deterministic finite state transducer which is used for generating programs for morphological analysis.

2. ROBRA - a tree-to-tree transducer which is used for multi-level analysis (syntax and partial semantics), for the structural transfer and also for syntactic generation in the target language.

3. TRANSF - a system for bi-lingual dictionary look-up. It is used for lexical transfer.

4. SYGMOR - a deterministic finite state transducer used for morphological generation.

This paper describes the modelling of this linguistic data to be supplied to ATEF for morphological analysis. No attempt has been made to describe in detail the usage nor the writing of the external artificial language for ATEF as the purpose of this model is to test the capability of handling B.M. morphological analysis under ATEF and also to provide a first working model for B.M.

In the sections that follow, a morphological description of B.M. will be given, followed by the morphological model.

## B.M. Grammar

The morphological description given here is taken mainly from 'The Morphology of Malay'[3].

There are three main morphological processes in B.M., that is, reduplication, compounding and affixation.

### Reduplication

Reduplication is the process whereby a base or some part of the base is repeated. In B.M. there are two types of reduplication: reduplication proper and rhyming and chiming. Reduplication proper may be partial or full. For partial reduplication, the duplicate is determined by the initial or final syllable of the base. In initial syllable reduplication, only the initial consonant of the base (provided it begins with a consonant) is repeated while the rest of the duplicate is of constant shape (i.e., -ek). In final syllable reduplication, the last syllable is repeated without any change. For example,

budak 'child' ⇒ bek-budak 'children' (initial syllable)

dak-budak 'children' (final syllable)

Partial duplication generally occurs only in colloquial B.M. whereas full duplication occurs in standard B.M.

In full duplication, the duplicate is identical to the whole base. For example,

budak 'child' ⇒ budak-budak 'children'

Rhyming and chiming is also called reduplication with phonetic change. A compound form is called rhyming if one syllable of the base is repeated in the duplicate, example,

kuih 'cake' ⇒ kuih-muih 'variety of cakes'

and chiming if all the consonants are repeated in the duplicate and only the vowels changes, example,

gunung 'mountain' ⇒ gunung-ganang
'variety of mountains'

Generally, proper reduplication of nouns gives a semantic category of heterogenity or indefinite plural while rhyming and chiming has the added feature of variety. As for verbals, proper reduplication may result in at least one of the following semantical features: repetition, continuity, habituality, intensity, extensiveness and resemblance. For example,

baca 'to read' ⇒ baca-baca
'to read repeatedly/ continually/always'

kuning 'to be yellow' ⇒ kuning-kuning
'to be very yellow/ yellowish/yellow all over/ always'

When occurring with dynamic verbs, rhyming and chiming gives the semantical function of repetition, example,

beli 'to buy' ⇒ beli-belah
'to buy again and again'

and that of intensification when occurring with stative verbs, example,

malu 'to be shy' ⇒ malu-malah
'to be very shy'

## Compounding

A compound form is a construction that has two or three free forms as its constituents where each of the constituent forms may be either a root or a derived form, example,

kayu_api 'firewood'
(kayu 'wood' and api 'fire')

suratkhabar 'newspaper'
(surat 'letter' and khabar 'news')

## Affixation

Affixation is the process whereby a base may be extended by one or more affixes. Affixation is the most common and widely used of the three morphological processes. Affixes may be classified as prefixes, suffixes, infixes and circumfixes.

### Prefixes

The more common prefixes in B.M. include di, ter, ber, per, se, sese, juru, me_N, pe_N, ke. This list of prefixes with the exception of pe_N and me_N do not result in any morphographemic changes. Prefixes me_N and pe_N take on different forms (its allomorphs) depending on the initial segment of the root-forms. For example, for the prefix me_N, its allomorphs are:

1. me - used with the letters l, m, n, ng, ny, r, w, y
2. mem - used with b, p, f, v (f and p dropped)
3. men - used with d, t, c, j, z (t dropped)
4. meng - vowels, g, h, k (k dropped)
5. meny - s (s dropped)
6. menge - for monosyllabic forms.

Examples:

1. me_N + lawat 'to visit' ⇒ melawat 'to visit'
2. me_N + pukul 'to hit' ⇒ memukul 'to hit'
3. me_N + cari 'to find' ⇒ mencari 'to find'
4. me_N + kacau 'to disturb' ⇒ mengacau 'to disturb'
5. me_N + sapu 'to sweep' ⇒ menyapu 'to sweep'
6. me_N + cat 'to paint' ⇒ mengecat 'to paint'

The same rules hold for the prefix pe_N.

### Suffixes

The addition of suffixes do not present such morphographemic changes. Suffixes are merely attached to the rootform without any changes being made to the suffix nor the root-form. Four 'layers' of suffixes are possible in B.M., that is, from the innermost layer outwards, we have:

1. an, wan, wati, man, is, isma
2. i, kan
3. mu, ku, kau, nya
4. lah, kah

There is no co-occurrence of suffixes in each 'layer' except for 'i' and 'kan'. As an example, from the word baharu 'new', we have, on affixation,

di + baharu + i + kan + nya + kah
⇒ dibaharuikannyakah 'is it renewed (by subject)?'

Infixes

There are three infixes in B.M. - '-el-', '-em-' and '-er-'. For example,

getar + el ⇒ geletar 'to tremble'

gilang + em ⇒ gemilang 'to be very splendid'

gigis + er ⇒ gerigis 'to be very uneven'

These three infixes are not productive and only account for a small number of wordforms only.

Circumfixes

Circumfixes are discontinuous combinations of prefixes and suffixes. The most common circumfix is 'ke_an'. Example, by circumfixation of the word banyak 'many', we have

ke + banyak + an ⇒ kebanyakan 'majority'

Syntactical and Semantical Features

Affixation also plays an important role in the setting of the syntactical as well as semantical features of a wordform. It can cause a change in the grammatical class of the wordform or even to change the meaning of the wordform. For example,

latih 'to train' + an ⇒ latihan 'training' (syntactical change)

pe_N + dapat 'to obtain' ⇒ pendapat 'opinion' (semantical change)

In B.M., the nominal affixes are those which cause the wordform it is attached to, to take on the grammatical category of nominals. These affixes include

pe-N, ke, an, wan, man, wati, is, isma, ke-an

Verbal affixes are those which result in a verbal category, example,

se, ke, an, ke-an, per, ber, di, ter, kan, i, me-N

Affixes also provide semantical features to the wordforms they are attached to. For example, the semantical features that can be set to the wordforms to which the prefix 'ber' is attached are

reflexive, possessive, or productive.

Which particular feature should be assigned depends on the wordform itself. For example,

ber + cukor ' to shave' ⇒ bercukor 'to shave oneself' (reflexive)

ber + kereta 'car' ⇒ berkereta 'to possess a car' (possessive)

ber + anak 'child' ⇒ beranak 'to give birth' (productive)

Multiple Affixation

Multiple affixation also occurs in B.M. Fortunately, not more than three layers of affixation can occur, e.g.,

ber + ke + se + orang 'person' + an ⇒ berkeseorangan 'to suffer loneliness'

Multiple affixation results in added complexities in the setting of semantical features. The syntactical category of the affixed wordform is the category set by the outermost prefix. But due to combinations of affixes, the semantical features may increase and sometimes even differ. For example,

pe-N + dapat 'to obtain' ⇒ pendapat 'opinion'

ber + pendapat ⇒ berpendapat 'possess opinion'

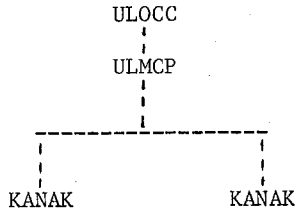
The Morphological Model for B.M.

The model described here covers the three morphological processes described in the above section though not all are included. Those excluded are:

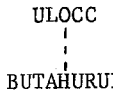
1. partial reduplication - this only occurs in colloquial B.M. and not in standard B.M.

2. infixes - the small number of occurrences do not justify its handling under ATEF. They are merely regarded as rootforms and set up as dictionary entries in this model.

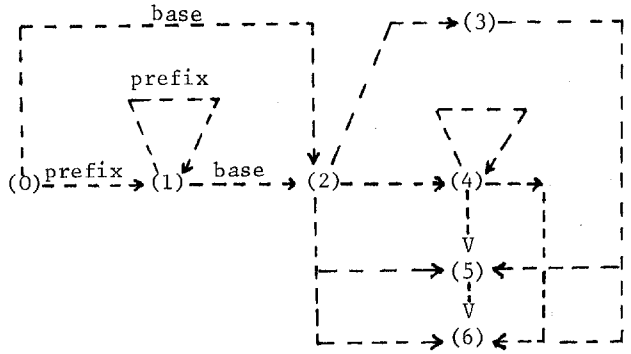
Reduplication and compounding are quite straightforward and do not present much of a problem in the morphological analysis of B.M. In this model, reduplicated words are treated as 'compound words', that is, a single node (ULMCP) is created with the duplicates as brother nodes. For example, the compound word kanak-kanak results in the following sub-tree:




Compound words are treated as idiomatic expressions and a single node is created for them. For example, the word buta huruf results in the following sub-tree:




As affixation is the most complex of the three processes, a Finite State Diagram is used to show the various stages in the de-segmentation of a given word-form:




Beginning with state 0, on encountering a prefix, the analysis proceeds to state 1. If a base is encountered, state 2 is reached. In state 1, on encountering another prefix, it still remains in state 1. This is due to the possibility of having multiple affixation in B.M. State 1 goes to state 2 when a base is encountered. From state 2, whichever state is reached depends on which suffixes are encountered:

state 3 - an, wan, wati, man, is, isma

state 4 - i, kan

state 5 - mu, ku, kau, nya

state 6 - lah, kah

This is due to the possible 'layering' in suffixation (as described above). States 2, 3, 4, 5 and 6 are all final states in this State Diagram.

In each state, information is added to the extracted rootform as the analysis proceeds, e.g. syntactical and semantical features set by affixation, reduplication, etc. (as described in the above section.) The set of semantical features used in this model can be obtained from the Appendix. In this model, no decision is made as to which of the possible semantical features should be added when the affix/affixes are deleted from the wordform. Instead, the whole set of possible features are added. At the time of writing, further research is being carried out concerning the extraction of the correct semantical feature and not include the whole set.

A sample of B.M. morphological analysis of a text using this model is included in the Appendix.

## Limitations of ATEF

While testing the model under ATEF, it was found that ATEF could not handle two aspects of B.M.:

1. Affixation of Proper Nouns

In B.M., it is possible to attach the particles kah and lah to proper nouns, example,

Ahmad 'name of person' + lah ⇒ Ahmadlah
'Ahmad prt.'

Penang 'name of state' + kah ⇒ Penangkah
'Penang prt.'

2. Affixation of Idiomatic Expressions

Idiomatic expressions in B.M. can also be subjected to affixation, example,

ber + buta huruf 'illiterate' ⇒ berbuta huruf 'to be illiterate'

These two problems have been communicated to the GETA group who are looking into these problems.

## Conclusion

This paper has attempted to provide a possible model for the morphological analysis of B.M. under ATEF but is not the complete model as yet. More research work is being done to refine this model with the ultimate objective of providing as much information as possible in the morphological analysis stage to be passed to the next stage in ARIANE-78, that is, the multi-level analysis stage under ROBRA. One aspect which is being looked into is the possibility of 'layering' prefixes just as has been done for suffixes. Another area of research is the possibility of extracting the

exact semantical feature set by affixation of the word-form, instead of including the whole set of possible semantical features. This would remove much ambiguity and would also simplify the analysis in the next stage.

## Appendix

### Semantical Features on Affixation

#### Prefixes

Pe_N = (agentive, qualitative, instrument, abstract, unit of measure, object of action, profession).

Se_ = (similar, singular).

Ber_ = (reflexive, possessive, productive.

Ter_ = (unintentional, superlative, capability, past).

Per_ = (causative passive).

Juru_ = (profession).

#### Suffixes

_i = (causative with locative benefactive complement).

_kan = (causative benefactive).

_an = (resultative, locative, collective, variety, repetition).

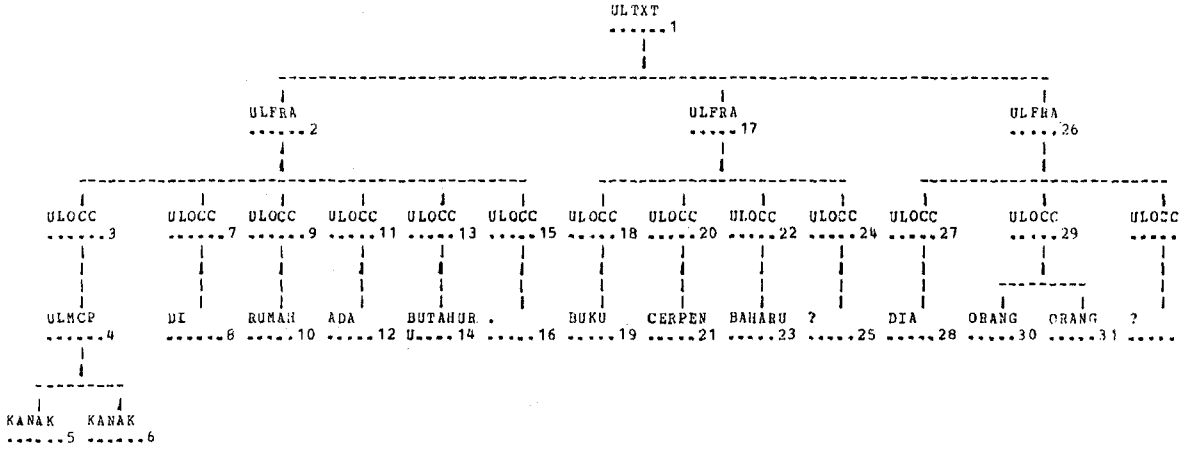
#### Circumfixes

pe_an = (process).

se_an = (abstract, locative, resemblance, passive).

ESSAI TIGA 7 AVRIL 1980 8H 42MN . 6S

KANAK-KANAK DI RUMAHNYA ADALAH BUTA HURUF . BUKU CERPEN
DIBAHARUIKANNYAKAH ? DIA BERKESEORANGANKAH ?




```
SOMMET 1   : UL(ULTXT).
SOMMET 2   : UL(ULFRA).
SOMMET 3   : UL(ULOCC).
SOMMET 4   : UL(ULMCP).
SOMMET 5   KANAK-KANAK: UL(KANAK),PETAT(1),CAT(N),SUBN(NC),HYPHEN(1),CASE(ANI).
SOMMET 6   KANAK-KANAK: UL(KANAK),PETAT(2),DRV(NN),CAT(N),SUBN(NC),HYPHEN(1),CASE(ANI).
SOMMET 7   : UL(ULOCC).
SOMMET 8   DI: UL(DI),CAT(P),SUBP(PL).
SOMMET 9   : UL(ULOCC).
SOMMET 10  RUMAHNYA: UL(RUMAH),PETAT(5),CAT(N),SUBN(NC),PERSON(3),CASE(LOC).
SOMMET 11  : UL(ULOCC).
SOMMET 12  ADALAH: UL(ADA),PETAT(2),CAT(V),SUBV(VB),TYPE(DCL).
SOMMET 13  : UL(ULOCC).
SOMMET 14  HURUF: UL(BUTAHURUF),CAT(N),SUBN(NC),LGID(2).
SOMMET 15  : UL(ULOCC).
SOMMET 16  .: UL(.),CAT(Z).
SOMMET 17  : UL(ULFRA).
SOMMET 18  : UL(ULOCC).
SOMMET 19  BUKU: UL(BUKU),PETAT(2),CAT(N),SUBN(NC),CASE(INANI).
SOMMET 20  : UL(ULOCC).
SOMMET 21  CERPEN: UL(CERPEN),PETAT(2),CAT(N),SUBN(NC),CASE(INANI).
SOMMET 22  : UL(ULOCC).
SOMMET 23  DIBAHARUIKANNYAKAH: UL(BAHARU),PETAT(5),DRV(VV),CAT(V),SUBV(PAS),PERSON(3),SEG(DI,KAN,I),TYPE(
    SEM(CAUSBEN,CLOCBEN).
SOMMET 24  : UL(ULOCC).
SOMMET 25  ?: UL(?),CAT(Z).
SOMMET 26  : UL(ULFRA).
SOMMET 27  : UL(ULOCC).
SOMMET 28  DIA: UL(DIA),PETAT(2),BIL(SING),CAT(N),SUBN(PRON),PERSON(3).
SOMMET 29  : UL(ULOCC).
SOMMET 30  BERKESEORANGANKAH: UL(ORANG),PETAT(3),CAT(V),SUBV(VB),SEG(KE,SE,BER,AN),TYPE(INTRR),
    SEM(ABSTR,LOCAT,SAME,SING,RESEMB,PASS,REFL,POSS,PROD,WORK).
SOMMET 31  BERKESEORANGANKAH: UL(ORANG),PETAT(3),DRV(NV),CAT(V),SUBV(VB),SEG(KE,SE,BER,AN),CASE(ANI),
    SEM(ABSTR,LOCAT,SAME,SING,RESEMB,PASS,REFL,POSS,PROD,WORK).
SOMMET 32  : UL(ULOCC).
SOMMET 33  ?: UL(?),CAT(Z).
```

## Reference


1. 'Pedoman Umum Ejaan Bahasa Malaysia'(1977). Kuala Lumpur: Dewan Bahasa dan Pustaka.

2. Vauquois, B. (1979). The Evolution of Oriented Software and Formalized Linguistic Models for Automatic Translation or Machine-Aided Translation of Natural Languages. Doc GETA, Grenoble, France.

3. Abdullah, Hassan (1974). The Morphology of Malay. Kuala Lumpur: Dewan Bahasa dan Pustaka.

4. Chauche, J., Guillaume, R., Quexel-Ambang, M. (1972). Le Systeme ATEF Doc GETA, Grenoble, France.