Journal of Modern Languages
Vol. 23, (2013)

The Symmetrical Architecture of Malay Simplex Forms

Zuraidah Mohd Don
University of Malaya

Abstract
This paper examines the structure of Malay words, concentrating on simplex forms with no affixation, and analyses over 6,400 simplex forms extracted from a corpus of about 3M words of written Malay. Over 80% of simplex forms are found to fit a template of the structure CV(C)CVC. Examination of the templates links the widespread view that Malay has a 'simple phonology' to an inadequate separation of abstract lexical representations and acoustic events in speech waveforms, and the claim is made that this separation is the key to understanding contemporary dialect variation.

Keywords: Malay, simplex forms, templates, corpus linguistics, phonology

1. Introduction

One of the most recognisable characteristics of Malay is the CVCVC structure at the heart of words. Complex words are built up by adding affixes before and after the stem. When all affixes have been removed, what remains is the **simplex form**, which in practice often has the structure CVCVC. We refer to this structure as a **template**[1]. It contains a number of slots which are filled by phonological elements which can be provisionally thought of as 'phonemes'. The phonological elements are used to predict the entries in a pronouncing dictionary, which is a table with a key role in connecting written texts to acoustic events in waveforms: given a phonological string corresponding to a lexical representation, it is possible to predict the acoustic events that are likely to occur when a word is uttered in some dialect of Malay.

CVCVC templates are not only phonologically interesting objects, but they are also characteristic of the Austronesian languages as a whole. In these circumstances, it is surprising to find very little literature describing them. If templates are mentioned at all, they are referred to indirectly as syllable structures. Even descriptions of the phonological elements tend to echo by implication that Malay is a 'simple' language, and so of little theoretical interest. The evidence for the *prima facie* case that Malay has a simple phonology is that words are made up of CV(C) strings, the vowels are /i e a o u ə/, and the main

[1] The term *template* is used here in a general sense for an instrument to replicate instances of a particular pattern. This does not necessarily conform to any of the objects called templates in generative phonology.

consonants are /p b t d c j k g s h m n ñ ŋ r l y w/[2]. Such simplified descriptions miss the point that Malay has quite distinctive phonological patterns for words.

This paper focuses on the template itself. It begins with a review of some previous work, and goes on to make a systematic investigation of the properties of the template, and show how these properties have changed in response to the assimilation of large numbers of loan words.

2. Templates and Phonological Domains

The structure of Malay words is usually described in terms of strings of phonemes or of CV(C) or even (C)V(C) syllables. It is argued that while these descriptions may be accurate as far as they go, they are insufficient to capture significant generalisations. The onus is therefore upon us to demonstrate that the establishment of the template as a phonological domain enables us to elucidate parts of the structure of Malay that would otherwise be unaccounted for. These include (1) patterns of distribution of sounds, (2) the classification of sounds, and perhaps most importantly, (3) regular correspondences.

The clearest description we have been able to find is actually in Hendon (1966, p. 23ff). Hendon makes an excellent analysis of Ulu Muar Malay in a strict structuralist framework in which the stems of complex words are treated as strings of vowels or consonants. He describes "allomorphs with two vowels" (p. 23) in which vowels and consonants are ordered right to left C_3 V_2 C_2 V_1 C_1, and he examines in detail possible combinations and co-occurrence restrictions in strings containing one or two vowels. However, he has nothing to say about the domain within which these combinations and restrictions operate. Phonemes cannot have zero realisations, and this leads to a loss of generalisations. For example, *jil* [jɪl] 'jail' has [ɪ] as expected before a final consonant (p. 35), but if [l] is not pronounced, [ɪ] loses its conditioning environment and so is promoted to phonemic status. The correspondences between Ulu Muar forms and those of standard Malay can in general be expressed by a set of simple rules, but these rules have no place in the phonological model. It is ironic that the very quality of Hendon's work exposes the shortcomings of the theoretical model he adopted.

Verguin (1967) is concerned with the frequencies of items in a small corpus, and his claim (p. 40) translates as 'the canonical form of the primary lexeme [= 'simplex form'] in Malay is disyllabic'. He gives tables of figures showing how relative frequencies depend on structural positions in what is in practice a CVCVC structure. However, he follows conventional phonemics in grouping sounds into phonemes irrespective of structural position. For example, he regards initial [k] and final "glottal stop" as members of the same phoneme (p. 33), and (interestingly) treats prenasalised stops as separate phonemes (pp. 27-34). The independence of structural positions suggests that the data might have been better analysed in terms of sounds and prosodies (Firth, 1948) than by

[2] In accordance with established practice in Austronesian linguistics, [c] is used to represent a voiceless palatal stop, [j] a voiced palatal stop, [ñ] a palatal nasal, and [y] a palatal glide. The examination of spectrograms shows that the palatal stops are affricated as expected, and that pre-vocalic [ñ] is followed by formant transitions consistent with a palatal glide.

conventional taxonomic phonemics, already beginning to be old fashioned in 1967 (Knowles, 2005).

Lapoliwa (1981) takes a generative approach to the phonology of Indonesian. He confirms previous findings that stems are typically bisyllabic (pp. 41-42), and notes that nearly half of the bisyllabic type conform to the structure CVCVC. He lists eleven bisyllabic types (pp. 46-49), but treats them as autonomous entities, and makes no attempt to treat them as variants of the basic CVCVC pattern. This falls short of the declared aim to 'capture phonological generalisations of the language' (p. 4).

Teoh (1994, pp. 15-23) gives a long list of syllable combinations and vowel patterns, but uses a phonemic analysis based on old spellings that distorts the vowel patterns he attempts to describe. For example, he uses "e" to transcribe [ɪ] in *bilik* [bilɪk] 'room' and assigns it to /e/, and similarly uses "o" to transcribe [ʊ] in *hidup* [hidʊp] 'life' and assigns it to /o/ (p. 19). This has the effect of distorting patterns of vowel harmony involving close and mid vowels, a property of Malay so well known ('sistem keselarasan') that it is referred to in non-technical discussions of the orthography (Ismail & Manshoor, 2001, pp. 9-12). Teoh is not writing in a conventional phonemic paradigm, but seeks to prove (p. 1) the superiority of a non-linear approach with syllable trees and metrical structures over standard generative phonology. But his approach is insufficiently flexible either to describe the Malay vowel system, or to identify higher level phonological structures formed out of syllables.

The conclusion to be drawn from the study of previous work in this area is that there is a mismatch between the phonological theories adopted and the theoretical tools required to describe Malay. Theoretical statements do not go far beyond describing stems as strings of syllables. But the task facing the phonologist is to account for the distribution of sounds and their co-occurrence restrictions, including patterns of vowel harmony, and to show how varieties of Malay are related to each other, and how the modern forms have evolved. It cannot be a matter of chance that a high proportion of Malay simplex forms fit the CVCVC template and its variants, or that the same structure accounts for a considerable range of phonological patterns and rules. This is our motivation for focusing on the template itself and describing it in detail in its own right.

3. Method

The work reported here builds on previous work in Malay corpus linguistics (Knowles & Zuraidah Mohd Don, 2006), and exploits the MALEX database extracted from about 3M words of naturally produced Malay written texts. The sample of the Malay lexicon contained in this database is intended to be a representative sample. All words do not have an equal chance of being included, for they are more likely to be included the more frequent they are in running text.

Automated procedures are used to identify words the first time they are encountered in texts. Each new word is analysed grammatically, and stored in a table of lexical items. Affixes are stripped off, leaving the simplex form, which is also used as the label for the lemma. Lemmas are stored in a related table. A

specially designed spelling-to-phoneme algorithm is used to generate phonological representations for simplex forms, and these are parsed automatically to populate phonological templates. A statistical description of a large number of simplex forms provides us with the justification to make claims about what is rare or frequent.

The methodology of corpus linguistics involves storing large amounts of data systematically in annotated textfiles, vertical files, arrays, tables or xml files. What really matters is the structure of the data. Notation is also important in so far as it correctly reflects the structure of the data.

An advantage of general-purpose data storage is that it enforces logical discipline in data handling, and strict data typing. An important distinction is made between abstract symbolic representations at a phonological level typically stored in lexical tables, and phonetic patterns such as release bursts and formant transitions, which are found in waveforms and recorded in annotation files. These are of course clearly distinct from orthographic representations. Abstract representations and phonetic events are connected using a series of related tables. A table of simplex forms, for example, uses "C1" for the first consonant and "V2" for the second vowel. These are fieldnames, and so they are not of the same type as the feature [consonantal] or [vocalic]. In our database, symbols such as "t" and "a" are used in phonological representations in the lexicon and also as identifiers in a common field to join related tables, and so are very clearly separate from acoustic events in waveforms or annotation files.

Most of the little that has been done on Malay, including the work reviewed above, has typically started with a theory, and with data selected to defend one theoretical approach against another. The problem with this approach is that the theory decides in advance what kind of data is relevant and how it is going to be analysed. In this case, we start with large amounts of carefully described and annotated data, and allow theoretical claims to arise naturally from the examination and manipulation of the data. As we hope to show below, interesting patterns emerge when large amounts of language data are considered as a whole and allowed to tell their own story.

These points are important, because the assertion that Malay has a simple phonology follows from certain theoretical positions, and is inconsistent with strict data typing. If one takes an intuitive approach to data, including an informal view of data types, then it appears superficially obvious that the phonology is simple. But if a rigorous approach is taken to data handling, then it quickly becomes clear that the phonology is not so simple after all.

4. The Phonological Template

We here represent the template as a sequence of structural positions C1 V1 C2 V2 C3.

4.1. Medial (C)C

The C2 position can contain a maximum of two consonants, and so could be written (C)C. In native Malay words, the optional (C) is typically, and perhaps historically always, a homorganic nasal in a cluster of the type /mb, nt, ns, ŋk/,

e.g. *lembu* 'ox'. The only other frequent type has /r/ followed by another consonant, e.g. *harga* 'price'.

Homorganic nasals and /r/ have a special privilege of occurrence in C2. Other combinations are possible, but infrequent. For example, *tanpa* 'without' with medial /np/ is the only case of its kind in the database. There is a strong case for recognising the homorganic nasal as an autonomous phonological entity, here written /N/, including /Nt/ in *hantar* 'send', /Ng/ in *tinggal* 'stay', and /Nc/ in *puncak* 'peak'. Before another consonant in C2, /r/ occurs as an alveolar tap, which is quite different from its form in other positions.

4.2. Null Vowels and Consonants

A large number of words can be shown to conform to the template if we allow consonant positions to be filled by null, here represented "_". Words beginning with a vowel letter in the spelling can be deemed to have a null C1, e.g. *ubat* /_ubat/ 'medicine'. Similarly, words ending in a vowel letter can be deemed to have a null C3, e.g. *gigi* /gigi_/ 'tooth'. There are also words in which no medial consonants are indicated in the spelling, e.g. *baik* 'good', *laut* 'sea', and these are possible candidates for null C2. These examples do not sound like English *bike* and *lout*, because they are generally pronounced with medial glides of varying degrees of prominence, thus [bayik, lawut]. In each case, /a/ fills V1 or V2, and [y] is associated with an adjacent [i], and [w] with an adjacent [u]. There are also words like *maaf* [maʔaf] 'pardon' – typically deriving from Arabic words with a medial glottal stop or pharyngeal approximant – in which V1 and V2 are both filled with /a/, and separated by a glottal stop or pharyngeal approximant.

Finally in connection with null, there are words which appear to end in final diphthongs and so have a null C3. Examples include *kedai* [kədai] 'shop' and *limau* [limau] 'lime, lemon'. The problem is that diphthongs are not generally a feature of (standard) Malay at all. These words are better analysed with glides in the C3 position, thus /kəday, limaw/.

4.3. Final "a"

A problem arises with two different final vowels both spelt "a" in e.g. *duta* 'ambassador' and *baca* 'read'. The first of these is always pronounced [a], but it occurs only in words of foreign origin. The second is also pronounced [a] in a range of varieties of Malay, including the official *sebutan baku* (roughly) 'received pronunciation' (Ismail & Manshoor, 2001). But this vowel varies according to dialect. In everyday educated Kuala Lumpur Malay it is pronounced [ə], while in Kelantanese, a phonologically advanced variety of Malay spoken in the north east of peninsular Malaysia and in southern Thailand, it is [ɔ]. On phonetic grounds, and to distinguish it from the *duta* vowel, this second vowel is represented /ə/. The phonetic forms used in related varieties are all fully predictable from the representations /duta_, bacə _/, e.g. /bacə_/ corresponds in different varieties to [bacə] or [baca] or [bacɔ].

4.4. The Two Halves of the Template

The template is made up of two halves, which we shall call H1 and H2. H1 includes C1, V1 and the first of two consonants in C2. H2 includes the second or only consonant in C2, together with V2 and C3. Words like *lembu* /ləNbu_/ 'ox' and *harga* /hargə_/ 'price' are thus divided /ləN|bu_/ and /har|gə_/ respectively. It might seem rather unusual to assert the presence of a high level boundary between a homorganic nasal and the following consonant, and this would indeed be a bizarre analysis were it not for several kinds of evidence that independently and consistently point in the same direction.

4.4.1. Short Words

There is a small number of short words like *wap* 'water vapour' and *cat* 'paint' which behave exactly like the sequence C2 V2 C3, i.e. as H2. This suggests these words have a null C1 followed by a null V1, thus /_'wap/ and /_'cat/, where the apostrophe represents null V1. However, null is in complementary distribution with [ə] in the V1 position. In a word like *beras* [bəras] 'hulled rice', the [b] and [r] can be separated by [ə], the [ə] and [r] can be run together as a syllabic consonant, or the [b] and [r] can be run together without [ə] as a cluster much as in English *brass*. In this way, null and [ə] appear to be variants of a single phonological entity. In the case of words like *wap* and *cat*, the first vowel is null. But if C2 is complex, e.g. *erti* 'meaning', or *empat* 'four', V1 is filled with [ə]. There are just a few cases in which [ə] (formally written "e") occurs before a single nasal or liquid, e.g. *erang* 'groan', *enam* 'six' or *elus* 'caress', and in several of these it is optional in speech and in informal writing, thus *(e)mas* [(ə)mas] 'gold'; *(e)nam* [(ə)nam] 'six'; *(e)mak* [(ə)maʔ] 'mother'. There are no contrasts brought about by the presence or omission of [ə], but since *elus* does or can have an initial [ə] while other words such as *lap* 'wipe' do not, representing all these cases in the same way would lead to a loss of information. We therefore followed the phonetics in the representations /_ələs/ and /_'lap/. These representations can in any case be modified globally if the occasion should ever arise.

4.4.2. Vowel Harmony

The phonetic value of /i/ and /u/ in V2 depends on whether C3 is null or filled with a consonant. Before null, both vowels are close [i] and [u] respectively, as in *sini* [sini] 'here' and *kuku* [kuku] '(finger)nail'. Let us call this type 1. But before a consonant, these vowels are more open [ɪ] and [ʊ], as in *balik* [balɪk] 'return' and *hidup* [hidʊp] 'live'. This is type 2.

There is a third type which results in mid [e] and [o], represented in the orthography by the spellings[3] "e" and "o", e.g. *leher* [leher] 'neck' or *kosong*

[3] Chronic confusion in the spelling of the vowels we here call types 2 and 3 was addressed in the reformed spelling adopted jointly by Malaysia and Indonesia in 1972 (see Asmah, 1993a, pp. 43 - 96). The new spellings would appear to have solved the problem, and we know of no evidence to the contrary.

[kosoŋ] 'empty'. As these examples show, there is a vowel harmony rule[4] (Asmah, 1983) that requires a preceding /i/ and /u/ also to take the form [e] and [o] respectively, as in *besok* [besoʔ] 'tomorrow', *boleh* [boleh] 'can, be able'.

There are some restrictions in that [e, o] never follow /ə/ in V1, and they are very rare after /a/. On the other hand, [e] and [o] are not always motivated by vowel harmony, and each has an independent phonological status, e.g. *merah* /merah/ 'red', *tongkat* /toŋkat/ 'walking stick'. It would appear that the mid allophones [e, o] have merged with the mid vowels /e, o/, but that the slightly closer [ɪ, ʊ] remain allophones of /i, u/.

4.4.3. Affixes and Pseudo-Affixes

Malay has a rich derivational morphology (Asmah, 1993b). It so happens that prefixes are structurally identical to H1, the first half of the template, and that suffixes are identical to H2, the second half of the template. Native Malay prefixes (with the exception of the passive prefix *di*) also have an additional restriction, in that the vowel is always /ə/.

Some prefixes such as *pe-* in *petani* 'farmer' and *ke-* in *ketua* 'director, boss' pattern like any H1 before a simple C2, as in *besar* 'big' or *keduk* 'dig'. Others such as *meng-* /məN/ and *peng-* /pəN/ pattern like any other H1 ending in a homorganic nasal, while a third group containing *ber-* /bər/, *per-* /pər/ and *ter-* /tər/ pattern like other H1s ending in *–r*. Other prefixes such as *dwi-* 'two', *pra-* 'pre-' and *neo-* are rather obvious borrowings from other languages. Proclitics such as the first person *ku* and the second person *mu* are written solid with the stem, and these too do not fit the general pattern. Suffixes, including enclitics, have exact counterparts in H2 forms which are beyond doubt part of the simplex form itself.

4.4.3.1. Allomorphs

An important property of C2 is that although it allows sequences of two consonants, these are never geminates. If the addition of a prefix would create a potential geminate in C1, it is simplified. For example, when the verbal prefix *ber-* is added to *renang* 'swim', the result is *berenang* /bərənaŋ/. We might say that *ber-* has the allomorph /bə/ in this case. Note that just by looking at the form, it is impossible to tell whether it derives from *ber-* + *renang* or from *ber-* + **enang*. This creates problems for a stemmer using an automatic procedure to strip affixes from complex words.

A homorganic nasal produces a potential geminate before another nasal, and this too is simplified. Thus *meng-* + *naik* 'go up' produces *menaik*, and *meng-* + *masak* 'cook' produces *memasak*. In this case, there is another pattern that is unusual in that it is quite different from internal C2. A homorganic nasal copies the place of articulation of a following voiceless obstruent[5], but the

[4] The subtle differences between [ɪ, ʊ] and [e, o] are inconvenient for some theoretical approaches, and so they are simply ignored (see e.g. Teoh, 1994). This obscures the vowel harmony completely, despite its salient role in the phonology of native Malay words.

[5] The voiceless palatal /c/ is an exception.

obstruent itself disappears. Thus *meng-* + *pilih* 'choose' forms *memilih*; *meng-* + *tulis* 'write' forms *menulis* and *meng-* + *kecil* 'small' forms *mengecil*. Most cases involve stops, but /s/ also disappears after a palatal nasal, so that *meng-* + *sapu* 'sweep' produces *menyapu* /məñapu_/[6]. Again, just by examining the complex form, it is sometimes not possible to infer the simplex form, e.g. *memasak* could derive from either *masak* or **pasak*. *Mengecil* could similarly derive from *kecil* or **ecil*. Forms that really do derive from simplex forms with null C1, e.g. *mengalir* deriving from *alir* 'flow', give us the clue that the velar place of articulation is actually the default, so that when there is no place of articulation for the homorganic nasal to copy, it is velar by default. The same applies before /h/, as in *menghantar* 'send'.

4.4.3.2. Pseudo-Affixes

A defining characteristic of affixed forms is that they are semantically related to the simplex form. The connection may be irregular, tenuous or idiosyncratic, as in the case of *mata-mata* 'policeman', a reduplicated form of *mata* 'eye'. There are also many simplex forms which are too big to fit the template, and which look as though they begin with a prefix. In this case there is no semantic connection at all. For example, *seluar* looks as though it might have the same structure as *seorang*, which is made up of *se-* 'one' and *orang* 'person'. But *luar* means 'outside' and *seluar* means 'trousers', so that treating *se-* as a prefix would in this case introduce an element of absurdity into the morphology. At the same time, over 10% of lemmas classed as of Malay origin are of this type, and they cannot be ignored in a description of Malay words.

Let us call *se-* in *seluar* a "pseudo-prefix". Pseudo-prefixes, like real ones, always have the vowel /ə/, spelt "e". But stripping off a consonant + /ə/ can leave a stem beginning with a consonant cluster, such as /nd/ or /rt/. In these cases, the initial consonant is added to the pseudo-prefix, thus *pendata* /pən+datə/ 'scholar' or *pertama* /pər+tamə/ 'first'. These are remarkably similar to the real complex forms *penduduk* 'inhabitant', where *duduk* means 'sit, dwell', or *pertanian* 'agriculture', where *tani* means 'farm'. Unlike real affixes, pseudo-affixes do not have a stem to which they are attached. In the case of *peraduan* 'contest' and *perasaan* 'feeling', the prefix *per-* and the suffix *-an* are added to *adu* 'compete' and *rasa* 'feel' respectively; in the latter case the potential /rr/ is simplified to /r/, represented by a single "r" in the spelling. But for *perabot* 'furniture' there is no simplex form **rabot* or **abot* in any case. In these cases, we arbitrarily took *pe-* as the pseudo-prefix.

Most pseudo-affixes are similar in form to real ones, but in allowing them to end with a homorganic nasal, we also allowed forms such as /kən/, as in *kenduri* 'feast' (which has no connection with *duri* 'thorn'), and also /sən/, as in *senduduk* 'a kind of shrub' (which has no connection with *duduk* 'dwell') . The latter form is also found in the anomalous *sendiri* 'oneself' in which the pseudo-

[6] This development is not easily explained synchronically. When the rule was operative, [s] was presumably palatal or at least laminal and post-alveolar.

prefix *sen-* is added to the simplex *diri* 'self' to create a form which is closely related in meaning.

There are many cases of consonant + /ə/ which pattern like pseudo-affixes. One of these, namely *ge-*was so frequent (55 cases) that we treated it as a pseudo-affix, even though there is no real prefix beginning with *g-* /g/. The initial *le-* of *lelemak* /lələmak/ 'a kind of climbing plant' (which has nothing to do with *lemak* 'fat') corresponds exactly in form to the alliterative reduplication found *lelaki*, which derives morphologically from *laki* 'man'.

4.5. The Structure of Malay Words

We are now in a position to define the structure of typical Malay words. The structure is

H1* H2+

where the star means 'zero or more' and the plus sign 'one or more'. Words like *cat* have null C1, null V1 and a simple C2, which also means that they have a null H1. H1 is therefore optional. H2 on the other hand, is obviously obligatory. The number of H1s and H2s in a word can be increased by affixation, or by pseudo-affixation. It is for this reason that the structure given above is not just a paper formula.[7]

The possibility of one or more H1s raises the question of multiple prefixing. In fact both the active *memper* (*meng-* + *per-*) and its passive *diper-* (*di-* + *per-*) are frequent combinations. The word *memerangsangkan* 'stimulate, encourage' has the structure meng(per(rangsang))kan. There are interesting minimal pairs such as *pinta* and *minta*, both meaning 'ask', where *minta* is identical to the stem derived from *pinta* when it is modified by the prefix *meng-*. The active form *meminta* could derive from either *pinta* or *minta*, and both passive forms *dipinta* and *diminta* are used. It is as though having undergone mutation after *meng-*, *minta* is treated as a simplex form, ready for further prefixation.

Another set of patterns which are explained by the word structure, and which are difficult to explain otherwise, concerns the prefixation of words like *cat* 'paint'. A prefix is added to H1, but in *cat* H1 is null. As noted above, null and [ə] are in complementary distribution in V1. In this case, null is replaced by [ə], /_'cat/ becomes /_ əcat/, and *peng-* + /_ əcat/ becomes *pengecat* /pəŋəcat/ 'painter'.

4.6. Matching Templates

We now consider the degree to which individual simplex forms match the template. Table 1 describes criteria to place simplex forms on a scale from 1 (perfect match) to 7 (no match).

[7] In discussing Madurese reduplication following Weeda (1986), McCarthy and Prince (1999, p. 274) refer to the "left branch" of a structure that remains unidentified. This structure would appear to be a template of the kind described here. Madurese is closely related to Malay and has templates similar but not identical to those of Malay.

Table 1. The scale from perfect match to no match

1	Perfect match
2	Vowel harmony rules ignored
3	Non-native consonants used
4	Non-native CC combinations in C2
5	Final vowel
6	Pseudo-affixes used
7	No match

All the simplex forms were assessed on this scale, and the numbers were broken down according to source language and are here presented in Table 2.

Table 2. Degrees of match by language

Language	**Fit1**	**Fit2**	**Fit3**	**Fit4**	**Fit5**	**Fit6**	**Fit7**	**total**
Malay	3030	8	44	131	86	401	86	3786
Sanskrit	153	2	0	47	89	57	19	367
Tamil	72	2	0	4	1	9	3	91
Chinese	22	2	1	3	1	2	1	32
Arabic	110	0	78	143	19	77	150	577
Portuguese	16	1	0	4	7	6	3	37
Dutch	10	2	0	6	0	4	2	24
English	159	54	61	103	57	78	927	1439
Others	23	4	6	18	3	6	16	76
Total	3595	75	190	459	263	640	1207	6429

First consider the figures for Malay. It is unlikely that native speakers of Malay will spontaneously create new words which do not match the phonological patterns of their own language. We have to assume that since words are classed as Malay by default, the Malay figures include borrowings the true source of which has not yet been identified. As it happens, the non-matching residue (86 /3786) amounts to just over 2%.

Most words from most languages fit the template to a greater or lesser extent. The Sanskrit and Tamil residues are 5% and 3% respectively, while the Arabic residue is 26%, and the English residue 64%. At the other extreme, 80% of Malay forms fit perfectly, followed by Tamil (79%), with Arabic (19%) and English (11%) in the tail. Vowel harmony rules do not seem to make much difference (Fit 2), except (perhaps surprisingly) for English. Words like *lori* /lori_/and *blok* /bəlok/ fit the template consonant requirements but violate the vowel harmony rules, and while *stamp* is cut to size by the removal of the final consonant, the resulting *setem* /sətem/ still breaks the vowel harmony rules.

Allowing foreign consonants (Fit 3) such as /f/ and /z/ enables a large number of Arabic and English words to be included, while relaxing the rules for C2 to allow any pair of consonants (Fit 4) brings in a large number of words from Arabic and Sanskrit. Allowing a final vowel (Fit 5) brings in a large proportion (24%) of Sanskrit words.The recognition of pseudo-affixation (Fit 6)

accounts for over 10% of the Malay forms. Given such a high proportion, it would be difficult to deny pseudo-affixation a role in Malay word formation.

5. Discussion

The analysis of simplex forms including words of foreign origin, extracted from a corpus, led to the description of the symmetrical architecture of typical Malay simplex forms. Words grow symmetrically outwards from the centre, first the syllables of the simplex form, then any pseudo-affixes, and then real affixes, and finally clitics. The conventional practice (Anderbeck, 2008) is to reference syllables from the end of the word, e.g. "penult"; but where the penult is depends on how many suffixes and enclitics there are. The junction between H1 and H2 is the only fixed point of reference. Malay words clearly have phonological structure above the level of the syllable, and different patterns are associated with different structural positions. Similar discoveries have been made in the past (Twaddell, 1935), and it is central to the 'prosodic' approach of Firth (1948); but in the mainstream words are still seen as strings of syllables or even phonemes.

We now turn to the significance of the findings, and how they affect our understanding of Malay phonology. We start with the nature of the phoneme-like objects that fill the slots in the template, and then discuss the special case of the so-called "glottal stop".

5.1. Templates and Phonemes

The template has five structural slots labelled C1, V1, C2, V2, C3, with fillers presented in forward slashes, e.g. /t/, /ə/, and which might superficially appear to be classical or 'taxonomic' phonemes (Knowles, 2005). Like phonemes they form strings that constitute words, and like phonemes they enable us to predict the phonetic form. Different phonemes are associated with contrasting phonetic forms. Since these phoneme strings contain all and only the unpredictable information required to define the phonological composition of a simplex form, they clearly constitute lexical representations.

However, Malay also has another set of phoneme-like entities at a less abstract level. In some varieties of Malay, /r/ in C3 has no articulation of its own[8], but merely continues the articulation of the vowel in V2. Although Malay is often said not to make any distinction of vowel length, there are actually five taxonomic long vowel phonemes /ɪː, eː, aː, oː, ʊː / in *pasir* 'sand', *leher* 'neck', *besar* 'big', *ekor* 'tail', and *bubur* 'porridge'. Malay may be said not to have diphthongs, but formant transitions unmistakably characteristic of diphthongs are found in words like *kedai* and *limau*. In fact, Malay phonology cannot be described at a single level of representation, and a distinction has to be made between the abstract and the phonetic, reminiscent of the systematic phonemic and systematic phonetic levels of Chomsky and Halle (1968).

The connections between lexical representations and phonetic form can be expressed in the form of ordered triples, e.g. (V2, u, ʊ) which means that /u/

[8] For the sake of clarity and simplicity, we assume that C3 is also word final, unless otherwise stated.

fills the V2 position and takes on the form [ʊ]. The formant configuration in the waveform is of course to be predicted from [ʊ], not from /u/.

5.2. The Status of the "Glottal Stop"

Ordered triples are needed to explain a set of patterns associated with the "glottal stop"[9]. These patterns involve the movement of the vocal folds, and they are therefore part of the phonation system. The popular term "glottal stop" implies a phoneme-like entity related to velar stops or alveolar stops. This is where the problem starts.

Farid Onn (1980) denies the glottal stop phonemic status on the grounds of its high level of predictability, and writes a glottal formation rule to account for glottal stops associated with "k", and a complex glottal insertion rule to account for alleged glottal stops in other positions. Teoh (1994, p. 59) treats our null C1 as a case of an obligatory glottal stop, and uses this as evidence that Malay is a language in which syllables must begin with an onset, and so a Type 3 language according to the classification of Clements and Keyser (1983). He sets out to demonstrate the advantages of an autosegmental approach over the more conventional linear approach of Yunus Maris (1980), and also (pp. 60-61) criticises the attempts by Farid Onn to explain the role of the glottal stop; but in the event he proves unable to describe crucial aspects of the data. Tajul (2000) asserts that Malay has no underlying glottal stops. Yong (2008, pp. 47-48) asserts counterfactually that a glottal stop is obligatory between the prefix *ber-* and a following vowel, and denies the claim (Tadmor, 2004; Tajul, 2000) that words spelt with a final "k" have a glottal stop in pronunciation. The problem here is that these are attempts to find a place for a segment thought of as a "glottal stop", when what is needed is a systematic theoretical framework that includes a theory of phonation. These claims are based on a misunderstanding of acoustic events in waveforms, and the conclusion has to be drawn that conventional phonological theories cannot account adequately for this part of the Malay phonological system.

Like speakers of many other languages, Malay speakers tend to begin an initial vowel with a closed glottis. This produces what is popularly known as a "sharp attack", in which the vowel formants rapidly reach their target values, possibly following an audible release corresponding to a vertical striation on the spectrogram. This is the pattern typically associated with what we have called a null C1. In a word like *ini* 'this', (C1,_,NULL) is followed by (V1,i,ʔi). A slightly different pattern is associated with a null C2 bounded on either side by [a], as in a word like *maaf* /ma_af/ 'pardon'. There may be an actual silence brought about by the closure of the vocal folds, but if the closure is incomplete, the adduction of the vocal folds produces a glottal constriction in the middle of a continuous [a] vowel, which shows up on the spectrogram as an irregular series of striations superimposed on the quasi-periodic voicing striations. In this case, we have (C2,_, ʔ).

[9] Another set of patterns which can only be explained satisfactorily using templates and ordered triples involves nasalisation (see also Anderbeck, 2008; Blust, 1997).

A third related pattern is associated with final voiceless stops. As in some other Asian languages, final voiceless stops in Malay, as in *hidup* 'live', *dapat* 'get' and *masak* 'ripe, cook', tend to be cut short. This means that they are unreleased, and the preceding vowel may also be produced with glottal constriction. Glottal constriction is a property of the vowel, and so in *dapat* we have (V2,a,aʔ). In the special case of the entity spelt "k" in *masak*, there is no velar articulation at all in some varieties of Malay, leaving only the glottal constriction, which can be represented (V2,a,aʔ) followed by (C3,k,NULL). Before the nominalising suffix *–an* in *hidupan*, *dapatan* and *masakan*, the item in C3 is grouped with the following vowel and so is fully released; but with the causative affixation *meng..kan* in *menghidupkan*, *mendapatkan* and *memasakkan*, the item in C3 is unreleased as in final position, and the preceding vowel is given glottal constriction. More generally, if the voiceless stop is followed immediately by a vowel it is treated as the onset to the vowel, and otherwise the stop is unreleased and the preceding vowel is given glottal constriction.

6. Conclusion

Every linguist who has ever worked on Malay or related languages must surely have been aware of the template described here, and of its general symmetry. As our database has grown, we have observed similar patterns appearing literally thousands of times, and we have accordingly set out to describe the templates and seek to understand their properties. To say that Malay words consist of strings of syllables may be true, but it is not enough.

All the patterns we have described can be analysed drawing on concepts from different phonological theories, and this is essentially what has been done in previous work. But this approach encounters the problem of the blind men and the elephant. What we have done is to start with a phonological structure whose reality, in view of the thousands of examples included in the database, cannot seriously be denied. We have used the observed properties of this structure to make explicit sets of relationships within a wide range of phenomena which might otherwise not be seen to be related at all. We have done this using standard techniques and procedures for storing and manipulating data, and this has brought with it the great advantage of discipline in data typing. Our solution of the glottal stop problem, for example, makes perfect sense at the level of lexical representation, and reflects a formal link between lexical representations and acoustic events readily to be observed in Malay speech waveforms.

What has been exciting about the present research is that what started out as an attempt to describe templates has led to the discovery of a set of tools for tackling phonological problems in Malay more generally, ranging from sound change and dialect variation to the assimilation of loan words. Since we have been using standard techniques and procedures, there is no reason to suppose they only work for Malay and other Austronesian languages. The overall approach developed here could no doubt be adapted for work on any language,

and has particular potential for use in connection with undocumented and endangered languages.

References

Anderbeck, K. R. (2008). Malay dialects of the Batanghari river basin (Jambi, Sumatra). Available from http://www.sil.org/silepubs/abstract.asp?id=50415

Asmah, H. O. (1972). Some rules for the coining of technical terms in Bahasa Malaysia. *Nusantara, 1*, 44-59.

Asmah, H. O. (1983). *The Malay peoples of Malaysia and their languages*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah, H. O. (1993a). *Essays on Malay Linguistics*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah, H. O. (1993b). *Nahu Melayu mutakhir*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Blust, R. A. (1997). Nasals and nasalization in Borneo. *Oceanic Linguistics 36*(1), 149-179.

Blust, R. A. (1999). Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In E. Zeitoun & P. J.-K. Li (Eds.), *Selected Papers from the 8th International Conference on Austronesian Linguistics*. Taipei: Academica Sinica.

Chomsky, N., & Halle, M. (1968). *Sound Patterns of English*. New York: Harper & Row.

Clements, G. N., & Keyser, S. J. (1983). *CV phonology: a generative theory of the syllable*. Cambridge, MA: MIT.

Coedès, G. (1930). Les inscriptions malaises de Çrīvijaya. *Bulletin de l'Ecole française d'Extrême-Orient, 30*, 29-80.

Coedès, G., & Damais, L.-C. (1992). *Sriwijaya: history, religion and language of an early Malay polity*. Kuala Lumpur: Malaysian Branch, Royal Asiatic Society.

Collins, J. T. (1989). *Antologi kajian dialek Melayu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Damais, L.-C. (1968). La langue B des inscriptions de Śrī Wijaya. *Bulletin de l'Ecole française d'Extrême-Orient* 54, 523-566.

Farid, M. O. (1980). *Aspects of Malay phonology and morphology - a generative approach*. Bangi: Universiti Kebangsaan Malaysia.

Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, 127-152.

Hashim, H. M. (1999). *Sejarah perkembangan tulisan Jawi*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Hendon, R. S. (1966). *The phonology and morphology of Ulu Muar Malay*. New Haven, CT: Yale University Press.

Ismail, b. D., & Manshoor, b. H. A. (2001). *Daftar kata Bahasa Melayu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Jones, R. (2009). *Chinese Loan-words in Malay and Indonesian*. Kuala Lumpur: University of Malaya.

Knowles, G. (2005). Taxonomic phonemics. In K. Brown (Ed.). *The Encyclopedia of Language and Linguistics*: Elsevier.

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

Lapoliwa, H. (1981). *A generative approach to the phonology of Bahasa Indonesia*. Canberra: Australilan National University.

McCarthy, J. J. & Prince, A. S. (1999). Prosodic Phonology. In J. A. Goldsmith (Ed.). *Phonological Theory* (pp. 238-288). Malden, MA: Blackwell.

Tajul, A. K. (2000). *The phonological word in standard Malay*. Newcastle: University of Newcastle Press.

Teoh, B. S. (1994). *The sound system of Malay revisited*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Tryon, D. T. (Ed.). (1994). *Comparative Austronesian dictionary: An introduction to Austronesian Studies*. Berlin: Mouton de Gruyter.

Twaddell, W. F. (1935). *On defining the phoneme*. Baltimore: Linguistic Society of America.

Verguin, J. (1967). *Le Malais: essai d'analyse fonctionelle et structurale*. Paris: Mouton.

Weeda, D. (1986). Formal properties of Madurese final syllable reduplication. In A. Bosch & al. (Eds.), *Papers from the 23rd Annual Regional Meeting of the Chicago Linguistic Society, part 2* (pp. 403-417). Chicago: Chicago Linguistic Society.
Wells, J. C. (1982). *Accents of English.* Cambridge: Cambridge University Press.
Yong, Y. E. J. (2008). *The phonetic properties of words in Malay*. University of Malaya.
Yunus Maris, M. (1980). *The Malay Sound System*. Kuala Lumpur: Fajar Bakti.

About the Author

Zuraidah Mohd Don is Professor of Linguistics in the Faculty of Languages and Linguistics at the University of Malaya. Her research interests include investigating language structure, in particular the structure of Malay, and the use of language in the context of professional communication.

E-mail: zuraida@um.edu.my