NEO-MALAY LEXEME SYNTHESIZER

Neo-malay here means that the language to be used in this study is not exactly any extant Malay language variety, but rather is a constructed language inspired by the Malay language.

Lexeme Synthesizer is a procedure whereby lexemes are synthesized according to the phonological and phonotactical rules of a language.

Goal — to synthesize new lexemes that visually and aurally resemble extant native Malay words.

Significance — a lexeme synthesizer may be useful for building a *lorem ipsum* generator localized for the Malay language. It may also be useful for inventing new names for writing fictions.

Methods

1. Analyse the phonology of Malay, then derive a phonology for Neomalay
2. Analyse the phonotactic rule of Malay, then derive a phonotactic rule for Neomalay
3. Write a computer program that uses the rules set out in Step 1 & 2 to generate new lexemes in Neomalay

1 Phonology

The phonology of the Malay language is analysed, and a phonology is devised for the Neomalay language. In this paper, the character ⟨ỹ⟩ is used instead of ⟨ny⟩, and ⟨g̃⟩ instead of ⟨ng⟩. The symbols ⟨j⟩, ⟨c⟩ and ⟨y⟩ are used according to the Malay orthographic convention instead of the IPA standard.

1.1 Consonants

The consonant register seems to be mostly consistent across the Standard Malay languages in the Nusantara. Most analysis recognise the distinction between the native malay phonemes and the foreign phonemes. For this synthesis, only the native consonant phonemes are used, consisting of 18 consonants. The following is one analysis of the Malay consonant register by Adelaar.

	labial	dental	alveolar	palatal	velar
Voiceless stop	p	t		c	k
Voiced stop	b		d	j	g
Nasal	m		n	ỹ	g̃
Semivowel	w			y	
Others			s, h, l, r		

Clynes and Deterding presented an alternate consonant patterning based on the active articulator, instead of the passive articulator as it is usually done.

	Labial	Apical	Laminal	Dorsal
Plosive	p b	t d	c j	g k
Fricative			s	h
Nasal	m	n	ỹ	g̃
Others	w	r l	y	

1.2 Vowels

Standard Malay is generally described as having six vowel: front-close, front-mid, central-mid, central-open, back-close and back-mid. However, there is a situation of degeneracy between the close and mid varieties of the front and back vowels in the final syllable. Additionally, Bruneian Malay is described as having only three vowels {a, i, u}. The following is a simplified vowel register devised for Neomalay, consisting only of four vowels.

	Front	Central	Back
Close	i		u
Mid		e	
Open		a	

The merging of the close-mid vowel pairs aims to simplify the synthesis process. The phonetic distinction can be regenerated afterwards according to their own rules.

2 Phonotactics

> "More than 90% of the native lexicon is based on disyllabic root morphemes, with small percentages of monosyllabic and trisyllabic roots." (Adelaar 1992)

The root lexeme prototype to be used in this synthesis is given below, consisting of two syllables, each having the C1–V–C2 structure.

$$\text{C1} - \text{V1} - \text{C2} - \text{C3} - \text{V2} - \text{C4}$$

All but one of the consonants are optional, with the minimal lexeme having the structure V1–C3–V2: ara, apa, api, ura, ubi, itu, etc. The following rules are adapted from Adelaar and Clynes & Deterding.

C1 & C3, syllable onset – Any consonant

C1 & C3 homorganic rule – When both C1 and C3 are plosives, they cannot have the same articulation but different voicing. Thus, /b-b-/, /b-k-/ and /b-g-/ are allowed but not /b-p-/ and /g-k-/.

C2 & C4, syllable coda – Can't be laminals {c, j, ỹ} and voiced plosives {b, d, g}

C2 – Also can't be voiceless plosives and /h/. If C2 is /s/, then C3 must be a voiceless plosive: puspa, pasti, pasca, laskar.

C2 nasal & C3 homorganic rule – When C2 is a nasal, it must have the same articulation with C3, and C3 can only be plosives and /s/, but not ỹ. Thus, /-nd-/ is allowed, but not /-mg-/. For the laminals, /c/ and /j/ are preceded by /n/ while /s/ is preceded by /g̃/.

C2 /r/ and & C3 – When C2 is /r/, C3 can't be /h/, /w/ and /j/.

No gemination – C2 and C3 must not be the same consonant.

V1 & V2 – Vowels

V1 and V2 can be monopthongs only. V1 can be any vowel, while V2 can't be /e/.

The syllable's vowel puts a restriction on the possible semivowels that can occupy the coda. Standard Malay has been described as having only monophtongs, and what seems like diphtongs (ked**ai**) can instead be analysed as a

vowel-semivowel sequence (k-e-d-a-y). This is supported by the fact that no consonant ever follow the phonemes /ai/, /au/ and /ui/ in the same syllable, thus it can be argued that the semivowel occupies the single consonant coda position.

Semivowel	y	w
a	ay	aw
e	-	-
i	-	iw
u	uy	-

A semivowel cannot follow /e/, nor precede it. The vowel-semivowel mostly occurs in the root-final position (V2–C4). The rare instances where it occurs in the root-initial position (V1–C2) are loanwords: 'haiwan', 'kailan'. For this synthesis, I decided to prohibit root-initial vowel-semivowel. /iw/ seems to be very rare, but some examples include the place name 'Setiu' and the folk hero 'Hang Lekiu'.

Some other sequences that look like diphtongs are analysed as V-SV-V sequences.

Word	Analysis				
		V	SV	V	
kuat	k	u	w	a	t
buah	b	u	w	a	h
liar	l	i	y	a	r
laung	l	a	w	u	g̃
air		a	y	i	r

This implies that V1 also restricts the semivowel in C3 if C2 is absent. Put in another way, if C3 is a semivowel, then C2 must be null.

That said, there are the /ian/ and /uan/ sequences, found in 'kalian', 'haruan', 'durian', 'tebuan'. The literature that I've found so far made no mention of these sequences, which appear to be legitimate diphtongs, although they could be analysed as trisyllabic roots as well, or disyllabic roots with a fossilized suffix. They will be ignored for now and are assumed to be forbidden.

3 Lexeme formation

The Neomalay language described above has 18 consonants (19 if counting null) and 4 vowels. Before the phonotactic rules are considered, the number of words that can be generated using the CVCCVC prototype is

$$19 \times 4 \times 19 \times 18 \times 4 \times 19 = 1\,975\,392$$