MORPHOLOGICAL STRUCTURE OF MALAY: USING PSYCHOLINGUISTIC ANALYSES OF RATED FAMILIARITY

Lat Choo Lee, Susan J. Rickard Liow & May-Ling Olivia Wee
National University of Singapore
<swksusan@nus.edu.sg>

0 Introduction
Psycholinguistic databases, listing properties of English words such as frequency and concreteness, have proved useful for the design of language assessment tools and experimental investigations of cognitive processes. These databases are now widely available for several other languages, including French, German, Dutch and Spanish, thereby facilitating work on bilingualism and cross-linguistic research. In this paper, we describe how and why we are developing a psycholinguistic database for Malay, which takes into account the relationship between morphological structure and rated word familiarity.

We started the database by asking 35 Malaysian first language speakers of Malay to rate a corpus of 4,328 words, comprising stem and affixed forms, on a 7-point familiarity scale. From this main database, we then extracted the mean ratings for 36 stem verbs and compared them to the mean ratings of their affixed forms: teR-, meN-, di-, beR-, -kan, meN-....-kan and di-....-kan. Our results provide support for several predictable patterns: stems are rated as more familiar than affixed forms, active forms are rated as more familiar than passive forms, and informal forms are rated as more familiar than formal forms. The utility of these kinds of analyses and the main psycholinguistic database is then illustrated with particular reference to the design of spelling and reading tests in Malay, and empirical studies of written language processing.

1 Psycholinguistic Analyses of Writing Scripts
Words have many properties that can be quantified and stored in a psycholinguistic database, e.g., frequency, number of syllables, number of letters, number of phonemes, and type of morpheme. These databases are useful for the development of clinical and educational assessments as well as for the design of psycholinguistic experiments. By manipulating one property of a word with precision, while holding the others relatively constant, the clinician or researcher can make a more accurate assessment of a person’s language processing skills.

For English, and many other alphabetic scripts, psycholinguistic databases have already provided the foundation for reliable diagnostic tools and coherent models of cognitive processing. The importance of calibrating word properties has led to computerized compilations, e.g., Coltheart’s (1981) M.R.C. Psycholinguistic Database for English, and the Max Planck Institute’s CELEX Lexical Database for Dutch, English, French and German (Center for Lexical Information 1995). Before these computerized

© Lat Choo Lee, Susan J. Rickard Liow, and May-Ling Olivia Wee.
resources became available, databases comprised a large number of printed pages which required painstaking word-matching by hand. Carroll, Davis and Richman’s (1971) detailed work is a good example. These authors assembled a sizeable corpus of 5,088,721 English words, with different affixed forms listed separately. The resulting 805 pages of text, entitled *The American Heritage Word Frequency Book*, lists objective word frequencies for 86,741 different forms from 1,045 published materials. Other more modest databases, such as Toglia and Battig’s (1978) *Handbook of Semantic Word Norms*, have been compiled using subjective 7-point ratings, rather than objective word counts, for a range of word properties including concreteness, number of associations, and familiarity.

Although databases utilizing subjective ratings are often smaller in size and they are not founded on word counts from printed matter, they can incorporate a broader range of properties that are salient for cognitive processing, such as imageability, and they retain the scaling used by a specific sample of readers. In a recent study of Chinese character processing, Rickard Liow, Tng and Lee (1999) used Toglia and Battig’s method to calibrate semantic and phonetic regularity for Mandarin across subject pools from China, Singapore and Taiwan. Their results suggested that empirical research on the processing of Chinese characters demands more precision than has been observed to date, and that ratings are best obtained from local subject pools.

Given some of the advantages of subjective ratings, we reasoned that Toglia and Battig’s method of scaling word properties would be a good starting point for a psycholinguistic database of the Malay language. In what follows, we describe how and why we collected and analyzed familiarity ratings.

2 Towards a Psycholinguistic Database for Malay

Malay languages belong to the Austronesian family and are used by more than 100 million people in South East Asia. Their respective standard forms are amongst the national languages for Indonesia, Malaysia, Singapore and Brunei, and there are also a large number of colloquial versions of Malay spoken throughout the region. The standard, or formal version\(^1\) of Malay, is more commonly used for writing. Compared to English, formal Malay is very heavily inflected and the relationship between the orthography and phonology is much more predictable, both in the Arabic script and the more common Romanized script.

Despite the widespread use of informal and formal Malay and the potential for interesting cross-linguistic studies, no psycholinguistic database is yet available. Inevitably this has limited the development of standardized language assessment tools and the scope of empirical research on cognitive processing in Malay. For undergraduate projects, some pioneering students at the National University of Singapore gauged familiarity for 530 words (see Pereira, Rickard Liow & Mohd Saniff, 1992) and a search of the international literature yielded just one psycholinguistic study of Malay language processing published in English. This was a Stroop experiment conducted by Baluch and Abdullah-Darlan (1991) which compared reading in the Arabic and Romanised forms of the script. The scarcity of psycholinguistic research papers published in English on the Malay language suggests a brief review of the script would be useful here.

\(^1\) We have sometimes used the terms standard and formal interchangeably but recognize that there are formal and informal forms of particular words within standard Malay.
3 The Malay Writing Script
Both forms of the Malay writing script are alphabetic, but they are used in different domains. The Arabic form (Jawi) is now used almost exclusively for the teachings of Islam, whilst the newer Romanized form (Rumi) is used for all other kinds of education and almost all types of printed materials. For this reason, and because comparisons with English are more feasible, we restricted our study to the Romanized form.

Orthography and Phonology. We noted earlier that the relationship between Malay orthography and phonology is more predictable (or regular) than it is for English. In Singapore and Malaysia, the grapheme-phoneme (letter- to- sound) correspondence is perfect for non-loan words with the exception of the letter ‘e’, which carries two phonological forms in standard Malay (but only one Indonesia). Although the grapheme-phoneme mappings are sometimes different from those in English (e.g., ‘c’ sounds like ‘ch’ in church), the same set of vowels and consonants are in use (except the letter ‘x’ which is found in foreign words).

Unlike English, there are only three diphthongs in Malay (ai, au and oi) and most syllables are very short. Possible syllable structures include v, vc, vcc, cv, cvv, cvc, ccv, cccv, and cvcc, but when the vowels ‘a’ and ‘i’ or ‘a’ and ‘u’ are adjacent in a closed syllable, they are pronounced with a syllable boundary between them, e.g., lain (other) and laut (sea) are both two syllable words.

Morphology. We also noted earlier that affixation is more common in Malay than in English. At least nine prefixes (meN², beR-, teR-, di-, peN-, peR-, se-, ke-, and menpeR-), three suffixes (-kan, -i, -an), four infixes (-el-, -er-, -em-, and -in-) and thirteen circumfixes (meN-...-kan, di-...-kan, meN-...-i, di-...-i, beR-...-an, beR-...-kan, peN-...-an, peR-...-an, ke-...-an, menpeR-...-kan, diper-...-kan, menpeR-...-i, and diper-...-i) are widely used. Like affixation, reduplication³ is also common in Malay, and so the nature of the morphology makes the number of letters and syllables per word much higher than in English: average word length is about seven letters, with about three syllables, but some words are much longer. For example, the 18-letter, 8-syllable word membahagi-bahagikan (to divide into smaller parts!) comprises a prefix, a reduplicated stem, and a suffix. See Karim, Onn, Musa and Mahmood (1996) for further details. Each of these affixes serves at least one semantic function depending on the word class (noun, adjective or verb) of the stem, and the word class often changes after affixation. For example, when the prefix beR-is attached to a noun payung ‘umbrella’, the resulting affixed word berpayung is a verb ‘using the umbrella’.

This kind of morphological structure suggests that the meaning of a relatively unfamiliar affixed word such as menderas ‘move quickly can usually be predicted by putting the meanings of its more familiar stem, derais ‘quick’ and its affix together, and conversely, the meaning of a relatively unfamiliar stem such as sarap ‘have breakfast’ can be predicted by splitting the more familiar affixed word sarapan ‘breakfast’ into its stem and affix. However, exceptions occur when the meaning of a stem or affixed form cannot

² Depending on the initial letter of the stem to which the affix is attached, there may be morphological assimilation, e.g., meN + buka = membuka. In some instances, the initial letter of the stem word is dropped after affixation, e.g., peN + tari = penari. For more details on morphological assimilation, refer to Koh (1978).

³ Reduplication is used to express indefinite plurality, intensity and repetition. Refer to Heah (1989), for more details.
be deduced by adding or subtracting the meanings of the components, e.g., salin ‘to change’ and bersalin ‘to give birth’.

These patterns, and the exceptions, strongly suggest that the relationship between familiarity and morphology in Malay is likely to be particularly important for the development of language assessment tools and empirical research (see Lee, 2000). For this reason, we decided to start developing a psycholinguistic database by focusing on familiarity and morphology. In this paper, we describe how we collected mean familiarity ratings from a large, representative sample (N=4,328) of stem and affixed words, and then for illustrative purposes, we analysed the subset of 36 stems which carried the greatest number of different inflectional forms.

4 Method
Using Toglia and Battig's methodology (Ibid.), a 7-point familiarity rating scale was constructed in questionnaire form, in which 1 = very unfamiliar and 7 = very familiar. All the Malay words were typed in lower case and order of presentation in the list was randomised.

Participants. Thirty-five Malaysian first language speakers of Malay between the ages of 18 and 39 participated in this study and were paid RM 50.00 (USD $14) each for their help.

Materials. To ensure that the participants would have been exposed to most of the words at some time, a corpus of 4,328 different lemma were compiled from the Primary 1 to Primary 3 textbooks used in Malaysian schools plus and from a Malay-English dictionary (Kamus Harian Federal, Daud 1945) that is used by Primary 4 pupils to lower Secondary school level (i.e., vocabulary of 6 – 14 year olds). In the final corpus, 2,510 items were stem words, 1,625 were stem words with affixes, 123 were reduplicated words, and 70 were reduplicated words with affixes. Word length ranged from two to eighteen letters (mean = 6.924; s.d. = 2.237), and from one to eight syllables (mean = 2.77; s.d. = 0.869).

Procedure. Participants were asked to rate each of the 4,328 words in the corpus on the 7-point familiarity scale using the following instructions (written in Malay) which are similar to those of Toglia and Battig:

For every individual, there are words that are frequently encountered and experienced in daily life. The aim of this experiment is to collect ratings on the following list of words according to their familiarity. For words that you frequently encounter or experience, a high rating should be given. Conversely, for words that you seldom encounter or experience, a low rating should be given. Circle one of the numbers given on the 7-point scale to indicate your own familiarity with the words given in the list. If you do not recognize a given word, you are required to cancel it out. Please try to use all 7 numbers on the scale whilst giving your familiarity ratings. Do not limit yourself to using only 3 and 4, or 1 and 7. There are no correct answers in this familiarity rating task.

5 Results
For each of the 4,328 words, the ratings given by all 35 subjects were pooled together, and means and standard deviations were calculated. A subset of 116 mean ratings, comprising the 36 stem verbs with their affixed forms, was then extracted to enable a more detailed