Behavior Research Methods
2010, 42 (4), 992-1003
doi:10.3758/BRM.42.4.992

The Malay Lexicon Project: A database of lexical statistics for 9,592 words

Melvin J. Yap and Susan J. Rickard Liow
National University of Singapore, Republic of Singapore

Sajlia Binte Jalil
Changi General Hospital

and

Siti Syuhada Binte Faizal
Washington University, St. Louis, Missouri

Malay, a language spoken by 250 million people, has a shallow alphabetic orthography, simple syllable structures, and transparent affixation—characteristics that contrast sharply with those of English. In the present article, we first compare the letter–phoneme and letter–syllable ratios for a sample of alphabetic orthographies to highlight the importance of separating language-specific from language-universal reading processes. Then, in order to develop a better understanding of word recognition in orthographies with more consistent mappings to phonology than English, we compiled a database of lexical variables (letter length, syllable length, phoneme length, morpheme length, word frequency, orthographic and phonological neighborhood sizes, and orthographic and phonological Levenshtein distances) for 9,592 Malay words. Separate hierarchical regression analyses for Malay and English revealed how the consistency of orthography–phonology mappings selectively modulates the effects of different lexical variables on lexical decision and speeded pronunciation performance. The database of lexical and behavioral measures for Malay is available at http://brm.psychonomic-journals.org/content/supplemental.

There is very little empirical work on the cognitive processes involved in Malay reading and spelling, even though mutually intelligible forms are spoken by about 250 million people living in Indonesia, Malaysia, Brunei, and Singapore (Tadmor, 2009). *Rumi*, the most prevalent form of written Malay, has a relatively shallow alphabetic orthography, simple syllable structures, and transparent affixation. For English, and for many other European languages, it is clear that each of these characteristics influences both the development of literacy in children (Caravolas, 2004; Ellis & Hooper, 2001; Seymour, Aro, & Erskine, 2003) and skilled processing in adults (see Balota, Yap, & Cortese, 2006, for a review; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001, on English; Ziegler, Perry, & Coltheart, 2000, on German). To support future research aimed at separating language-specific from language-universal processing, and to develop a better understanding of word recognition in orthographies that also have relatively consistent mappings to phonology, we compiled a database of lexical and behavioral measures for a large set of Malay words. In the present article, we will discuss the background and motivation for this database, as well as compare and contrast the effects of different lexical variables on English and Malay word recognition performance.

Background

Visual word recognition studies have played a central role in experimental cognitive psychology over the past 4 decades, informing domains as diverse as psycholinguistics (Andrews, 2006), computational modeling (Coltheart et al., 2001; Perry, Ziegler, & Zorzi, 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996), automatic versus attentional processing (Neely, 1977), and the neural correlates of language processing (Petersen, Fox, Posner, Mintun, & Raichle, 1989). Thus far, the vast majority of the empirical work has been directed at word recognition in English, but there is a growing interest in crosslinguistic work because of the potential pedagogical implications (e.g., Caravolas & Bruck, 1993, on Czech; Durgunoglu & Oney, 1999, on Turkish; Lervåg, Bråten, & Hulme, 2009, on Norwegian; see Ziegler & Goswami, 2005, for a review). Although many critical issues have been addressed by studying the effects of different psycholinguistic variables on visual word recognition tasks, such as lexical decision and speeded pronunciation, it re-

M. J. Yap, melvin@nus.edu.sg

mains unclear whether the findings are specific to English orthography, because its mappings to phonology are atypical (see Share, 2008, for a review). For this reason, many important questions about cognitive–linguistic processing can be satisfactorily answered only by making comparisons between alphabetic languages whose lexical characteristics contrast with those of English.

In an early example of this type of crosslinguistic work, Frost, Katz, and Bentin (1987; see also Frost & Katz, 1992) investigated how reliance on different reading mechanisms can be modulated by the depth of a language's orthography (Lukatela, Popadić, Ognjenović, & Turvey, 1980). According to the dual-route model of reading for English (Coltheart et al., 2001), there are two mechanisms mediating word recognition: a serial, frequency-insensitive nonlexical mechanism that assembles pronunciations of words via a limited set of spelling-to-sound rules (i.e., assembled phonology), and a parallel lexical mechanism that maps orthographic strings onto lexical representations (i.e., addressed phonology) and is sensitive to word frequency and semantic variables.

In *shallow* orthographies (such as Finnish and Serbo-Croatian), there is an isomorphic relationship between spelling and sound—that is, the mappings between orthography and phonology are transparent and predictable. In *deep* orthographies (such as French and English), this mapping is more complex—that is, the same graphemes represent different sounds across different contexts. Frost et al. (1987) designed a study that treated orthographic depth as a continuum, with Hebrew as a deep orthography, Serbo-Croatian as a shallow orthography, and English in between. With adult native speakers of these three languages as participants, they found that semantic priming and word-frequency effects in naming times were positively correlated with the depth of the orthography. Specifically, priming and frequency effects were strong in Hebrew, moderate in English, and absent in Serbo-Croatian, leading them to suggest that phonology is predominantly assembled via the nonlexical mechanism in shallow orthographies such as Serbo-Croatian, but is predominantly addressed in deep orthographies such as English.

For any language, a basic assumption is that the degree of adherence to the alphabetic principle determines whether the underlying processing is similar when readers decode print into sound during reading. This relationship between orthography and phonology varies considerably across European languages, ranging from isomorphic single-grapheme to single-phoneme mappings to higher order correspondences involving orthographic sequences (Borgwaldt, Hellwig, & De Groot, 2005; Seymour et al., 2003). These sequences sometimes represent larger orthographic units (rimes or syllables) that are neither predictable from simple grapheme–phoneme rules nor consistent across words. For example, in English, *-eigh* has different pronunciations in *weight* and *height* (see Spencer, 2009, for more discussion of sonograph metrics). Furthermore, although morphology is said to be preserved at the expense of orthography–phonology consistency (e.g., mean, meant), and larger orthographic sequences can mark word class (e.g., *-tion* usually denotes a noun in English), many plurals and past tense verbs are irregular, or are less transparent in terms of meaning (e.g., *foot–feet*; *sleep–slept*).

The relationship between orthography and phonology, and between orthography and morphology, determines how many rules children need to learn and apply to become proficient readers and spellers. Word frequency is a potential confound in direct comparisons across orthographies, but Ellis and Hooper (2001) matched Welsh and English target words in terms of written exposure to create parallel reading tasks. As predicted, their results showed that the Year 2 Welsh-educated children read more accurately than the English-educated children in their respective languages. This is consistent with the fact that the grapheme–phoneme mappings for Welsh are much more transparent than those for English, but more importantly, they also found that word length determined 70% of the variance in Welsh reading latencies, but only 22% in English. This, together with the higher proportion of nonword responses for the Welsh-educated children, indicated that they were relying more heavily on serial nonlexical decoding, (i.e., assembled phonology, which produces strong length effects), whereas the English-educated children were obliged to rely more on addressed phonology.

Similarly, as part of a crosslinguistic study of lexical variables that might influence spelling development in alphabetic orthographies, Caravolas (2004) compared English, French, and Czech. She found that learning to spell was less demanding in Czech than in French and less demanding in French than in English. This is consistent with the pattern of results found by Seymour et al. (2003), who explored the development of reading in native speakers of 13 different European orthographies and found that the rate of acquisition for common words was up to 2 years slower for children learning to read English than for those learning shallow orthographies such as Finnish.

Concepts such as grapheme–phoneme consistency (Ellis & Hooper, 2001), syllabic complexity (Seymour et al., 2003), and grain size (Ziegler & Goswami, 2005) have been invoked to capture the complexity of processing in English in comparison with that in other languages. However, these notions are difficult to quantify objectively. For example, when Seymour et al. rank ordered 13 European alphabetic languages with respect to their orthographic depth, this hierarchy was based on subjective ratings that were obtained using a questionnaire in which co-authors estimated the orthographic depth of their respective languages. Although it is clear that learning to read English words makes more demands on processing than do most other alphabetic orthographies, more objective benchmarks are needed for reliable crosslinguistic comparisons. The simplest unambiguous metric that can be used as a proxy for the number of different mappings is the ratio of the number of letters to the number of phonemes in that language. There are regional variations in phonology for most languages, but this ratio provides a reasonably objective index for making predictions about the balance that skilled readers might strike between nonlexical rule-based decoding (assembled phonology) and lexical mappings (addressed phonology) for a specific orthography.

Table 1 lists the letter–phoneme ratios for Malay and eight other alphabetic languages. Although it is possible that languages with letter–phoneme ratios of 1.00 could be deep orthographies (many-to-many), the data in Table 1 suggest they are more likely to be shallow. All of the lexical statistics for Malay were computed using the 9,592-word database that is described in the present article, but the ratios shown in Table 1 were based on data collated from published papers (Dutch, English, French, German, Hungarian, and Italian, Borgwaldt et al., 2005; Czech, Caravolas, 2004; Finnish, Seymour et al., 2003; Serbo-Croatian, Frost et al., 1987).

As Table 1 shows, when the letter–phoneme ratios for these nine orthographies are considered, there is evidence of different clusters of languages for both vowel and consonant mappings, rather than a simple shallow–deep dichotomy. Except for Italian, the vowel letter-to-phoneme ratios appear to determine orthographic depth, and the rank order (see first column of Table 1) concurs with previous claims that English is an atypically deep alphabetic orthography (see Share, 2008), whereas Serbo-Croatian (Frost et al., 1987) and Finnish (Leppanen, Niemi, Aunola, & Nurmi, 2006) are shallow as compared with French, German, and Dutch (e.g., Seymour et al., 2003). This suggests that Malay might be a particularly useful language for contrasting against both Romance (Spanish, French, and Italian) and Germanic (Dutch, German, and English) languages.

The main point of such a comparison is that vowel letter-to-phoneme number ratios provide a simple but relatively objective heuristic for scaling orthographic depth, clustering languages, and making predictions about processing demands. Given Ellis and Hooper's (2001) data comparison of processing in Welsh and English, Table 1 suggests that reading might require more lexical support in English, German, and Dutch than in French, Malay, or Spanish, and that Finnish and Serbo-Croatian can be read aloud using only nonlexical rules. The critical corollary of these differences in processing demands is that the relative importance of the lexical variables germane to models of reading in deeper orthographies, such as English, might be less salient for readers of shallower orthographies such as that of Malay. Obviously, the letter-to-phoneme ratio is at best a crude metric for quantifying orthographic depth, and more sophisticated and potentially superior approaches exist, in principle. For example, one could compute the ratio of number of graphemes to number of phonemes, or the consistency of grapheme-to-phoneme and phoneme-to-grapheme mappings. However, the definition of a grapheme, as compared with a letter, is more ambiguous, and so identifying the number of graphemes for most of the languages featured in Table 1 would be unreliable.

In addition to orthographic depth, Seymour et al. (2003) also classified the 13 European orthographies according to their syllable complexity. This dimension refers to whether syllables are usually open with few consonant clusters (e.g., Italian) or are closed with initial and/or final consonant clusters (e.g., German and English), and it was used to better predict the likely processing demands for beginning readers. Grain size, a similar construct, refers to the size of the sublexical representations that readers need to learn and use for optimal processing (see Wydell & Butterworth, 1999; Ziegler & Goswami, 2005), and ranges from small units, such as phonemes, to larger units, such as rimes or syllables. Both grain size and syllable complexity appear to be related to orthographic depth. Readers of English (and of other deeper orthographies) are likely to rely on larger units for sublexical processing, since these larger units possess more consistent mappings to phonology (see Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995, for more discussion).

Although syllable complexity and grain size are potentially useful dimensions for crosslinguistic comparisons, they, like orthographic depth, are also difficult to quantify objectively. This is because phonotactic constraints differ across languages, and reader skill and experience will determine performance on graphemic parsing tasks. We therefore computed a proxy for syllable complexity/grain size that was based on the ratio of number of letters to number of syllables for the subset of languages with acces-

Table 1
Estimated Letter–Phoneme Ratios for Nine Alphabetic Orthographies As a Function of Vowels, Consonants, and All Letters

Language	Ratio of Vowels/ Phonemes		Ratio of Consonants/ Phonemes		Ratio of Letters/ Phonemes	
Serbo-Croatian	5/5	1.00	25/25	1.00	30/30	1.00
Finnish	8/8	1.00	18/20	0.90	26/28	0.93
Malay	**6/7**	**0.86**	**19/27**	**0.70**	**25/34**	**0.74**
Spanish	6/7	0.86	21/18	1.17	27/25	1.08
French	12/16	0.75	21/20	1.05	33/36	0.92
Italian	5/7	0.71	18/43	0.42	23/50	0.46
Dutch	10/19	0.53	20/22	0.91	30/41	0.73
German	9/19	0.47	20/24	0.83	29/43	0.67
English	**6/20**	**0.30**	**20/24**	**0.83**	**26/44**	**0.59**
M	7.4/12.0	**0.72**	20.2/24.8	**0.87**	27.7/36.8	0.79
SD	2.5/6.3	0.24	2.1/7.4	0.22	3.1/8.3	0.20

Note—The ratio that is based on the number of vowel letters to the number of vowel phonemes is used to rank-order the languages, so that orthographically deeper languages have a lower rank.

sible databases: English (English Lexicon Project, Balota et al., 2007), French (Lexique 2, New, Pallier, Brysbaert, & Ferrand, 2004; see also the French Lexicon project, Ferrand et al., 2010), German and Dutch (CELEX, Baayen, Piepenbrock, & van Rijn, 1993), and Malay (database described in the present article).

Table 2 shows the mean number of letters per word, the mean number of syllables per word, and the mean ratio of number of letters to number of syllables across words, for the five alphabetic orthographies. Again, the ratio of number of letters to number of syllables (see final column of Table 2) vastly simplifies the concept of syllable complexity. For example, it does not take into account the number of morphemes in each word. However, because the metric reflects the presence of consonant clusters and vowel digraphs, the number of letters to number of syllables ratio does provide a means of objectively quantifying an important aspect of the processing demands faced by readers. More importantly for our purposes, it is also clear, again, that Malay orthography provides a contrast to English because its words have more syllables but fewer letters per syllable. This is partly because of differences in the relationship between English and Malay in terms of the syllabic structure of root and bound morphemes. Malay, unlike English, is agglutinative, and affixes are simple orthographic units such as CVC, CV, or VC syllables (e.g., per-, ber, di-, -kan; see below). In other words, the nature of Malay morphology contributes to the low ratio of number of letters to number of syllables.

Although the body of knowledge based on crosslinguistic research is growing rapidly, there are still no systematic empirical comparisons across alphabetic orthographies that are based on megastudy corpora that include behavioral data for lexical decision and speeded pronunciation (e.g., English Lexicon Project, Balota et al., 2007). Thus, the Malay Lexicon Project reported presently provides potentially important insights into the processing characteristics of shallower orthographies (see Table 1) with simpler syllable structures (see Table 2).

THE MALAY LANGUAGE

Malay, or *Bahasa Melayu*, refers to a group of languages that belong to the Malayo-Polynesian branch of the Austronesian family of languages. Unlike English, Malay has shallow orthography–phonology mappings, transparent morphology, and simple, short syllabic structures.

Table 2
Means and Standard Deviations for the Number of Letters Per Word, and Syllables Per Word, and the Ratio of Letters to Syllables, for Five Alphabetic Orthographies

Language	Sample Size	Letters/Word		Syllables/Word		Letters/Syllable	
		M	*SD*	*M*	*SD*	*M*	*SD*
Malay	9,592	7.56	2.51	3.00	0.97	2.54	0.41
Dutch	117,867	10.60	3.54	3.37	1.28	3.27	0.69
German	50,658	10.12	3.26	3.21	1.14	3.27	0.75
English	38,477	7.98	2.46	2.55	1.11	3.41	0.99
French	38,335	8.23	2.29	2.54	0.89	3.45	1.00

We now turn to a description of Malay's various lexical characteristics.

Orthography

Malay is usually written in Rumi, a Latin alphabet, although Jawi, an adapted Arabic system, is used for Islamic teaching and some cultural proceedings. Rumi has five simple vowels and 20 consonants; <x> is not used, and <q> and <v> are found only in foreign loanwords. There are five digraphs (<gh>, <kh>, <ny>, <ng>, and <sy>) and three diphthongs (<ai>, <au>, <ua>) (see Awang, 2004; Lee, 2008).

Phonology

We examined three pronunciation systems initially—Standard Baku, Singapore Baku, and Nonstandard (colloquial) Malay—but the data we report are based on Standard Baku, which is serviceable across Southeast Asia. The three forms are mutually intelligible, but the drive toward Standard Baku (Mohamed, 2003) suggests that it will be the dominant form in the future. Regardless of the type of phonology, there are six vowel phonemes (/a/, /ə/, /e/, /i/, /o/, /u/), three diphthongs (/ai/, /au/, and /oi/), and 28 consonant phonemes (e.g., /b/, /d/, /m/, /x/-<kh>, /ɣ/-<gh>, /ʃ/-<sy>, etc.) in Malay.

Orthography–Phonology Mappings

Most syllables are very short, and possible structures include V, CV, VCC, CV, CVV, CVC, CVCC, and CCVC, the most common being CV, CVC, and CVCC. There are 25 letters and 34 phonemes in Malay, as compared with 26 letters and 44 phonemes in English, but it is the ratio of vowel letters to vowel phonemes that is likely to influence performance in naming tasks. Malay and English both have six vowel letters, but Malay has only 7 vowel phonemes, as compared with 20 in English (see Table 1). The only exception for vowels is <e>, which takes two phonemic forms (/ə/ or /e/), and whenever <a> and <i> or <a> and <u> are adjacent in a closed syllable, they are pronounced with a syllable boundary between them instead of as a diphthong; for example, *ka-in* ("cloth") and *ba-u* ("smell") are both two-syllable words. The simple syllable structures (mostly CVC or CV) are also easy to decode during oral reading and to encode during spelling (see Rickard Liow & Lee, 2004, on spelling), because there are so few consonant clusters and no vowel digraphs.

Morphology

There are three main morphological processes in Malay: affixation, reduplication, and compounding (see Karim, Onn, Haji Musa, & Mahmood, 2008, for a detailed overview). Malay is an agglutinative language; prefixes and suffixes are used extensively to express grammatical relationships and to form new words. Malay affixation is relatively transparent and rule based. Affixes for deriving nouns include <peN->, <pe->, <-an>, <ke-an>, <per-an>, and <peN-an>, and verb inflections include <meN->, <beR->, <ter->, <peR->, <-I>, and <-kan> (Hassan, 1974). Some affixes change according to the stem that follows; for example, <meN-> changes to <menge->

for monosyllabic words, and to <meng->, <mem->, <meny->, and <me-> for other stems. In general, prefixes and suffixes are appended to the morphemic stem; for example, the noun prefix <pe-> is simply added to *nyanyi* ("to sing") to form the affixed word *penyanyi* ("singer"). However, phonotactic constraints reduce the transparency of some affixed words whenever the morphemic stem's initial letter is changed to facilitate pronunciation. More specifically, all stems starting with the letters <p>, <k>, <t>, or <s> are changed when a prefix is added to the noun; for example, when the noun prefix <peN-> is added to the stem word *pegang* ("to hold"), the suffix changes to <peM->, whereas the first letter <*p*> of the stem is dropped to form the affixed word *pemegang* ("holder"). Other examples include <men-> + *kenal* → *mengenal* ("recognize"), and <men-> + *sapu* → *menyapu* ("sweep"). Reduplication, sometimes used to mark plurals, is also common in Malay—for example, *kupu-kupu* ("butterfly") and *kanak-kanak* ("children"). Examples of compounding, in which words contain more than one stem, include *tanggungjawab* ("responsibility"), which is made up of *tanggung* ("to bear the burden of") and *jawab* ("to answer"), literally meaning to bear the burden of answering for an action, and *kereta api* ("train"), which is made up of *kereta* ("car") and *api* ("fire"), literally meaning a car that runs on fire.

Malay's simple syllabic structure could be an indication that syllable boundaries are salient in word recognition. However, some affixes pull the first/last letter of the stem word, whereas others push the first/last letter to the stem word. For example, *makanan* ("food") is made up of *makan* ("to eat") and the suffix <-an>, but the word is syllabified as *ma-ka-nan* (suffix pulls the last letter of stem word), thus changing the derived form and breaking the stem word (Koh, 1978). So, although Malay has very consistent orthography–phonology mappings, and all affixes consist of discrete syllables, the syllable structure may make the phonology less accessible than the orthography seems to indicate. At the same time, Malay readers must also be sensitive to the irregular <e> and how the vowel digraph phonology is split across a syllable boundary.

DEVELOPING A MALAY LEXICAL DATABASE

Despite Malay's large user base of about 250 million people (Tadmor, 2009) and its interesting linguistic properties, published experiments on the cognitive processing of Malay are relatively sparse and mostly are concerned with children's spelling (e.g., Jalil & Rickard Liow, 2008; Lee, 2008; Rickard Liow & Lee, 2004; Rickard Liow, Yap, Lee, & Ramos, 2008; Winskel & Widjaja, 2007). One major impediment to the development of research on skilled adult reading, and hence to reliable crosslinguistic comparisons with English, has been the lack of psycholinguistic norms (e.g., word frequency, neighborhood density, length) that are needed to support the design of empirical studies. Over the past decade, this limitation has been ameliorated for some of the major European languages besides English (English Lexicon Project, Balota et al., 2007; CELEX for Dutch and German, Baayen et al., 1993; Lexique 2 for French, New et al., 2004; Greek, Ktori, Van Heuven, & Pitchford, 2008; and Spanish, Algarabel, Ruiz, & Sanmartín, 1988). These repositories of language-specific lexical variables have enabled researchers to extend models of word recognition (e.g., DRC Model for German, Ziegler et al., 2000).

To support research on an alphabetic Asian language that contrasts with English and with many European languages (see Tables 1 and 2), we started with a list of candidate words that were identified through examining corpora based on the two main Malay-language daily newspapers published in Malaysia and Singapore.

1. *Berita Harian*, published by The New Straits Times Press (Malaysia) Berhad. The corpus (n = 2.14 million) is based on a database compiled by Dewan Bahasa dan Pustaka, including 2 years' worth of articles (2001–2002) from three sections of the Berita Harian: *hiburan* ("entertainment"), *ekonomi* ("economy"), and *sukan* ("sports").

2. *Berita Harian*, published by Singapore Press Holdings. The corpus (n = 5 million) compiles 2 years' worth of articles (2006–2007) from all sections of the Berita Harian: *dalam negeri* ("local news"), *luar negeri* ("world news"), *ekoniaga* ("business"), *hiburan* ("entertainment"), and *sukan* ("sports").

Both lists underwent a preliminary screening process by two native speakers of Malay whereby typographical errors (which resulted in nonwords) and letter strings with nonalphabetic characters were removed. This left us with a master list of 14,849 words. We then eliminated proper nouns (n = 407, 7.7%) and foreign loanwords (n = 1,637, 31.1%), which are often cognates (e.g., *produktif*, *stadium*, *zoo*, *sirap*, *industri*). Words with the following four affixes (n = 3,098, 58.9%) were also eliminated: the suffixes <-nya> (denoting "his/her") and <-kan> (used to formally tell someone to do something), and the prefixes <di-> (the passive form) and <se-> (denoting "one"). The morphology of these affixed forms is transparent, and they represent phrases rather than single words. For example, *bukunya* = *buku* ("book") + *nya* ("his/her") = his/her book, *bukakan* = "open" (for me), *dilakukan* = "was done" (by), and *sepotong* = (a) "slice." A small number of onomatopoeic words (n = 28, 0.53%), slang/truncations (n = 59, 1.12%), and conjunctions/prepositions (n = 28, 0.53%) were also removed to reduce ambiguity. Thus, in what follows, we describe the various lexical characteristics that were computed for the remaining master list of 9,592 words.

Database Descriptives

The descriptive statistics for the measures (described below) computed in Malay (see top panel of Table 3) are provided with equivalent data for English for comparison (see bottom panel of Table 3). The reference English data set is made up of approximately 35,000 words from the English Lexicon Project (Balota et al., 2007), which has values for the variables of interest. Zero-order correlations between the measures for Malay are presented in Tables 4A and 4B (see Yap & Balota, 2009, for a similar correlation coefficient matrix for English monomorphemic multisyllabic words).

Table 3
Means and Standard Deviations for the Lexical Variables in the Malay and English Databases

Variables	*M*	*SD*
Malay (n = 9,592)		
Frequency (Singaporean): Observations/million	79.81	410.71
Log frequency (Singaporean)	1.07	0.77
Frequency (Malaysian): Observations/million	59.74	308.09
Log frequency (Malaysian)	0.96	0.71
Letter length	7.56	2.51
Syllable length	3.00	0.97
Phoneme length	7.33	2.39
Morpheme length	1.59	0.52
Orthographic neighborhood size	1.70	2.71
Phonological neighborhood size	1.78	2.86
Orthographic Levenshtein distance 20	2.45	0.85
Phonological Levenshtein distance 20	2.48	0.84
English (n = 35,043)		
Frequency (SUBTL): Observations/million	25.55	470.75
Log frequency (SUBTL)	0.44	0.53
Letter length	7.82	2.40
Syllable length	2.49	1.07
Phoneme length	6.56	2.23
Morpheme length	2.06	0.84
Orthographic neighborhood size	1.43	2.88
Phonological neighborhood size	3.09	6.21
Orthographic Levenshtein distance 20	2.78	0.99
Phonological Levenshtein distance 20	2.76	1.18

Note—The measures for English were obtained from the English Lexicon Project (Balota et al., 2007), and the frequency measure for English was based on Brysbaert and New's (2009) subtitle (SUBTL) frequency estimate.

Word frequency (Malaysian). These frequency counts are based on a 2.14-million-word corpus that is based on articles in a major daily Malaysian newspaper *Berita Harian* (http://bharian.com.my/), published in Kuala Lumpur. Raw and log-transformed counts are provided.

Word frequency (Singaporean). These frequency counts are based on a 5-million-word corpus that is based on articles in a major daily Singaporean newspaper, also called *Berita Harian* (http://cyberita.asia1.com.sg/), but with different text and published in Singapore. Raw and log-transformed counts are provided.

Pronunciation. Pronunciations for Standard Baku, Singapore Baku, and Nonstandard Malay are provided in the database. For most words, pronunciations are identical across the three systems, but the analyses refer to Standard Baku.

Length measures. Number of morphemes, syllables, letters, and phonemes are provided for each word.

Orthographic neighborhood size. This is a measure of orthographic distinctiveness and reflects the number of words that can be obtained by changing one letter while preserving the identity and positions of the other letters (Coltheart, Davelaar, Jonasson, & Besner, 1977; Davis, 2005). For example, the neighbors of *aba* ("command") include *apa* ("what"), *asa* ("hope"), and *abu* ("ashes").

Phonological neighborhood size. This is the phonological analogue of orthographic neighborhood size and reflects the number of words that can be obtained by changing one phoneme while preserving the identity and positions of the other phonemes (Yates, 2005; Yates, Locker, & Simpson, 2004). Again, separate counts are provided for each of the three pronunciation systems, but the analyses refer to Standard Baku.

Orthographic Levenshtein distance 20. Orthographic Levenshtein distance 20 (OLD20) is a relatively new measure of orthographic distinctiveness that is optimized for longer words (Yarkoni, Balota, & Yap, 2008). This measure is based on Levenshtein distance (LD), a computer science metric that is based on the minimum number of substitution, insertion, or deletion operations required to convert one string of elements (either letters of phonemes) into another. For example, the LD from *kitten* to *sitting* is 3, reflecting two substitutions ($k \rightarrow s$, $e \rightarrow i$) and one insertion (insert g at the end). We used the LD calculator (www.talyarkoni.org/materials.php) to generate OLD20 measures for our words. Essentially, the program first computed the LD from each of our 9,592 words to every other word in the database, and this was used to generate OLD20 values for each word, defined as the mean LD from a word and its 20 closest neighbors. Words with higher mean LD values are further from their closest neighbors, implying that they are more orthographically or phonologically distinct. This metric may be particularly useful for models of word recognition in Malay, given that it is an agglutinative language and that a large number of words are quite long, as compared with those in English. Traditional neighborhood size measures have limited utility for long words, which have few or no neighbors. Measures that are based on LD circumvent this limitation by providing estimates of distinctiveness for even very long words. Interestingly, Yarkoni et al. have reported that LD-based measures of orthographic distinctiveness provide a significant advantage over traditional density-based measures in predicting performance on English word recognition tasks, particularly for longer words.

Phonological Levenshtein distance 20 (PLD20). The OLD20 measure (described previously) is based on the mean LD from the orthographic form of a word and the spellings of its 20 closest neighbors. We also computed the phonological Levenshtein distance 20 (PLD20) for each word (Yap & Balota, 2009), which is essentially like OLD20, except that this captures the mean LD from the phonological form of a word and the pronunciations of its 20 closest neighbors. Separate estimates are provided for each of the three pronunciation systems.

BEHAVIORAL DATA

Predictive Power of Lexical Measures

Using multiple regression analyses, we then assessed the validity of these measures by using them to predict speeded lexical decision and speeded pronunciation performance for a large set of 1,510 mono- and multisyllabic Malay words. To the extent that these measures predict variance in word recognition performance, we can be more confident that they are tapping task-general processes of interest. We now turn to a description of how the behavioral data were collected.

Table 4A
Correlations Between the Lexical Variables in the Malay Database

Variables for Malay	1	2	3	4	5	6
1. Frequency (Singaporean): Observations/million	–	.431***	.898***	.415***	−.106***	−.070***
2. Log frequency (Singaporean)		–	.393***	.809***	−.096***	−.038***
3. Frequency (Malaysian): Observations/million			–	.472***	−.098***	−.063***
4. Log frequency (Malaysian)				–	−.091***	−.041***
5. Letter length					–	.891***
6. Syllable length						–
7. Phoneme length						
8. Morpheme length						
9. Orthographic neighborhood size						
10. Phonological neighborhood size						
11. Orthographic Levenshtein distance 20						
12. Phonological Levenshtein distance						

***$p < .001$.

Table 4B
Correlations Between the Variables in the Malay Database

Variables for Malay	7	8	9	10	11	12
1. Frequency (Singaporean): Observations/million	−.099***	−.089***	.065***	.067***	−.075***	−.073***
2. Log frequency (Singaporean)	−.081***	−.069***	.039***	.040***	−.101***	−.093***
3. Frequency (Malaysian): Observations/million	−.091***	−.073***	.065***	.066***	−.075***	−.073***
4. Log frequency (Malaysian)	−.077***	−.049***	.064***	.065***	−.116***	−.108***
5. Letter length	.968***	.751***	−.518***	−.512***	.768***	.766***
6. Syllable length	.934***	.752***	−.493***	−.496***	.725***	.768***
7. Phoneme length	–	.774***	−.527***	−.525***	.747***	.780***
8. Morpheme length		–	−.447***	−.450***	.471***	.503***
9. Orthographic neighborhood size			–	.962***	−.536***	−.554***
10. Phonological neighborhood size				–	−.527***	−.563***
11. Orthographic Levenshtein distance 20					–	.959***
12. Phonological Levenshtein distance						–

***$p < .001$.

METHOD

Participants

Forty-four skilled Malay speakers (mean age = 22.7 years) were recruited from two university communities in Singapore to take part in both the lexical decision and speeded pronunciation tasks. There were a total of three 2-h sessions, spaced at least 1 week apart. All participants spoke English in addition to Malay, had normal or corrected-to-normal vision, and were paid SGD20 after each session.

Apparatus

Stimuli were presented on a 17-in. Viewsonic CRT monitor with a refresh rate of 85 Hz and a resolution of 1,024 × 768 pixels; this was controlled by a Pentium 4 3-GHz PC. Stimuli were presented in lowercase in 20-point Arial font, and they appeared as white characters on a black background. A microphone (Audio Technica ATR20) was connected to the PST serial response box (Schneider, Eschman, & Zuccolotto, 2001), which served as a voice key. Vocal responses were also recorded by a Samsung MP3 digital voice recorder so that the accuracy of participants' responses could be verified offline by an experimenter.

Stimuli

The word stimuli for both tasks consisted of the same 1,520 words selected from the master list of 9,592 words. Of these 1,520 items, 570 items were made up of three groups of 190 words each that consisted of morphemic stems in three different forms: the morphemic stems, the same stems with noun affixation "*pe . . . an*," and the same stems with verb affixation "*me . . . kan*." These triplets, naturally occurring in the language, were counterbalanced across groups of stimuli to avoid stem repetition. The remaining stimuli consisted of 570 stems, 190 noun-affixed "*pe . . . an*" words, and 190 verb-affixed "*me . . . kan*" words that were randomly selected. Pronounceable nonwords for the lexical decision task (n = 1,520) were then constructed by replacing a letter in a corresponding target word with another. Care was taken to ensure that the nonwords were legal and true nonwords by asking a native Malay speaker to check the words against a Malay dictionary, *Kamus Dewan* (4th ed.; Dewan Bahasa, 2007). The stimuli for the speeded pronunciation task were limited to words.

Procedure

Both lexical decision and speeded pronunciation tasks were administered during each session, but counterbalancing ensured that participants were not given identical lists of stimuli within a given session. All participants were tested individually in sound-attenuated cubicles, and each experimental session comprised 18 to 20 blocks of 80 trials (1,020 trials for lexical decision and 510 trials for speeded pronunciation). Each block of 80 trials was followed by a mandatory 3-min rest break. At the end of the first task, participants rested for 10 min before starting on the second task. There were 16 practice trials for the lexical decision task and 8 practice trials for the speeded pronunciation task. For both the lexical decision and speeded pronunciation tasks, E-Prime 1.2 and the PST serial response box (Schneider et al., 2001) were used for stimuli presentation and data collection.

Lexical decision task. In the lexical decision task, participants were instructed to indicate as quickly and as accurately as possible via a buttonpress on the response box whether a presented letter string formed a word or a nonword. Each trial started with a centered fixation point "+" for 500 msec, followed by a dark interval for 200 msec, followed by the target letter string, which stayed on until the participant responded or until 3,000 msec had elapsed. The intertrial interval was 750 msec. A tone was presented if the trial had timed out or if the response was incorrect. This tone was accompanied by the word "Salah" ("incorrect") if the participant pressed the wrong button.

Speeded pronunciation task. In the speeded pronunciation task, participants were instructed to read the target word aloud as quickly and as accurately as possible. Stimulus presentation conditions were

identical to the lexical decision task (see above), except that participants had to read words aloud into the microphone, which was connected to the voice key on the PST serial response box. Detection of the voice onset erased the word from the screen, and participants then left-clicked on the mouse if they thought that they had pronounced the word correctly, or right-clicked if they had mispronounced the word or if there was some technical issue (e.g., inadvertent triggering of the voice key). All vocal responses were also recorded on the MP3 digital voice recorder and were checked for accuracy.

RESULTS

Error trials (9.6% for lexical decision and 3.9% for speeded pronunciation) were first excluded. This was followed by a two-stage procedure in which response times (RTs) faster than 200 msec or slower than 3,000 msec were first removed, followed by trials that were 2.5 *SD*s above and below each participant's mean RT (see Balota et al., 2007, who used a similar trimming procedure). This removed an additional 2.6% of data for lexical decision and 3.8% for speeded pronunciation. Mean item latencies for speeded pronunciation and lexical decision were then computed on the basis of the remaining trials.

Hierarchical regression analyses were then computed, with the articulatory features of the initial phoneme (e.g., voiced, bilabial, alveolar, etc.) entered as dichotomous variables in the first step (see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Chateau & Jared, 2003; Treiman et al., 1995), followed by log-transformed word frequency (Singaporean), number of letters, orthographic neighborhood size, and orthographic Levenshtein distance.

For Malay, these variables accounted for a very substantial proportion of variance in lexical decision (55.5%) and speeded pronunciation (74.3%) latencies, and these estimates are comparable to, or larger than, the variance accounted for in English megastudy data (see Balota et al., 2004; Treiman et al., 1995; Yap & Balota, 2009). Length was clearly the best predictor of Malay word recognition times, followed by word frequency; shorter, more frequent words were recognized faster. This result is consistent with the idea that word recognition in Malay, as in Welsh (Ellis & Hooper, 2001), relies far more heavily on a serial, sublexical procedure than on a frequency-sensitive lexical procedure. More interestingly, effects of neighborhood structure (i.e., neighborhood density and orthographic Levenshtein distance) were much smaller and were reliable only in lexical decision. Specifically, words with fewer orthographic neighbors and more distant Levenshtein neighbors were responded to faster. This indicates that orthographically distinct words were recognized better, consistent with a competitive lexical identification process in Malay lexical decision, although it must be pointed out that these neighborhood effects were relatively small, as compared with length effects. Equally important, a parallel hierarchical regression analysis on 35,039 words[1] from the English Lexicon Project (see Table 5) revealed substantial differences in the relative contribution of the predictor variables to performance on lexical decision and speeded pronunciation tasks. Frequency effects were far more influential in English, and length effects were not reliable in lexical decision (see New, Ferrand, Pallier, & Brysbaert, 2006, for a discussion of length effects in English). The effect of Levenshtein neighbors was also strong and reliable across both tasks in English, whereby words with closer Levenshtein neighbors (i.e., words that are less distinct) were responded to faster. Although a detailed discussion of task-specific and language-specific effects is outside the scope of the present article, it is evident that models of word recognition in Malay, and perhaps of other shallow orthographies, would need to place more emphasis on assembled phonology and less emphasis on addressed phonology. This pattern is consistent with the letter–phoneme ratios (Table 1) and letter–syllable ratios (Table 2) for Malay and English.

Although the differences between the English and Malay behavioral results cannot be attributed to differences in sample size (see note 1), they may be driven by the specific lexical characteristics of the words in each data set. For example, as compared with the English words, the Malay words were much higher in word frequency and

Table 5
Standardized Response Time Regression Coefficients of the Item-Level Regression Analyses for Speeded Lexical Decision and Naming Performance for 1,510 Malay Words

	Lexical Decision		Speeded Pronunciation	
	Malay	English	Malay	English
Predictor Variables	n = 1,510	n = 35,039	n = 1,510	n = 35,043
	Surface Variables (Onsets)			
Adjusted R^2	**.176*****	**.020*****	**.500*****	**.046*****
	Lexical Variables			
Log frequency	−.470***	−.511***	−.157***	−.376***
Number of letters	.655***	−.005	.720***	.065***
Orthographic neighborhood size	.058*	.088***	−.014	.015**
Orthographic Levenshtein distance	−.072*	.424***	.010	.369***
Adjusted R^2	**.555*****	**.555*****	**.743*****	**.480*****
R^2 change	**.379*****	**.535*****	**.243*****	**.434*****

Note—The measures for English were obtained from the English Lexicon Project (Balota et al., 2007), and the frequency measure for English was based on Brysbaert and New's (2009) subtitle (SUBTL) frequency estimate. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 6
Descriptive Statistics for Lexical Variables for English and Malay With Statistics for the Full Set and Matched Subset of Words Reported Separately

	English						Malay					
	Full			Subset			Full			Subset		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Frequency	0	4,1857	25.55	0.49	1,500	**57.57**	0	15,136	79.81	0	12,513.4	**111.33**
Log frequency	0	4.62	0.44	0.17	3.18	**1.36**	0	4.18	1.07	0	4.1	**1.36**
Number of letters	1	20.00	7.82	6.00	14.00	**7.94**	2	22.00	7.56	3	15	**8.04**
Number of syllables	1	8.00	2.49	1.00	6.00	**2.45**	1	9.00	3.00	1	6	**3.15**
Number of phonemes	1	17.00	6.56	2.00	14.00	**6.45**	2	20.00	7.33	3	14	**7.83**
Number of morphemes	1	6.00	2.06	1.00	5.00	**1.97**	1	4.00	1.59	1	3	**1.51**
Orthographic neighborhood size	0	25.00	1.43	0.00	11.00	**1.24**	0	23.00	1.70	0	20	**1.77**
Phonological neighborhood size	0	48.00	3.09	0.00	30.00	**2.71**	0	23.00	1.78	0	20	**1.89**
Orthographic Levenshtein distance	1	7.94	2.78	1.30	5.30	**2.53**	1	9.30	2.45	1	5.4	**2.35**
Phonological Levenshtein distance	1	10.45	2.76	1.00	6.25	**2.48**	1	9.35	2.48	1	5.5	**2.41**

also contained more syllables and phonemes (see Table 3). Differences such as these reflect the general properties of Malay (i.e., Malay's smaller vocabulary, greater spelling–sound regularity, and highly agglutinative morphology) rather than a bias in sampling. To explore these aspects further, we conducted another regression analysis and selected 1,510 English words that were yoked as closely as possible to their Malay counterparts on the same range of lexical variables (see Table 6 for descriptive statistics).

The matching of the stimuli was automatically carried out using Van Casteren and Davis's (2007) match program. However, it is important to note that although the program managed to match the two sets of words closely on critical characteristics, such as frequency and number of letters, other variables could not be matched because of intractable language constraints. For example, the Malay words still contain more syllables, because of Malay's simple syllabic structure, and the critical differences in the mapping between orthography and phonology (English is deep and Malay is shallow) are not captured.

Nevertheless, the results of the second regression analysis (see Table 7) with the matched subset of words showed that word length was a stronger predictor than frequency for both English lexical decision and speeded pronunciation. This pattern of performance is consistent with our results for Malay performance (see Table 5) but is inconsistent with the extant English word recognition literature. The exaggerated length effects also contrast with the analyses based on the full set of English words (see Table 5), where frequency was clearly a stronger predictor.

This finding is interesting because it illustrates how observed effects yielded by regression analyses depend on the specific properties of the stimulus set. Matching English words to a representative sample of Malay words resulted in a biased sample for the English words that does not faithfully reflect the general characteristics of English orthography (see Table 6). Specifically, these English words are much higher in frequency than the norm. The nonlinear sigmoidal relationship between word frequency and recognition times (Plaut et al., 1996) implies that frequency effects plateau for very high frequency words. Length effects continue to be robust because long words still engage serial processing and multiple fixations. Hence, it is unsurprising that length effects overshadow frequency effects for these very high frequency English words. Taken together, the regression analyses

Table 7
Standardized RT Regression Coefficients of the Item-Level Regression Analyses for Speeded Lexical Decision and Naming Performance for 1,510 Malay Words and 1,510 English Words Matched on Word Frequency, Number of Letters, Orthographic Neighborhood Size, and Orthographic Levenshtein Distance

	Lexical Decision		Speeded Pronunciation	
	Malay	English	Malay	English
Predictor Variables	n = 1,510	n = 1,510	n = 1,510	n = 1,510
Surface Variables (Onsets)				
Adjusted R^2	**.176*****	**.049*****	**.500*****	**.174*****
Lexical Variables				
Log frequency	−.470***	−.286***	−.157***	−.209***
Number of letters	.655***	.372***	.720***	.381***
Orthographic neighborhood size	.058*	.068**	−.014	−.048†
Orthographic Levenshtein distance	−.072*	.241***	.010	−.009
Adjusted R^2	**.555*****	**.469*****	**.743*****	**.393*****
R^2 change	**.379*****	**.420*****	**.243*****	**.219*****

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

with matched stimuli suggest that the way people process this atypical set of English words may be similar to the way that people process typical Malay words. Such item-selection effects underscore the importance of having the language, rather than the experimenter, define the word properties being investigated (see Balota et al., 2004, for more discussion on megastudies and the limitations of factorial designs).

DISCUSSION

To support future research aimed at separating language-specific from language-universal processing, and to develop a better understanding of word recognition in orthographies with consistent mappings to phonology, we compiled a new database of lexical variables and behavioral (lexical decision and speeded pronunciation) measures for Malay, a shallow orthography with simple syllable structures that embodies alphabetic principles more faithfully than English does (see Share, 2008). We also explored how the ratio of number of vowel letters to number of vowel phonemes (proxy for orthographic depth) and ratio of number of letters to number of syllables (proxy for syllable complexity) modulate the effects of lexical measures such as word length, word frequency, neighborhood size, and LD on word recognition.

The relative influence of these lexical measures on lexical decision and speeded pronunciation performance have informed and constrained models of word recognition in English, which is by far the most popular language for models of word recognition (see Balota et al., 2006, for a review). The equivalent data we have compiled for Malay have implications for models of word recognition in shallow orthographies and further underscore the importance of objective and systematic crosslinguistic comparisons.

The findings in our study make three main contributions (see Table 5). First, word length—a marker of serial sublexical processing—predicted both lexical decision and speeded pronunciation performance far better than did word frequency in Malay. The reverse pattern was seen in English, a deep orthography, where frequency predicted recognition times better than did any other lexical variable. Length effects were also larger in speeded pronunciation, as compared with lexical decision, consistent with the pronunciation task's reliance on serial sublexical processing. Second, although frequency effects were attenuated in Malay, they were nevertheless reliable, indicating that lexical processing is implicated even in a very shallow orthography. Frequency effects were considerably larger in lexical decision than in speeded pronunciation, in line with lexical decision's emphasis on familiarity-based information for discriminating between familiar words and unfamiliar nonwords (see Balota & Chumbley, 1984). Finally, meshing well with the previous observation, LD and orthographic neighborhood effects were significant in lexical decision but not in speeded pronunciation in Malay. More interestingly, lexical decision times were faster for words with fewer orthographic neighbors and further Levenshtein neighbors, pointing to a competitive lexical selection procedure.

To summarize, we have generated and provided measures of frequency, length, orthographic distinctiveness, and phonological distinctiveness for a set of 9,592 Malay words, along with behavioral measures for 1,520 words. To our knowledge, this represents the first such database for Malay—an unusually shallow alphabetic orthography with simple syllabic structures and relatively transparent affixation. This resource, which will be made freely available (http://brm.psychonomic-journals.org/content/supplemental), should be useful for researchers who are studying Malay lexical and memory processing. More generally, it can be used by investigators who are interested in making systematic crosslinguistic comparisons that allow better delineation of language-specific and language-general processes in word recognition. The results of methodologically sound crosslinguistic research, across orthographies varying in depth and syllable complexity, have major implications for unilingual literacy programs and for the increasing number of bilinguals who are now learning English as a second language. Such work will also help facilitate the design and testing of universal models that account for the relationship between psycholinguistic variables and cognitive processing.

AUTHOR NOTE

This research was supported by National University of Singapore SRSS Grant R-581-000-076-133 to M.J.Y. We thank Gregory Francis and two anonymous reviewers for their useful comments on a previous version of the manuscript. We are greatly indebted to the staff of Dewan Pustaka (Kuala Lumpur) and Sebastian Chow of Singapore Press Holdings for access to the newspaper corpora, without which this resource could not have been developed. We also acknowledge the help of Choong Yea Jye, Maswati Mashoed, and Ivan Rickard Liow in the preparation of the database. Address correspondence to M. J. Yap, Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Block A, #02-07, Singapore 117570, Republic of Singapore (e-mail: melvin@nus.edu.sg).

REFERENCES

ALGARABEL, S., RUIZ, J. C., & SANMARTÍN, J. (1988). The University of Valencia's computerized word pool. *Behavior Research Methods, Instruments, & Computers*, **20**, 398-403.

ANDREWS, S. (2006). *From inkmarks to ideas: Current issues in lexical processing*. Hove, U.K.: Psychology Press.

AWANG, S. (2004). *Teras pendidikan bahasa Melayu: Asas pegangan guru* [Core of Malay language education: Teachers' foundational beliefs]. Bentong, Pahang: PTS Publications Sdn Bhd.

BAAYEN, R. H., PIEPENBROCK, R., & VAN RIJN, H. (1993). *The CELEX lexical database*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

BALOTA, D. A., & CHUMBLEY, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, **10**, 340-357. doi:10.1037/0096-1523.10.3.340

BALOTA, D. A., CORTESE, M. J., SERGENT-MARSHALL, S. D., SPIELER, D. H., & YAP, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, **133**, 283-316. doi:10.1037/0096-3445.133.2.283

BALOTA, D., YAP, M. J., & CORTESE, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In

M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285-375). Amsterdam: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods*, **39**, 445-459.

Borgwaldt, S. R., Hellwig, F. W., & De Groot, A. M. B. (2005). Onset entropy matters: Letter-to-phoneme mappings in seven languages. *Reading & Writing*, **18**, 211-229. doi:10.1007/s11145-005-3001-9

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, **41**, 977-990. doi:10.3758/BRM.41.4.977

Caravolas, M. (2004). Spelling development in alphabetic writing systems: A cross-linguistic perspective. *European Psychologist*, **9**, 3-14. doi:10.1027/1016-9040.9.1.3

Caravolas, M., & Bruck, M. (1993). The effects of oral and written language input on children's phonological awareness: A cross-linguistic study. *Journal of Experimental Child Psychology*, **55**, 1-30. doi:10.1006/jecp.1993.1001

Chateau, D., & Jared, D. (2003). Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory & Language*, **48**, 255-280. doi:10.1016/S0749-596X(02)00521-1

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256. doi:10.1037/0033-295X.108.1.204

Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic characteristics. *Behavior Research Methods*, **37**, 65-70.

Dewan Bahasa dan Pustaka (2007). *Kamus Dewan Edisi Keempat* (4th ed.). Kuala Lumpur, Malaysia: Author.

Durgunoglu, A.Y., & Oney, B. (1999). A cross-linguistic comparison of phonological awareness and word recognition. *Reading & Writing*, **11**, 281-299. doi:10.1023/A:1008093232622

Ellis, N. C., & Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics*, **22**, 571-599. doi:10.1017/S0142716401004052

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., et al. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, **42**, 488-496. doi:10.3758/BRM.42.2.488

Frost, R., & Katz, L. (Eds.) (1992). *Orthography, phonology, morphology, and meaning*. Amsterdam: Elsevier.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 104-115. doi:10.1037/0096-1523.13.1.104

Hassan, A. (1974). *The morphology of Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Jalil, S., & Rickard Liow, S. J. (2008). How does home language influence early spellings? Phonologically plausible errors of diglossic Malay children. *Applied Psycholinguistics*, **29**, 535-552.

Karim, N. S., Onn, F. M., Haji Musa, H., & Mahmood, A. H. (2008). *Tatabahasa Dewan* (*Edisi Ketiga*). Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Koh, B. B. (1978). *The teaching of Malay affixes*. Kuala Lumpur, Malaysia: Fajar Bakti.

Ktori, M., van Heuven, W. J. B., & Pitchford, N. J. (2008). GreekLex: A lexical database of modern Greek. *Behavior Research Methods*, **40**, 773-783. doi:10.3758/BRM.40.3.773

Lee, L. W. (2008). Development and validation of a reading-related assessment battery in Malay for the purpose of dyslexia assessment. *Annals of Dyslexia*, **58**, 37-57. doi:10.1007/s11881-007-0011-0

Leppanen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2006). Development of reading and spelling Finnish from preschool to grade 1 and grade 2. *Scientific Studies of Reading*, **10**, 3-30. doi:10.1207/s1532799xssr1001_2

Lervåg, A., Bråten, I., & Hulme, C. (2009). The cognitive and linguistic foundations of early reading development: A Norwegian latent variable longitudinal study. *Developmental Psychology*, **45**, 764-781. doi:10.1037/a0014132

Lukatela, G., Popadič, D., Ognjenovič, P., & Turvey, M. T. (1980). Lexical decision in a phonologically shallow orthography. *Memory & Cognition*, **8**, 124-132.

Mohamed, P. G. (2003). *Pragmatik bahasa Melayu Baku di media massa Singapura* [The pragmatics of "Baku" Malay language in the Singapore mass media] (Working Paper, Special Teaching Programme, Malay). Singapore: National Institute of Education.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, **106**, 226-254. doi:10.1037/0096-3445.106.3.226

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, **13**, 45-52.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**, 516-524.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, **114**, 273-315. doi:10.1037/0033-295X.114.2.273

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1989). Positron emission tomographic studies of the processing of single words. *Journal of Cognitive Neuroscience*, **1**, 153-170. doi:10.1162/jocn.1989.1.2.153

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, **103**, 56-115. doi:10.1037/0033-295X.103.1.56

Rickard Liow, S. J., & Lee, L. C. (2004). Metalinguistic awareness and semi-syllabic scripts: Children's spelling errors in Malay. *Reading & Writing*, **17**, 7-26. doi:10.1023/B:READ.0000013833.79570.de

Rickard Liow, S. J., Yap, M. J., Lee, L. C., & Ramos, S. D. S. (2008, November). *Influence of lexical–orthographic variables in children's spelling skills.* Poster presented at the 49th Annual Meeting of the Psychonomic Society, Chicago.

Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools, Inc.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, **94**, 143-174. doi:10.1348/000712603321661859

Share, D. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, **134**, 584-615. doi:10.1037/0033-2909.134.4.584

Spencer, K. A. (2009). Feedforward, -backward, and neutral transparency measures for British English. *Behavior Research Methods*, **41**, 220-227. doi:10.3758/BRM.41.1.220

Tadmor, U. (2009). Malay–Indonesian. In B. Comrie (Ed.), *The world's major languages* (2nd ed., pp. 791-818). London: Routledge.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, **124**, 107-136. doi:10.1037/0096-3445.124.2.107

Van Casteren, M., & Davis, M. H. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, **39**, 973-978.

Winskel, H., & Widjaja, V. (2007). Phonological awareness, letter knowledge, and literacy development in Indonesian beginner readers and spellers. *Applied Psycholinguistics*, **28**, 23-45. doi:10.1017/S0142716407070026

Wydell, T. N., & Butterworth, B. (1999). A case study of an English–Japanese bilingual with monolingual dyslexia. *Cognition*, **70**, 273-305. doi:10.1016/S0010-0277(99)00016-5

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory & Language*, **60**, 502-529. doi:10.1016/j.jml.2009.02.001

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Moving beyond

Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, **15**, 971-979. doi:10.3758/PBR.15.5.971

YATES, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 1385-1397. doi:10.1037/0278-7393.31.6.1385

YATES, M., LOCKER, L., & SIMPSON, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, **11**, 452-457.

ZIEGLER, J. C., & GOSWAMI, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, **131**, 3-29. doi:10.1037/0033-2909.131.1.3

ZIEGLER, J. C., PERRY, C., & COLTHEART, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology*, **12**, 413-430. doi:10.1080/09541440050114570

NOTE

1. To ensure that the differences in the results of the regression analyses between English and Malay were not driven by the disparity in sample sizes, we also conducted an additional analysis in which we randomly sampled 1,510 words from the full English database. The results from the analysis of this randomly sampled subset of words were qualitatively identical to the results for the full set of items.

SUPPLEMENTAL MATERIALS

The database of Malaysian word norms discussed in this article may be downloaded from http://brm.psychonomic-journals.org/content/supplemental.

(Manuscript received September 17, 2009;
revision accepted for publication May 23, 2010.)