*Kajian Malaysia, Vol. 32, Supp. 1, 2014, 151–166*

# THE DEVELOPMENT AND APPLICATION OF AN ONLINE MALAY LANGUAGE CORPUS-BASED LEXICAL DATABASE

**Lay Wah Lee[1*] and Hui Min Low[2]**

[1,2]School of Educational Studies, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia
[*]Corresponding author: lwah@usm.my

*The Online Malay Language Corpus-based Lexical Database for Primary Schools discussed in this paper is a long-term research project by the School of Educational Studies, Universiti Sains Malaysia. Based on a corpus of Malay language textbooks used in primary schools, the project aims to develop a lexical database of Malay words commonly encountered by elementary school children in Malaysia. Available online, the database has an interactive interface that allows users to search in real-time primary linguistic features such as word frequency, word length, phoneme length, number and type of syllables, as well as word category. The database has proven to be a useful resource for both researchers and practitioners who use it to identify linguistically and culturally appropriate sets of word stimuli for material development, language assessment, language teaching, and language remediation. These applications facilitate and promote evidence-based teaching and research practices pertaining to reading acquisition in the Malay language. The system has undergone initial validation and is continuously being revised and expanded to improve on its usability, readability, and accuracy.*

Keywords: reading, online lexical database, Malay language, teaching, assessment

## INTRODUCTION

The origin of word corpora based on children's reading materials could be traced back to the Dolch list compiled by Edward William Dolch in 1936 and other lists compiled in the 60's and 70's (Kučera and Francis, 1967). Many of these lists were developed based on words extracted from dictionaries, best-selling children books and magazines, and have been used widely by teachers to develop word lists for assessment and teaching purposes. In the recent decade, advances in technology have made it possible to develop and compile large scale children's word databases for teaching and research purposes.

Today, most online corpus-based lexical databases are in European languages such as French (Peereman, Lete and Sprenger-Charolles, 2007), Spanish (Corral, Ferrero and Goikoetxea, 2009) and English (Masterson et al., 2010). One such database is an online comprehensive database known as the Children's Printed Word Database (henceforth CPWD) (http://www.essex.ac.uk/psychology/cpwd/) developed by Masterson et al. (2010). The CPWD contains 12,193 word-types and 995,927 occurrences of these word-types (i.e., word-tokens) based on data collected from over one thousand popular books used in British schools. The development of this database, which is fully interactive and accessible via the web, has stimulated better-designed research approaches to child English language and writing skills (Masterson et al., 2010). Another big-scale corpus-based lexical database is called MANULEX (Lete, Sprenger-Charolles and Cole, 2004), a web-based French database. It contains 1.9 million words collected from over 50 books used in French primary schools. Chosen based on sales record in 1996, the books cover a range of genres from novels, technical writings, and poetries to theatre plays. MANULEX has information on word frequencies for four levels based on grades, namely first grade (G1), second grade (G2), third to fifth grades (G3–5), and all grades (G1–5), and provides information on number of letters in both the word and syntactic categories. As pointed out by (Lete, Sprenger-Charolles and Cole, 2004: 156), the database "is intended to be a useful tool for studying language development through the selection of stimuli based on precise frequency norms." The information is also useful for language instruction, syllabus design, vocabulary grading and material writing (Lete, Sprenger-Charolles and Cole, 2004).

The success of these large databases, among many others, is testament to Gardner's (2007) observation of the influence of corpora and corpus-based research on educational theories and practices in both first language and second language settings[1]. Besides impacting theory and practice, the research and educational potential of these databases has encouraged researchers who work on languages other than English and European languages to develop their own corpus-based lexical databases for teaching and research purposes. One such database is the Online Malay Language Corpus-based Lexical Database for Primary Schools which is currently being developed by researchers at the School of Educational Studies, Universiti Sains Malaysia.

Partly inspired by earlier corpus-based lexical database projects, particularly the CPWD, the Online Malay Language Corpus-based Lexical Database for Primary Schools may be considered as the first of its kind available in the Malay language. In Malaysia especially, where corpus–based education research is still rather new, the development of the Malay lexical database will have important implications for educational research and practice in Malay language primary education in Malaysia. This paper discusses the development of the database and its application as a repository of information on Malay words, and considers the main challenges that have emerged as well as improvements to the database in realising its full potential.

## THE MOTIVATION FOR AN ONLINE MALAY LEXICAL DATABASE

The acquisition of reading skills is fundamental to literacy development. For reading in alphabetic languages like the Malay language, children learn to recognise the alphabet, the sounds of the letters in the alphabet, and the common words before they can read connected texts (Wagner and Torgesen, 1987). The acquisition of these fundamental prerequisite reading skills typically takes place in the pre-school years. Yet, this is not always necessarily true because in communities where reading is not inculcated in children at a very young age, particularly at home, exposure to the task of reading really only begins when children attend kindergarten or primary school. In such cases, teachers play an important role in teaching them to read.

In addition to the prerequisite reading skills which children need, the literature shows that cognitive abilities, oral language, print knowledge, vocabulary, and phonological sensitivity/phonological awareness are also important precursors of reading (Lonigan, Burgess and Anthony, 2000; Lonigan et al., 2009; Skebo et al., 2013). In connection to this, the stage theory of reading development proposed by Chall (1983) estimates that children acquire phonological awareness and other pre-reading skills by the end of first grade, and are able to read fluently by the end of third grade (Ehri, 2012). While children are commonly able to master the skills and are able to read relatively fast when given the basic instructions and suitable reading materials, some children require more help than others. These are essentially children who have problems remembering letters, associating sounds with printed letters or combinations of letters, identifying sound borders such as syllables in words (Lee, 2008; Lee and Wheldall, 2011) and so on. These problems affect their recognition of single words and the reading of connected texts. If the problems persist, an assessment of their reading ability would be carried out and if necessary reading remediation would be administered. Early identification and remediation of reading difficulties are important, particularly for children with dyslexia.

Research shows that of the five precursors of reading stated earlier, phonological sensitivity, or better known as phonological awareness has been widely accepted as predictive of later reading achievement (Hulslander, et al., 2010; Zhang et al., 2013). Trainings on phonological awareness have also proven to be effective in remediating children with reading difficulties (Slavin et al., 2011). Indeed research (e.g., Spencer, 2007; 2010; Birsh, 2011; Blachman et al., 2013) has shown that children with reading difficulties can experience long-term positive changes in their reading abilities using suitable training materials and intervention techniques that focused on phonological-related reading skills. This means that the word stimuli that are used (Spencer, 2007; 2010) in teaching, evaluating and remediating children with reading difficulties are pertinent because children are sensitive to the characteristics of printed texts that they read.

Different orthographies are known to elicit different reading strategies (Farrington-Flint et al., 2008). For alphabetic languages like English and Malay, children need to decode the sound units in words in order to read successfully (Lee and Wheldall, 2009; 2011). The sound units refer to the basic sound units (also known as phonemes) and sound combinations (also known as syllables). Examples of phonemes are /b/, /m/ and /i/ and examples of syllables are /bi/ and /mi/. The decoding of these sound units is known as phonological decoding. This skill, coupled with reading comprehension, lays the foundation for reading (Goswami, 2008; Wagner and Torgesen, 1987). In relation to this, different word structures and word meanings impose different levels of reading challenges to beginning readers (Spencer, 2010). Words with simple structures, such as those composed of single-letter graphemes (e.g., "h", "n" and "g") and open syllables (e.g., V, CV) would be easier to decode compared with words with multi-letter graphemes (e.g., "ng", "ny", and "gh") and closed syllables (e.g., CVC) (Lee and Wheldall, 2011). This comparison is illustrated in Table 1 using two Malay words *mata* (eye) and *ghairah* (strong desire).

Table 1: Comparison of word structures

| Word examples | Syllable units | Sound units |
|---|---|---|
| *Mata* | "ma" + "ta"<br>Two open syllables, CV + CV | "m", "a", "t", "a"<br>Four single-letter graphemes |
| *Ghairah* | "ghai" + "rah"<br>Open syllable (CV) + closed syllable (CVC) | "gh" (C)*, "ai" (V)*, "r", "a", "h"<br>Two multi-letter graphemes, and three single letter graphemes |

*"gh" (pronounced as /gh/) is one of the very few multi-letter graphemes, represented by C, in the Malay orthographical repertoire. The other multi-letter graphemes, in the category of digraphs are "kh", "ng", "ny", and "sy", and in the category of diphthongs, represented by V, are "ai", "au", and "oi" (Awang, 2004). These multi-letter graphemes, each represents the sound of a phoneme. For instance, the combination "ng" in Malay represents the phoneme /ŋ/.

Besides the complexity of the multi-letter grapheme ("gh"), the word *ghairah* also imposes more challenge to beginning readers compared to the item *mata* semantically. This is because beginning readers would find it easier to access the semantic meaning carried by *mata* which is concrete nominal term compared with *ghairah* which is abstract adjective term [refer to Danielsson (2001) for a discussion on this]. Thus, for beginning readers in Malay, the word *mata* as a whole, imposes less of a reading challenge compared to the form *ghairah*.

Besides phonological decoding and semantic accessibility, there are other word properties that might influence the relative ease of decoding and understanding words. These include word frequency, word length, grain-size of the sound units, inflectional patterns, and word categories (Spencer, 2010; Ziegler and Goswami, 2005). Given this, it is pertinent that teachers understand

and are aware of the distinctions when choosing word sets for teaching, evaluating, and remedial purposes. Indeed a thorough consideration of these word properties is of prime importance when teachers choose the word stimuli for beginning readers. Similarly, teachers need to consider the relevance of these word properties when designing assessment and remedial materials for slow readers (Ehri and McConnick, 1998).

While it is important for teachers to be aware of these issues, it must also be noted that the selection of word stimuli for reading, assessment and remedial tasks is not a simple matter. Whether teachers use printed materials such as dictionaries, books and newspapers or their introspective knowledge of words, the task of sourcing and selecting words can be tedious and demanding. Also, such methods of selecting word stimuli can be highly subjective, and raise the issue of whether the words selected are the most suitable word sets for beginning and slow readers.

These concerns and the need to provide resource and solutions for teachers in helping beginning/slow readers are the motivation for the development of the Online Malay Language Corpus-based Lexical Database for Primary Schools discussed in this paper.

## DEVELOPMENT OF THE ONLINE MALAY LANGUAGE CORPUS-BASED LEXICAL DATABASE FOR PRIMARY SCHOOLS

The development of a Malay corpus-based lexical database which is accessible to teachers, academics and researchers is not only timely but is also a step towards gearing up corpus-based research in Malaysia that can contribute towards Malay language educational theories and practices in the country. And importantly, establishing a Malay language corpus-based lexical database is in line with the current educational development in Malaysia which places currency on literacy acquisition at the primary school level. Literacy intervention programmes, such as the reading and writing early intervention class known as Kelas Intervensi Awal Membaca dan Menulis (KIA2M) programme, have been developed and implemented. The KIA2M programme which emphasises literacy acquisition in Malay and English languages from the Year 1 to Year 3 (Sharifah Nor and Juti, 2009) programme has evolved into the current Literacy and Numeracy Screening (LINUS) programme which, as the name suggests, includes a numeracy component (Ministry of Education, 2011). These literacy intervention programmes are available in all primary schools throughout Malaysia and remedial education teachers are entrusted with the responsibility of ensuring children's response to intervention.

In line with such interest and emphasis on literacy development, the Online Malay Language Corpus-based Lexical Database for Primary Schools described here constitutes a valuable educational resource. It may be considered as the first of its kind available in the Malay language, especially in Malaysia. Based at the School of Educational Studies, Universiti Sains Malaysia, the database comprises data that are generated from a corpus of Malay language written materials used in Malaysian primary schools. An open-access database system is chosen to host the word corpora because a web-based system is more sustainable and accessible to users at any time and from any location. Also, a web-based learning resource is more cost-effective in terms of database expansion in the long run (McKimm, Jollie and Cantillon, 2002).

The chief objective of the project is to develop a lexical database comprising printed words extracted from Standard 1 (Year 1) to Standard 6 (Year 6) Malay language textbooks. As it stands, the database contains 4,035 words from Standard 1 (Year 1) and Standard 2 (Year 2) Malay language textbooks (Lee and Low, 2011). At this juncture, it should be noted that the preliminary word analysis works conducted prior to the development of this online database are reported in Lee, Low and Abdul Rashid (2012), and a prototype database system has been developed to house the results of this analysis (available online at http://mybaca.org/).

Inspired by the CPWD system, the Online Malay Language Corpus-based Lexical Database for Primary Schools is equipped with interactive features that allow users to generate word lists according to certain linguistic features. The words in the database are arranged according to seven linguistic properties, i.e., frequency of occurrence, word length, phoneme length, number of syllables, types of syllable, and word category. Thus, as a repository of information on words that are accessible online, the database can be used to study word-level linguistic structures to improve teaching and learning. The interactive features of the online lexical database allow teachers and users to generate the most suitable word lists according to their intended teaching objectives and learning outcomes. For primary school teachers in Malaysia especially, the database is a readily-accessible literacy resource because it contains quantitative descriptions of Malay word properties that are useful for teaching reading to young readers. Such additional linguistic information is also useful for remedial educators in planning their language teaching, remediation, and evaluation tasks. For instance, a teacher with beginning students who have basic reading difficulties can design a simple teaching material by accessing the database for Year 1 vocabularies or for words that begin with a particular phoneme or syllable structure. The database can facilitate the teacher's job and empower users to create materials that are appropriate for specific learning needs.

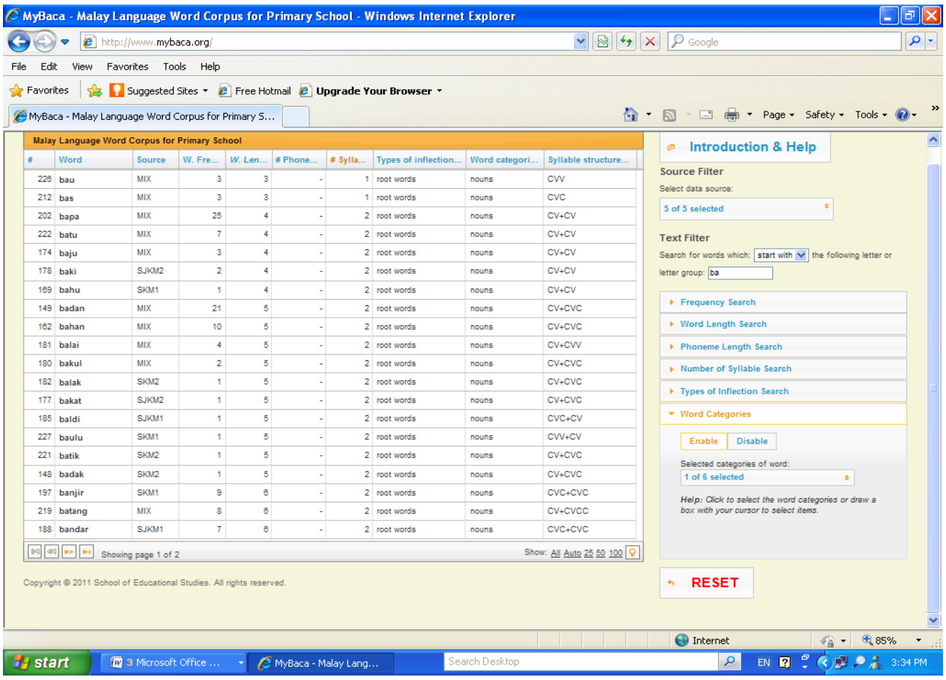
In addition to its major function as a resource for educational aid, the availability of a Malay language online lexical database also provides useful information to researchers in the field of linguistics and psychology. Given that the lexical database is based on words extracted from reading materials used in classrooms in Malaysia, the data can be used for evidence-based research in the areas of language development and psycholinguistics, particularly with regard to the Malay language. The use of the lexical database for research has far-reaching implications. Firstly, by making use of the available data, researchers can make informed decisions about the selection of words which are familiar to children in Malaysia. Secondly, the availability and accessibility of the lexical database can facilitate work by research groups locally and internationally. In relation to this, it should be noted that the transparency in the development protocols allows researchers from within and outside the country to conduct cross-linguistic lexical studies by comparing the information available in the current lexical database with other web-based databases such as the CPWD (Masterson et al., 2010) and MANULEX (Lete, Sprenger-Charolles and Cole, 2004).

## APPLICATION: PRELIMINARY EXAMPLES

As mentioned earlier, the online Malay corpus-based lexical database allows users to search for word- or vocabulary-related information for teaching, especially in the teaching and remediation of reading at the word level, and for research in the Malay language. To provide some insight into how these can be done, case examples of how to develop reading materials and words lists are illustrated below. These examples reflect on the teaching and remediation of reading, in line with the original purpose of the database design. And to illustrate how the database can facilitate research, a case example of research using the different types of data is also presented.

### Developing Reading Materials

The linguistic information contained in the Online Malay Language Corpus-based Lexical Database for Primary Schools is a useful resource for individuals, story writers, and teachers who are interested to develop reading materials for young children. Using the online search filters in the lexical database, users can look for words with controlled linguistic properties. For instance, a teacher who plans to create a story that contains two-syllable words which begin with the syllables "ba" and "bu" can set these criteria using the search filters in the database. For words beginning with the syllable "ba", the search will generate words such as *bapa* (father), *batu* (rock), *baju* (clothes), *baki* (remnants) and *bahu* (shoulder), and for words beginning with the syllable "bu", the system will produce a different list with words such as *buku* (book), *budi* (gratitude) and *bumi* (earth). These words would be drawn from the corpus of primary school textbooks. An example of a search outcome for words that begin with syllable "ba" from the lexical database is presented in Figure 1.



Malay Language Word Corpus for Primary School

| # | Word | Source | W. Fre... | W. Len... | # Phone... | # Sylla... | Types of inflection... | Word categori... | Syllable structure... |
|---|---|---|---|---|---|---|---|---|---|
| 226 | bau | MIX | 3 | 3 | - | 1 | root words | nouns | CVV |
| 212 | bas | MIX | 3 | 3 | - | 1 | root words | nouns | CVC |
| 202 | bapa | MIX | 25 | 4 | - | 2 | root words | nouns | CV+CV |
| 222 | batu | MIX | 7 | 4 | - | 2 | root words | nouns | CV+CV |
| 174 | baju | MIX | 3 | 4 | - | 2 | root words | nouns | CV+CV |
| 178 | baki | SJKM2 | 2 | 4 | - | 2 | root words | nouns | CV+CV |
| 169 | bahu | SKM1 | 1 | 4 | - | 2 | root words | nouns | CV+CV |
| 149 | badan | MIX | 21 | 5 | - | 2 | root words | nouns | CV+CVC |
| 162 | bahan | MIX | 10 | 5 | - | 2 | root words | nouns | CV+CVC |
| 181 | balai | MIX | 4 | 5 | - | 2 | root words | nouns | CV+CVV |
| 180 | bakul | MIX | 2 | 5 | - | 2 | root words | nouns | CV+CVC |
| 182 | balak | SKM2 | 1 | 5 | - | 2 | root words | nouns | CV+CVC |
| 177 | bakat | SJKM2 | 1 | 5 | - | 2 | root words | nouns | CV+CVC |
| 185 | baldi | SJKM1 | 1 | 5 | - | 2 | root words | nouns | CVC+CV |
| 227 | baulu | SKM1 | 1 | 5 | - | 2 | root words | nouns | CVV+CV |
| 221 | batik | SKM2 | 1 | 5 | - | 2 | root words | nouns | CV+CVC |
| 148 | badak | SKM2 | 1 | 5 | - | 2 | root words | nouns | CV+CVC |
| 197 | banjir | SKM1 | 9 | 6 | - | 2 | root words | nouns | CVC+CVC |
| 219 | batang | MIX | 8 | 6 | - | 2 | root words | nouns | CV+CVCC |
| 188 | bandar | SJKM1 | 7 | 6 | - | 2 | root words | nouns | CVC+CVC |

Showing page 1 of 2 Show: All Auto 25 50 100

Copyright © 2011 School of Educational Studies. All rights reserved.

Source Filter
Select data source: 5 of 5 selected

Text Filter
Search for words which: start with the following letter or letter group: ba

Selected categories of word: 1 of 6 selected

*Help: Click to select the word categories or draw a box with your cursor to select items.*

Figure 1: Search outcome for words beginning with the syllable "ba"

The list of words that is generated can be used by a teacher to plan a story or a comprehension exercise for various teaching and testing purposes. Also importantly, the use of the lexical database to search for words with specific linguistic features allows for the development of reading materials that are more structured and systematic. Consider the set of sentences in Figure 2 which are built from the data (beginning with the syllable "ba") generated by the lexical database.

* ***ba****pa beli* ***ba****ju* (father bought a shirt)
* ***ba****ju corak* ***ba****tu* (the shirt has rock print)
* *corak* ***ba****tu di* ***ba****hu* (the rock print is on the shoulders)
* ***ba****pa suka* ***ba****ju corak* ***ba****tu* (father likes the shirt with the rock print)

Figure 2: Connected texts using words beginning with "ba" generated from the database

The sentences in Figure 2 exemplify how the database can facilitate the teacher's job. In remedial and intervention cases for instance, the emphasis on the initial syllable "ba" in the text and the ample opportunities for repetitive practices can be very helpful for beginning and challenged readers. As research shows, the linguistic features of single words contribute greatly towards Malay word recognition acquisition (Lee and Wheldall, 2009). Thus, such an exercise can be a highly effective intervention activity particularly for slow readers.

## Developing Word Lists

Reading difficulties in children, as stated earlier, may be due to their inability to recognise certain letters or pronounce certain sounds. For example, some children have the tendency to confuse between the letters "b" and "d". One way to help them to overcome this problem is to use discrete instructions and repetitive reading practices. For this purpose, the Online Malay Language Corpus-based Lexical Database for Primary Schools can be useful for generating word lists that contain words beginning with the letter "b" and those beginning with the letter "d" (see Table 2) for young children.

Table 2: Words beginning with "b" and "d" generated from the database

| Words that begin with "b" | Words that begin with "d" |
|---|---|
| *bapa* | *dada* |
| *batu* | *dagu* |
| *baju* | *dadu* |
| *buku* | *duku* |

Glossary:
*bapa* (father), *batu* (rock/stone), *baju* (shirt/dress), *buku* (book),
*dada* (chest), *dagu* (chin), *dadu* (dice), *duku* (a type of local fruit)

While teachers are able to source words from books and other printed materials to create word lists, the use of a lexical database, such as the Online Malay Language Corpus-based Lexical Database for Primary Schools, increases the efficacy and accuracy of developing structured resources. The two lists in Table 2 for instance allow opportunities for differential learning of the two orthographically similar letters and can be used for both evaluation and remedial purposes. Students can be presented with the opportunity to practice letter-sound correspondence rules which encourage encoding and retrieval of information from long-term memory. Also, such word lists can be utilised to investigate the contribution of visual perception error in reading problems. The availability of such database also facilitates teachers in terms of speeding up the process of creating diagnostic and intervention resources and allows them to focus on the more important aspect of individualised diagnosis and instruction.

## Analysing Linguistic Properties of Words for Research

The Online Malay Language Corpus-based Lexical Database for Primary Schools, as stated earlier, provides information on seven linguistic properties of words, namely, frequency of occurrence, word length, phoneme length, number of syllables, types of syllable, and word category (see Figure 1). These different linguistic details of words are crucial for language researchers who aim to study the linguistic properties of Malay language words. In language acquisition and education research, information on the linguistic properties of words, as found in the database, could be vital in analysing words that are easier to be accessed by children or beginning language learners.

With the sub-lexical information also, Malay language researchers and educators can access information on words and word categories frequently encountered by primary school children who are learning Malay language in Malaysia. For instance, users can access data from the Online Malay Language Corpus-based Lexical Database for Primary Schools to study word patterns and frequency of occurrence. A simple search of the database shows that there are 1,535 nominal vocabularies (nouns) and 1,639 action vocabularies (verbs) in Year 1 and Year 2 Malay textbooks. And interestingly, the database shows that nouns occur 8,321 times while verbs occur 8,127 times. This finding is rather intriguing as it suggests that while Year 1 and Year 2 students are presented with more verb types, they encounter nouns slightly more often than verbs in the Malay texts they read. Why this is the case is something that researchers can further explore by analysing data generated by the lexical database. For example, an analysis of the verbs in the database will reveal that there are more inflected words in the category of verbs than in the category of nouns. One interesting case involves the inflected words for the verb *dapat* (get/obtain). In the database, these include *mendapat* (occurs 21 times), *mendapati* (occurs 1 time) and *mendapatkan* (occurs 6 times). With the information provided by the Online Malay Language Corpus-based Lexical Database for Primary Schools, the patterns and varieties of inflection attached to words from different word categories could be compared and analysed more systematically.

The lexical information from the database can be used for theory testing and also to conduct comparative analyses. For instance, research on cross-linguistic comparisons between English and Malay children's story books (Lee, Low and Abdul Rashid, 2013) reveals that, unlike English, multi-syllabic words are dominant in Malay texts. This finding has important pedagogical implications with regard to reading in Malay. Chiefly, it foregrounds the necessity to focus on multi-syllabic words in teaching reading in Malay. In addition, the research also found significant cross-language differences between Malay and English words in terms of phoneme-grapheme correspondence, syllabic structure and types of inflectional morpheme. These emerging findings bring to the fore more research issues that can be addressed using data from the Online Malay Language Corpus-based Lexical Database for Primary School. Thus, further research between Malay and English could be carried out to provide more insight into cross-linguistic issues that can impact Malay language teaching and research.

Aside from the stated applications, the lexical information contained in the database could also benefit teachers in teaching other literacy skills, such as speaking and grammatical skills. Thus, the utilisation of the lexical database greatly depends on the needs and creativity of the users. In connection to this, feedback from users of the database has been invaluable in providing the developers with some insight into how the lexical information is being used and the potential applications of the database. From the enquires and feedback from a wide range of users since the launch of the system prototype, it is heartening to see that the actual functions of the Online Malay Language Corpus-based Lexical Database for Primary Schools have extended beyond its initial aim of contributing to teaching and reading acquisition research. Thus far, users' feedback suggests that its applications range from choosing the words to be taught in language lessons to building word sets for psychological assessments. To date, the Online Malay Language Corpus-based Lexical Database for Primary Schools has been used in several projects, including projects to develop a screening tool for reading difficulties, multimedia stories, a verbal memory test in an IQ evaluation, as well as a reading list for adults with acquired language disorders.

## FUTURE DEVELOPMENT AND CHALLENGES

The Online Malay Language Corpus-based Lexical Database for Primary Schools is a long-term research project which aims to compile texts in Malay language textbooks used in the Malaysian primary schools. In realising this aim, there are certain issues that must be addressed, namely (1) expansion of the corpus-based lexical database; (2) increased utilisation of the database; and (3) improvement of its online features.

In terms of both database expansion and increased utilisation, a working model that combines training with expansion has been adopted. The expansion of this corpus-based lexical database is currently being carried out in conjunction with the training of pre-service special education teachers. The combination of expanding the corpus while training these teachers is motivated by research which found that content knowledge of language structure is the missing foundation in teacher education of pre-service special educators (Moats, 2009). Special education teachers with rudimentary knowledge of language structure would not be able to help their students. Taking this into consideration, the Online Malay Language Corpus-based Lexical Database for Primary Schools is currently used as a resource for Malaysian pre-service teachers to learn content knowledge of language structure. More importantly, pre-service special education teachers are being trained to analyse words from the Malay language textbooks according to phonological units, morphological structures, and word categories, and the linguistic information obtained from the analyses that they carry out are uploaded into the online corpus (Lee and Low, under review). Thus, the lexical database is not only a repository of information for the development of teaching and learning materials, but is also a teacher training resource. This essentially enhances the potential of the database as a Malay language resource.

Besides data expansion and increased utilisation, the development of the corpora must also involve improvements of its online features such as the search and retrieval features. In relation to this, one interactive feature which could be added is information on lexeme frequency. Lexemes are a set of words derived from a single word such as the example of the inflected word forms of the verb *dapat* (i.e., *mendapat*, *mendapati* dan *mendapatkan*) discussed earlier. As highlighted by the developers of the French lexical database, MANULEX (Lete, Sprenger-Charolles and Cole, 2004), information on lexeme frequency is important in word processing. There is evidence to suggest that apart from surface word frequency, word processing also draws upon lexeme frequency. In other words, it might be easier for learners to acquire the word *dapat* compared to other root words with lower or no occurrence of lexemes; even though the original root word *dapat* may not occur as frequently as other root words.

The development of the online corpus-based lexical database also brought to light a number of fundamental challenges in such a project. One that is worth mentioning here is the importance of validating the construct "word" in the Malay language. Gardner (2007) discusses the negative implications of a weak theoretical construct of "word" in computer-processed corpora. He highlights three problematic areas in relation to this, chiefly (1) morphological relationships between words, (2) homonymy and polysemy, and (3) multi-word items. These problems became apparent in the course of developing the current lexical database. And questions remain on how to better account for multiword items such as *ibubapa* (from *ibu* (mother) + *bapa* (father) meaning "parents") and *matapelajaran* (from *mata* (eye) + *pelajaran* (lesson) meaning "subject") as a word, and homonyms such as *agak* which could serve as a verb (meaning "guess") or a functional word (meaning "rather").

In addition to the construct of "word", issues concerning morphological and phonological constructions of words have also emerged. For example, in the word *makanan* (food), the morphological construction is "CV+CVC+VC" while the phonological construction is "CV+CV+CVC". Such distinctions have yet to be made transparent in the current database and are essentially the challenges that the researchers need to overcome in the process of developing the lexical database.

## CONCLUSION

Today, compared to English and European languages, lexical databases for Asia-pacific languages are still scarce. It is hoped that the online Malay language lexical database discussed in the paper will inspire the development of databases in other languages in the Asia-pacific context.

Beyond its original aim of developing a database of words from a corpus of primary school textbooks, the current Malay lexical database will be further expanded with the inclusion of other children's reading materials, such as novels and magazines, to achieve the project's ultimate aim of promoting evidence-based practices in Malay language teaching and language research. In relation to this also, the database users' feedback will continue to be valuable to the developers of the Online Malay Language Corpus-based Lexical Database for Primary Schools in enhancing its usability, readability, and accuracy.

## ACKNOWLEDGEMENT

The study was funded by Universiti Sains Malaysia Short-Term Grant. Design and development of a Malay language dyslexia reading program. Grant number: 304.PGURU.6312002.

## NOTE

1. In Malaysia, corpus–based education research is still rather new but is gaining ground with the formation of a collaborative network of corpus-based researchers throughout Malaysia to spearhead this line of investigation into both first and second language use (Malaysia Corpus Research Colloquium, 5 April 2012, School of Humanities, Universiti Sains Malaysia).

## REFERENCES

Awang Sariyan. 2004. *Teras pendidikan bahasa Melayu: Asas pegangan guru (Core of Malay language education: Teachers' foundation beliefs).* Bentong: PTS Publications Sdn. Bhd.

Birsh, J. R. 2011. *Multisensory teaching of basic language skills*. Baltimore: Brookes Publishing Company.

Blachman, B., C. Schatschneider, J. M. Fletcher, M. Murray, K. A. Munger and M. G. Vaughn. 2013. Intensive reading remediation in Grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology.* DOI: 10.1037/a0033663.

Chall, J. 1983. *Stages of reading development*. New York: McGraw Hill.

Corral, S., M. Ferrero and E. Goikoetxea. 2009. LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods* 41(4): 1009–1017.

Danielsson, K. 2001. Beginning readers' sensitivity to different linguistic levels: An error and correction analysis at the lexical, syntactic and semantic levels. *Reading and Writing: An interdisciplinary Journal* 14: 395–421.

Dolch, E. W. A. 1936. A basic sight vocabulary. *The Elementary School Journal* 36: 456–460.

Ehri, L. C. 2012. Why is it important for children to begin learning to read in kindergarten? In *Contemporary debates in childhood education and development*, eds. S. Suggate and E. Reese, 171–180. New York: Routledge.

Ehri, L. C. and S. McConnick. 1998. Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading and Writing Quarterly* 14(2): 135–163.

Farrington-Flint, L., E. Coyne, J. Stiller and E. Heath. 2008. Variability in children's early reading strategies. *Educational Psychology* 28(6): 643–661.

Fielding-Barnsley, R. 2010. Australian pre-service teachers' knowledge of phonemic awareness and phonics in the process of learning to read. *Australian Journal of Learning Difficulties* 15(1): 99–110.

Gardner, D. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics* 28(2): 241–265.

Gentner, D. 1982. Why nouns and learned before verbs: Linguistic relativity versus natural partitioning. In *Language development (Vol. 2): Language, thought and culture*, ed. S. A. Kuczaj, 301–334. Hillsdale: Erlbaum.

Goswami, U. 2008. The development of reading across languages. *Annals of the New York Academy of Sciences* 1145: 1–12.

Hulslander, J., R. K. Olson, E. G. Willcutt and S. J. Wadsworth. 2010. Longitudinal stability of reading-related skills and their prediction of reading development. *Scientific Studies of Reading* 14(2): 111–136.

Kučera, H. and W. N. Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Lee, L. W. 2008. Dyslexia: Different cultures, similar behavioral signs. *Dyslexia Review* 19(3): 19–25.

Lee, L. W. and H. M. Low. 2011. Developing an online Malay language word corpus for primary schools. *International Journal of Education and Development Using Information and Communication Technology* 7(3): 96–101.

______. Under review. Analysis of Malay word structure by pre-service special education teachers: Foundation knowledge for remedial instruction.

Lee, L. W. and K. Wheldall. 2011. Acquisition of Malay word recognition skills: Lessons from low-progress early readers. *Dyslexia* 17(1): 19–37.

______. 2009. An examination of the simple view of reading among beginning readers in Malay. *Reading Psychology* 30: 250–264.

Lee, L. W., H. M. Low and Abdul Rashid Mohamed. 2013. A comparative analysis of word structures in Malay and English children's stories. *Pertanika Journal of Humanities and Social Sciences* 21(1): 67–84.

______. 2012. Word count analysis of Bahasa Malaysia textbooks for the purpose of developing a Malay reading remedial program. *Writing Systems Research* 4(1): 103–119.

Lete, B., L. Sprenger-Charolles and P. Cole. 2004. MANULEX: A lexical database from French readers. *Behavior Research Methods, Instruments and Computers* 36(1): 156–166.

Lonigan, C. J., J. L. Anthony, B. M. Phillips, D. J. Purpura, S. B. Wilson and J. D. McQueen. 2009. The nature of preschool phonological processing abilities and their relations to vocabulary, general cognitive abilities, and print knowledge. *Journal of Educational Psychology* 101(2): 345–358.

Lonigan, C. J., S. R. Burgess and J. L. Anthony. 2000. Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. *Developmental Psychology* 36(5): 596–613.

Masterson, J., M. Stuart, M. Dixon and S. Lovejoy. 2010. Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology* 101: 221–242.

McKimm, J., C. Jollie and P. Cantillon. 2002. Web based learning. *British Medical Journal* 326(7394): 870–873. DOI: 10.1136/bmj.326.7394.870.

Ministry of Education. 2011. *Buku panduan dan pengoperasian program literasi dan numerasi (LINUS) [A guidebook on the operation of the literacy and numeracy program (LINUS)]*. Kuala Lumpur: Ministry of Education.

Moats, L. 2009. Still wanted: Teachers with knowledge of language. *Journal of Learning Disabilities* 42: 387–391.

Peereman, R., B. Lete and L. Sprenger-Charolles. 2007. Manulex-infra: Distributional characteristics of grapheme-phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods* 39(3): 593–603.

Sharifah Nor Puteh and Juti Ranti. 2009. Kelas intervensi awal membaca dan menulis (KIA2M): Satu kajian kes di Sekolah Rendah Kebangsaan Nanga Meluan [Early internvention class of reading and writing (KIA2M): A case study in Nanga Meluan Primary School]. *International Journal of Learner Diversity* 1(1): 189–206.

Skebo, C. M., B. A. Lewis, L. A. Freebairn, J. Tag, A. A. Avrich and C. M. Stein. 2013. Predictors of reading skills for students with and without speech sound disorders at three stages of literacy development. *Language, Speech and Hearing Services in Schools.* DOI: 10.1044/0161-1461.

Slavin, R. E., C. Lake, S. Davis and N. A. Madden. 2011. Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review* 6(1): 1–26.

Spencer, K. 2010. Predicting children's word-reading accuracy for common English words: The effect of word transparency and complexity. *British Journal of Psychology* 101: 519–543.

———. 2007. Predicting children's word-spelling difficulty for common English words from measures of orthographic transparency, phonemic and graphemic length and word frequency. *British Journal of Psychology* 98: 305–338.

Tardif, T., S. A. Gelman and F. Xu. 1999. Putting the "Noun Bias" in context: A comparison of English and Mandarin. *Child Development* 70(3): 620–635.

Wagner, R. K. and J. Torgesen. 1987. The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin* 101(2): 192–212.

Washburn, E. K., R. M. Joshi and E. Binks-Cantrell. 2011. Are preservice teachers prepared to teach struggling readers? *Annals of Dyslexia* 61(1): 21–43.

Zhang, Y., T. Tardif, H. Shu, H. Li, H. Liu, C. McBride-Chang, W. Liang and Z. Zhang. 2013. Phonological skills and vocabulary knowledge mediate socioeconomic status effects in predicting reading outcomes for Chinese children. *Developmental Psychology* 49(4): 665–671. DOI: 10.1037/a0028612.

Ziegler, J. C. and U. Goswami. 2005. Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin* 131: 2–29.