International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2011, Vol. 7, Issue 3, pp. 96-101.

Developing an online Malay Language Word Corpus for primary schools

Lay Wah Lee, Hui Min Low
Universiti Sains Malaysia

ABSTRACT

This paper reports the development of an online Malay Language Word Corpus for primary schools. This online system contains the common words that occur in the Malay language textbooks used in Malaysian schools and the related linguistic information. The targeted users include regular teachers, special education teachers, pre-service teachers, parents, researchers and interventionists. Users could conduct word search according to six primary linguistic features, which include frequency of occurrence, word length, phoneme length, number of syllable, type of syllable and word category. The listing of words according to these linguistic features allow the users to choose linguistically appropriate and culturally relevant words for the purpose of education and research involving young Malay-speaking children in Malaysia. This word corpus is also potentially useful for the broader Asia Pacific context which includes Singapore, Brunei and Indonesia.

BACKGROUND

Word corpus is the collection of words used for linguistic analyses and educational purposes. In the area of education, English word lists such as those compiled by Dolch (1936) and Kucera and Francis (1967) have been widely used by educators for developing suitable teaching and educational materials for young children of different ages. For example, Dolch (1936) compiled a basic list of 220 words which he recommended as essential for children's reading. This list had received widespread use amongst teachers in teaching and assessment until it was gradually replaced by contemporary lists from the 70s onwards. Today, the development of contemporary word list or word corpus is made easy with the advancement of information technology. The technology has allowed for a broad collection of words with readily identifiable linguistic characteristics. An example of a creditable word corpus is the online database, called Children's Printed Word Database (CPWD) developed by Masterson and colleagues (2010) in the United Kingdom. This word corpus compiled the reading vocabulary from 1011 books reportedly used by the school authorities in England for students aged five to nine. The word corpus contains 12,193 different words and 995,927 occurrences of these words. The advantages of such an online word corpus include the representativeness of its data and the flexibility of the search criteria, which hence ease the transferring and manipulation of data for the practical use.

Compared to the commonplace of English word corpuses, Malay language word corpuses are few. None to our knowledge has been built based on children's reading materials. The few available Malay language word corpuses include the word corpuses developed by Tan and Sh-Hussain (2009) in Malaysia and Yap, Richard Liow, Sajlia and Siti Syuhada (2010) in Singapore. Both word corpuses were built based on online news sources. However, being a different genre from children's reading materials, these existing word corpuses might not be validly representing the common words experienced by young children in reading. Besides that, research has also provided evidence that adult-based word data has weak associations with children's language skills, compared to data generated from children's reading materials (Masterson et al., 2010; Spencer 2007, 2010). Therefore, aimed for educational and research use in primary schools, it is important to have a Malay Language word corpus that is built based on reading materials that are most familiar to students in the primary schools.

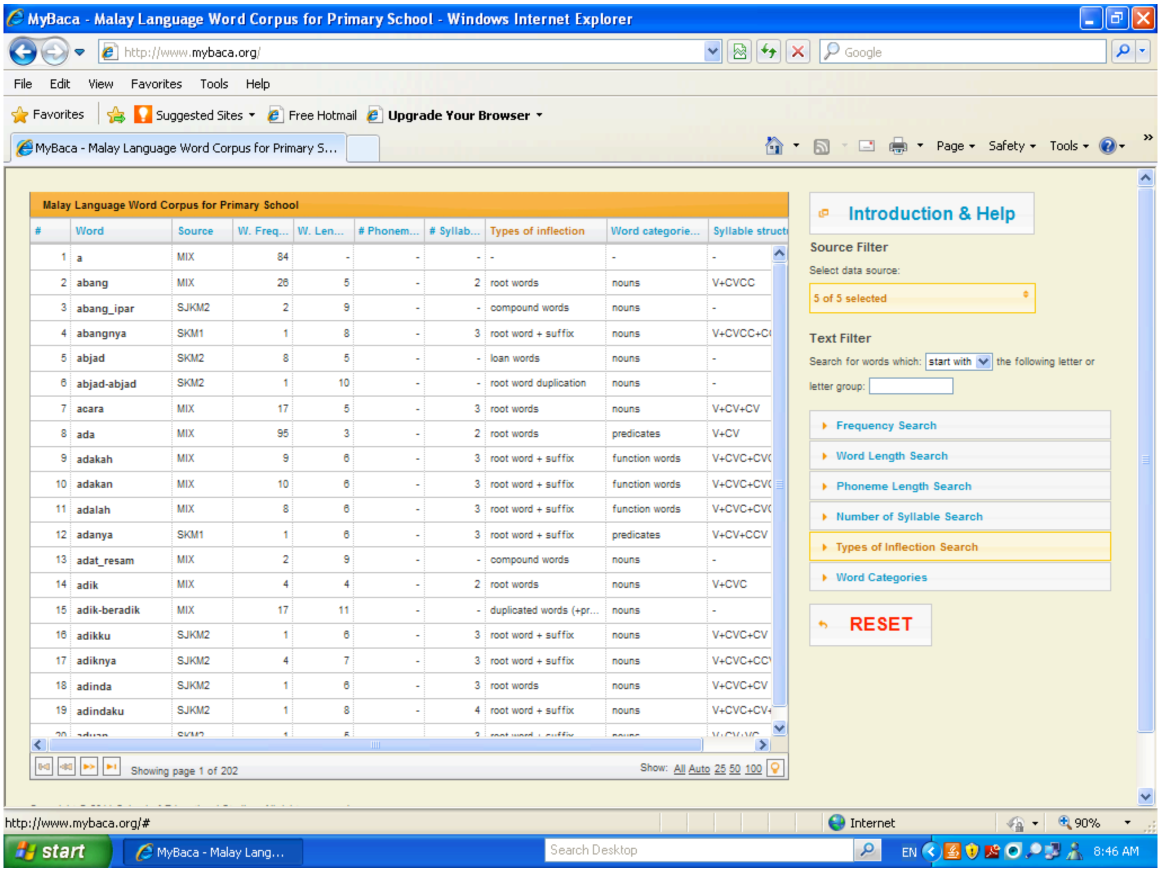
Some specific educational and research functions that can be served by this word corpus include (1) the provision of suitable word set for assessing reading and writing skills in primary-school students, (2) the selection of suitable word choice for teaching and intervention activities, and (3) the selection of appropriate vocabulary to be used in children's literature. The presence of this word corpus will contribute to evidence-based practices in the fields of education and intervention for school-age Malay-speaking students in Malaysia and in the broader Asia-Pacific region. Further, the presence of this word corpus also allows for comparison of word properties to be made across different languages. Such cross-linguistic comparison is possible in this modern time given the open access of different online word corpuses available on the internet.

For this purpose, we built a Malay language word corpus for primary schools based on two sources, the Malay language textbooks used in bilingual National schools and trilingual National-type schools in Malaysia. Until now, the Year 1 and Year 2 textbooks have been successfully processed, analyzed and added into the word corpus. In this project sheet, we will illustrate the methodology of development, the features available on the online word corpus and the future plan.

METHODOLOGY OF DEVELOPMENT AND THE PRIMARY FEATURES

The Malay language textbooks used in the Malaysian primary schools are the primary sources to build the word corpus. Differences in the textbook designs were noted across the textbooks used. For standardization, only words in the components of body text and activity were converted into digital format for data processing and analysis. The digital data was transferred to the CLAN transcription processing program (MacWhinney 2000) to generate a list of vocabulary, which comprised a total of 4035 words and 27671 occurrences of these words. The vocabulary list was then transferred back to the word processor for coding of linguistic properties, which include (1) frequency of occurrence, (2) word length, (3) phoneme length, (4) number of syllable, (5) type of syllable, (6) word category and (7) syllable structure. Two raters were involved in the coding process and discrepancies of coding were resolved in discussion.

The filtered data with the related linguistic information were uploaded onto the online system. The online system was developed with a focus to achieve balance in aesthetics and usability (Brady and Phillip 2003). Different from Children's Printed Word Database (CPWD) developed by Masterson and colleagues (2010) in the United Kingdom, colors were added to Malay Language Word Corpus for Primary Schools to make the website more appealing for the users. The search function was also empowered with multi-criteria search filters to produce an aggregated result of all the research filters (refer to Figure 1). To optimize the usability, an 'introduction and help' page was added to provide step-by-step instructions to the users. This online system is currently available at http://www.mybaca.org/.



Malay Language Word Corpus for Primary School

#	Word	Source	W. Freq...	W. Len...	# Phonem...	# Syllab...	Types of inflection	Word categorie...	Syllable struct
1	a	MIX	84	-	-	-	-	-	-
2	abang	MIX	26	5	-	2	root words	nouns	V+CVCC
3	abang_ipar	SJKM2	2	9	-	-	compound words	nouns	-
4	abangnya	SKM1	1	8	-	3	root word + suffix	nouns	V+CVCC+C
5	abjad	SKM2	8	5	-	-	loan words	nouns	-
6	abjad-abjad	SKM2	1	10	-	-	root word duplication	nouns	-
7	acara	MIX	17	5	-	3	root words	nouns	V+CV+CV
8	ada	MIX	95	3	-	2	root words	predicates	V+CV
9	adakah	MIX	9	6	-	3	root word + suffix	function words	V+CVC+CV
10	adakan	MIX	10	6	-	3	root word + suffix	function words	V+CVC+CV
11	adalah	MIX	8	6	-	3	root word + suffix	function words	V+CVC+CV
12	adanya	SKM1	1	6	-	3	root word + suffix	predicates	V+CV+CCV
13	adat_resam	MIX	2	9	-	-	compound words	nouns	-
14	adik	MIX	4	4	-	2	root words	nouns	V+CVC
15	adik-beradik	MIX	17	11	-	-	duplicated words (+pr...	nouns	-
16	adikku	SJKM2	1	6	-	3	root word + suffix	nouns	V+CVC+CV
17	adiknya	SJKM2	4	7	-	3	root word + suffix	nouns	V+CVC+CC
18	adinda	SJKM2	1	6	-	3	root words	nouns	V+CVC+CV
19	adindaku	SJKM2	1	8	-	4	root word + suffix	nouns	V+CVC+CV+
20	aduan	SKM2	1	5		3	root word + suffix	nouns	V+CV+VC

Showing page 1 of 202

Show: All Auto 25 50 100

Introduction & Help

Source Filter

Select data source: 5 of 5 selected

Text Filter

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

- Frequency Search
- Word Length Search
- Phoneme Length Search
- Number of Syllable Search
- Types of Inflection Search
- Word Categories

RESET

Figure 1: Main page of the online Malay Language Word Corpus for Primary Schools

As shown in Figure 1, this word corpus comes with information on seven word properties, which are:

(1) frequency of occurrence in the sources (i.e., the Malay language textbooks),
(2) word length (e.g., the word '*abang*' consists of 5 letters, 'a', 'b', 'a', 'n' and 'g')
(3) phoneme length (e.g., the word '*abang*' consists of 4 phonemes, /a/, /b/, /a/, /ŋ/)
(4) number of syllable (e.g., the word '*abang*' consists of two syllables, 'a' and 'bang')
(5) type of inflection (e.g., the word '*abang*' (meaning: brother) is a root word)
(6) word category (e.g., the word '*abang*' (meaning: brother) is a noun)
(7) syllable structure (e.g., the word '*abang*' has a CV+CVCC structure, C=consonant, V=vowel)

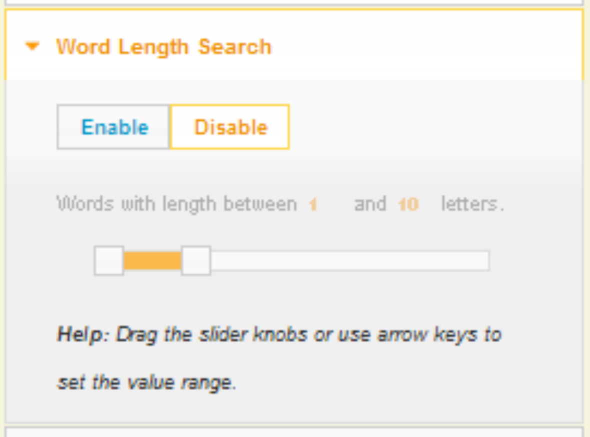
The users may actively search for words with the former six word properties in this online word corpus. For this purpose, the search is organized into different search filters. By default, all the search filters are disabled to allow the users to perform a new search. The users may then choose to enable either one (e.g., only word length) or more search filters (e.g., word length and syllable structure) according to their aim of search. To start a search by using a selected search filter (e.g., word length search), the steps are as follows:

(1) Click RESET to clear previous searches, if any

(2) Select the Source (tick on the boxes for source options) and the Text filter (e.g., search for a word which 'begin' with the letter 'a')
(3) Click to open the desired search filter(s) and then click on the Enable button
(4) Drag the slider knobs or use arrow keys to set the value range (refer to Figure 2)

For example

- 1-2 for words with letter length of 1 to 2
- 2-2 for words with letter length of 2
- 3-5 for words with letter length of 3, 4 or 5



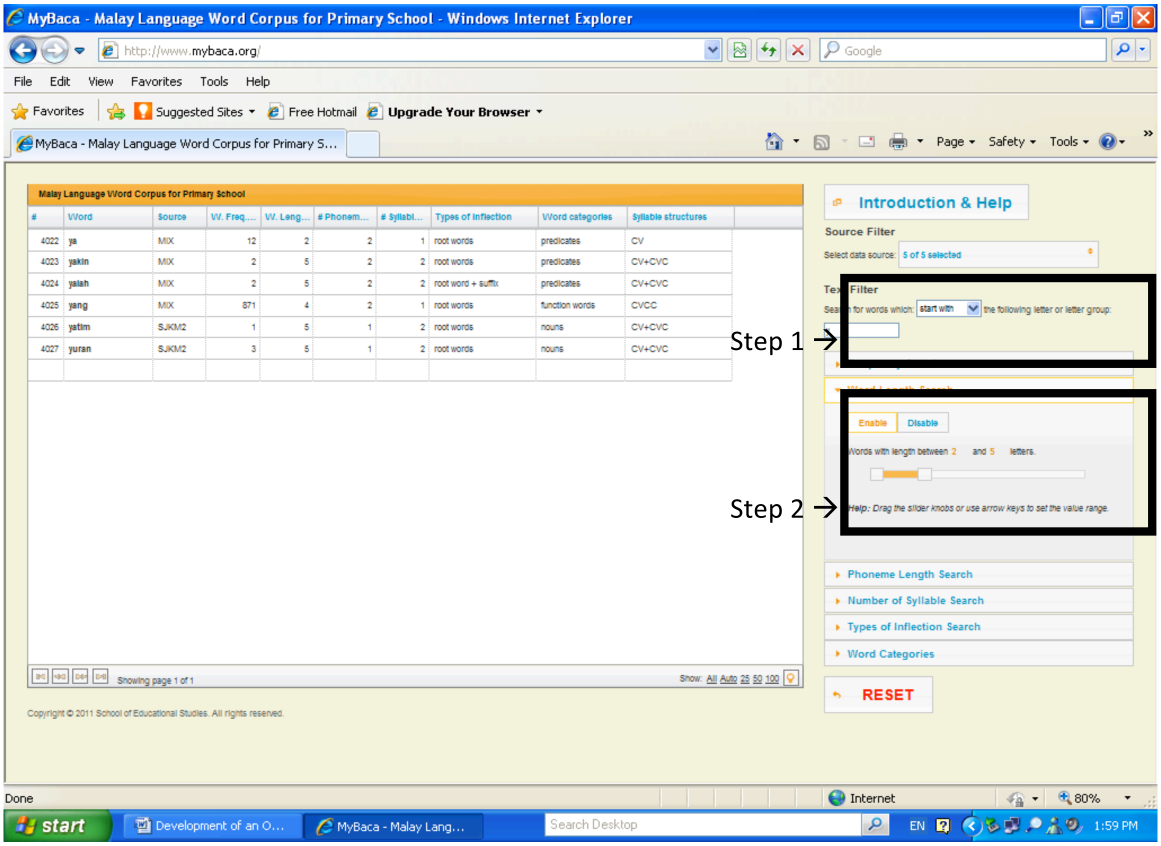
Words with length between 1 and 10 letters.

Help: Drag the slider knobs or use arrow keys to set the value range.

Figure 2: Filter display for word length search

An Example of a Word Search Result

As an example, a user can search for words beginning with the letter 'y', by entering 'y' into the box at Text Filter (refer to Figure 3). The user can narrow down the search by setting the values of the word length, e.g., words with length between 2 and 5 letters. As shown in Figure 3, the system produced a result of six words that matched the criteria, i.e., '*ya*', '*yakin*', '*yalah*', '*yang*', '*yatim*' and '*yuran*'. The other associated linguistic characteristics are also shown on the display table.



Malay Language Word Corpus for Primary School

#	Word	Source	W. Freq....	W. Leng...	# Phonem...	# Syllabl...	Types of Inflection	Word categories	Syllable structures
4022	ya	MIX	12	2	2	1	root words	predicates	CV
4023	yakin	MIX	2	5	2	2	root words	predicates	CV+CVC
4024	yalah	MIX	2	5	2	2	root word + suffix	predicates	CV+CVC
4025	yang	MIX	871	4	2	1	root words	function words	CVCC
4026	yatim	SJKM2	1	5	1	2	root words	nouns	CV+CVC
4027	yuran	SJKM2	3	5	1	2	root words	nouns	CV+CVC

Source Filter

Select data source: 5 of 5 selected

Text Filter

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

Step 1 →

Word Length Search

Words with length between 2 and 5 letters.

Help: Drag the slider knobs or use arrow keys to set the value range.

Step 2 →

Figure 3: Example of a word search attempt

RESEARCH APPLICATIONS AND FUTURE PLAN

The basic framework for the online Malay Language Word Corpus for Primary Schools has been established. This marked the beginning of its continual development where the word corpus will be expanded to include words from a wider range of reading materials for primary-school students, which include textbooks from the higher grades, complementary educational reading materials and leisure reading materials. The expanding of its database will further enhance the representativeness of its data and enrich its research and educational values.

This website has been made available online since July 2011. So far, the website had been referred for both teaching and research purposes by users in South East Asia. In the first month of this prototype launching, the developers had been contacted by a Malaysian user from the field of medical physiology who is in the process of developing a verbal memory test and a Singaporean user who conducts research in word analysis for Malay language. Both users reported their intentions to use this word corpus as a reference tool for their works.

At the same time, a group of pre-service teachers in the Special Education program, Universiti Sains Malaysia had referred to this word corpus to identify words with specific characteristics for developing multimedia talking books in a structured and cumulative manner. The books contained words with controlled linguistic characteristics such as words with only simple consonant and vowel structures (e.g., CVCV) to help students with dyslexia to read. These actual applications of this online word corpus provided the initial evidence towards proving the functionality of this online word corpus for both research and teaching purposes.

For the next step of this project, we aim to investigate the associations of the common words generated from this word corpus with the language learning skills of school-age students and also to assess the relations of these common words with the manifestations of reading and spelling difficulties experienced by low-performing students, such as those with dyslexia. We also aim to introduce this online word corpus and to promote its functions to a wider range of users, especially those in the field of education. We hope that more people could gain benefits from this product of information and communication technology. Most importantly, the users' feedbacks will be gathered for usability, content and readability evaluation to further improve the features available on this online system and the content structure of the corpus.

REFERENCES

Brady, L and C Phillip. 2003. "Aesthetics and usability: A look at color and balance." Usability News 5(1-4).

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

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

MacWhinney, B. 2000. The CHILDES project: Tools for analyzing talk (3rd ed). Mahwah: Lawrence Erlbaum Associates.

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.

Spencer, K. 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.

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.

Tan, T.S. and Sh-Hussain. 2009. "Corpus design for Malay corpus-based speech synthesis system." American Journal of Applied Sciences 6(4):696-702.

Yap, M.J., S.J. Richard Liow, Sajlia Binte Jalil and Siti Syuhada Binti Faizal. 2010. "The Malay lexicon project: A database of lexical statistics for 9,592 words." Behavior Research Methods 42(4):992-1003.