ORIGINAL ARTICLE

An Exploratory Corpus Study of Malay Word Structures in School-Aged Children

Fatin Syatirah Binti Fakrul Razi [1], Saiful Adli Bin Jamaluddin [2], Wan Aslynn Wan Ahmad [2,*]

[1]ReGen Rehab Hospital, 46200 Petaling Jaya, Selangor, Malaysia
[2]Department of Audiology & Speech-Language Pathology Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia

ABSTRACT

Introduction: Corpus-based research plays an essential role in understanding language structures relevant to education and clinical practice. This study provides an initial exploration of Malay word structures in school-aged children, focusing on grapheme-phoneme correspondences, syllable patterns, and inflectional morphemes. The findings aim to support linguistic research and inform speech-language pathology interventions. **Methods:** A descriptive analysis was conducted on a paediatric Malay corpus comprising the 1,000 most frequent words used by school-aged children. The words were analysed in terms of grapheme-phoneme relationships, syllable structures, and inflectional morphemes. Data were summarised descriptively. **Results:** The majority of words demonstrated direct grapheme-phoneme correspondence. Disyllabic words (50%) and CV syllable structures were the most frequent patterns identified (47%). Inflectional analysis revealed that root words (65%) predominated, with relatively fewer affixed or reduplicated forms. **Conclusion:** This preliminary study highlights key features of Malay word structures in school-aged children, confirming the language's transparent orthography and prevalence of simple syllabic forms. These findings provide a foundation for future research and have potential applications in educational resources and speech-language pathology, particularly in developing assessment and intervention tools.

Keywords:
Malay corpora; word structures; school-aged children; grapheme-phoneme correspondence; morphology

INTRODUCTION

A corpus is a collection of texts, conversations, or speeches used to describe and analyse a language (Sinclair, 2005). In Malaysia, the Institute of Language and Literature (Dewan Bahasa dan Pustaka) provides an online database (http://sbmb.dbp.gov.my/) that enables the construction of Malay corpora from a wide range of sources, including newspapers, novels, and magazines. This database offers flexibility, as corpora can be generated based on different text types, publication sources, or time periods, thus serving as a valuable tool for both linguistic research and educational practice.

The Malay corpus helps in constructing tests and assessments. For example, Jamaluddin (2016), used the most frequent Malay words in a text corpus in developing the Malay Matrix Sentence Test, one of the speech tests in audiological evaluation. Additionally, speech corpus helps in the study of Malay speakers with speech impairments (Rosdi et al., 2017). Rosdi et al. (2017) conducted a study that consisted of a speech corpus for 30 speech impaired children with different diagnosis and severity level and measure the intelligibility of speech production.

For Malay language, Zakaria et al. (2021), generated 1000 most frequent words used in Paediatric Malay Corpora specifically for school-aged children. All words have different word structures such as the grapheme-phoneme correspondence.

The corpora for adults and children typically differ, as the most common words used in daily communication are not the same across age groups in Malaysia. Zakaria et al.

* *Corresponding author.*
E-mail address: wanaslynn@iium.edu.my

Journal homepage: https://journals.iium.edu.my/ijahs/index.php/IJAHS
EISSN NO 2600-8491

(2021) for instance, demonstrated that the phoneme distribution in the first 1,000 most frequent Malay words used by children differs substantially from that reported by for adults in Jamaluddin (2016). These differences are evident in vocabulary selection, with adult corpora containing more complex and varied word forms, while children's corpora reflect simpler lexical patterns. Farwell (2009) further notes that children's language is generally shorter and less complex than that of adults, which tends to involve greater variation and ambiguity.

The Malay Word Structures

Malay exhibits a transparent orthography and relatively complex syllable structures compared to languages such as English (Lee et al., 2013). The language contains four basic syllable types: V, VC, CV, and CVC, where 'V' represents a vowel and 'C' represents a consonant. The word *emak* /ə.mak/ (mother) consists of two syllables. The first syllable follows a V structure, while the second syllable follows a CVC structure. For the word *amboi* /am.boi/ (an utterance of disbelief or surprise), there are two syllables, which are VC and CVC respectively. Loanwords, however, often introduce more complex forms such as CCV structures (Sariyan, 2004, as cited in Lee et al., 2014). An example of a loanword in Malay *anggur* /aŋgur/ (grapes, borrowed from Persian) in which there is a consonant cluster CCV in the word.

Syllabic and phonetic structures are central to characterising Malay words (Lee et al., 2011). For instance, words of varying lengths may display distinct syllabic patterns, with long-string and short-string words showing different structural configurations. Moreover, the language's agglutinative nature means that affixation plays a key role in extending meaning. Affixes in Malay occur in several forms; prefixes, suffixes, circumfixes, and infixes; which systematically modify root words to create new grammatical or semantic functions (Benjamin, 2009).

Text corpora have been widely recognised across disciplines as valuable sources of linguistic and educational data (Awang, 2020). For example, Lee et al. (2013) analysed Malay word structures in children's storybooks and compared them with English. However, there remains limited research focusing specifically on the words most frequently used by school-aged children. To address this gap, the present study examines word structures in a paediatric Malay corpus.

The primary aim of this study is to analyse the structure of Malay words in school-aged children's for 7-12 years of age, corpora compiled by Zakaria et al. (2021). This work is particularly relevant to the field of speech-language pathology (SLP) in Malaysia, as it provides insights that can help identify, elicit, and correct morphological errors in children with speech and language difficulties. Since children with language impairments often produce incorrect word structures, understanding the most frequent forms used by their peers can guide clinicians in recognising and addressing such errors. Previous findings have highlighted the value of corpora in clinical applications, for instance, Mazenan et al. (2014) demonstrated the use of Malay corpora in articulation disorder interventions, while Carlisle et al. (2010, as cited in Lee et al., 2014) noted the role of morphological knowledge in reading comprehension.

Accordingly, this study pursues four specific objectives:

1. To identify grapheme-phoneme correspondences.
2. To examine syllable counts and structures within the corpus.
3. To describe the types of inflectional morphemes; prefixes, suffixes, circumfixes, reduplications, and irregular forms; present in the corpus.

MATERIALS AND METHODS

This study employed a descriptive design to analyse the structure of Malay words in a paediatric corpus. The corpus, developed by Zakaria et al. (2021), comprises the 1,000 most frequent words used by school-aged children. These words were examined for grapheme-phoneme correspondences, syllabic structures, and inflectional morphemes.

The Corpus

The corpus (Zakaria et. Al, 2021) was derived from both written and spoken sources of the Malay language. The written component was extracted from Malaysian primary school textbooks, and the 1,000 most frequent words were identified using the frequency of occurrence function. The spoken component was compiled from popular Malay animated films and television series, including *Geng: Pengembaraan Bermula*, *Boboiboy: The Movie*, *Upin dan Ipin*, and *Ejen Ali*, after which all dialogue was transcribed into written text.

To construct the final Paediatric Malay Corpus, the 1,000 most frequent words from the written and spoken sources were tabulated, cross-checked, and merged into a single list. Redundant entries were removed, and the frequency of each word was retained to ensure representativeness.

Criteria for Study Characteristics

Only Malay words, fillers, and exclamations were retained in the corpus, as these categories reflect the natural language use of school-aged children. In contrast, English words and abbreviations or short forms were excluded to maintain the linguistic focus on Malay and to avoid distortions arising from code-switching or non-standard orthography.

Research Procedure

The analysis of word structures in the paediatric Malay corpus was carried out at three linguistic levels: grapheme-phoneme correspondence, syllabic structure, and inflectional morphemes.

Grapheme-Phoneme Correspondence

Grapheme-phoneme correspondence was examined to determine the relationship between written symbols and spoken sounds. A phoneme is defined as the smallest linguistically distinctive unit of sound, whereas a grapheme represents its written form (Tan et al., 2012). For each word in the corpus, the number of letters and phonemes was calculated. Words were then classified into two categories: (1) those with a direct one-to-one correspondence between graphemes and phonemes, and (2) those without such direct correspondence.

Syllabic Structure

The number of syllables in each word was calculated to identify overall syllabic patterns within the paediatric Malay corpus. For example, the word *kita* (we) consists of two syllables, whereas *aktiviti* (activity) comprises four. The structural composition of syllables was also analysed, such as the V.CVC pattern found in the word *ikan*. This analysis enabled the identification of the most frequent syllable types and distributions.

Inflectional Morphemes

The analysis of inflectional morphemes was conducted to determine the grammatical modifications of root words. Inflectional morphemes, such as prefixes, suffixes, circumfixes, and reduplications, were identified and categorised. The relative frequencies of root words, prefix–root word combinations, root word–suffix combinations, circumfixed forms, and reduplicated words were calculated. This analysis provided insight into the distribution and prevalence of morphological structures within the corpus. Figure 1 shows the flowchart of the procedure carried out.

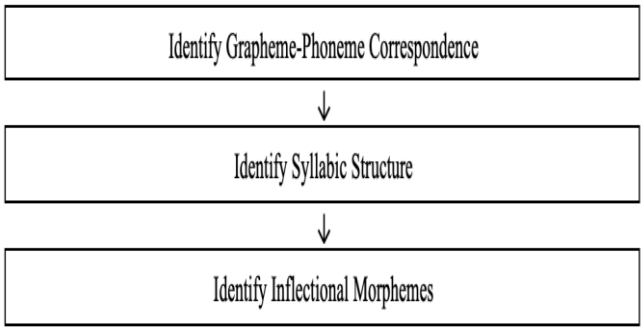

Figure 1: The Research Flowchart

RESULTS

The study data were analysed and presented descriptively. In addition, the results were organised into a summary table to provide a clear overview of the findings in line with the stated objectives and research questions.

Grapheme-Phoneme Correspondence

A total of 1,000 words from the paediatric Malay corpus (Zakaria et al., 2021) were analysed. As shown in Figure 2, 86% of the words displayed a direct one-to-one correspondence between graphemes and phonemes, reflecting the high transparency of Malay orthography.

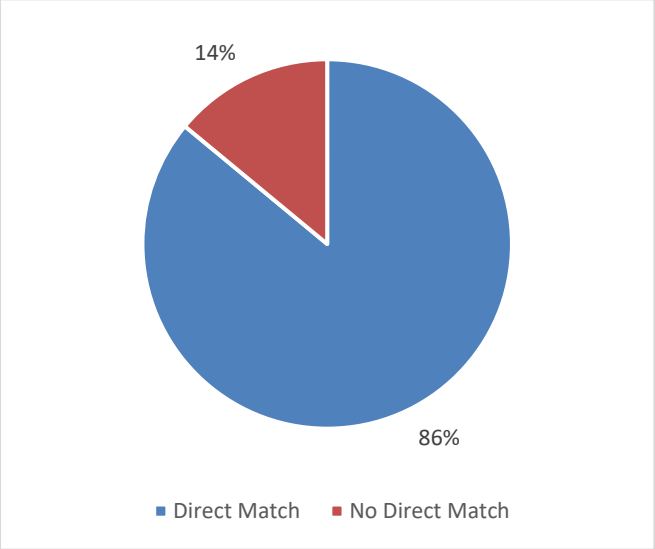

Figure 2. Grapheme-Phoneme Correspondence in the Corpus

Syllabic Structures

Syllabic structures were examined in two aspects: the number of syllables and the types of syllable patterns. As indicated in Figure 3, disyllabic words were the most frequent (50%), whereas words with five syllables represented the smallest proportion (3%).

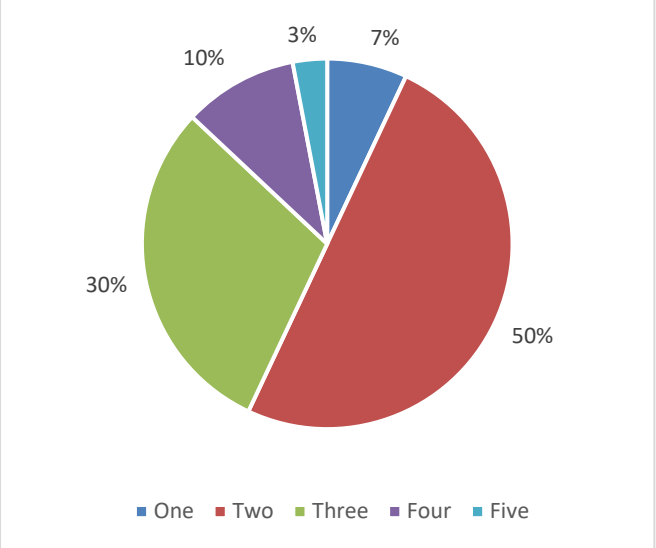

Figure 3. Distribution of Number of Syllables in the Corpus

Analysis of syllable types (Figure 4) revealed that the CV structure was the most common (47%), while the VC structure was the least common (7%).

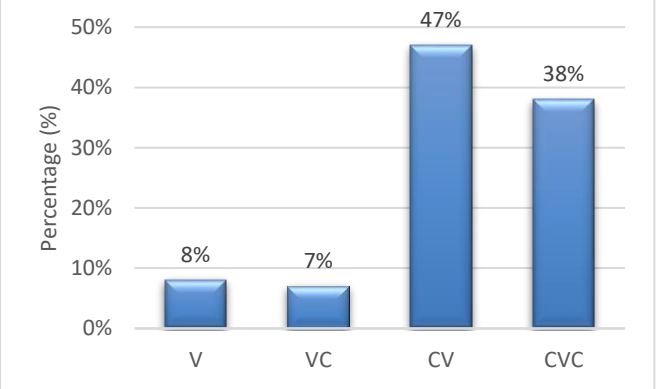

Figure 4. Distribution of Syllable Types in the Corpus

Inflectional Morphemes

Inflectional morphemes in the corpus included prefixes (e.g., *ber+jaya*), suffixes (e.g., *makan+an*), circumfixes (e.g., *ber+dasar+kan*), reduplications (e.g., *gambar-gambar*), and affixed reduplications (e.g., *men+jerit-jerit*). As presented in Figure 5, root words accounted for the majority of cases (65%), while reduplicated words were the least frequent (3%).

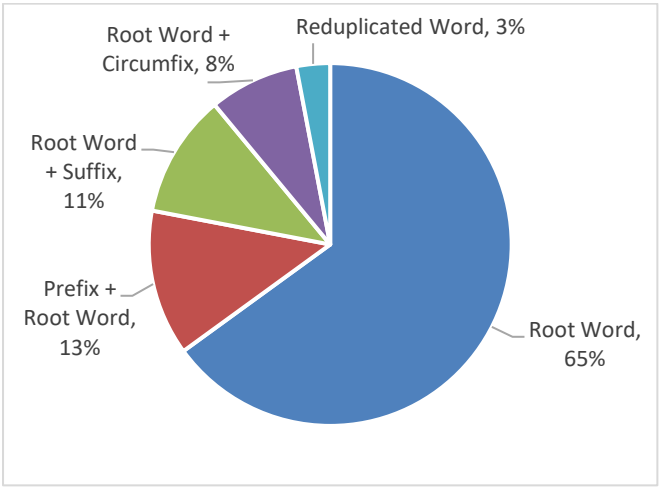

Figure 5. Distribution of Inflectional Morphemes in the Corpus

Overall Findings

Across all three analyses, the corpus revealed strong grapheme-phoneme transparency, a predominance of disyllabic words with CV structures, and a higher proportion of root words compared to affixed or reduplicated forms. These findings highlight the regularity and simplicity of Malay word structures commonly used by school-aged children.

DISCUSSION

This study conducted a descriptive analysis of word structures in the paediatric Malay corpus, focusing on grapheme-phoneme correspondences, syllabic structures, and inflectional morphemes. The results revealed clear patterns that align with initial assumptions: a predominance of words with direct grapheme-phoneme correspondences, a higher proportion of disyllabic words with CV syllable structures, and a greater frequency of root words compared to affixed forms.

Grapheme-Phoneme Correspondence

The analysis of grapheme-phoneme correspondences highlights the transparency of Malay orthography, where letters typically map directly onto sounds. This finding is consistent with Lee et al. (2013), who also reported strong grapheme-phoneme alignment in Malay. Such transparency supports word recognition and fluency, making Malay relatively accessible for early reading

acquisition. Previous research has shown that grapheme-phoneme knowledge is a strong predictor of decoding ability and reading fluency (Nation & Hulme, 1997; Vernon, 1993). Beal-Alvarez (2011) further demonstrated that grapheme-phoneme awareness is a critical early reading skill, predicting variance in children's literacy outcomes (as cited in Anthony & Lonigan, 2004; Lonigan, Burgess, & Anthony, 2000). For clinical practice, these findings provide important implications for Speech-Language Pathologists (SLPs). For example, children with speech sound disorders may produce errors such as final consonant deletion (*makan* → /maka/), which can be identified and addressed by comparing with the expected grapheme-phoneme structure.xxx

Syllabic Structures

The syllabic analysis revealed that most Malay words in the corpus are disyllabic and structured around CV patterns, reflecting the typical phonological organisation of the language. This supports earlier findings by Lee et al. (2013). Syllable awareness is a key component of phonological awareness, which plays an important role in literacy development (Ott, 1997; Wright & Jacobs, 2003). For SLPs, accurate knowledge of syllable boundaries is essential for detecting errors in children's speech production. Moreover, the prevalence of CV structures suggests that syllabic features should be integrated into early reading instruction. As Lee et al. (2013) noted, incorporating disyllabic word stimuli in intervention programmes enhances children's decoding skills and supports phonemic manipulation. These findings underscore the importance of addressing both syllabic and phonemic awareness in therapy and instructional contexts.

Inflectional Morphemes

The analysis of inflectional morphemes showed that root words predominated in the corpus, with fewer instances of affixed or reduplicated forms. This indicates limited morphological variation in the vocabulary most frequently used by school-aged children. Morphological knowledge is vital for spelling and literacy development. Pacton et al. (2013) emphasised the principle of root consistency, whereby roots often retain their spelling across related forms. For example, *makan* serves as the base for *makanan*. Strong morphological instruction from the early grades can therefore reinforce spelling accuracy and vocabulary expansion (Apel & Laurence, 2011). For SLPs, understanding morphological structures is crucial in assessing and intervening with children who present with morphological or language-related difficulties. As children are expected to manipulate word structures across different contexts, morphological awareness becomes an important tool for both reading and language learning. Deacon et al. (2017) provided further evidence that children's reading development is sensitive to the morphological structure of derivative words.

Overall Implications

Taken together, the findings demonstrate that Malay word structures at the graphemic, syllabic, and morphological levels contribute to literacy development and provide a foundation for clinical application in speech-language pathology. By highlighting the transparency of Malay orthography, the prominence of CV syllables, and the importance of root word morphology, this study underscores the need to integrate word structure knowledge into both educational and therapeutic practices for school-aged children.

CONCLUSION

This study analysed Malay word structures in school-aged children's corpora through three dimensions: grapheme-phoneme correspondences, syllabic structures, and inflectional morphemes. The findings confirmed the transparency of Malay orthography, the predominance of disyllabic CV structures, and the high frequency of root words relative to affixed forms. These results highlight the key orthographic and phonological features of Malay and provide practical implications for (SLPs. Specifically, knowledge of common word structures can assist SLPs in distinguishing between typical and erroneous productions, thereby improving intervention for children with speech and language difficulties. Early identification and remediation are crucial, as challenges in word structure can affect children's language, vocabulary, speech development, and reading ability.

Limitations and Suggestions

This study has several limitations. First, the analysis was based solely on the school-aged Malay corpora generated by Zakaria et al. (2021), without incorporating additional sources such as spontaneous speech samples directly produced by children. As a result, the corpus may not fully capture the natural variation present in children's everyday language use. Second, the analysis was time-intensive and required careful manual verification, which may have constrained the depth of exploration.

For future research, it is recommended that an updated version of the school-aged Malay corpora be developed using a broader range of sources, including children's

spoken data, as well as written materials such as newspapers, books, and magazines. A more diverse and representative corpus would provide a richer basis for understanding Malay word structures and better reflect children's actual linguistic output. Such advancements would strengthen both linguistic research and its clinical applications in speech-language pathology.

ACKNOWLEDGEMENT

This research was not funded by any grant.

REFERENCES

Abu Bakar, J., Khairuddin, K., Faidzul Nasrudin, M., & Zamri Murah, M. (2018). Pos-tagging malay corpus: A novel approach based on maximum entropy. *International Journal of Engineering & Technology*, *7(3.20),* 6. https://doi.org/10.14419/ijet.v7i3.20.18721

Alias, S., Mohammad, S. K., Keng Hoon, G., & Tien Ping, T. (2016). A Malay text corpus analysis for sentence compression using pattern-growth method. *Jurnal Teknologi, 78(8).* https://doi.org/10.11113/jt.v78.7413

Anderson, C. (2018). 6.3 inflectional morphology. Essentials of Linguistics. https://pressbooks.pub/essentialsoflinguistics/chapter/6-4-inflectional-morphology/#:~:text=Inflectional%20morphemes%20are%20morphemes%20that,a%20noun%20is%20inflectional%20morphology.

Awal, N. M., Ho-Abdullah, I., & Zainudin, I. S. (2014). Parallel corpus as a tool in teaching translation: Translating English phrasal verbs into Malay. *Procedia - Social and Behavioral Sciences, 112,* 882–887. https://doi.org/10.1016/j.sbspro.2014.01.1245

Bakar, N. S. (2020). The development of an integrated corpus for Malay language. *Lecture Notes in Electrical Engineering,* 425–433. https://doi.org/10.1007/978-981-15-0058-9_41

Beal-Alvarez, J. S., Lederberg, A. R., & Easterbrooks, S. R. (2011). Grapheme-phoneme acquisition of deaf preschoolers. Journal of Deaf Studies and Deaf Education, 17(1), 39–60. https://doi.org/10.1093/deafed/enr030

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112(4),* 417–436. https://doi.org/10.1016/j.jecp.2012.01.005

Chung, S.-F. (2011). Uses of ter- in Malay: A corpus-based study. *Journal of Pragmatics, 43(3),* 799–813. https://doi.org/10.1016/j.pragma.2010.10.004

Cohen-Mimran, R., Reznik-Nevet, L., Gott, D., & Share, D. L. (2022). Preschool morphological awareness contributes to word reading at the very earliest stages of learning to read in a transparent orthography. Reading and Writing. https://doi.org/10.1007/s11145-022-10340-z

Deacon, S. H., & Francis, K. A. (2017). How children become sensitive to the morphological structure of the words that they read. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01469

Farwell, C. B. (2009). The language spoken to children. Karger Publishers. https://karger.com/hde/article-abstract/18/4/288/156207/The-Language-Spoken-to-Children?redirectedFrom=fulltext

Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in Medical Informatics. *Journal of the American Medical Informatics Association, 13(1),* 16–23. https://doi.org/10.1197/jamia.m1749

Ibrahim, B., Yunus, K., & Ibrahim, B. (2018). Perspectives on corpus linguistics: The methodological synergy in Second language pedagogy and research. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3258776

Jamaluddin, S. A. (2016). Development and evaluation of the digit triplet and auditory visual matrix sentence tests in Malay.

Lee, L. W., Low, H. M., & Mohamed, A. R. (2012). Word count analysis of Malay language textbooks for the purpose of developing a malay reading remedial programme. *Writing Systems Research, 4(1),* 103–119. https://doi.org/10.1080/17586801.2012.690713

Lee, L. W., Low, H. M., & Mohamed, A. (2013). A

Comparative Analysis of Word Structures in Malay and English Children's Stories. Undefined. https://www.semanticscholar.org/paper/A-Comparative-Analysis-of-Word-Structures-in-Malay-Lee-Low/902612e86018adf08602df263ec949daee5f42f6

Lee, L. W., & Low, H. M. (2014). Analysis of Malay word structure by pre-service special education teachers: Foundation-Level Knowledge for remedial instruction. *Australian Journal of Learning Difficulties, 19(1),* 33–46. https://doi.org/10.1080/19404158.2014.891531

Leech, G. N. (2011). Principles and applications of corpus linguistics. *Perspectives on Corpus Linguistics,* 155–170. https://doi.org/10.1075/scl.48.10lee

Louleli, N., Hämäläinen, J. A., Nieminen, L., Parviainen, T., & Leppänen, P. H. T. (2020). Dynamics of morphological processing in pre-school children with and without familial risk for dyslexia. Journal of Neurolinguistics, 56, 100931. https://doi.org/10.1016/j.jneuroling.2020.100931

Maseri, M., & Mamat, M. (2018). Malay language speech recognition for preschool children using Hidden Markov model (HMM) system training. *Lecture Notes in Electrical Engineering,* 205–214. https://doi.org/10.1007/978-981-13-2622-6_21

Mazenan, M. N., Swee, T. T., & Soh, S. S. (2014). Recognition test on highly newly robust Malay corpus based on statistical analysis for Malay articulation disorder. *The 7th 2014 Biomedical Engineering International Conference.* https://doi.org/10.1109/bmeicon.2014.7017394

Omar, S., Bakar, J. A., Mohd Nadzir, M., Harun, N. H., & Yusoff, N. (2021). Text simplification for malay corpus: A Review. *2021 International Conference on Computer & Information Sciences (ICCOINS).* https://doi.org/10.1109/iccoins49721.2021.9497167

Pacton, S., Foulin, J. N., Casalis, S., & Treiman, R. (2013). Children benefit from morphological relatedness when they learn to spell new words. Frontiers in Psychology, 4. https://doi.org/10.3389/fpsyg.2013.00696

Ramli, I., Jamil, N., Seman, N., & Ardi, N. (2015). An improved syllabification for a better Malay language text-to-speech synthesis (TTS). Procedia Computer Science, 76, 417–424. https://doi.org/10.1016/j.procs.2015.12.280

Rosdi, F., Mustafa, M. B., & Salim, S. S. (2017). Assessing automatic speech recognition in measuring speech intelligibility: A study of Malay speakers with speech impairments. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI).* https://doi.org/10.1109/iceei.2017.8312396

Tan, T.-P., Xiao, X., Tang, E. K., Chng, E. S., & Li, H. (2009). Mass: A Malay language LVCSR Corpus Resource. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments.* https://doi.org/10.1109/icsda.2009.5278382

Tan, T.-P., Goh, S.-S., & Khaw, Y.-M. (2012). A Malay dialect translation and synthesis system: Proposal and preliminary system. 2012 International Conference on Asian Language Processing. https://doi.org/10.1109/ialp.2012.14

Zakaria, N. A., Jamaluddin, S. A., Aslynn, W. A. W., O'Beirne, G., Awatif. N., & Zawani. A. (2021). Development of Pediatric Malay Matrix Sentence Test (PeadMalayMST) materials through construction of Malay corpus.