Journal of Modern Languages
Vol. 32, No.2 (2022)
https://doi.org/10.22452/jml.vol32no2.1

The Development of Paediatric Malay Matrix Sentence Test (PaedMalayMST) Materials through the Construction of a Paediatric Malay Corpus

Nur 'Azzah Zakaria*[1]
azzah@iium.edu.my

[1]Saiful Adli Jamaluddin[2]
jsaiful@iium.edu.my

Wan Ahmad Wan Aslynn[3]
wanaslynn@iium.edu.my

Greg A. O'Beirne[4]
gregory.obeirne@canterbury.ac.nz

Nur Awatif Binti Zulkeflee[5]
awatif_zulkeflee@yahoo.com

Atiffah Zawani binti Abd Razak[6]
atiffahzawani@gmail.com

[1, 2,3,5,6] Department of Audiology and Speech-language Pathology, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia.

[4] Department of Communication Disorders, College of Science, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

Abstract

The selection of material is important in the development of a paediatric speech perception test. Even though several studies have documented various sources of speech testing

* corresponding author

materials, using corpus was recommended to develop a Matrix Sentence Test (MST). However, as available Malay corpus was limited to written words only, this research aims to construct a Paediatric Malay corpus from a combination of written and spoken words of children in Malaysia and to develop word-based matrix (WBM) for the Paediatric Malay Matrix Sentence Test (PaedMalayMST) based on the developed corpus. The WBM replicates the Malay grammatical structure in these three categories: number, object, and adjective. Based on these criteria, the words from the developed corpus were then filtered by considering the frequency of occurrences and its phonemic distributions to be included in the WBM. As a result, a matrix of three categories with six words for each category was constructed.

Keywords: development, Malay, paediatric, matrix sentence test, corpus, phoneme distribution

1. Introduction

The Matrix Sentence Test (MST) was first introduced as a Swedish adult speech test by Hagerman in 1982. Since then, MST is available in other languages such as German (Wagener et al., 1999), Russian (Warzybok, Zokoll, et al., 2015) and Malay (Jamaluddin, 2016). The MST uses five-word sentences as speech stimuli that was generated by a combination of 10 words from five different categories. The words in each category were fixed according to the structure of the language to form a word-based matrix (WBM). This random combination of the WBM can potentially produce up to 100,000 sentences that allow the test to be syntactically fixed but semantically unpredictable. The nature of this test gives it the advantage to be used repeatedly as a diagnostic and rehabilitation tool without any concerns of possible memorization by the patient (Kollmeier et al., 2015).

Currently, the MSTs available for the paediatric population are in German (Wagener & Kollmeier, 2005), Polish (Ozimek et al., 2012), Russian (Garbaruk et al., 2020), Finnish (Willberg et al., 2020), and Italian (Puglisi et al., 2021). These MSTs were developed by adapting and modifying their respective adult MST. The adaptation and modification involved reducing the number of categories in the generation of sentences from five to three to form a pseudo-sentence (Ozimek et al., 2012; Puglisi et al., 2021; Willberg et al., 2020). The based matrix is simplified to three categories after considering the possibility of children's limited memory load on the task and shorter auditory memory span (Case et al., 1982; Mendel, 2008; Wagener & Kollmeier, 2005).

While the MST for children has been accessible in some countries, the development of MST in Malay is still ongoing. The need for age and language-specific speech perception test material is important, as a speech test administered in non-native languages gave higher speech intelligibility score to non-native listeners especially in noise (Wijngaarden, 2001; Marinova-Todd, Siu, & Jenstad, 2011; Warzybok, Brand, Wagener, & Kollmeier, 2015). Factors such as experiences, language proficiency and differences in components and structure of one's first language with that second language influenced the listener's speech recognition results (Wijngaarden, Steeneken, & Houtgast, 2002; Warzybok, Brand, Wagener, & Kollmeier, 2015). A major limitation when the speech test material has effects on the perception and production of speech of non-native listeners is that there is a danger of underestimating a child's auditory ability.

The Paediatric Malay MST (PaedMalayMST) is not directly adapted from the adult Malay MST, but instead it is developed by following a guideline by the International Collegium of Rehabilitative Audiology (ICRA). The ICRA assembled a protocol in 2015 for the development of the MST to assure comparability of the findings between different languages (Akeroyd et al., 2015). This guideline included detailed instructions for selection of material, recording and resynthesis, optimization, evaluation, and validation phases involved in the development of the MST. In the selection of material phase, Akeroyd et al. (2015) emphasize that the words should be familiar especially to children and have similar phonemes distribution with the intended language. Whilst the Polish Paediatric MST (PPMST) has implemented this suggestion by selecting the words from a frequency dictionary, surprisingly, the Oldenburg sentence test for children (Oldenburger Kinder Satztest; Olkisa) (Wagener & Kollmeier, 2005), simplified Russian matrix sentence test (Simplified RUMatrix) (Garbaruk et al., 2020), Finnish simplified matrix sentence test (FINSIMAT) (Willberg et al., 2020) and simplified Italian matrix test (SilMax) (Puglisi et al., 2021) were developed only by adapting the sentence material in the adult version of the MST. In an attempt to follow the recommendation of ICRA, corpus was used to select frequent words to be included as sentence materials.

Corpus is defined as a compilation of writings, conversations and speeches that are used to characterize a language (Sinclair, 2005). The Institute of Language and Literature Malaysia website (http://sbmb.dbp.gov.my) has a database where a corpus of Malay words can be generated from different types of material such as newspapers, books, and magazines. However, when selecting words for children aged six to 12-year-old, the corpus was constructed in a way that did

not allow words to be selected from either primary or secondary textbooks. Thus, the existing word corpus might not be a valid representation of the frequently used words produced by young Malay-speaking-children in Malaysia. In addition, currently, the Malay corpus database only focuses on the written form (*Pangkalan Data Korpus*, 2017). This prioritization of written over spoken forms in a corpus may give imbalance representativeness of a language (Sinclair, 2005). Moreover, speech is the main means of communication and has features that are not presented in a written corpus such as diglossia (Čermák, 2009). Thus, including both written and spoken forms of a language in a corpus will provide an equitable representation of a language.

The primary aim of this research was to develop testing materials for the PaedMalayMST for Malay native speakers aged six to 12-year-old. This aim was achieved by two studies with the specific aims of 1) constructing a Paediatric Malay Corpus from the combination of written and spoken words that were commonly encountered by six to 12-year-old children in Malaysia and 2) developing a WBM by selecting frequent and familiar words from the Paediatric Malay Corpus. At the same time, with the construction of the corpus, phonemic distributions were considered when selecting the words to be included in the WBM for the PaedMalayMST.

2. Methodology

The development of PaedMalayMST began with the construction of the Paediatric Malay Corpus which was a combination of both written and spoken corpora. Based on this developed Paediatric Malay Corpus, the Word-Based Matrix (WBM) for the PaedMalayMST was then selected. Therefore, this paper was divided into two studies with Research Project 1 for the development of Paediatric Malay Corpus and Research Project 2 for the development of WBM.

2.1 Research Project 1: The Development of the Paediatric Malay Corpus

The construction of a corpus depends on the purpose that it will serve (Sinclair, 2005). In this research project, the corpus was developed with the intention to select words that are deemed familiar to Malay children to be the WBM for speech perception assessment application known as PaedMalayMST. In order to fit the purpose, both written and spoken language were included in the development of the Paediatric Malay Corpus. As there is no Malay corpus that uses materials for children younger than seven-year-old, the written and spoken corpora were developed first before both corpora were combined to form a Paediatric Malay Corpus. At the same time,

phonemic balance was considered as an important characteristic when selecting the words for a speech test (Akeroyd et al., 2015; Kollmeier et al., 2015). Therefore, to consider equal phonemic distribution of the based matrix with the Malay language, each word in the corpus was transcribed phonetically according to the most common phonemes in Malay. This method is similar to other available MST such as Russian (Warzybok, Zokoll, et al., 2015), Danish (Wagener et al., 2003) and Italian (Puglisi et al., 2015, 2021). As familiarity of words among Malay children is crucial, the first 1000 words that occurred in each corpus were included in the Paediatric Malay Corpus. One reason why the first 1000 words was included is because Laufer and Nation (1999) found that the most frequent 1000 words cover 75% of written text and 84% of informal speech. The following sections (2.1.1 until 2.1.4), as presented in Figure 1, described the steps taken to develop the Paediatric Malay Corpus.

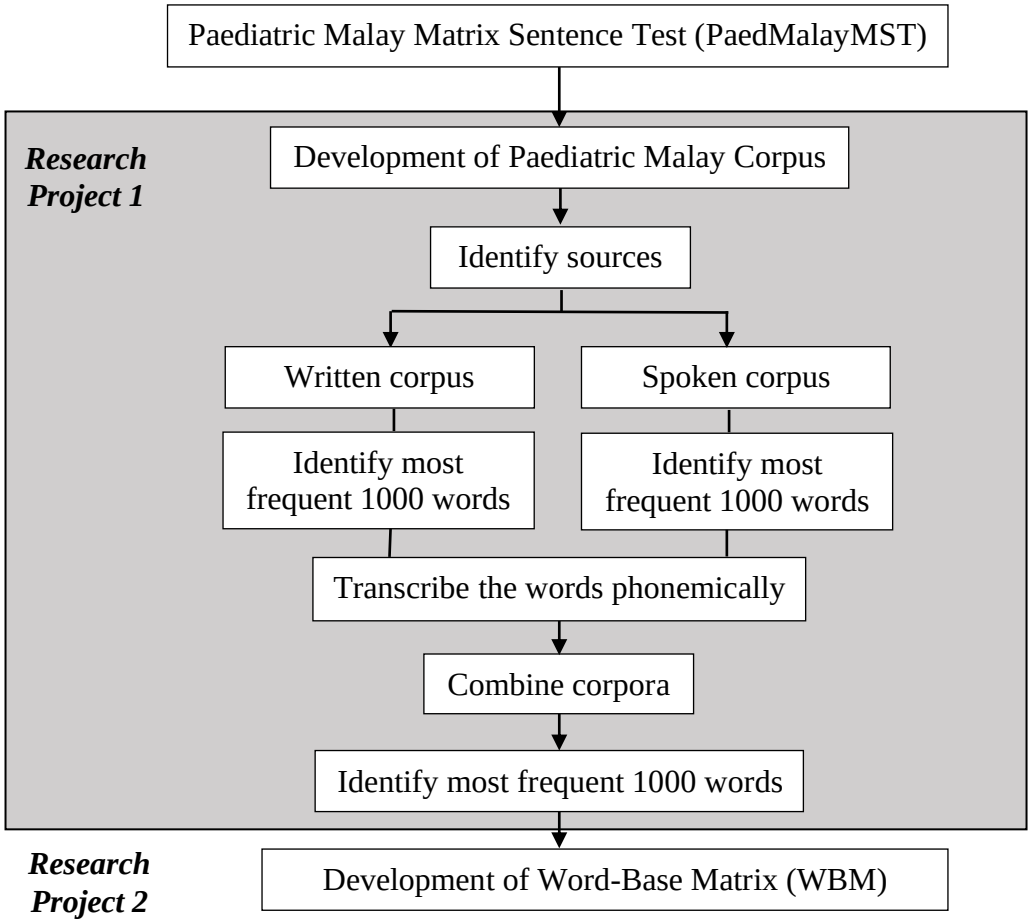

Figure 1: Flow chart of Research Project 1 for the development of Paediatric Malay Corpus

2.1.1 Written Corpus

The written corpus was based on the work of a team of researchers at the School of Educational Studies, Universiti Sains Malaysia. It is available as an open-access database at http://www.mybaca.org/. The word corpus was generated from a corpus of Malay language written materials used in Malaysian primary schools. The words were extracted from Malay language textbooks and at the time of data retrieval, the database produced a collection of 4,035 words from Year 1 and 2 primary school Malay language textbooks (Lee & Low, 2011). The words extracted from the printed text were coded according to seven linguistic properties: frequency of occurrence, word length, phoneme length, number of syllables, types of word, word category, and syllable structure in the database. Thus, the database allowed users to use the interactive features to generate word lists according to certain linguistic features. Using the frequency of occurrence feature in the database, a list of 1000 most frequent words for the written corpus were extracted. These words together with the number of occurrences in the database were then transferred to a spreadsheet for further analysis.

2.1.2 Spoken Corpus

Considering words that are familiar among children younger than primary school age (below 7 years old), a spoken corpus was also included in the development of the Paediatric Malay Corpus. The spoken corpus was developed based on stages recommended by Thompson (2005). The first stage was word retrieval from popular Malay animated movies and television series; *Geng: Pengembaraan Bermula*, *Boboiboy: The Movie*, *Upin dan Ipin* and *Ejen Ali*. These movies and television series were selected to be included in the corpus because they have been aired for a few times at free-to-air television channels. In addition, these movies and television series have received recognition and won several awards that suggest their popularity among Malaysian children. The *Upin and Ipin* animated movie *Geng: Pengembaraan Bermula* has won the viewer's choice award at the Kids Film Festival in 2009 while its television series were awarded 'The Most Popular Video' award during the Web TV Asia Awards in 2016 (Les' Copaque Production, n. d.). At the same time, both *BoBoiBoy* and *Ejen Ali* television series have gained popularity among children in Malaysia as each YouTube channel have reached five and two million subscribers respectively (Animonsta, n. d.).

The selected animated movies and television series then underwent the stage of construction where transcription of all the dialogues into written texts was carried out. The transcription was done in two stages by two researchers. Prior to the transcribing process, both researchers watched the movies and television series to be familiar with the storyline and dialogues. Subsequently, the dialogues were then transcribed into written text. This stage required the researchers to listen to the dialogue back and forth to be able to accurately transcribe all the dialogues. In the second stage of the process, both the researchers rechecked the transcriptions by watching the selected media again. If there were any discrepancies, both researchers listened and rectified it together. The transcriptions were then compiled into a text file for analysis. The compiled text file was referred to as spoken corpus. A corpus analysis software, AntConc software (Anthony, 2018), was used to analyse the spoken corpus. The software ranked the words in the compilation based on frequency of occurrence and a list of 1000 most frequent words were extracted and then transferred to a spreadsheet for further analysis.

2.1.3 Phonetic Transcription of Written and Spoken Malay Corpora

The 1000 words that were retrieved from each of the written and spoken corpora were phonetically transcribed according to Malay phonemes using the International Phonetic Alphabet (IPA) transcription to evaluate the distribution of phonemes. The first process involved transcribing each word in each corpus according to the IPA for the Malay phonemes by two researchers. Following the transcriptions by the researchers, a linguist from the research team reviewed the accuracy of transcriptions, and the accepted transcriptions were combined.

2.1.4 Combination of Corpora to Form the Paediatric Malay Corpus

The 1000 most frequent words from both written and spoken corpus were compiled into a spreadsheet together with each word frequency of occurrence in the corpus to develop the Paediatric Malay corpus. The frequency of occurrence was used as a basis to compile the words from both corpora. The redundant words were identified by cross checking the words between the written and the spoken corpus. Once a redundant word was identified and harmonized, the frequency of occurrence for that word in each corpus was compiled. In the end, the words were then ranked based on the number of occurrences in both corpora. The phonetic transcription of

the words included in the combined corpus were retrieved from each written and spoken corpora to account for the phoneme distribution of the Paediatric Malay corpus.

2.2 Research Project 2: The Development of Word-Based Matrix (WBM)

In this research project, the development of WBM for the PaedMalayMST is shown in Figure 2. The selection of words from the Paediatric Malay Corpus that formed a WBM consisted of three categories: number, object, and adjective are based on several inclusion and exclusion criteria. A word is included in the WBM if it is a root Malay word, disyllabic, spondee and falls under any of these categories; number, object and adjective. As the PaedMalayMST utilized a closed-set of response format, words that are difficult to be represented with illustration were excluded. The sentences for PaedMalayMST are generated by combining each word in the WBM randomly from each category by following the syntax of the Malay language. As it is anticipated that the number of words in each category could not be equal, additional word options are selected as alternatives WBM. The words are then selected by comparing the phonemes distribution between the possible sets of WBM with the Paediatric Malay corpus as phoneme distribution for each specific language is recommended to be taken into consideration for word selection (Akeroyd et al., 2015).

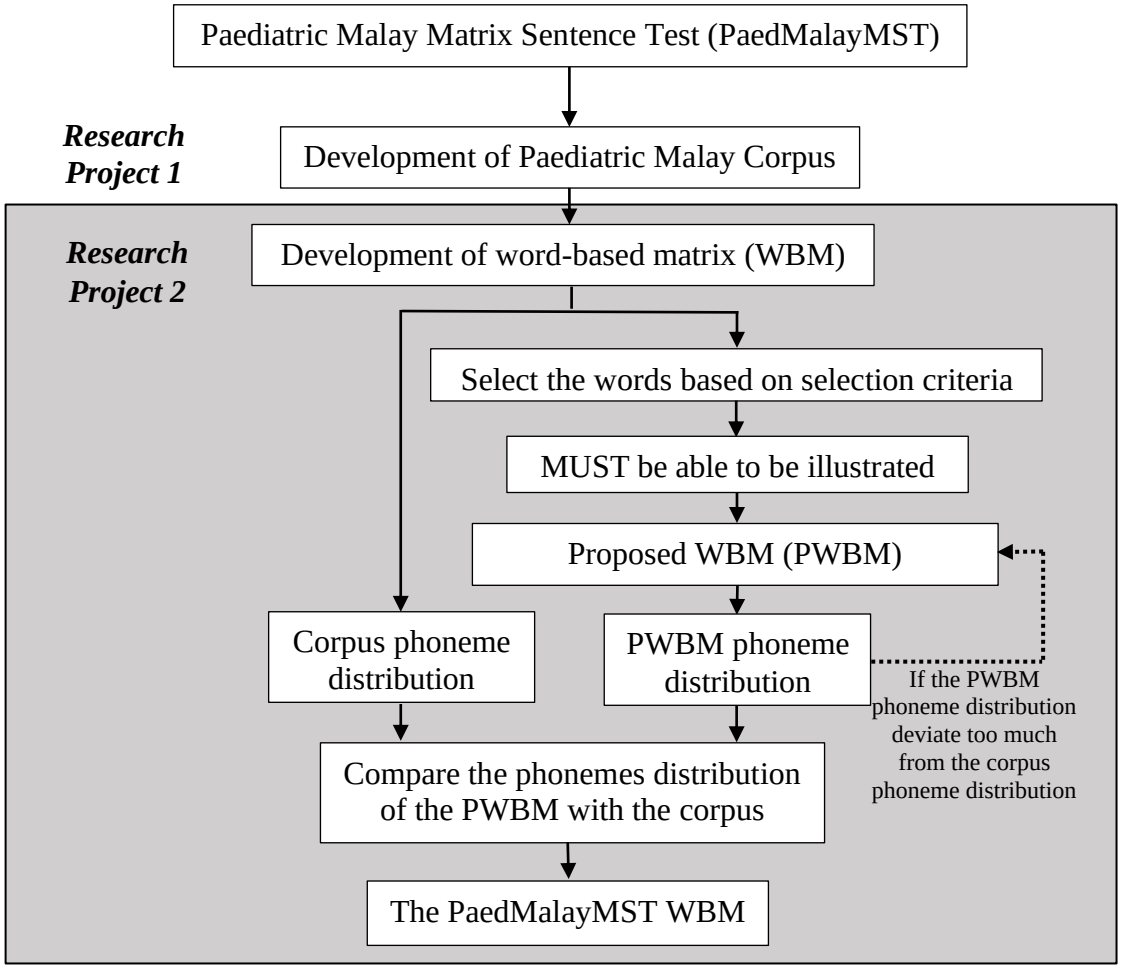

Figure 2: Flow chart of Research Project 2 for the development of Word-Based Matrix (WBM) for the Paediatric Malay Matrix Sentence Test (PaedMalayMST)

3. Results and Discussion

3.1 Research Project 1: The Development of Paediatric Malay Corpus

The written and spoken corpora yielded a total of 4035 and 1900 words, respectively. Due to a redundancy of 307 words within both corpora, the compilation gave a grand total of 1693 words. Ultimately, the Paediatric Malay corpus only accounted for the first 1000 words that were ranked according to the frequency of occurrences in both the written and spoken corpora.

The transcription of the words in each corpus allowed for a comparison of phonemes distribution as in Figure 3. Based on this comparison, the spoken corpus showed distinct

differences in the phonemes /v/, /ʃ / and /oɪ/ compared to the written and Paediatric Malay corpora. The phonemes /v/ and /oɪ/ were contributed by the presence of a lot of borrowed words from English language such as gravity (*graviti*) and activity (*aktiviti*) and proper nouns such as '*Boboiboy*' in one of the selected movies. At the same time, all the selected movies had a lot of filler words such as '*ishhh'*, '*shhh'* and *'haishhh*' in their scripts that led to high frequency of the phoneme /ʃ / in the spoken corpus. The presence of these filler words gave emotional meaning to the scripts by demonstrating anger, frustration, surprise, and action.

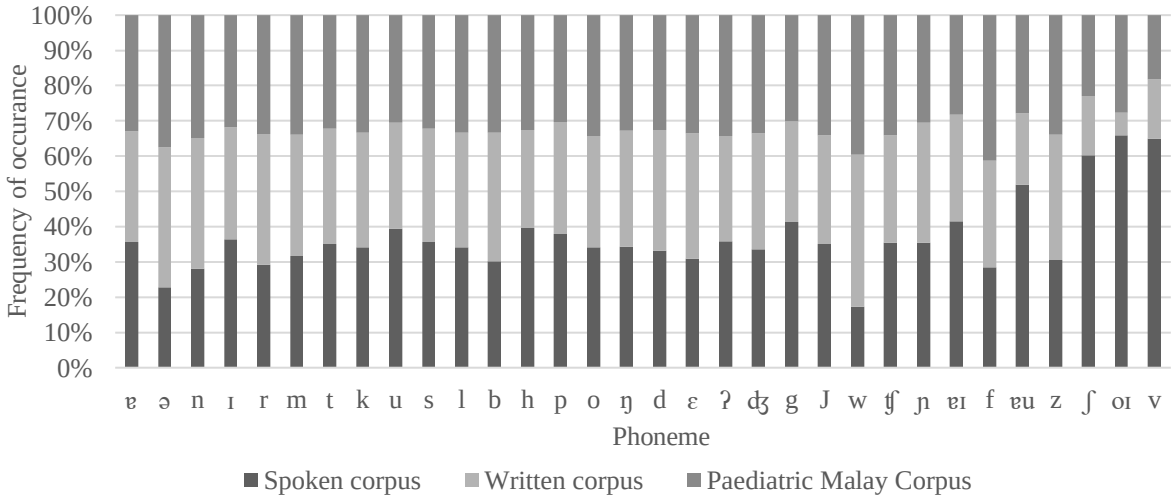

Figure 3: The percentage of phoneme distribution of spoken, written and combined corpora

Even though the frequency of phonemes /v/, /ʃ / and /oɪ/ in the spoken corpus did not represent the Malay phoneme distribution, exclusion of words with these less common occurring phonemes did not support the true purpose of developing the Paediatric Malay corpus. As the aim to develop the Paediatric Malay corpus is to identify words that are familiar among children, thus the inclusion of the words with less common occurring phonemes is valid to represent the language that are available, familiar, and natural in normal everyday speech among children. In addition, the combination of both spoken and written corpora in the end give phoneme distributions that represent the Malay language in the Paediatric Malay corpus. Thus, total exclusion of the spoken sample in the development of a corpus for any language affects the representativeness and balance of that language because speech is encountered as many times as writing (Sinclair, 2005). Jamaluddin (2016) also pointed out that informal words from the spoken

corpus that were not used in newspapers may add importance to the comparison between corpus of a language and the selected words for a speech material. Overall, these findings discussed above suggest that future researchers should take precaution when selecting movies or television series as a sample to represent the spoken corpus of any language.

3.2 Research Project 2: The Development of Word-Based Matrix (WBM)

In order to select words to be included in the WBM, the words in the Paediatric Malay corpus were categorized into three categories: number, object, and adjectives. Words in the object category have the highest count of 57 compared to adjective and number, being 34 and 11, respectively. Multisyllabic words and words that were difficult to illustrate were excluded causing the number of words in each category to be reduced to eight (number), 11 (object) and six (adjective). In the object category, 46 words were excluded from being considered into selection of WBM as these words have negative meanings (e.g., *anjing* (dog), *sampah* (thrash)) or the objects were unfamiliar or obsolete among children (e.g., *dodol* (a type of traditional dessert), *disket* (floppy disk)). Interestingly, in the end, the adjective category consisted of only six colours as words such as *besar* (big) and *kecil* (small) or *baru* (new) and *lama* (old) were excluded. The reasons for the exclusion were because these words need to be represented with an object and challenging to be portrayed. In comparison with Olkisa, FINSIMAT and SilMax that have similar sentence structure as PaedMalayMST, nearly half of the words in the adjectives category consists of colours. However, they also included words such as new, old, big and small because their MSTs were not designed for a closed-set response format (Puglisi et al., 2021; Wagener & Kollmeier, 2005; Willberg et al., 2020). As a result, the words selected in the adjective category for the other paediatric MSTs were not excluded simply because it is difficult to be illustrated.

As the number of words in the adjective category were limited to six compared to other categories, thus the words in the adjective category were fixed. Selection of words in the other two categories were dependent on the phonemic balancing of the WBM compared to the Paediatric Malay Corpus. In order to select the best WBM that closely represent the Paediatric Malay Corpus, six potential WBMs were proposed. These proposed WBMs (PWBM) have fixed words in the adjective category but different combinations of words in the number and object categories.

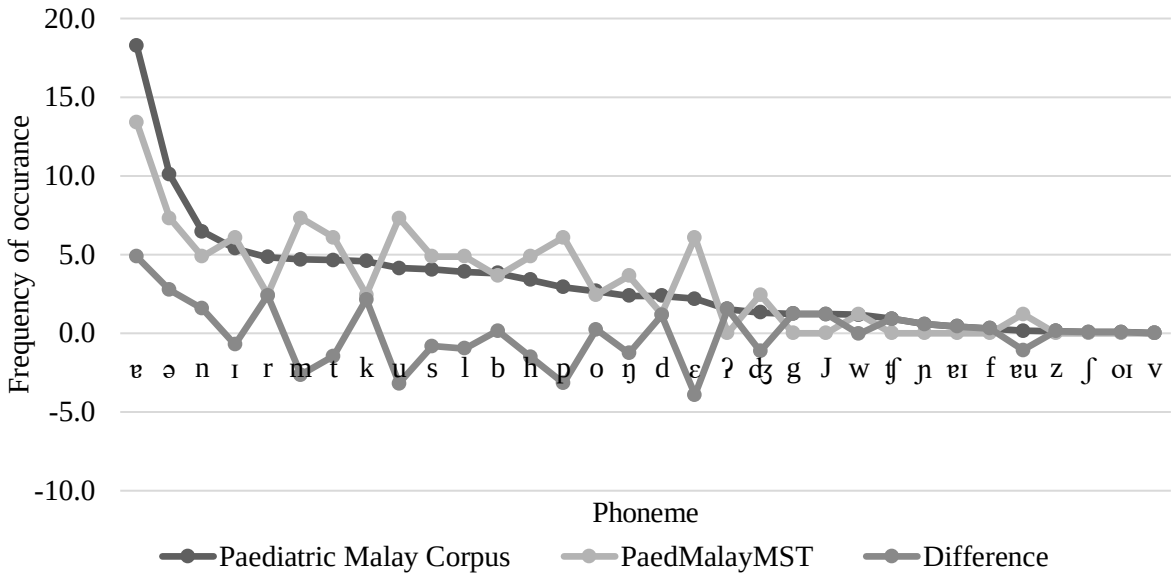

Figure 4: The comparison of phoneme distribution between Word-Based Matrix (WBM) and Paediatric Malay Corpus

Comparison of the phonemes distribution between the six PWBMs have identified that the 4th PWBM have similar phonemes distribution with the Paediatric Malay corpus as shown in Figure 4. Even though a few phonemes in the 4th PWBM such as / **ɐ /, /ɛ** / and /p/ have differences as large as 4.9%, it is still acceptable compared to other PWBM as the phonemes closely represent the Paediatric Malay corpus. These results were in accord with the recommendation outlined by ICRA (Akeroyd et al., 2015) and review of MST tests by Kollmeier et al. (2015) which highlighted the importance of selecting the based matrix according to the phonemes distribution of the language. Thus, the 4th PWBM was selected as the WBM. Table 1 below showed the words that were included in the final WBM from the developed Paediatric Malay Corpus for the PaedMalayMST. Although this six-by-six WBM is small in number of words, it is able to generate 216 unique sentences as the speech stimuli. In addition, it is worth noting that important factors such as frequency of occurrence, phonemic distribution that represents the language and familiarity to the children were taken into consideration when selecting words for this WBM.

Table 1: The Word-Based Matrix (WBM) of Paediatric Malay Matrix Sentence Test (PaedMalayMST) based on the selection process for words and its English translation

Number	**Object**	**Adjective**
Satu (One)	Pensel (n. Pencil)	Putih (White)
Dua (Two)	Bola (n. Ball)	Kuning (Yellow)
Empat (Four)	Pisang (Banana)	Merah (Red)
Lima (Five)	Kasut (n. Shoe)	Biru (Blue)
Enam (Six)	Meja (n. Table)	Hijau (Green)
Lapan (Eight)	Bunga (n. Flower)	Hitam (Black)

4. Conclusion

The early stage of the development of PaedMalayMST was achieved with the formation of the WBM as in Table 1. The current research project's second and third goals were made possible with the construction of Paediatric Malay Corpus from a combination of both spoken and written word corpus. As a result, the three-category WBM for the PaedMalayMST consists of six words in each category. However, this phase needs to be followed by a few crucial steps as recommended by ICRA (Akeroyd et al., 2015). As WBM is available, recording of the sentences from combinations of the words in the WBM will commence followed by optimizing of the recorded speech and pictorial materials, evaluating the list equivalence, and validating with hearing impaired listeners.

Acknowledgments

We are grateful to our former Audiology students at International Islamic University Malaysia; Nur Awatif Binti Zulkeflee and Atiffah Zawani Abd Razak for their assistance with the research. Authors acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2016/SKK06/UIAM/03/1).

References

Akeroyd, M. A., Arlinger, S., Bentler, R. A., Boothroyd, A., Dillier, N., Dreschler, W. A., Gagné, J. P., Lutman, M., Wouters, J., Wong, L., & Kollmeier, B. (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. *International Journal of Audiology*, *54*(November 2014), 17–22. https://doi.org/10.3109/14992027.2015.1030513

Animonsta Studios (n. d.). Boboiboy. Monsta. https://animonsta.com/boboiboy/

Anthony, L. (2018). *AntConc (Version 3.5.7) [Computer Software]* (3.5.7). Waseda University. http://www.antlab.sci.waseda.ac.jp/

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

Čermák, F. (2009). Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics*, *14*(1), 113–123. https://doi.org/10.1075/ijcl.14.1.07cer

Garbaruk, E. S., Goikhburg, M. V., Vazhibok, A., Tavartkiladze, G. A., Pavlov, P. V., & Kollmeier, B. (2020). Application of the Russian-language version of the matrix phrasal test in children. *Bulletin of Otorhinolaryngology*, 85(1), 34–39. https://doi.org/https://doi.org/10.17116/otorino20208501134

Jamaluddin, S. A. (2016). *Development and evaluation of the digit triplet and auditory-visual matrix sentence tests in Malay* (Publication No. 4633) [Doctoral dissertation, University of Canterbury]. UC Research Repository. http://dx.doi.org/10.26021/7303

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V. N., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54*(sup2), 3–16. https://doi.org/10.3109/14992027.2015.1020971

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Lee, L. W., & Low, H. M. (2011). Developing an online Malay Language Word Corpus for primary schools. *International Journal of Education and Development Using Information and Communication Technology (IJEDICT)*, *7*(3), 96–101. http://210.48.222.80/proxy.pac/scholarly-journals/developing-online-malay-language-word-

corpus/docview/1017894711/se-2

Les' Copaque Production. (n. d.). *Upin & Ipin*. Les' Copaque. https://lescopaque.com/v11/our-works/upin-ipin/

Marinova-Todd, S. H., Siu, C. K., & Jenstad, L. M. (2011). Speech audiometry with non-native English speakers: The use of digits and Cantonese words as stimuli. *Canadian Journal of Speech-Language Pathology and Audiology*, *35*(3), 220–227. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Providing+support+to+school+children+with+hyperacusis#0

Mendel, L. L. (2008). Current considerations in pediatric speech audiometry. *International Journal of Audiology*, *47*(9), 546–553. https://doi.org/10.1080/14992020802252261

Ozimek, E., Kutzner, D., & Libiszewski, P. (2012). Speech intelligibility tested by the Pediatric Matrix Sentence test in 3-6 year old children. *Speech Communication*, *54*(10), 1121–1131. https://doi.org/10.1016/j.specom.2012.06.001

Pangkalan Data Korpus. (2017). Dewan Bahasa dan Pustaka. http://lamanweb.dbp.gov.my/index.php/pages/view/76?mid=61

Puglisi, G. E., di Berardino, F., Montuschi, C., Sellami, F., Albera, A., Zanetti, D., Albera, R., Astolfi, A., Kollmeier, B., & Warzybok, A. (2021). Evaluation of Italian Simplified Matrix Test for Speech-Recognition Measurements in Noise. *Audiology Research*, *11*(1), 73–88. https://doi.org/10.3390/audiolres11010009

Puglisi, G. E., Warzybok, A., Hochmuth, S., Visentin, C., Astolfi, A., Prodi, N., & Kollmeier, B. (2015). An Italian matrix sentence test for the evaluation of speech intelligibility in noise. *International Journal of Audiology*, *early onli*(September), 1–7. https://doi.org/10.3109/14992027.2015.1061709

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne, *Developing Linguistic Corpora : A guide to good practice*, 1998–2001. https://users.ox.ac.uk/~martinw/dlc/

Thompson, P. (2005). Spoken language corpora. In M. Wynne, *Developing Linguistic Corpora : A guide to good practice*, 1998–2001. https://users.ox.ac.uk/~martinw/dlc/

Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication*, *35*(1–2), 103–113. https://doi.org/10.1016/S0167-6393(00)00098-4

Wijngaarden, S. J., Steeneken, H. J. M., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America*,

111(4), 1906–1916. https://doi.org/10.1121/1.1456928

Wagener, K. C., Brand, T., Kollmeier, B., Kuhnel, V., Kollmeier, B., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests Development and evaluation of a German sentence test Part 11 : Optimization of the Oldenbu rg sentence test. *Z Audiol*, *38*(2), 44–56.

Wagener, K. C., & Kollmeier, B. (2005). Evaluation des Oldenburger Satztests mit Kindern und Oldenburger: Evaluation of the Oldenburg sentence test with children and the Oldenburg children's sentence test. *Z Audiol*, *44*(3), 134–143.

Wagener, K., Josvassen, J. L., & Ardenkjær, R. (2003). *Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido*. https://doi.org/10.3109/14992020309056080

Warzybok, A., Brand, T., Wagener, K. C., & Kollmeier, B. (2015). How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *International Journal of Audiology*, *54*(early online), 88–99. https://doi.org/10.3109/14992027.2015.1063715

Warzybok, A., Zokoll, M. A., Wardenga, N., Ozimek, E., Boboshko, M., & Kollmeier, B. (2015). Development of the Russian matrix sentence test. *International Journal of Audiology*, *2027*(October), 1–9. https://doi.org/10.3109/14992027.2015.1020969

Willberg, T., Kärtevä, K., Zokoll, M., Buschermöhle, M., Sivonen, V., Aarnisalo, A., Löppönen, H., Kollmeier, B., & Dietz, A. (2020). The Finnish simplified matrix sentence test for the assessment of speech intelligibility in the elderly. *International Journal of Audiology*, *59*(10), 763–771. https://doi.org/10.1080/14992027.2020.1741704