JCSHD
COGNITIVE SCIENCES AND HUMAN DEVELOPMENT

Using Entropy to Measure Text Readability in Bahasa Malaysia for Year One Students

Mohamad Hardyman Barawi*[1], Siti Nabilah Mohamed Osman[1], Noor Fazilla Abd Yusof[2], Ebuka Ibeke[3] & Muhibuddin Fadhli[4]

[1]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

[2]Faculty of Information Communication Technology, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

[3]School of Creative and Cultural Business, Robert Gordon University, AB10 7AQ, United Kingdom

[4]Education Technology Department, Faculty of Education, Universitas Negeri Malang, 65145 East Java, Indonesia

ABSTRACT

Text readability is essential for effective learning and communication, especially for beginner readers. However, there are no known measures to calculate the readability of Bahasa Malaysia, the national language of Malaysia. This research proposes a new method based on entropy, a measure of information and uncertainty, to assess the readability of Bahasa Malaysia texts for Year One students. An experiment was conducted with six Year One students to determine the relationship between entropy and readability. The results indicated a positive correlation, suggesting that higher entropy values corresponded with lower readability for this age group. This study also revealed the need for beginner readers to focus on the text difficulty level to enhance learning.

Keywords: readability, reading, text analysis, text difficulty

ARTICLE INFO
Email address: bmhardyman@unimas.my (Mohamad Hardyman Barawi)
*Corresponding author
https://doi.org/10.33736/jcshd.6817.2024
e-ISSN: 2550-1623

Manuscript received: 20 March 2024; Accepted: 27 March 2024; Date of publication: 31 March 2024

1 INTRODUCTION

One of the ways humans gain information about the world is through reading. Reading has been described as a combination of skills of decoding information from written materials into a mental representation of letters and words (Just & Carpenter, 1980). The relations between components of texts have led to the revision of text readability measures. The primary purpose is to make texts more accessible (McNamara & Kintsch, 1996). Individuals with low literacy skills struggle to comprehend written material (Snowling, 2013). At their initial stages of achievements, they focus more on processing the meaning rather than the form of the written texts (Farrokhi et al., 2008).

A text with more frequent words and shorter sentences is more appropriate for beginner readers (Crossley et al., 2007a). In Malaysia, most teachers rely on textbooks to convey information and as instructional guides in teaching. Thus, textbooks play an essential role in the education of students (Ball & Cohen, 1996; Kulm et al., 1999; Rockinson-Szapkiw et al., 2013). The definition of a textbook, as put forth by Brammer (1967), describes a textbook as a book that contains established principles in a specific subject and is primarily designed for use in classroom instruction or student-book-teacher scenarios. Given the widespread use of textbooks in the classroom, it is essential to carefully select textbooks to ensure they are effective for all students (Bruhn & Hasselbring, 2013). A good textbook must meet benchmarks and standards (Kulm et al., 1999). This helps ensure that the textbook's material is relevant, appropriate, and high-quality for students. Educators can create a positive learning environment by selecting textbooks that meet these criteria and help students achieve their full potential.

According to Chall's Stages of Reading Development (Chall, 1983), children aged seven to eight years are at the second stage, the confirmation and fluency stage. At this stage, children can only read simple and familiar texts. To help children develop strong reading skills by the end of this stage, it is essential to focus on building their basic decoding abilities, expanding their sight vocabulary, and enhancing their understanding of context when reading familiar stories. A solid foundation in these elements is essential for compelling reading and will empower children to become confident and skilled readers. By focusing on these critical areas, educators and parents can support children in developing the skills they need to become successful readers. At the end of this stage, children should be able to read about 3000 words and understand about 9000 words when listening (Chall, 1983).

Despite the problems and criticisms levelled at conventional readability formulas, they remain popular and are still used in research. However, significant technological advances in the last two decades have enabled the streamlining and automation of traditional readability formulas and developed more modern methods for measuring text difficulty (Crossley et al., 2019). In addition, researchers, administrators, and policymakers in the education field may require guidance on which methods are helpful in research studies and classrooms.

Since the rapid introduction of newly designed approaches for analysing text difficulty and matching readers and texts, educational researchers may become overwhelmed by the numerous options or be tempted to stick with the conventional methods used for decades. In contrast,

different methods may be appropriate to select suitable texts to accommodate various populations of readers with different reading levels and understanding of diverse texts.

Researchers have significantly advanced text readability across various disciplines over the past two decades. As a result, researchers must consider a broad range of past research when determining their research direction. By considering the advancements and findings from previous studies, researchers can build upon and contribute to the ongoing development of knowledge in text readability. Doing this will ensure that future research is insightful and impactful. As a result, there is a need to evaluate methods developed in the more conventional style and emerging methods of assessing text difficulty level and matching appropriate readers to texts. This study was motivated by the need for awareness of the rapidly expanding field of text readability analysis.

The main contributions of this paper can be summarised as follows: (i) We review and evaluate the effectiveness of text readability methods that have emerged in the last couple of decades. We achieve this by describing the foundation blending elements within "types" of methods, identifying their strengths and weaknesses, and pointing out their fundamental differences. (ii) We propose a new text readability method for analysing Bahasa Malaysia texts based on entropy. Because our method is language-independent, we believe it may also be applied to other languages. (iii) We identify the relationship between the entropy of text and the difficulty of Bahasa Malaysia texts. To achieve this, we calculated entropy values from various chapters in a textbook. We verified our results with respondents to understand the entropy values with the difficulty of Bahasa Malaysia texts.

The following sections discuss readability approaches developed in the last few decades and summarise recommendations for using these methods in education practice and research. Finally, we describe and evaluate currently available tools and approaches in text readability research. These tools and approaches are divided into three types: (1) Conventional Readability, (2) Cognitive Inspired Readability, and (3) Statistical based Readability. All approaches discussed in this paper are quantitative to simplify the evaluation and comparison of methods. The discussion for each text readability approach concludes with recommendations for research directions. Following the discussion of approaches, a section provides suggestions for the extensive use of existing text readability approaches.

2 RELATED WORK

Years of investigation suggest that reading is intricate and comprises various components, categorised into lower-level and higher-level processes. Lower-level processes encompass word recognition, syntactic parsing, and semantic proposition encoding (Tiffin-Richards & Schroeder, 2015; Duke & Cartwright, 2021). On the other hand, higher-level processes involved in comprehension consist of updating, inferencing, inhibition, and strategic processing, including metacognition (Wylie et al., 2018; Barzillai et al., 2018).

The most efficient method for enhancing reading comprehension is exposing learners to materials slightly above their reading proficiency and engaging with overly simplistic texts, which results in

monotonous, unproductive efforts. Conversely, if the text proves too challenging, language learners may experience a decline in confidence and interest in language acquisition (Jian et al., 2022).

More than 50 readability formulas have been proposed since 1920 with the expectation of measuring text difficulty more precisely and effectively (Crossley et al., 2007b). Advancements in artificial intelligence have led to the exploration of machine learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), for text readability assessment (Azpiazu & Pera, 2019; Martinc et al., 2021; Zulqarnain & Saqlain, 2023). Most readability formulas are based on lexical or semantic features and sentence or syntactic difficulties (Crossley et al., 2023; Arshad et al., 2023). The most famous readability formula is the Flesch Reading-Ease formula. The formula solely depends on the number of words and sentences to measure a text's readability (Rafatbakhsh & Ahmadi, 2023).

Advances in computational linguistics and discourse processing have enabled automating languages and text processing mechanisms (Khurana et al., 2023). Advanced readability formulas explore deeper attributes of language as the analysis of textual coherence is automated, allowing more precise and detailed analyses to occur (Crossley et al., 2023). The Lexile Framework for Reading is a scientific approach to reading and text measurement. It matches the reader's ability and text difficulty (Orellana et al., 2024). Lexile scale is a developmental scale ranging from 200L for beginner readers to 1700L for advanced readers (Graden, 2023). This makes it easier for educators to provide appropriate reading material for learners according to their capabilities, as every student in the same grade may have a different Lexile scale score (Orellana et al., 2024). In information theory, information and uncertainty are closely related. This means that the more predictable a text is, the less information it contains and the easier it is to read. Entropy remains a valuable concept in readability assessment, with some recent studies exploring its effectiveness when combined with other features or machine-learning approaches (Hovious & O'Connor, 2023; Hadfi & Ito, 2024).

Text Readability

Text readability formulas estimate difficulty based on factors like sentence and word length (e.g., Davison and Kantor (1982)). Despite criticism (e.g., Bamford (1984); Brown (1997); Greenfield (2004)), these formulas remain popular (e.g., Agrawal et al. (2011); Zamanian & Heydari (2012)). Technological advancements allow researchers to automate these approaches and explore new variables (Crossley et al., 2023).

While conventional approaches focusing on sentence length, familiar words, and word length are still being developed (Crossley et al., 2019), they consider shorter, shorter, and frequent words more straightforward to read. This assumption is based on correlations between readability scores and reader comprehension (Makebo et al., 2022). However, this might be a loose estimate, as nonsensical text with frequent short words and short sentences could still be considered readable by the formula.

Recent methods are gaining traction due to their use of conventional features like word frequency in new ways (Martinc et al., 2021; Wright & Stenner, 2022). Examples include the New Dale-Chall Readability Formula (Crossley et al., 2022), Lexile framework (Stenner, 2022), and ATOS formula (Makebo et al., 2022). A lesser-known method, Read-X (Orellana et al., 2024), highlights potential future directions for these approaches.

Conventional Readability

In ancient Greece, philosophers such as Socrates, Aristotle and Plato introduced two main philosophical views: rationalism and empiricism (Austin et al., 2001). In the Roman era, learning focused on skills that could contribute to society, such as building a house or road. During the Roman Catholic era, learning started in a formal institution such as the church or university. In this era, the education system is introduced (Austin et al., 2001).

More than 50 readability formulas have been proposed since 1920 with the expectation of measuring text difficulty more precisely and effectively (Crossley et al., 2007a). Most readability formulas are based on lexical or semantic features and sentence or syntactic difficulties (Chall & Dale, 1995). The most famous readability formula is the Flesch Reading-Ease formula. The formula solely depends on the number of words and sentences to measure a text's readability (Marnell, 2008).

Cognitive-Inspired Readability

In the modern approach of learning theories, the researchers are most concerned about the most effective strategies for learning. Many learning theories have been proposed to clarify the most effective strategies for learning. The first theory of the modern world is behaviourism. The researchers focus on the behaviour to explain learning as behavioural responses to physical stimuli (Fosnot & Perry, 1996). Later, researchers proposed a theory based on a cognitive perspective. This cognitive perspective focuses on understanding concepts and theories such as reasoning, problem-solving and planning (Ojose, 2008). The leading theory is Piaget's Theory of Cognitive Development (Papalia et al., 2007).

Advances in computational linguistics and discourse processing have enabled the automation of languages and text-processing mechanisms (Graesser et al., 2004). Advanced readability formulas explore more profound attributes of language as the analysis of textual coherence is automated, allowing more precise and detailed analyses to occur (Crossley et al., 2007a).

Statistical Readability

Shannon (1951) defined information as a measure of one's freedom of choice when selecting a message. In information theory, information and uncertainty are intricately linked. Information refers to the amount of uncertainty inherent in an event. The more uncertainty a message resolves, the stronger the correlation between the input and output of a communication channel. This means

that a more elaborate message that reduces uncertainty transmits more information. Conversely, uncertainty is related to the concept of predictability. When a message is predictable, it is said to be specific. Thus, it contains less information MacKay (2003).

Support Vector Machine (SVM)

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm that can be used to classify text into various categories, such as readability levels. The goal of using SVM in text readability is to train a model to identify the characteristics of text that make it easy or difficult to read and then use this model to predict the readability level of new text (Maqsood et al., 2022).

To use SVM for text readability, a dataset of texts with their corresponding readability level must be collected. This dataset is then used to train the SVM model. After training, the model can classify new text into the appropriate readability level based on the characteristics it learned from the training dataset. SVMs are particularly useful in text readability because they can handle high-dimensional data, such as text, and can effectively identify complex patterns and relationships in the data. Additionally, SVM models can handle linear and non-linear relationships, making them well-suited for text readability classification tasks.

SVMs are used for text readability to classify texts based on their level of readability, thus helping to identify whether a text is appropriate for a specific audience. They can be used in many applications, such as educational content, news articles, scientific papers, etc.

Flesch–Kincaid

The Flesch–Kincaid readability test is a commonly used method for measuring the readability of text. It uses a formula to calculate the readability level of a text-based on two factors: the average number of words per sentence and the average number of syllables per word (Eleyan et al., 2020). The resulting score is then converted into a grade level, with lower scores indicating texts that are easier to read.

The Flesch–Kincaid readability test can be applied in various contexts to evaluate the readability of texts. Some of the main applications include:

Educational materials: Teachers and educators can use the Flesch–Kincaid test to evaluate the readability of textbooks, workbooks, and other educational materials to ensure they are appropriate for the intended audience.

News articles: News organisations can use the Flesch–Kincaid test to evaluate the readability of news articles and ensure that they are accessible to a broad audience.

Legal documents: Legal professionals can use the Flesch–Kincaid test to evaluate the readability of legal documents, such as contracts and agreements, to ensure they are easy to understand for non-experts.

Medical documents: Medical professionals can use the Flesch–Kincaid test to evaluate the readability of medical documents, such as patient education materials, to ensure that they are easy to understand for patients.

Technical documents: Technical writers can use the Flesch–Kincaid test to evaluate the readability of technical documents, such as user manuals and product documentation, to ensure they are easy to understand for non-experts.

The Flesch–Kincaid readability test is widely used because it is simple and easy to use. It can be integrated into various software and does not require a large dataset of labelled text.

Lexile

The Lexile Framework for Reading is a widely used method for measuring the readability of texts. It uses a formula to calculate the readability level of a text-based on two factors: the text's complexity and the reader's abilities (Baker, 2020). The resulting Lexile measure score is presented as a number, e.g. 880L, and can match readers with texts appropriate for their reading level.

The Lexile Framework can be applied in a variety of contexts to evaluate the readability of text; some of the main applications include:

Education: The Lexile Framework can be used in educational settings to match students with texts appropriate for their reading level. This can help ensure students are challenged but not overwhelmed by the texts they read.

Libraries and bookstores: The Lexile Framework can be used by librarians and bookstore staff to match readers with books appropriate for their reading level. This can help ensure that readers are engaged and motivated to read more.

Online content: The Lexile Framework can be used to evaluate the readability of online content, such as news articles and blog posts, to ensure that it is appropriate for the intended audience.

Test and assessment: The Lexile Framework can be used to evaluate the readability of texts used in standardised tests, such as the SAT and ACT, to ensure that they are appropriate for the intended audience.

Business: The Lexile Framework can be used by businesses to evaluate the readability of marketing materials, such as brochures and websites, to ensure that they are appropriate for the intended audience.

One of the main advantages of the Lexile Framework is that it is widely used. An extensive database of books and articles is already labelled with their Lexile measure. The Lexile measure can also be used to track a reader's progress over time, making it a valuable tool for monitoring the development of reading skills.

Coh-Metrix Psycholinguistics

Coh-Metrix Psycholinguistics is a tool that uses computational methods to evaluate the readability of text by measuring various psycholinguistic and text characteristics (Ryu & Jeon, 2020). It is based on the idea that readability is determined by the text's complexity and the reader's ability to understand it. The Coh-Metrix Psycholinguistics tool can be applied in a variety of contexts to evaluate the readability of text; some of the main applications include:

Education: The tool can be used in educational settings to evaluate the readability of textbooks, workbooks, and other educational materials to ensure they are appropriate for the intended audience.

News articles: News organisations can use the tool to evaluate the readability of news articles to ensure that they are accessible to a broad audience.

Legal documents: Legal professionals can use the tool to evaluate the readability of legal documents, such as contracts and agreements, to ensure they are easy to understand for non-experts.

Medical documents: Medical professionals can use the tool to evaluate the readability of medical documents, such as patient education materials, to ensure that they are easy to understand for patients.

Technical documents: Technical writers can use the tool to evaluate the readability of technical documents, such as user manuals and product documentation, to ensure they are easy to understand for non-experts.

The Coh-Metrix Psycholinguistics tool can also be used to compare the readability of different texts, such as the readability of different versions of the same document or the readability of diverse types of texts. It can also be used to identify specific areas of a text that may be difficult to understand and to make recommendations for improving the readability of a text. The tool measures various characteristics of text, such as lexical, syntactic, semantic, discourse, and pragmatic features, which makes it a powerful tool to evaluate the readability of text. However, it requires many computational resources and an elevated level of expertise.

Entropy

Entropy is a fundamental term in information theory Ben-Naim (2019). The term entropy frequently refers to Shannon entropy, which was introduced by Shannon (1951). To quantify the idea of "information content," Shannon (1951) adapted the concept of "entropy" from physics (Fossum, 2013). In physics, entropy measures the amount of unpredictability in a physical system. For example, boiling water in molecules moving randomly and colliding randomly is known as high entropy. At the same time, ice contrasts with boiling water as its state of entropy is low. The molecules are fixed in an orderly pattern and move little. Shannon (1951) uses this concept of entropy in language by remarking that a sentence complete of random words is like boiling water: chaotic, unpredictable, and disorganised. This will cause difficulty in predicting the next word in a sentence. Below is an example adapted from (Fossum, 2013): *friend's The reflects feelings setting my sun old.* Here is the same set of words, presented in an ordered sentence that follows the pattern of a typical English sentence: *The setting sun reflects my old friend's feelings.* Drawing on these intuitions, Shannon (1951) developed a measure of entropy for languages that assign high entropy to the disordered, random first sentence and low entropy to the ordered, patterned second sentence. In the above example, Shannon entropy illustrated that a random, unorganised sentence has high entropy, which contains more information. On the contrary, a systematic, organised sentence has low entropy and less information. Some might disagree by saying that only the second sentence conveys information, while the first sentence is nonsense and does not convey any information. This is where Shannon's entropy mathematical definition differs from the information content. The definition of Shannon entropy is that if a sentence is predictable, it does not convey much current information; it just explains something already known and expected. For example, if a person lives in the Sahara, the weather forecast is likely to predict "Sunny and 40 Celsius" daily. Thus, viewing weather forecasts in the Sahara is very predictable, so the added information gained is less as the event is predictable. Shannon (1951) reasoned that an unorganised sentence has exceedingly high information content because it is difficult to predict; on the contrary, an organised sentence has low information content because it is easy to predict the following sequence of letters. When measuring entropy, the main concern is text predictability (Fossum, 2013). Shannon's (1951) approach is to ignore semantic aspects of the message and focuses on the physical and statistical constraints limiting the transfer of the message, despite context meaning (Lesne, 2014). According to Cherry (1953), it was proved that Shannon's (1951) entropy could be identified as a measure of the average number of choices required to detect each message symbol from the alphabet (Titchener, 2000).

Shannon (1951) shows that entropy is equivalent to potential information gained once the learner learns the outcome of an event. The process of gaining information is equivalent to the process of losing uncertainty. The formula of Shannon's entropy is as shown below:

$$H = -p\,(x)\,\log_2 p\,(x)$$

The entropy (H) can be defined as a discrete random variable of X (Cover, 1999). It quantifies the disorganisation of the probability distribution p (Lesne, 2014). The smaller the redundancy of text, the more challenging it is to predict it, thus the more complex it is (Kontoyiannis, 1997). Based on

the entropy formula in 1, H is the average number of binary digits required per alphabet (symbol) of the original language. This is applied when the concerned language is translated into binary digits (0 or 1).

3 METHODS AND EXPERIMENTAL SETUP

Research Design

This research is qualitative and uses a mixture of experimental designs. The rationale for using experimental designs is to test cause-and-effect relationships. It represents the most valid approach to solving educational problems, both practical and theoretical (Gay et al., 2011). This research design is used to identify the relationship between the entropy value of texts and their difficulty. Moreover, there are no known studies done using Bahasa Malaysia; thus, not much is known about the phenomenon.

Data Collection Procedure

The procedure for data collection started by approaching SK Laksamana's principal to conduct the research. After presenting the research purpose and the planned experiment in detail to the school principal and receiving their positive feedback, a mutually agreeable date was set for experimenting. On the date of the experiment, Year One's teacher randomly selected students as the selection criteria for the participants. The school has provided a room for experimenting. Then, the experiment was conducted. The experiment was done one-on-one with every participant.

Instrument

The instrument used in this research is test questions. The instrument needs participants to read and answer the questions. There are two (2) sets of test questions with four (4) questions each. There are two sets of questions, one for the lowest entropy value text and the other for the highest entropy value text. The first two questions (Question 1 and Question 2) need participants to guess the alphabet to complete the missing letter in a sentence. The following two questions (Question 3 and Question 4) are comprehension questions to test understanding based on the text.

Text 1

HUDA DI TAMAN, (HUDA IN THE PARK,) HUDA MAIN BUAIAN. (HUDA PLAYS ON THE SWING.) HUDA MAIN BUAIAN DI TAMAN. (HUDA PLAYS SWING IN THE PARK) DIA MAIN BUAIAN DI TAMAN DENGAN DEVI. (SHE PLAYS SWING IN THE PARK WITH DEVI.) MEREKA MAIN BUAIAN DI TAMAN PADA WAKTU PETANG. (THEY PLAY SWING IN THE PARK IN THE EVENING.)

Question 1

1. HUDA D_ T_M_N

2. HUDA _AI_ B_AIA_
3. Huda bermain buaian dengan siapa? (With whom does Huda play swing?)
4. Mereka bermain pada waktu bila? (When do they play?)

Text 2

SINGGAH DI GERAI BERSAMA AMAR, (STOP BY THE STALL WITH AMAR,) JAGALAH SUNGAI JANGAN TERCEMAR. (KEEP THE RIVER FROM BEING POLLUTED.) SUBUR BERCAMBAH DAUN SEMALU, (THRIVE LIKE THE MIMOSA LEAVES,) KUTIP SAMPAH JANGANLAH MALU. (DON'T BE SHY TO PICK UP TRASH.) TANAM SIRIH WAKTU PETANG, (PLANT BETEL IN THE EVENING,) SUNGAI BERSIH CANTIK DIPANDANG. (CLEAN RIVER IS PLEASANT TO BEHOLD.) TUMBUH DI SAWAH POHON ARA, (GROWING IN THE RICE FIELDS, THE FIG TREE,) SUNGAI INDAH HATI GEMBIRA. (BEAUTIFUL RIVER BRINGS JOY.) DI CELAH BATU IKAN KELISA, (AMONG THE ROCKS, AROWANA FISH,) BANTU-MEMBANTU HIDUP SENTOSA. (HELPING EACH OTHER LIVE IN PEACE.)

Question 2

1. SUBUR BERCAMBAH DAUN SEMALU, K_T_P S_MPA_ _AN_ANLAH _A_ _.
2. DI CELAH BATU IKAN KELISA, BA _ _U - M_ _B_N_U HI_UP S_ _TO_A.
3. Pohon ara tumbuh di mana? (Where does the fig tree grow?)
4. Berikan dua (2) pengajaran yang terdapat dalam pantun di atas. (State two (2) lessons inthe poem above.)

The design of the test questions is adapted based on Shannon's first paper in 1951 titled Prediction and Entropy of Printed English. Besides using Shannon's method as a primary reference, another similar method was used by Moradi et al. (1998). Thus, both methods are adapted and used in designing the test question for the experiment. Both methods were used to experiment with adults. In this research, the participants are six children aged seven. Therefore, the method is modified to suit seven years old children. The difference between the method used by this research and previous research is that guessing the alphabet is not continuous from beginning to end. In this research, guessing the alphabet is done as a fill-in-the-blank method. This is because the continuous concentration level of seven-year-old children when completing a sentence or text is not as high as that of adults.

Text Pre-processing

Several steps were needed to design the test question to achieve the texts with high and low entropy values. Due to research limitations, the pre-processing was done manually and with the help of an off-the-shelf tool Wang and Hu (2021).

Text Selection

Appropriate texts need to be identified before any extraction is carried out. The text is chosen based on the characteristics listed below:

- The text should be in Bahasa Malaysia (Malay language). This research aims to identify the entropy of Bahasa Malaysian text, so only Bahasa Malaysia text has been chosen.

- The texts chosen must only be from Year One texts. This is to identify the entropy of text read by a beginner reader. Most children start formal learning in Year One; thus, knowing the entropy of Year One texts is essential.

- The text must be selected from any textbook published or released by the Ministry of Education Malaysia for students' use. This is because the texts published by the Ministry will be the same for the whole country.

- The text must be in a complete sentence. This is because this research will find the relationship between entropy and the difficulty of a text. The text chosen must be understandable by students to measure the difficulty of the text.

Text Extraction

The textbook used in this research was borrowed from the Year One students. This is because textbooks published or issued by the Ministry of Education are unavailable in any bookstore. The first step is to scan the text of Bahasa Malaysia. This is because a written text in a book cannot be tested using a computer directly. Besides that, the Year One Bahasa Malaysia textbook comprises texts and images. In this research, only texts are needed. The texts that are considered were manually typed in Microsoft Word as .doc.

Evaluation Task

The 60 texts from the entire Year One Bahasa Malaysia textbook were chosen as the sample for this research and then analysed using Microsoft Excel. Microsoft Excel is a spreadsheet application developed by Microsoft for Microsoft Windows and Mac OS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. This software was used to calculate the average entropy of the textbook.

The students were given two texts each: one with low entropy and the other with high entropy. The students must read the passage loudly to ensure they know how to read each word. After that, they must answer some questions based on the text given. This will show whether the texts are difficult for their level of literacy.

4 FINDINGS

The Standard One Bahasa Malaysia textbook contains six (6) units. The accepted texts from each unit are calculated based on their respective entropy. The average entropy value of each chapter and the average entropy value of the textbook are calculated as shown in Figure 1. Based on the result, the average entropy value is not consistent. The result also showed that the average entropy value differs significantly from each unit. The average of the whole textbook (orange bar) is 3.918 bits. The lowest average entropy value is 3.902 bits in Unit 2. The unit with the highest average entropy value is Unit 4, with 3.928 bits. As shown in Figure 1, four units are above average entropy values of the textbook, and two units are below average (3.918 bits). Out of the four units above the average entropy value of the textbook, two (2) units differ significantly. Unit 4 (3.928 bits) and Unit 6 (3.927 bits) are the two units.

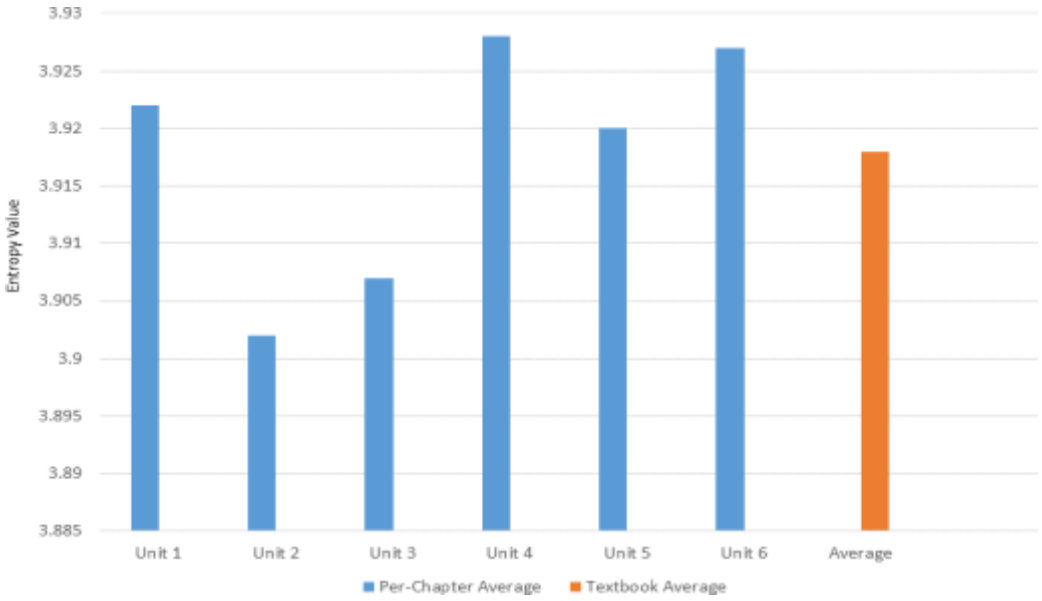

Figure 1. Average entropy value for each unit and the whole textbook.

Each unit has different numbers of text. To measure the average entropy value of each unit, the total number of entropy values is divided by the total number of texts in the unit to obtain the average entropy value. For example, in Unit 1, there are seven (7) texts. Thus, the total number of all texts in Unit 1 is divided by seven to obtain the average entropy value of Unit 1.

Table 1. Entropy values of texts for each unit in the textbook.

Text \ Unit(bits)	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6
Text 1	4.074	3.946	3.925	3.907	3.944	3.828
Text 2	3.931	3.992	4.009	3.888	3.940	4.084
Text 3	3.686	4.005	3.999	3.908	3.966	3.950
Text 4	3.783	4.061	3.850	3.900	3.949	3.840
Text 5	3.993	4.070	3.819	3.897	3.786	3.941

Text 6	4.043	3.552	3.912	3.941	3.799	3.939
Text 7	3.944	3.899	3.901	3.954	3.908	3.972
Text 8	-	3.694	3.899	3.855	3.993	3.935
Text 9	-	-	3.965	4.034	3.993	3.894
Text 10	-	-	3.790	3.959	-	3.889
Text 11	-	-	-	3.966	-	3.960
Text 12	-	-	-	3.973	-	3.892
Text 13	-	-	-	3.949	-	-
Text 14	-	-	-	3.860	-	-
Total	27.454	31.219	39.070	54.991	35.278	47.124
Average	3.922	3.902	3.907	3.928	3.920	3.927

Table 2. Result for Text 1: lowest entropy value text with 3.552 bits.

Respondents (R) / **Questions**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**
Question 1	2,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0
Question 2	1,3,0,0	4,0,0,0	0,0,0,0	0,0,0,0	0,0,0,0	0,0,0,0
Question 3	Easily answered	Easily answered	Easily answered	Easily answered	Easily answered	Easily answered
Question 4	Easily answered	Easily answered	Easily answered	Easily answered	Easily answered	Easily answered

The text with the lowest entropy is text six from Unit 2, with an entropy value of 3.552 bits. The text with the highest entropy is text two from Unit 6, with an entropy value of 4.084 bits. The text with the lowest entropy**, Text 1, with 3.552 bits,** has redundancy of the same word or alphabet occurring in the text. In text six from Unit 2, the alphabet 'A' has a frequency of 30 occurrences, which occurs most of the time. The letters 'R,' 'V,' and 'W' occur once in each text. The exact words that occur frequently, such as 'TAMAN,' 'MAIN' and 'BUAIAN,' are why the text has a lower entropy value than others.

The text with the highest entropy**, Text 2, with 4.084 bits,** has little redundancy of the same word or alphabet occurring in the text. Compared to the text with the lowest entropy value, the text with the highest entropy value has less redundancy in the occurrence of words. The words keep repeating in texts with the lowest entropy value, while in texts with the highest entropy value, the words rarely repeat themselves. Besides that, the text is also longer than the text with the lowest entropy value.

Text with Lowest Entropy Value Analysis

Table 2 shows that students can quickly answer the questions based on the text with the lowest entropy value. The test questions are available in the appendix (Appendix AB). The test questions are divided into two (2) sections: guessing the alphabet and comprehension questions. In Questions 1 and 2, respondents are given a complete sentence with a few missing letters. They are required to guess the alphabet until they complete the sentence. The number of guesses of the alphabet is recorded. Question 1 has three (3) missing alphabets, while question 2 has four (4) missing

alphabets. In Table 2, the row Question 1 and column Respondent 1 indicate that Respondent 1 made two guesses on the first blank before getting the correct alphabet. In the second and third blanks of Question 1, Respondent 1 answered them straightaway without the need to guess.

The result in Table 2 shows that five (5) respondents out of six (6) respondents (83.33%) answered Question 1 with ease. They do not need to guess and can know the answer straight away. The result of Question 2 shows that 66.67% of respondents (four out of six respondents) can answer the question without any guesses of the alphabet. Respondent 1 needed to guess once on the first blank and thrice on the second blank, and no guesses were made on the third and fourth blanks, as the respondent answered them immediately. Respondent 2 made four guesses on the first blank, and no guesses were made on the consequent blanks as the respondent answered them immediately.

Question 3 and Question 4 are comprehension questions based on the text given. Table 2 shows that all respondents (100%) answered the given text quickly. During the experiment, all respondents were told to read Question 3 and Question 4 aloud. After reading the questions, they can answer them immediately.

Text with Highest Entropy Value Analysis

Table 2 shows the result of the text with the highest entropy value answered by respondents. The text is attached in Appendix C. The method of testing Text 2 is the same as Text 1. The respondents used in the second text are the same as in the previous text. The division of the section in Text 2 is identical to Text 1, where Question 1 and Question 2 are guessing the alphabet. Question 3 and Question 4 are comprehension questions.

In Text 2, the text is longer and has more alphabets; thus, the blank space is more. In Question 1, there are nine (9) blank spaces, while in Question 2, there are ten blank spaces. In Question 1, the respondents each have trials of guesses made at least once in the whole sentence. All respondents (100%) made at least one guess of the alphabet to complete the sentence. The one blank space in Question 1 with the highest guesses is blank space number 5. In Table 3, respondents 1, 2, 3, 4, 5, and 6 made 14, 7, 10, 15, 3, and 5 guesses, respectively.

Table 3. Result for Text 2: Highest entropy value text with 4.082 bits.

Respondents (R) / **Questions**	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**
Question 1	3|7|1|5|14|1|3|2|4	10|0|0|0|7|0|4|0|0	0|0|0|0|10|2|4|3|0	5|0|0|0|15|0|8|12|5	0|0|0|0|3|0|0|0|0	0|2|0|0|5|2|0|0|0|0
Question 2	6|4|13|4|0|0|1|5|1|2	3|0|0|0|0|0|8|0|5|0	17|0|0|18|2|0|0|0|0|0	12|0|0|0|0|0|3|7|10|5	0|0|0|0|0|0|0|0|0|0	0|0|0|0|0|0|0|0|0|0
Question 3	Success in third trial (find an answer in the text)	Success in third trial (read the text twice)	Success in third trial (reread the text)	Success in fifth trial (find an answer in the text)	Success in third trial (find an answer in the text)	Success in second trial (reread the text)
Question 4	Kutip sampah; bantu-membantu (read again and with help)	Kutip sampah; bantu-membantu (read again and with help , fourth trial)	Kutip sampah; Bersihkan sungai (read again and with help, second trial)	Kutip sampah; bantu-membantu (read again and with help)	Jangan buang sampah (because of the kutip sampah); Bersihkan sungai (answered easily)	Kutip sampah; bantu-membantu (answered easily)

In question 2 in Table 2, 66.67% of respondents (four out of six respondents) made guesses of the alphabet to fill in the blank to complete the sentence. 33.33% of respondents (two out of six) do not make any guesses and cannot answer the text immediately. The significant guesses were made at the third blank by Respondent 1 (13 guesses) and Respondent 3 (18 trials of guesses).

Question 3 and Question 4 are comprehension questions based on the text given. All respondents (100%) successfully answered Question 3 in at least two (2) trials. Table 2 shows that 16.67% of respondents answered the text in the second and fifth trials, respectively. 66.67% of respondents (four out of six) answered correctly on their third trial. Question 4 required respondents to give two answers based on the text given. The result of Question 4 is unanimously the same for the first part of the questions, as all the respondents (100%) answered the same answer. The second part of the answer is not unanimous, as 16.67% of respondents (one out of six respondents) succeeded in the second and fourth trials. 33.33% of respondents (two out of six) must reread the text to find the answer. Four of six respondents (66.67%) needed help answering the second part of the question. The other 33.33% of respondents (two out of six respondents) can quickly answer the second part of the questions.

5 CONCLUSION AND LIMITATIONS

This study investigated the relationship between entropy and text difficulty in Bahasa Malaysia textbooks for beginner readers (age 7). The findings show that Bahasa Malaysia text exhibits higher entropy than English (around 3.9 bits per letter vs 2.3 bits). Entropy values within the textbook varied, with some units having significantly higher entropy than others. The study also revealed a correlation between entropy and text difficulty. Texts with lower entropy had a higher frequency of familiar words, facilitating sight word reading and comprehension. Conversely, texts with higher entropy had fewer frequent words, requiring more effort for decoding and potentially hindering comprehension, especially for young readers. These findings suggest that considering text entropy can be beneficial when selecting or creating educational materials for beginner readers. Focusing on high-frequency words can improve readability and comprehension, particularly for students at the initial reading and decoding stage. The study also identified limitations, such as not accounting for sentence structure complexity. Future research could explore how factors like sentence structure interact with entropy to influence reading difficulty.

ACKNOWLEDGEMENTS

This research received no specific grant from public, commercial, or not-for-profit funding agencies.

REFERENCES

Agrawal, R., Gollapudi, S., Kannan, A., & Kenthapadi, K. (2011). Identifying enrichment candidates in textbooks. In *Proceedings of the 20th International Conference Companion on World Wide Web* (pp. 483–492).

Arshad, M., Yousaf, M. M., & Sarwar, S. M. (2023). Comprehensive readability assessment of scientific learning resources. *IEEE Access.*

Austin, K., Orcutt, S., & Rosso, J. (2001). How people learn: Introduction to learning theories. *The Learning Classroom: Theory into Practice–A Telecourse for Teacher Education and Professional Development*.

Azpiazu, I. M., & Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, *7*, 421–436. https://doi.org/10.1162/tacl_a_00278

Baker, J. R. (2020). Going beyond the readability formula: How do titles contribute to the readability of essays? *International Journal of TESOL Studies, 2*(1), 119–132. https://doi.org/10.46451/ijts.2020.06.08

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6–14.

Bamford, J. (1984). Extensive reading by means of graded readers. *Reading in a Foreign Language*, *2*(2), 218–260.

Barzillai, M., Broek, P., Schroeder, S., & Thomson, J. (2018). *Learning to read in a digital world.* John Benjamins Publishing Company.

Ben-Naim, A. (2019). Entropy and information theory: Uses and misuses. *Entropy*, *21*(12), 1170.

Brammer, M. (1967). Textbook publishing. *What happens in book publishing* (pp. 320–349).

Brown, J. D. (1997). An EFL readability index. *University of Hawai'i Working Papers in English as a Second Language, 15*(2), 85–119.

Bruhn, A. L., & Hasselbring, T. S. (2013). Increasing student access to content area textbooks. *Intervention in School and Clinic*, *49*(1), 30–38.

Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.

Cover, T. M. (1999). *Elements of information theory.* John Wiley & Sons.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007a). Toward a new readability: A mixed model approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society, p. 29.*

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007b). Toward a new readability: A mixed model approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society,* p. 29.

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3-4), 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, S., Skalicky, S., Berger, C., & Heidari, A. (2022). Assessing readability formulas in the wild. In *Conference on Smart Learning Ecosystems and Regional Development,* 91–101, Springer. https://doi.org/10.1007/978-981-19-5240-1_6

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scale corpus for assessing text readability. *Behavior Research Methods, 55*(2), 491–507. https://doi.org/10.3758/s13428-022-01802-x

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17,* 187–209. https://doi.org/10.2307/747483

Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly, p. 56,* S25–S44. https://doi.org/10.1002/rrq.411

Eleyan, D., Othman, A., & Eleyan, A. (2020). Enhancing software comments readability using flesch reading ease score. *Information, 11*(9), 430. https://doi.org/10.3390/info11090430

Farrokhi, F., Ansarin, A. A., & Mohammadnia, Z. (2008). Preemptive focus on form: Teachers' practices across proficiencies. *Linguistics Journal, 3*(2), 150–157.

Fosnot, C. T., & Perry, R. S. (1996). Constructivism: A psychological theory of learning. *Constructivism: Theory, Perspectives, and Practice, 2*(1), 8–33.

Fossum, V. (n.d.). Entropy, compression, and information content. *Unpublished article. https://a2957a73-a-62cb3a1a-ssites.googlegroups.com/site/vfossum/entropy_explanation.pdf.*

Gay, L. R., Mills, G. E., & Airasian, P. W. (2011). *Educational Research: Competencies for*

Analysis and Applications. Pearson Higher Education.

Graden, I. C. (2023). The effects of research-based strategies on reading achievement among English language learners. Doctoral Dissertations and Projects. 4640. https://digitalcommons.liberty.edu/doctoral/4640

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-matrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal, 26*(1), 5–24.

Hadfi, R., & Ito, T. (2024). Structural complexity predicts consensus readability in online discussions. *Social Network Analysis and Mining, 14*(1), 51.

Hovious, A. S., & O'Connor, B. C. (2023). The reader as subjective entropy: A novel analysis of multimodal readability. *Journal of Documentation, 79*(2), 415–430.

Jian, L., Xiang, H., & Le, G. (2022). English text readability measurement based on convolutional neural network: A hybrid network model. *Computational Intelligence and Neuroscience, 2022*,1–9. https://doi.org/10.1155/2022/6984586)

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications, 82*(3), 3713–3744.

Kontoyiannis, I. (1997). *The complexity and entropy of literary styles*. Department of Statistics, Stanford University.

Kulm, G., Roseman, J., & Treistman, M. (1999). A benchmarks-based approach to textbook evaluation. *Science Books & Films*, *35*(4), 147–153.

Lesne, A. (2014). Shannon entropy: A rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science, 24*(3), e240311. https://doi.org/10.1017/S0960129512000783

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge University Press.

Makebo, T. H., Bachore, M. M., & Ayele, Z. A. (2022). Investigating the correlation between students' reading fluency and comprehension. *Journal of Language Teaching and Research, 13*(2), 229–242. https://doi.org/10.17507/jltr.1302.02

Maqsood, S., Shahid, A., Afzal, M. T., Roman, M., Khan, Z., Nawaz, Z., & Aziz, M. H. (2022). Assessing English language sentences readability using machine learning models. *PeerJ Computer Science, p. 8*, e818. https://doi.org/10.7717/peerj-cs.818

Marnell, G. (2008). Measuring readability, part 1: The spirit is willing, but the Flesch is weak. *Southern Communicator*, *14*(1), 12–16.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, *47*(1), 141–179. https://doi.org/ 10.1162/coli_a_00398

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288.

Moradi, H., Grzymala-Busse, J. W., & Roberts, J. A. (1998). Entropy of English text: Experiments with humans and a machine learning system based on rough sets. *Information Sciences*, *104*(1-2), 31–47.

Ojose, B. (2008). Applying Piaget's theory of cognitive development to mathematics instruction. *The Mathematics Educator*, *18*(1), 26.30

Orellana, P., Silva, M., and Iglesias, V. (2024). Students' reading comprehension level and reading demands in teacher education programs: the elephant in the room? *Frontiers in Psychology, 15*, 1324055. https://doi.org/10.3389/fpsyg.2024.1324055

Papalia, D., Olds, S., & Feldman, R. (2008). *Human development*. McGraw-Hill.

Rafatbakhsh, E., & Ahmadi, A. (2023). Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education, 8*(1), 41. https://doi.org/10.1186/s40862-023-00214-4

Rockinson-Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education*, *63*, 259–266. https://doi.org/10.1016/j.compedu.2012.11.022

Ryu, J., & Jeon, M. (2020). An analysis of text difficulty across grades in Korean middle school English textbooks using coh-metrix. *Journal of Asia TEFL*, *17*(3), 921. https://doi.org/10.18823/asiatefl.2020.17.3.11.921

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50–64.

Snowling, M. J. (2013). Early identification and interventions for dyslexia: A contemporary view. *Journal of Research in Special Educational Needs*, *13*(1), 7–14. https://doi.org/10.1111/j.1471-3802.2012.01262.x

Stenner, A. J. (2022). Measuring reading comprehension with the Lexile framework. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement*: Selected papers by A. Jackson Stenner (pp. 63–88). Springer.

Tiffin-Richards, S. P., & Schroeder, S. (2015). The component processes of reading comprehension in adolescents. *Learning and Individual Differences*, *42*, 1–9. https://doi.org/10.1016/j.lindif.2015.07.016

Titchener, M. R. (2000). A measure of information. In *Proceedings DCC 2000. Data Compression Conference* (pp. 353–362). IEEE.

Wang, M., & Hu, F. (2021). Applying the nltk library for Python natural language processing in corpus research. *Theory and Practice in Language Studies*, *11*(9), 1041–1049. https://doi.org/10.17507/tpls.1109.09

Wright, B. D., & Stenner, A. J. (2022). Readability and reading ability. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement*: *Selected Papers by A. Jackson Stenner* (pp. 89–107). Springer.

Wylie, J., Thomson, J., Leppänen, P., Ackerman, R., Kanniainen, L., & Prieler, T. (2018). Cognitive processes and digital reading. In M. Barzillai, J. Thomson, P. van den Broek, & S. Schroeder (Eds). *Learning to read in a digital world* (pp. 57-90). John Benjamins.

Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. *Theory & Practice in Language Studies*, *2*(1), 43–53. https://doi.org/10.4304/tpls.2.1.43-53

Zulqarnain, M., & Saqlain, M. (2023). Text readability evaluation in higher education using CNNs. *Journal of Industrial Intelligence, 1*(3), 184–193. https://doi.org/10.56578/jii010305