PROCEEDING
Indonesian Conference Series of Applied Linguistics
Volume 1 Issue 1 August 2023
E-ISSN: xxxx-xxxx

A CORPUS REVIEW ON LITERACY READING MATERIALS 2019

Dewi Nastiti Lestariningsih[1] , Nanik Sumarsih[2], Mutiara[2], Dody Kristanto[2], Prima Duantika[2], M. Yusuf[2] , Herlinda[2], Wedya Dhaneswara[2], and Mutiya[2]
[1]*Badan Riset Inovasi Nasional,*
[2]*Badan Pengembangan dan Pembinaan Bahasa*
dnastitilestari@gmail.com, dewi041@brin.go.id

Abstract

Language structure and context, which are suitable for the target reader, are very necessary to make a good children's storybook. In his regard, a tool is required to detect word consistency and to support structure and context-building. A tool to detect this requirement is the vocabulary children's book corpus. This study aims at determining a suitable book level for target children. This has an impact on children's storybook users who see the story as an appropriate parameter on their reading level. In this research, the corpus used refers to how the texts are treated in the annotated corpus. By analyzing sentences and classifying words through the corpus, authors and users of children's books can see the word lists that have a high-medium-low frequency. The data were gained from literacy reading materials in Badan Pengembangan and Pembinaan Bahasa in 2019. The information was gathered using antcon-derived document analysis. The vocabulary levels for each child and early graders can be seen in the appearance frequencies of the vocabulary for both levels. Based on the 13 data points included in the top order, nouns for elementary school are no longer found and the use of quotation marks for early readers has been introduced.

Keywords*:* Literacy; picture book; liguistics corpus

INTRODUCTION

In recent years, the Language Development and Development Agency has produced enrichment books in the category of non-text books to broaden children's knowledge at formal and non-formal levels from early childhood to high school. This non-text book consists of picture story books, narrative fiction, and fiction. This book has been launched to the public and can be accessed through the official website of the Language Agency, namely budi.kemdikbud.go.id. and also agency.kemdikbud.go.id. The number of books produced by the Language Agency makes the perpetrators of books, especially children's book writers, have to be creative in creating children's story books. In addition to observing the creative process, children's book writers and also children's book publishers must also know the criteria that exist in children's story books. As users, the targets of books produced by the Language Agency are quite diverse. There are various levels that become priority targets. The focus of this research examines children's story books which are literacy reading materials in 2019 with a hyperfocus on early childhood and early grade levels.

The process of providing literacy reading materials in 2019 began with a story board competition and selection for the improvement of children's story scripts. After the manuscript was obtained, then a writer's meeting was held for the coaching process between the author and the jury who were consultants for writing children's books. This process also pays attention to the substance of the assessment carried out by the Book Center as a book standardization institution in Indonesia. There are four parameters discussed in the assessment of children's story books by the Book Center, such as substance, graphics, presentation, and language. Especially for 2019 literacy reading materials, the resulting book is a picture book which has its own peculiarities with the involvement of graphics and simple text by prioritizing showing not telling techniques.

In the book tier map issued by the Book Center in 2019, it can be seen that there are seven levels in the hierarchy, such as pre-reading, early reading, early reading, fluent reading, advanced reading, advanced reading, and critical reading. The focus of the target for that year was at the early reading level for early childhood and early reading for children attending elementary school grades 1, 2, and 3.

Clay (1993) revealed in his book, readers must recognize 95% of the words in the text to be able to understand the text material independently. With the help of teachers and parents, 90-94% of readers are able to recognize words. A good children's story book can at least be understood by them well. The criteria for a good children's story book include things that children like and make children motivated to read books. Many cases of

children's closeness to books are found not because they naturally like books, but simply fulfill their reading obligations.

To make a good children's story book, a language structure and context are needed that are in accordance with the target audience, in this case referring to the early age and early grade levels. In this regard, a tool is needed to detect the suitability of words that support the building structure and context. A powerful detection tool for this need is the vocabulary corpus of children's books. Classification of word classes is needed by analyzing the sentence structure in children's books. Unfortunately, there is no corpus of children's book vocabulary in Indonesia. Therefore, this year, the effort made is to analyze children's story books that fall into the category of literacy reading materials to make annotations in order to classify word classes based on the text.

The purpose of this research is very basic as literacy development. So far, there is no corpus derived from children's story books, so it is very difficult to determine the level of a book. This has an impact on children's storybook users who see the story text as a parameter of suitability for their reading level.

The limitation in this study is the 2019 literacy reading material produced by the Center for Language and Literature Development, the Language Development and Development Agency which is located on the http://badanlanguagekemdikbud.go.id page.

To see the vocabulary mastered by children from their reading books, a tool is needed to detect the use of these words in a sentence structure. The tool is the corpus. Budiwiyanto, 2014 states that the corpus is a collection of natural texts, both spoken and written, which are systematically arranged. It is said to be "natural" because the texts collected are texts that are produced and used fairly and not artificially. These texts include novels, academic books and papers, newspapers, magazines, broadcast recordings of talks and interviews, blogs, online journals, and discussion groups, and more. It is said to be "systematic" because the structure and content of the corpus follow certain extralinguistic principles, particularly the principle of sampling, which is the basic principle in selecting texts to be included in the corpus. The corpus linguistic method is used to analyze linguistic symptoms using the corpus (Puspita, 2016). This method has been used as a means to analyze actual patterns of language use and also as a tool to develop language teaching materials in the classroom (Reppen & Simpson-Vlach, 2020).

In this study, the corpus used refers to how the texts are treated. That type of corpus is an annotated corpus (vs. orthographic corpus). In it, several types of linguistic analysis have been carried out on the text, such as sentence analysis and word class classification. By analyzing sentences and classifying words through the corpus, authors and users of children's books can see a list of words that have high-medium-low frequencies. This is very useful for the preparation of books based on the frequency of use of words in sentences and their designation for target users.

METHODS

The method used in this research is qualitative with literature study technique. Sugiyono (2005) states that the literature study is related to theoretical studies and other references related to values, culture, and norms that develop in the social situation under study. In other words, literature study is very important in research, because research cannot be separated from scientific literature. Meanwhile, according to Nazir (1998), literature study is an important step after the research topic is determined to further conduct theoretical studies and collect information from related libraries.

Collecting data in this study using a corpus approach through documentation techniques. After the data is collected, the data is then converted into a txt file with UTF-8 encoding code and then uploaded for analysis using the AntConc Windows (3.5.8) corpus processing program developed by Laurance (2004).

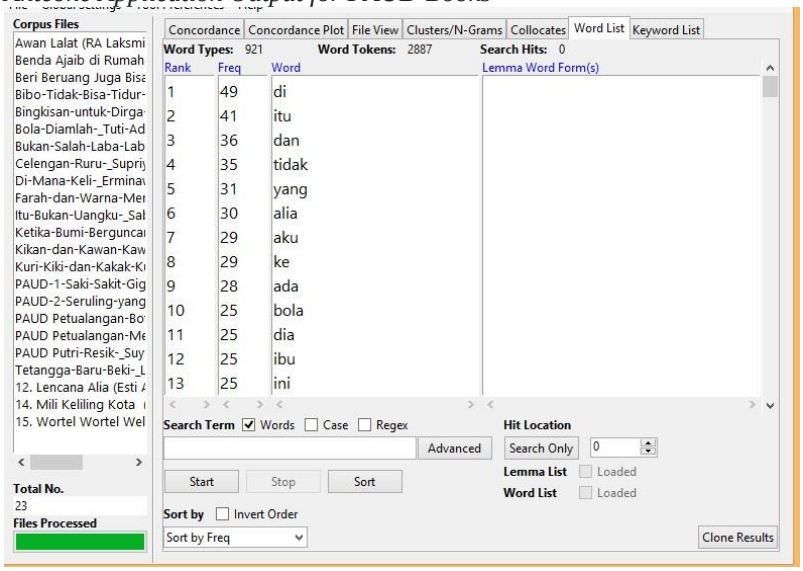
The population of this study is data sourced from children's story books which are classified as non-text books. The sample of this research is data on literacy reading materials produced by the Language Development and Development Agency in 2019.

The research technique is done by documenting and identifying the corpus based on word class and word frequency counting through annotations or through the antconc application. The use of antconc to see the occurrence of word class and frequency calculation is done by converting the file format into txt. and enter it into the antconc application to further see the word class and frequency.

FINDINGS AND DISCUSSION

The classification of word classes based on book data literacy reading materials and finally the frequency or frequency of word appearances for early readers will be shown. Based on the antconc application, it can be seen what words appear and the frequency with which they appear.Below is the output of the antconc application. There are 921 types of words (word classes) from 23 PAUD level books. In Figure 1, it can be seen that the top 13 sequences are dominated by particle, pronoun, verb, noun, and adverb word classes.

Figure 1.
Antconc Application Output for PAUD Books

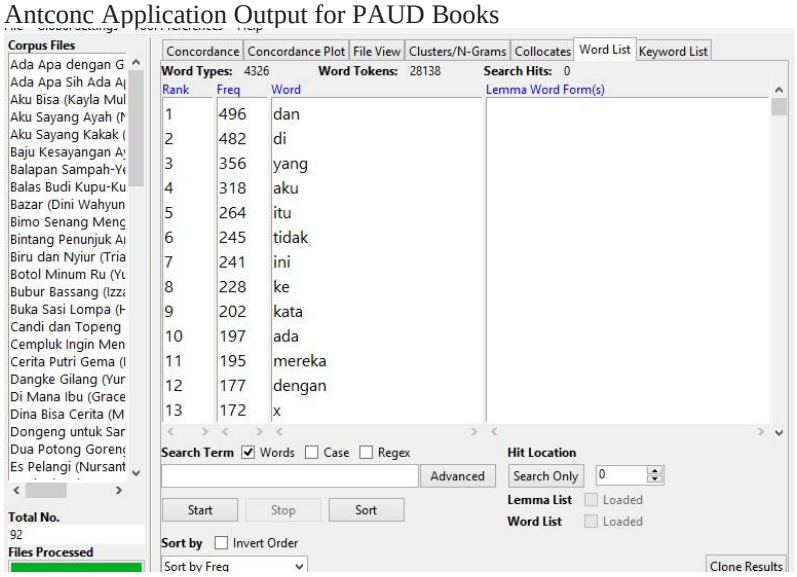


For particles such as those in the table there are words ‘di’ (49 times), ‘dan’ (36 times), ‘yang’ (31 times), and ‘ke’ (29 times). Then for pronouns, as seen in the table, there are the word ‘itu’ (41 times), ‘aku’ (29 times), ‘alia’ (30 times), ‘dia’ (25 times), ‘ibu’ (25 times), and ‘ini’ (25 times). Next for nouns, there are ‘bola’ (25 times). There are verbs ‘ada’ (28 times) and adverb word classes in the form of words ‘tidak’ (35 times) are also seen.

Classification of word classes based on data from books of literacy reading materials and finally the frequency or frequency of word appearances for early readers (SD Grades 1,2,3) will be shown. Based on the antconc application, it can be seen what words appear and the frequency with which they appear.

Below is the output of the antconc application. There are 4326 types of words (word classes) from 92 elementary school level books. In Figure 2, it can be seen that the top 13 sequences are dominated by particle word classes, pronouns, verbs, nouns, adverbs and punctuation marks. For particles such as those in the table there are words ‘dan’ (496 times), ‘di’ (482 times), ‘yang’ (356 times), ‘ke’ (228 times), ‘dengan’ (177 times). Then for pronouns, as shown in the table, there are the words ‘aku’ (318 times), ‘itu’ (264 times), ‘ini’ (241 times), ‘mereka’ (195 times). Furthermore, the verb ‘ada’ (197 times) and the class of adverb words in the form of the word ‘tidak’ (245 times) were also seen. In addition to word classes, the use of punctuation marks in the form of quotation marks as much as symbolized by the letter x for ‘tanda petik’ (172 times) in the antconc is also seen in Figure 2.

Figure 2.
Antconc Application Output for PAUD Books



CONCLUSION

Vocabulary gradations in story texts for early childhood and early graders can be seen in the frequency with which vocabulary appears for both levels. Based on the 13 data that are included in the top order, nouns for elementary school are no longer found and the use of quotation marks for early readers has been introduced. A unique finding for early readers is that the word alia as a pronoun is found 30 times. This indicates that the child's name is very familiar and is often used by the author.

In addition to the use of antconc, corpus identification is done by looking at the word class in each reading book. In this identification, it is found that the use of mother tongue is found in several book titles so that readers can be carried away by positive emotions and can improve the identity (identity) of children. In addition, there is also the use of foreign words and scientific words.

RECOMMENDATION

The recommendation from this study for bookkeepers is the use of vocabulary variations in children's books based on a corpus containing a list of word class identities and the number of word class frequencies. For the Language Agency agency, this effort is the starting point for documenting the corpus of children's books. Furthermore, for the Book Center, this effort is a simple classification in determining vocabulary according to the level of student readers. Furthermore, other efforts that have not been carried out are documenting works of other literacy reading materials that have been prepared by the Language Agency and other publishers that have been declared to have passed the Book Center and documentation of Latin names in the glossary contained in literacy reading materials

REFERENCES

Nazir, M.. (1998). *Metode penelitian.* Ghalia Indonesia.

Puspita, D. (2016). Pemanfaatan korpus dalam analisis makna kata bersinonim mau, ingin, hendak, dan akan. Dalam Prosiding *Seminar Leksikografi Indonesia: Tantangan Leksikografis Bahasa-Bahasa Daerah di Indonesia* (pp. 31-40).

Sugiyono. (2005). Memahami penelitian kualitatif. Alfabeta.

Suhardijanto, T., & Dinakaramani, A. (2018). *Korpus beranotasi: Ke arah Pengembangan korpus Bahasa-Bahasa di Indonesia*. http://repositori.kemdikbud.go.id/10088/

.