



Journal of Pragmatics and Discourse Research

journal homepage: <https://jurnal.ppjb-sip.org/index.php/jpdr/index>



Development of Graded Vocabulary Parameters for Chinese BIPA based on Frequency

Lu Yingxuan¹, Vismaia S. Damayanti², Andoyo Sastromiharjo³, Yeti Mulyati⁴

^{1,2,3,4}Universitas Pendidikan Indonesia, Bandung, Indonesia

ARTICLE INFO

Article history:

Received: 13 – May – 2023

Revised: 24 – June – 2023

Accepted: 29 – July - 2023

Keywords:

High-frequency words,
Graded Vocabulary Parameters,
Chinese BIPA,

ABSTRACT

This study addresses three main inquiries: 1) How to effectively classify Indonesian words based on frequency? 2) What are the distinctive characteristics of words across various proficiency levels? 3) How is vocabulary utilized in BIPA teaching materials? The article aims to create a graded vocabulary list to assess lexical complexity by examining word distribution in texts. Using a quantitative descriptive method, high-frequency words are chosen from eight widely-used Chinese BIPA language sources and categorized into frequency tiers using the %WF>=P method. The graded word list is then applied to BIPA textbooks to assess discourse difficulty. A total of 5,700 word families are organized into five proficiency grades based on cumulative frequencies. Analysis reveals that high-frequency words show a balanced distribution across materials, with a prevalence of polysemous words and concentrated word lengths. Low-frequency words are more domain-specific, emphasizing specific meanings and displaying a broader range of lengths. Examining word usage in different texts indicates alignment between Indonesian teaching materials, academic journals, newspapers, and audio-visual materials with BIPA textbooks. Speech scripts and public documents are more suitable for advanced teaching, while social media language differs notably from BIPA textbooks.

©2023 PPJB-SIP. All rights reserved

1. Introduction

Indonesia ranks among the top ten global economic powers, positioning it as a key focus for international investments (Badan Pengembangan dan Pembinaan Bahasa, 2015). Bahasa Indonesia holds the potential to evolve into an international language, particularly in the wake of the inauguration of the ASEAN Economic Community (MEA) in January 2015 (Chabibah & Kisyani, 2018). Furthermore, with its exceptional tourism potential, Indonesia consistently attracts a substantial influx of tourists and foreign workers. This influx serves as a strategic opportunity for the development of the Bahasa Indonesia untuk Penutur Asing (BIPA) program (Faznur, 2021). According to data from the Ministry of Education, Culture, Research, and Technology (2021), there are 355 institutions registered as organizers of BIPA program. These institutions have catered to a total of 72,746 BIPA learners spanning various countries. Notably, 146 institutions across 29 countries have received support from The Language Development and Fostering Agency for the execution of BIPA programs (Dwinitia, 2021). Furthermore, it is noteworthy that 17 universities in China currently offer courses in the Indonesian language.

¹Corresponding author:

Email addresses: luyingxuan.9321@gmail.com (Lu Yingxuan)

<https://doi.org/10.51817/jpdr.v3i2.712>

©2023 PPJB-SIP. All rights reserved (2828-4755)

The creation and formulation of instructional materials stand as routine responsibilities for Chinese teachers in the BIPA context (Yani, 2019). An initial inquiry, conducted through a questionnaire distributed to Chinese BIPA instructors concerning the crafting of teaching materials, unveiled a prevalent challenge: the complexity assessment of language materials during the design phase emerged as the most commonly reported issue. Nevertheless, the questionnaire findings indicated that over 78% of teachers identify vocabulary as the primary factor directly impacting the complexity of teaching materials. Then, how can words serve as criteria for evaluating teaching materials? According to Bin's research (2022), the gradual introduction of words positively influences language learners. One method for classifying word proficiency levels involves using a tiered word list. For example, the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) has introduced the English Vocabulary Profile (EVP), categorizing vocabulary into six CEFR levels (Yosuff, 2022). Similarly, for Mandarin, a word list exists with three segments encompassing nine levels (Ministry of Education of China, 2021). However, as of now, the authors have not encountered a graded vocabulary list for the Indonesian language.

How can a graded vocabulary list be systematically constructed? Essentially, the development of a word list is rooted in the theoretical framework of word frequency (Mingfu, 2011 & Jingan, 2016). As underscored by Roca-Varela (2013), numerous researchers have extensively explored high-frequency words, accentuating their pivotal role in second language (L2) instruction. Xiaobin and Yali (2012) contend that a meticulous examination of word frequency and repetition in instructional materials is crucial for crafting high-quality resources. Furthermore, both Siagian (2020) and Dang (2020) advocate prioritizing the acquisition of high-frequency words (HFW) for beginners, citing its potential to enhance the learning process and instill confidence in language acquisition. The utilization of a graded vocabulary list holds the potential to significantly enhance language acquisition. Variations in BIPA learners, encompassing factors such as age, culture, social backgrounds, and individual conditions, exert notable influence on the learning process and subsequent outcomes (Dwinitia, 2021). Additionally, the linguistic materials encountered by learners may vary based on their specific objectives. For the sake of result precision, this article predominantly centers its focus on the linguistic materials most frequently employed by Chinese BIPA scholars.

Therefore, the study aims to create a graded vocabulary list in the Indonesian language specifically designed for Chinese learners in BIPA programs, organized based on word frequency. This graded word list is intended to function as a resource for identifying vocabulary usage patterns within educational materials, assisting Chinese BIPA teachers in resolving challenges encountered during the development of teaching materials. Furthermore, the research endeavors to validate the effectiveness of multilevel word lists in various texts earmarked for integration into BIPA instructional materials. Advancements in technology are facilitating the gradual evolution of research on high-frequency words. Within the Indonesian academic community, scholars such as Siagian (2020) and Rofiuddina (2021) underscore the significance of incorporating high-frequency words in BIPA learning. Siagian introduced a methodology for selecting high-frequency words, yet their investigation was confined to novice BIPA learners. Likewise, Rofiuddina compiled a high-frequency word dictionary tailored for beginner BIPA learners. However, it's noteworthy that both studies were specifically centered on BIPA beginners.

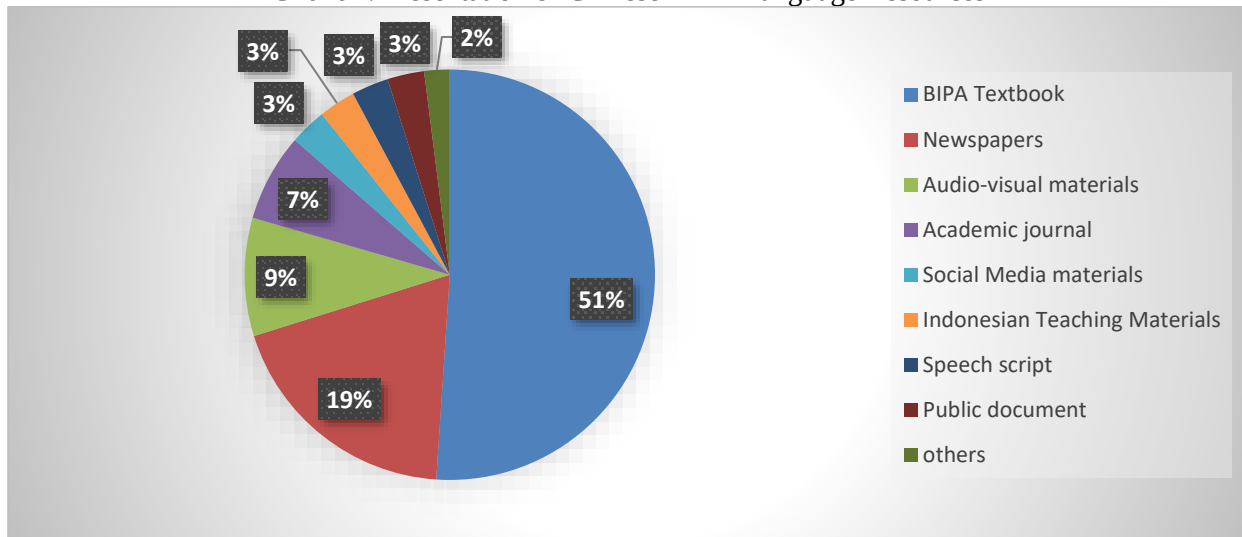
This research is centered on constructing a high-frequency word list designed for Chinese learners engaging in BIPA studies. The primary objectives include categorizing Indonesian words based on frequency, establishing grading parameters for each level, and using the graded vocabulary list as a reference to evaluate vocabulary usage in Chinese BIPA teaching materials. The central research questions address the grouping of words based on frequency, the distinguishing characteristics within each group, and the patterns defining vocabulary usage in BIPA teaching materials.

2. Method

This research employs a quantitative descriptive analysis methodology. As outlined by Listiani (2017), the quantitative descriptive method entails the numerical description, testing, and interpretation of phenomena. Furthermore, Sulistyawati (2022) underscores that quantitative descriptive research employs numerical data to depict, examine, and elucidate a phenomenon without aiming to test a specific hypothesis. The quantitative word frequency data will undergo filtration and management using Antconc and Excel software. The data for this research originates from a selection of eight widely utilized materials employed by Chinese teachers specializing in BIPA. At the core of these materials are fundamental and extensively used BIPA teaching materials. Specifically, the chosen BIPA textbooks comprise three series, collaboratively developed or organized by the Indonesian and Chinese governments. These series include Sahabatku Bahasa Indonesia, encompassing the 2016 and 2019 editions, and The Bahasa Indonesia Kuliah Intensif Tingkat Dasar series. The foundational data for this study is drawn from a subset of sixteen books

within this three-book collection. The utilization of the seven additional language sources is advocated for BIPA learning materials in alignment with the 2017 Curriculum Framework (SKL) outlined by the Indonesian Ministry of Education, Culture, Research, and Technology. Preliminary research findings depict the distribution of various language sources, as illustrated in Chart 1 below.

Chart 1. Presentation of Chinese BIPA Language Resources



The author's aim is to compile a word list based on their frequencies, tailored to the preferences of Chinese instructors specializing in Bahasa Indonesia untuk Penutur Asing (BIPA). This word list will consist of 51% of words sourced from BIPA textbooks, while the remaining 49% will be drawn from seven distinct language sources. Within this 49%, 19% will be derived from newspapers, 9% from audio-visual materials, 6.9% from academic journals, and so forth. Upon confirming the data proportions, the subsequent step involves determining the number of words in the vocabulary list. As per Hua's definition (2020), high-frequency words refer to those accounting for 90% of all words in an article when sorted from high to low frequency. Given that BIPA teaching materials are fundamental and widely used, the vocabulary primarily relies on BIPA teaching books. Subsequently, data from the other seven language sources were collected based on distribution percentages derived from prior research, as illustrated in the table below.

Tabel 1. Data from various language sources

Language Source	Number of Types	Number of Tokens	Proportion of Total Tokens
BIPA textbook	14436	324654	49.80%
Newspapers	5793	133865	20.53%
Audio-visual materials	3167	65078	9.98%
Academic Journal	1863	49518	7.60%
Indonesian teaching materials	859	19087	2.93%
Social media materials	1468	19236	2.95%
Speech scripe	632	19782	3.03%
Public document	478	20718	3.18%

All sources in the table will be consolidated, and subsequently, Antconc will be employed to arrange all words based on their frequency of occurrence. Words with a cumulative frequency of 90% will constitute the foundational data for the word list in this study. Not all words found in the sources will be included in the list. The BIPA textbook series commonly used in China has provided a grammatical understanding of the nature and use of affixes (Gangqin, 2010). This enables students to infer the meanings and word classes, contributing to their lexical comprehension. To enhance the validity of this study's results, the authors intend to calculate the frequency of words based on word family units—a method frequently employed by linguists

(Schmitt, 2000; Nation and Waring, 1997). For instance, if the word '*ketika*' occurs 5 times (5 tokens) in a text, and its derivative word '*seketika*' appears 2 times (2 tokens) in the same text, then the token count for the word family '*ketika*' will be 7 times, represented by the root word '*ketika*'. Word tokens in the corpus reflect the total number of words, encompassing the sum of the frequencies of each word (Arum, 2020).

Several word categories require manual adjustment during the data filtering process, outlined as follows:

- a. Words undergoing inflectional processes, such as '*tugasnya*', '*bahasamu*', '*kuncinya*'; and words with incomplete affixes, such as '*laporkan*', '*duluan*'.
- b. English words resembling Indonesian loanwords, particularly prevalent in scientific works. For instance, the word '*education*' has appeared 516 times in various scientific works, and this word will be used as "*edukasi*" in Indonesian, similar to its original form.
- c. Words with slang usage similar to the original words, often observed in social media language. For example, the word "*slalu*" stands for "*selalu*", the word "*emang*" is slang for "*memang*", and "*bener*" is an allophone form of "*benar*".

Words like the examples above will be entered as derivative words. However, there are also several types of words that are not chosen, even if they are of high frequency.

- a. Interjection. For example, words such as "*tuh*", "*ha*", "*yo*" in audiovisual material, or words like "*deh*", "*yaa*", "*ih*" in social media.
- b. Personal name. Notably, personal names such as the President's name "Jokowi," occurring 237 times in audiovisual materials, and the name "Dilan," which appears 112 times in the same language source.
- c. Brand name. "Tokopedia", "Ovo", "Sedaap", and "Tribunnews" are all well-known Indonesian brands.

3. Result

The key achievements of this study are comprised of three components, intended to tackle the aforementioned research questions. These components include: 1) Basis of word categorization that contains the criteria for grouping words. 2) The characteristics of each level's vocabulary, which describing the language features of words in each level. 3) The vocabulary usage in BIPA materials.

a. Five Levels of Vocabulary Based on Frequency of Use

As previously noted, this study determined the proportion of frequently employed language materials among BIPA teachers through a questionnaire, establishing the parameters for word selection in the vocabulary list. Utilizing the "Word Frequency" and "Lemma List" functions of the Antconc software, all word families were ranked based on their usage frequency. Subsequently, employing Excel's statistical function, words were systematically chosen until the cumulative frequency reached 90%, forming the foundational data for the vocabulary list. Finally, through meticulous word-by-word screening, words deviating from Indonesian language norms, as well as interjections, personal names, and brand names, were systematically excluded.

This systematic process resulted in 5,700 word families, each represented by their root words. How can these 5700 word families be classified into different grade? According to the results presented in Xingru's (2014) article, there are generally four main methods for grading words based on word frequency. a. T grade; b. Top N; c. $WF \geq M$; and d. $\%WF \geq P$. The T value was proposed by Donohue J. C in 1973, and is a

continuation of Zipf's Law. The formula is $T = \frac{1}{2} - 1 + \sqrt{I_1 * 8 + 1}$, where I_1 is the number of words that only appear once, and T is considered as the boundary value between HFWs and low-frequency words (Qiuyu, 2020). Top N refers to words selected by the researcher based on their frequency order. The researchers organized the words in descending order of frequency and chose the top N words with the highest frequency. Typically, N is denominated in units of 10, 50, 100, and so forth. $WF \geq M$ and $\%WF \geq P$ closely resemble Top N but represent the subsequent stages. The author arranges words according to their frequency, where WF denotes the word frequency, and M is a specified number. Words with a frequency (WF) $\geq M$ are then identified as high-frequency words (HFW). $\%WF \geq P$ provides further way of presenting cumulative frequency, with P representing a certain percentage of the cumulative frequency. If the frequency presentation of a word equals or surpasses P, these words are considered high-frequency words (HFW) (Xingru, 2014).

Based on the specific data derived from this study, the author has opted to employ the Percentage of Word Frequency ($\%WF$) method for classifying the 5,700 word families into five levels. The first group encompasses the top 140 highest-frequency words, constituting 50% of the entire dataset. This choice is congruent with the theory posited by Nation & Crabbe (1991), which asserts that a tourist should be

acquainted with approximately 120 items for essential "life needs," such as ordering food, shopping, and understanding basic signs.

The second word group is classified using the 90% cumulative threshold, as highlighted by Hua (2020), who asserts in his book that high-frequency words typically constitute 90% of the texts. This category encompasses a total of 1,125 words. Together, the first two groups comprise 1,265 word families, comprising 22.19% of the word list, collectively representing 90% of the total frequency. The third set of words is established using the 95% cumulative threshold, in line with Laufer's (2010) contention that to comprehend a text, one must grasp 95% of its entirety. This set encompasses 755 word families, accounting for 5% of the entire word list. Combined with the initial two sets, the total reaches 2,020 words, aligning with Schmitt's (2000) proposition that acquiring 2,000 words marks the foundational level for language learners and constitutes a pivotal target for language acquisition. Although mastery of these 2,000 words may not facilitate in-depth discussions with the local community, it does enable satisfactory interaction with native speakers on everyday topics.

The fourth set of words is determined by the 98% cumulative threshold, an extension of Laufer's theory (2010), further elucidated by Vilkaite-Lozdienė and Schmitt (2020). Their assertion posits that achieving mastery over 98% of the words in a text leads to enhanced comprehension. This set encompasses 1,062 word families, constituting 3% of the entire word list. The initial three sets comprise a total of 3,081 words, mirroring the recommendation by Nation & Waring (1997) that learners should adeptly handle a minimum of 3,000 word families to comprehend typical reading materials. The concluding category comprises 2,618 words characterized by the lowest frequencies in the word list, constituting 2% of the overall inventory. This corresponds with the principles outlined by Nation & Waring (1997), advocating that individuals acquainted with approximately 3,000 to 5,000 word families can proficiently grasp typical reading materials. Additionally, Schmitt (2000) underscores that word frequency alone does not suffice as the sole yardstick for evaluating an individual's vocabulary proficiency. He observes that upon mastering about 5,000 word families, students often transition to acquiring domain-specific vocabulary. The classification of the 5,700 word families is as shown in the following table.

Table 2. The Distribution of Words in Each Group

Group	Word Number	Word count	%WF	%WF Cumulative
Group 1	1-140	140	50%	50%
Group 2	141-1265	1125	40%	90%
Group 3	1266-2020	755	5%	95%
Group 4	2021-3081	1062	3%	98%
Group 5	3082-5700	2618	2%	100%

b. Distribution of Stratified Word Frequency from Word Characteristics

Beyond word frequency, linguists routinely evaluate word difficulty by considering various factors. Milton's (2009) comprehensive work delineates the primary criteria for gauging word difficulty, with a particular focus on four key facets: 1) Similarity to the learner's native language; 2) Meaning; 3) Word length; and 4) Word class. In instances where Indonesian employs the alphabet script and Chinese utilizes a logographic script (Maulana, 2020), there is negligible relevance in terms of linguistic similarity. Consequently, relying on the available data, the author opts to conduct an analysis grounded in 1) domain of use; 2) word class; 3) meaning; and 4) word length.

Domain of Use

As previously indicated, this research gathered data from eight distinct language sources. The inclusion of these 5700 word families from various sources does not imply exclusivity to a single origin. For instance, within the initial dataset, 120 word families are present in all surveyed language sources. This signifies that roughly 86% of the data in the first group is distributed across all eight sources, while the

remaining 20 word families are found in five, six, or seven sources. The detailed distribution percentages for each dataset across varying numbers of language sources are presented in Chart 2 below.

Chart 2. The proportion of Word Groups in Various Materials

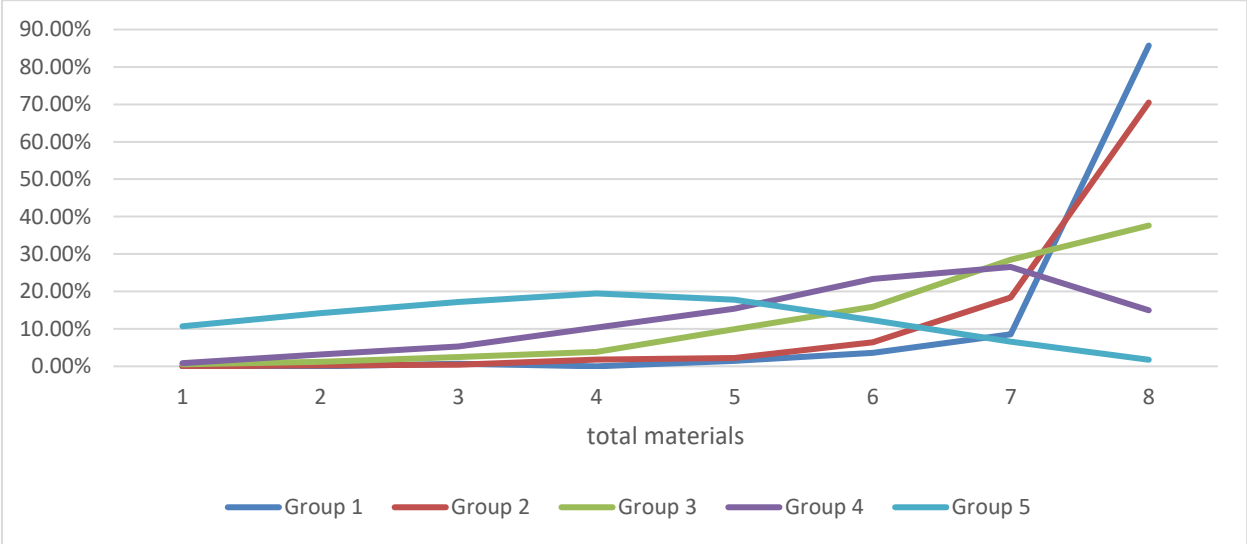
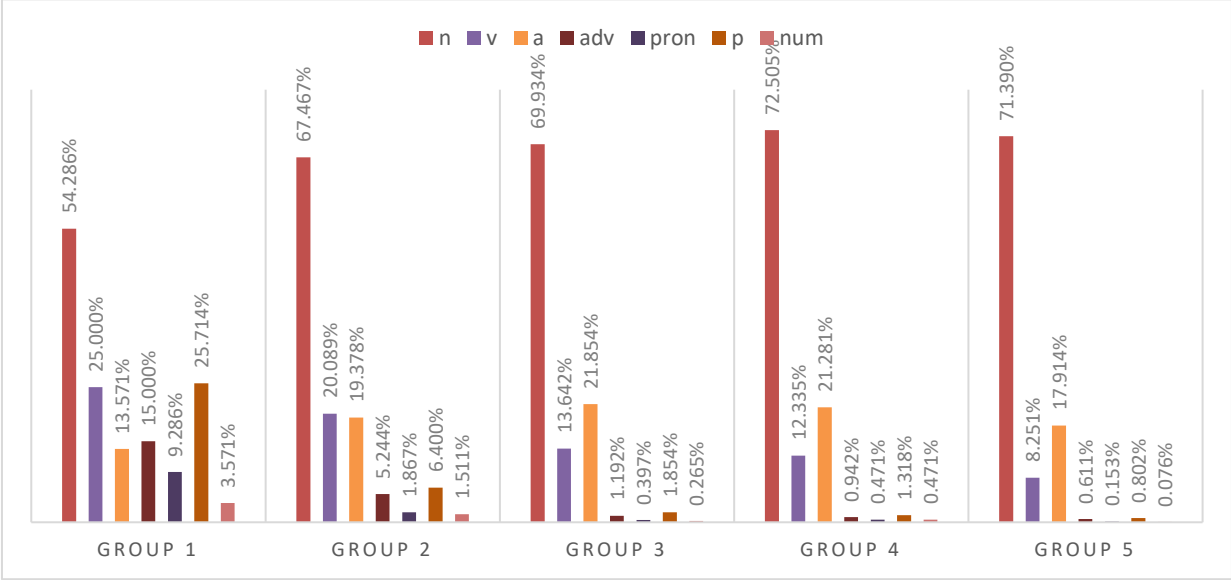


Chart 2 illustrates that words from the first, second, and third groups are present across all text ranges, while the fourth group exhibits a concentration of words in ranges six or seven. The fifth group, encompassing the highest number of words, demonstrates a distribution across all ranges, with a notable prevalence in the third, fourth, and fifth ranges. The implication is that words with higher frequencies are pervasive across diverse text types. For instance, the first group comprises the 140 most frequently occurring words from various language sources, with over 85% found in all types of text collected by the author. A similar trend is observed in the second group, where more than 70% of the words are present in all texts. In contrast, the fourth and fifth groups display a focus on the seventh and fourth ranges, indicating that less frequent words tend to be more specific to certain text types than their more commonly used counterparts.

Word Class

The classification of words into different classes is a factor that can influence their level of difficulty. Insights from Read (2000) and Milton (2009) distinctly suggest that nouns generally pose less difficulty compared to other word classes. In this study, word counts are based on the root forms of word families, and comparisons have been made exclusively within their respective base word classes. The findings for each proportion are detailed in Chart 3.

Chart 3. The Proportion of Word Classes in Each Group



Nouns emerge as the predominant word class in the depicted data, constituting over half of each group. This inclination consistently intensifies, starting at 54% in the first group and progressively reaching around 72% in the fourth and fifth groups. The representation of verb words exhibits a declining trend, dwindling from 25% to 8%. In contrast, the proportion of adjectives remains relatively stable, fluctuating between ten to twenty percent. Meanwhile, adverbs, pronouns, and particles undergo a substantial decline from the first to the third group, plummeting below one percent in the fourth and fifth groups. The numeric word category constitutes the smallest proportion of the dataset. Chart 3 illustrates that word classes in the first group exhibit a more balanced distribution than those in the other groups. The data further indicates that root words in the first group encompass more than one word class compared to base words in the subsequent groups. This phenomenon is mitigated by the reduced use of commonly employed words.

Word Meaning

In line with word classes, Read (2000) emphasizes that word meaning plays a crucial role in determining the difficulty of a word. The author gathered data from KBBI, and the results from a frequency group perspective are depicted Chart 4.

Chart 4. Distribution of the Proportion of Word Meanings in Each Group

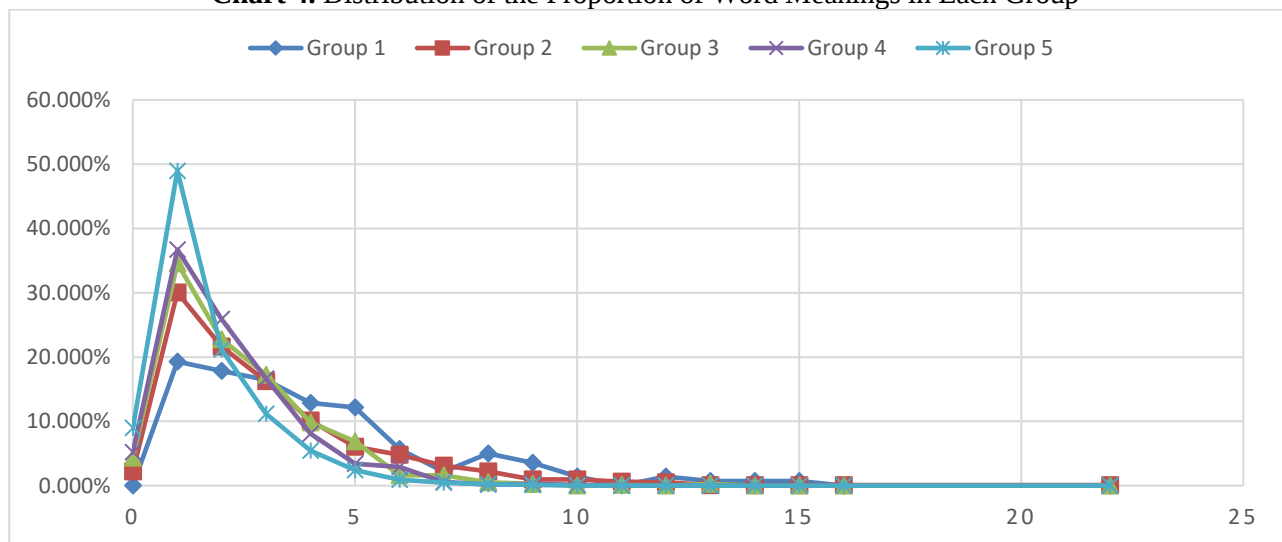


Chart 4 illustrates the distribution of word proportions within each group, revealing a negative correlation between the number of words and their meanings. Words with a solitary meaning are predominant in both high and low-frequency groups. Consequently, the impact of word frequency appears minimal from a trend perspective. However, examining the proportion values unveils that the first group comprises words with a single meaning at approximately 20%. This percentage escalates to 30% in the second group and steadily advances across subsequent groups, culminating in nearly 50% of words with a singular meaning in the fifth group. This observation suggests that words with relatively lower frequency tend to concentrate more on a single meaning.

Word Length

Milton (2009) posited that shorter word forms are more memorable than longer ones, implying a correlation between word length and difficulty. To investigate this, the author conducted a letter count analysis of productive words, revealing that the majority consist of four or five letters. The distribution is outlined below.

Chart 5. Distribution of Word Length in Each Group

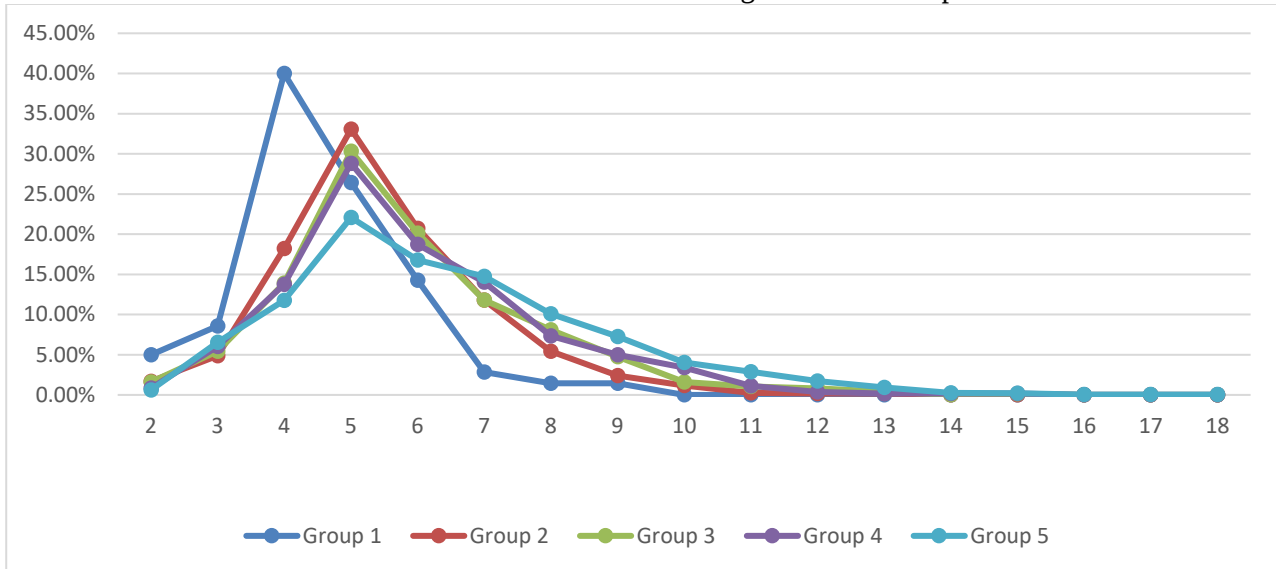


Chart 5 illustrates the distribution of word lengths in each group. In the first group, 40% of the words consist of four letters, with the longest words, "Indonesia" and "bagaimana," containing nine letters. The occurrence of five-letter words increases from the second group to the fifth group, albeit with a decline in proportion from 33% in the second group to 22% in the fifth group. Notably, the data in the fifth group exhibits greater dispersion compared to the other groups. Given that Indonesian words are formed by syllables, the author aims to examine the distribution of syllables among the 5700 selected words.

Chart 6. Distribution of Word Groups with Multiple Syllables

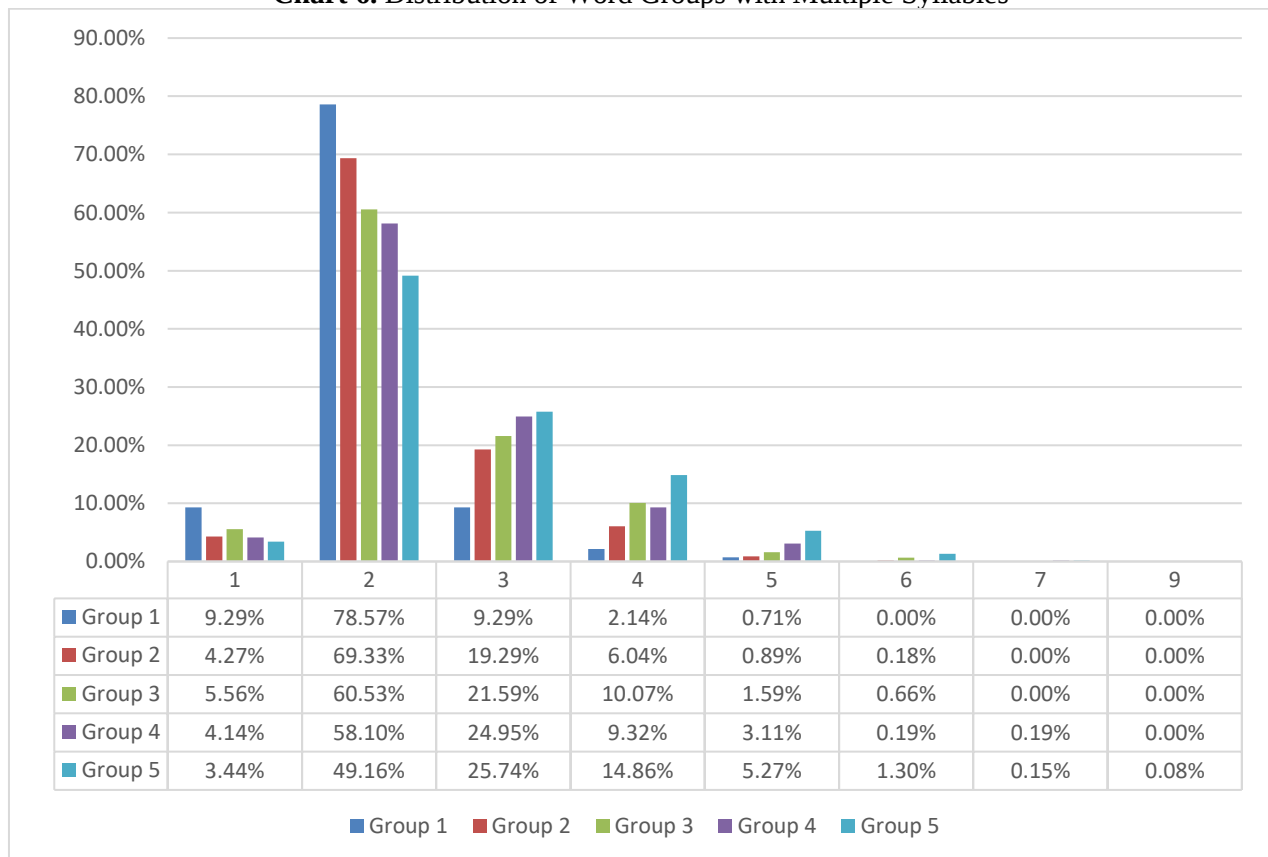


Chart 6 illustrates a prevailing trend where words with two to four syllables constitute the majority. Upon closer examination, it is revealed that nearly 80% of the words in the first group are disyllabic, and this proportion steadily decreases, reaching less than 50% in the fifth group. Conversely, the trend reverses when

considering an increase in syllables. For instance, in the three-syllable word category, the fifth group comprises more than a quarter of the data, surpassing all other groups, and this pattern persists throughout the study, extending to the final group.

c. Distribution of Word in Use

The author conducted a preliminary test on the usage of words in BIPA textbooks and seven other language sources, obtaining word distribution patterns from each of these language materials.

The Use of Grouping Words in the BIPA Textbook

How are the above 5 groups of words used? The following is the use of grouped vocabulary in BIPA textbooks.

Table 3. Distribution of word groups in the BIPA Textbook

Book	Group 1	Group 2	Group 3	Group 4	Group 5	Other words	New words
Chinese textbook	42.5%	39.5%	5.8%	3.7%	3.0%	5.6%	1.3%
Indonesian textbook 2016	44.3%	37.4%	5.2%	3.4%	2.3%	7.4%	3.2%
Indonesian textbook 2019	43.1%	39.3%	4.4%	3.6%	2.8%	6.8%	2.8%

Table 3 shows the proportion of word groups from the three BIPA book series, and the data from these three book series are not much different. The first and second groups account for the largest proportion of words, with around 80% of the words in the three series being from the first and second groups. At the same time, this figure drops sharply from the third group, with a gradual decline to 2% or 3% from the third to the fifth. The proportion of “other words”, which include numbers, people’s names, interjections cannot be ignored either. The difference between Indonesian books and Chinese books in this category of words is the acronym more commonly used in Indonesian books. Apart from numbers and names, there is less than 4% of truly new words. According to the opinion of Laufer, Vilkaite-Lozdiene and Schmitt (2020), this figure should below 5% for understanding the text.

The Use of Grouping Words in Various Materials

Preliminary findings indicate that teachers also prefer to use a variety of language materials in their teaching, in addition to BIPA materials. Based on their percentage of use, the authors also took 100 random language texts to see the distribution of the 5700 words, with the average values below.

Table 4. Distribution of HFW in Various Language Materials

Language Material	Group 1	Group 2	Group 3	Group 4	Group 5	Other words	New words
Newspapers	35.52%	36.99%	5.73%	3.99%	3.76%	14.13%	4.26%
Audiovisual Material	44.57%	37.24%	4.93%	2.22%	2.01%	9.04%	5.67%
Academic journal	46.64%	35.00%	3.66%	2.26%	3.16%	9.28%	5.38%
Social Media materials	31.50%	31.08%	5.48%	12.73%	3.40%	24.10%	14.03%
BI Teaching Materials	42.23%	39.02%	3.85%	3.18%	2.42%	9.15%	3.97%

Speech script	48.63%	33.77%	2.23%	3.67%	13.62%	9.50%	6.23%
Public document	44.85%	32.72%	4.58%	4.27%	2.38%	11.25%	8.27%

As can be seen from Table 4, apart from social media, the new word categories from other sources are all below 10%, and language materials taken from news and other Indonesian language teaching materials are below 5%. Meanwhile, the largest number of new words came from social media, around 14%, other data in social media materials are also different from other language sources.

4. Discussion

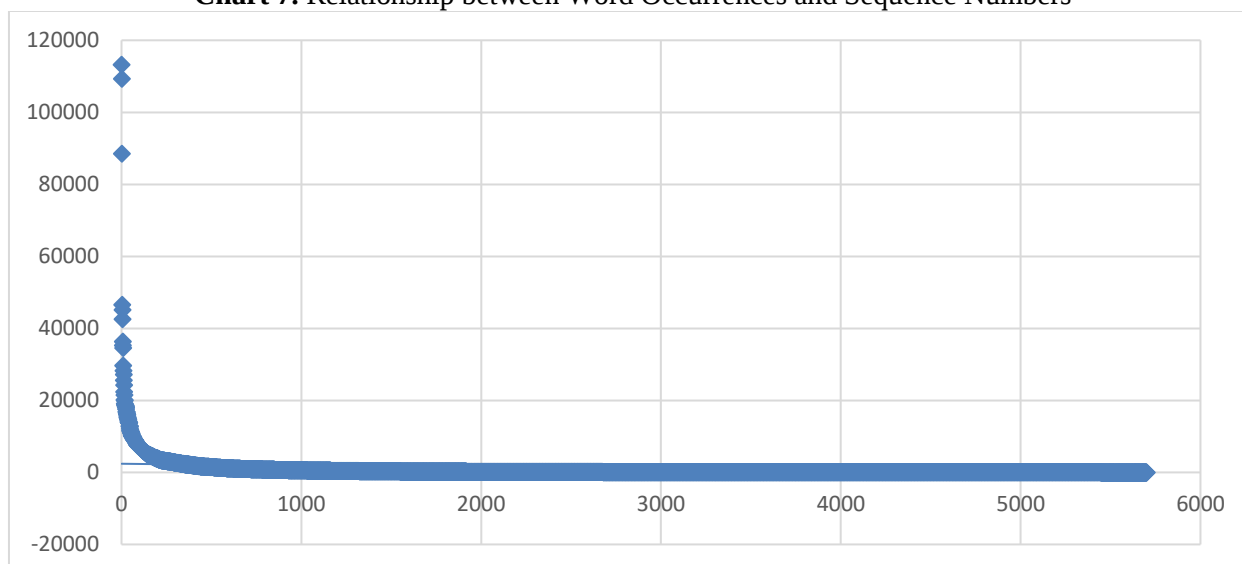
As previously indicated, this section will undertake analysis from two distinct perspectives: the categorization of words and the attributes of vocabulary.

a. Stratified Word Frequency Distribution

The classification of word levels in this study primarily utilizes a combination of the Top N and %WF methods. The 5700 word families are distributed into five groups based on word frequency. To assess the alignment of this level division with the BIPA curriculum requirements, a comparison is made between the vocabulary needs for different levels outlined above and the BIPA curriculum. The proposed levels by experts broadly cover distinct language proficiency ranges consistent with the language proficiency requirements of BIPA levels 1 to 5. As per the theories of Nation & Crabbe (1991), Nation & Waring (1997), and Schmitt (2000), it becomes evident that words in Levels 1 to 3 primarily address the daily life expression needs of language learners, akin to the requirements of BIPA 1 and BIPA 2 in the 2017 Indonesian Language Learning Standards (SKL2017, Kemendikbud, 2017). According to Nation & Waring's theory (1997), the vocabulary in the fourth group of the graded word list necessitates learners to have the ability to read short articles, aligning with the requirements of BIPA level 3. The final group of words requires learners to comprehend general articles, consistent with the reading requirements of BIPA levels 4 and 5. Schmitt's theory (2000) suggests that to comprehend texts in a specific field, one needs to master at least 5,000 word families, corresponding to the reading content of BIPA levels 6 and 7, where learners are expected to understand discourse in a specialized domain.

However, due to differing criteria for word selection, SKL 2017 provides detailed descriptions of the types of words that students at each level should master based on themes, text types, and affixes. This differs from the approach in this study, where word roots were selected as representatives of word families based on their frequency of occurrence. Consequently, a direct match between the selected words and the word requirements for each level of SKL BIPA is not feasible. Nevertheless, the overall language requirements and the general grading align with SKL 2017's expectations for BIPA students. Therefore, dividing the 5700 word families into five levels based on cumulative frequencies of 50%, 90%, 95%, 98%, and 100% is in line with the requirements for BIPA students in SKL 2017. The outcomes of word categorization reveal a noteworthy pattern where the highest-frequency words dominate the overall distribution. This is exemplified by the first group, encompassing a mere 140 words, yet contributing to 50% of the total. In stark contrast, the final group encompasses 2,618 word families, representing nearly 46% of the 5,700 word families in the dataset. However, its contribution to the cumulative frequency in the entire dataset is a mere 2%. This pattern aligns with Zipf's law, recognized as the power-law distribution. A key characteristic of this theory is the prevalence of events with very small magnitudes, while only a few events attain relatively large scales. In simpler terms, power-law distributions depict highly uneven distributions (Liu, 2017). This theory is commonly referred to as the "long-tailed distribution." The data collected in this study illustrate the word ranking and frequency of occurrence, portraying a resemblance to the pattern observed in Zipf's Law.

Chart 7. Relationship between Word Occurrences and Sequence Numbers



Upon attempting to formulate this relationship in mathematical terms, the investigation yielded the equation $y = 5E+07x-1.709$, accompanied by an R^2 value of 0.4401. In general, an R^2 value below 0.75 is deemed unacceptable (Haitao, 2020). Does this research data align with Zipf's Law? It appears that in this study, the data was organized based on word occurrences, revealing that for words falling below the 1539th position, $R^2 \geq 0.75$, and the formula $y = 858826x-1.056$. In essence, these findings indicate adherence to Zipf's Law only for words of high and medium frequency, where each frequency number (r) corresponds to a single word frequency (fr) in a given corpus. However, this observation does not hold true, particularly for words of relatively low frequency (Haitao, 2020).

Practically, instances of words sharing the same frequency within a corpus are commonplace. Consequently, the Zipf formula mentioned above is more suited to an ideal scenario involving words of medium to high frequency. Analyzing the study's data, the usage frequencies of the first and second groups conform to Zipf's law. This underscores the limitations of Zipf's law in practical applications, as individual variations emerge depending on the language and selected corpus. Nevertheless, when examining the overall distribution, the pattern of word usage still aligns with Zipf's law. This implies that in a language, a small subset of words is used extensively and frequently, while the majority of words are used less often. Consequently, it remains imperative for foreign students to focus on learning high-frequency words. As highlighted by Siagian (2020) and Dang (2020), mastering high-frequency words can significantly bolster beginners' confidence and enhance their learning efficiency.

b. Distribution of Grouping Words in Word Characteristics

The analysis of data distribution reveals that high-frequency words, utilized across diverse domains, exhibit a higher number of tokens and a more substantial presence within texts. Put differently, words commonly employed across varied sources tend to accumulate more tokens, classifying high-frequency words as general terms. Drawing from the data on total occurrence proportions, it becomes apparent that the majority of tokens in a given text fall under the category of general words. In contrast, lower-frequency words demonstrate a propensity for heightened usage within specific sources. The 5700 words selected in this study encompass frequently occurring words from diverse texts, indicating that they embody not only general terminology but also domain-specific general terms from distinct language sources.

Analysis of data across various word classes indicates that low-frequency words tend to be predominantly associated with a specific word class. Overall, basic words have more nominal meanings, with verbs, adverbs, pronouns, auxiliaries and counters decreasing significantly to less than 1% as they are used less frequently. Meanwhile, there is little change in the proportion of adjectives used. It can be said that nouns have a high proportion, which is related to relatively large number of nouns, a more balanced distribution of classes in common words, and a more representative word focused on nouns. On the other hand, in addition to verbs and adjectives, adverbial words, pronouns and auxiliary words are also common words. Examination of the word classes data reveals that while selected words typically exhibit single meanings, the first group maintains a better balance compared to the subsequent groups. The fifth group, in particular, demonstrates an even more restricted range of meanings. In essence, infrequently used words tend to concentrate on specific

meanings, whereas commonly used words often encompass multiple meanings, each falling within distinct word classes. Less frequently used words exhibit a more consistent association with a particular word class in terms of meaning.

In the comprehensive breakdown, this study reveals a prevailing dominance of nouns, constituting approximately 65% of the content. Adjectives closely trail at around 18%, while verbs follow suit with an approximate 11.5%. Prepositions represent a mere 2.5%, adverbs 1.8%, and both pronouns and numerals each fall below 1%. It is noteworthy that these findings slightly diverge from those of Chabibah (2018) and Wahyuningtyas (2021). In both studies, which classify all words in "Sahabatku Bahasa Indonesia" by their parts of speech, significant variations emerge. Particularly, verbs emerge as the second most common category, with Chabibah (2018) reporting a percentage of around 30%, a marked contrast to the 11.5% observed in this study. Wahyuningtyas (2021), on the other hand, highlights particles as the third most frequent word type, with only 52 occurrences in total, despite appearing 1,756 times in the 2019 version of "Sahabatku Bahasa Indonesia." These discrepancies can be attributed not only to variations in research material but also to the focus of this study on word roots. Given the agglutinative nature of Indonesian (Hasana, 2022), this is particularly significant, as affixes directly influence the part of speech of words, potentially exerting a substantial impact on the research outcomes.

Similarly, as word lemmas constitute the focal data in this study, they exert a direct influence on the word length. The outcomes of the analysis on word length reveal that the most frequently used words typically encompass four to five letters with two syllables. This phenomenon is attributed to the nature of Indonesian as an agglutinative language, wherein roots are intentionally kept concise to facilitate the seamless addition of affixes. Regarding usage, words that are phonetically straightforward are inherently more memorable and user-friendly, aligning with Zipf's law governing low-frequency words—the principle of energy conservation. In essence, people tend to employ fewer words to convey more, and frequently used words are generally easier to articulate (Haitao, 2020).

c. The Use of Word in Various Sources

The results of word usage in BIPA textbooks and various language materials demonstrate that the vocabulary in Indonesian language textbooks remains consistent, whether the material is intended for BIPA students or not. This trend is also observed in audio-visual materials and academic works, where a more formal language style is employed, aligning closely with the standards set by BIPA materials. Analyzing data from audio-visual materials reveals that it does not significantly deviate from the content found in BIPA textbook series. This alignment can be attributed to the nature of audio-visual language, characterized by its mass appeal, universality, demonstrative qualities, diversity, and aesthetics (Ruiling, 2021). Therefore, the language used in audio-visual materials caters to a wider audience and subtly influences societal language usage.

In addition, the collected newspaper language data demonstrates similarities in word usage when compared to BIPA textbooks. This observation can be readily understood when considering Sumadiri's (2011) summary of key features of news language, which emphasizes attributes such as conciseness, directness, clarity, engagement, democracy, active voice usage, and the avoidance of specialized terminology or jargon. Moreover, since 1975, the Indonesian Journalist Association has issued guidelines for the use of language in journalism, emphasizing the importance of adhering to the Indonesian language spelling guidelines consistently. On the other hand, the utilization of words in news, speech texts, and public documents tends to concentrate more on categories three to five. This indicates a higher prevalence of representative words from these three sources. The focus on using appropriate language for speeches is essential, as such texts serve not only in the school environment but also in various community settings. When learning to write speech texts, the vocabulary used plays a crucial role in persuading, inviting, and influencing the audience. Therefore, it is necessary to employ clear, logical, and systematic sentences while enriching the content with accurately spelled and suitable vocabulary (Yudha, 2019).

Government work reports, on the other hand, represent highly formalized texts with relatively fixed topics, writing styles, syntax, and vocabulary, reflecting a wealth of information across various sectors like politics, economics, agriculture, diplomacy, education, culture, health, and sports, which forms their distinct discourse characteristics (Shengran, 2014). Thus, the precision of word choice is of particular importance in speech drafts and government documents, as they come with distinct themes and, consequently, higher language requirements for students. Regarding the language used in social media, a distinct form of communication emerges, characterized by the use of slang, which varies from the language used in BIPA textbooks. A substantial portion of the language employed on social media platforms is in the form of local dialects or slang. This can be attributed to the source of the materials, which often include comments from

platforms like Youtube, Tiktok, and Tokopedia. Irawan & Hidayat (2020) have previously emphasized the informality of spoken language, which is often tailored to specific social groups. Kholifah (2020) has indicated that language usage on social media platforms is frequently accompanied by spelling errors.

Furthermore, Hilaliyah (2010) has noted that a prominent feature of spoken Indonesian language is its transient nature. Given the rapid evolution of spoken language, which can pose a challenge for language learners, educational materials tend to focus on more stable and standardized vocabulary. This implies that the primary focus of the study on 5700 HFW is the use of formal and official language, which is typically found in the Great Indonesian Dictionary (KBBI). The author applied graded vocabulary to the BIPA textbooks, resulting in an analysis of the vocabulary usage across different textbook series. This enabled the observation of variations in word choice among these series, serving as a reference parameter for teachers when designing and selecting teaching materials in terms of word usage. Additionally, the author extended the application of graded vocabulary to seven other commonly used BIPA class. This analysis of vocabulary difficulty levels revealed the suitability of different learning resources for students at varying proficiency levels, aligning with the recommended corpus usage sequence in the 2017 Curriculum (SKL).

5. Conclusion

The researcher curated a list of the 5,700 most frequently employed words by Chinese BIPA teachers, drawing from eight diverse language sources. These words were subsequently organized into five distinct groups according to their frequencies, employing cumulative thresholds of 50%, 90%, 95%, and 98%. In order to discern the distribution patterns of these 5,700 selected words, a comprehensive analysis was conducted, focusing on both word characteristics and patterns of word usage. The trends in the distribution of word characteristics reveal the following insights: 1) High-frequency words exhibit broader usage across diverse language materials, while relatively low-frequency words tend to be concentrated in specific contexts; 2) Word roots with high frequency in Indonesian tend to center around one or two meanings, whereas less frequently used words often focus on a specific meaning; 3) More than 90% of basic words have two to four syllables, while words of relatively low frequency display greater variability in length; 4) The word classes in the first group demonstrate greater diversity, corresponding to the number of meanings represented, with less frequently used words typically belonging to a single word class. Through an examination of the usage patterns of 5,700 vocabulary families across diverse materials, notable variations emerge in vocabulary selection. This dataset serves as a valuable resource for BIPA teachers in the design of teaching materials, ensuring alignment with the current needs of students.

The purpose of this research is to overcome the difficulty that Chinese BIPA teachers have been facing, namely the lack of indicators to assess self-developed teaching materials. Words are the most frequently mentioned elements by teachers as indicators that have a direct effect on the difficulty of teaching materials. Thus, the author hope that the results of this study will provide a reference for Chinese BIPA teachers in evaluating teaching materials as they design them.

Declaration of Conflicting Interest

The author stated that there is no conflict of interest concerning the publication of this paper.

References

- A T Wahyuni. (2021). Pengembangan Bahan Ajar Kosakata dalam Pembelajaran Teks Bahasa Indonesia Kelas VIII Semester Gasal Sekolah Menengah Pertama. *Metafora*. 8 (2) 179-193. DOI: 10.30595/mtf.v8i2.11996
- A Rofiuddina. (2021). Pengembangan Bahan Ajar BIPA Daring Tingkat Pemula Rendah. *Ranah: Jurnal Kajian Bahasa*. 10(1). 153—169 DOI: <https://doi.org/10.26499/rnh.v10i1.3376>
- A Xingru. (2014). The Research on the Threshold of High - Frequency Words Based on the Normal Distribution in Word Frequency Analysis. *Journal Of Intelligence*. 33(10), 129-136. DOI: 10.3969/j.issn.1002-1965.2014.10.022
- A Umiyati, BS Pratama, N Aini, WA Kesumastuti. (2021). Afiks Derivasional Ber- pada Media Massa Indonesia. *Jurnal Hasta Wiyata* 4(2). DOI 10.21776/ub.hastawiyata.2021.004.02.01
- B. Laufer (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* April 2010, Volume 22, No. 1 ISSN 1539-0578. Pp. 15-30
- Chabibah, S., & Kisyani, D. (2018). Perkembangan kosakata dan fonotaktik pada buku ajar bipa tingkat a1—c2 Terbitan Kemendikbud. *Jurnal: Unesa*. <http://jurnalmahasiswa.unesa.ac.id/index.php/bapala/article/view/25793/23638> (diakses 30-Maret-2018, pukul: 21: 17).

- CM Zellaatifanny & B Mudjiyanto. (2018). Tipe Penelitian Deskripsi dalam Ilmu Komunikasi Jurnal Diakom, 1(2), 83-90 <https://media.neliti.com/media/publications/278580-tipe-penelitian-deskripsi-dalam-ilmu-kom-a74b7b08.pdf>
- Dewi, M. (2014). Gaya Bahasa Berita Media Online di Indonesia: Judul Menarik Tidak Harus Tidak Baku. *Humaniora*, 5(2), 1015-1022.
- Dwinitia, S., & Anggraini, R. (2021, September). Penggunaan Metode Langsung dalam Pembelajaran Bahasa Indonesia bagi Penutur Asing (BIPA). In Seminar Nasional SAGA# 4 (Sastra, Pedagogik, dan Bahasa) (Vol. 3, No. 1, pp. 101-110).
- D Purnanto, (2006). Kajian Morfologi Derivasional dan Infleksional dalam Bahasa Indonesia. *Kajian Linguistik dan Sastra*, 18(35), 136-152. ISSN 2541-2558. link: <<https://journals.ums.ac.id/index.php/KLS/article/view/5058>>.
- E N Siagian. (2020). Kata Berfrekuensi Tinggi dalam Pembelajaran BIPA Pemula. *Ranah: Jurnal Kajian Bahasa*, 9(2). DOI: <https://doi.org/10.26499/rnh.v9i2.2320>
- Ermanto. (2016). Proses Morfologi Infleksi pada Adjektiva Bahasa Indonesia. *Humanus*. 15(1), DOI: <https://doi.org/10.24036/jh.v15i1.6411>
- ER Arum, W Winarti. (2020). Penggunaan Linguistik Korpus dalam Mempersiapkan Bahan Ajar English for Specific Purpose di Bidang Radiologi. *Jurnal Teras Kesehatan*. 3(1), Januari. 2020. ISSN (p): 2622-2396, ISSN (e): 2622-3805 DOI: <https://doi.org/10.38215/jutek.v2i2.39>
- Faznur, L. S., Fadly, A., & Nufus, F. A. (2021). Pengembangan Buku Digital BIPA Berbasis Nilai Moderasi Islam. *Pena Literasi*, 4(2), 92-98.
- Hasana, H. (2022). Fungsi dan Peran Bahasa Indonesia dalam Penulisan Ilmiah. *Jurnal Literasiologi*, 8(4). <https://doi.org/10.47783/literasiologi.v8i4.391>
- Hilaliyah, H. (2010). Maraknya Penggunaan Bahasa Gaul di Kalangan Pelajar Sekolah Menengah Atas. *Deiksis*, 2(01), 13-26.
- H Kridalaksana. (2009). *Pembentukan Kata dalam Bahasa Indonesia*. Gramedia Pustaka Utama: Jakarta. ISBN 978-979-403-678-5
- Irawan, S., Sudika, I. N., & Hidayat, R. (2020). Karakteristik Bahasa Gaul Remaja sebagai Kreativitas Berbahasa Indonesia pada Komentar Status Inside Lombok di Instagram: Characteristics of Teenage Slang as Indonesian Language Creativity on Status Comments Inside Lombok on Instagram. *Jurnal Bastrindo*, 1(2), 201-213. <https://doi.org/10.29303/jb.v1i2.44>
- J. Read. (2000). *Assessing Vocabulary* (Cambridge Language Assessment). Cambridge University Press: Cambridge DOI <https://doi.org/10.4324/9780429291586>
- J Milton. (2009). *Measuring Second Language Vocabulary Acquisition* (Second Language Acquisition, Multilingual Matters Limited: Bristol. DOI <https://doi.org/10.21832/9781847692092>
- Kementerian Pendidikan dan Kebudayaan Republik Indonesia. (2017). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 27 Tahun 2017* Jakarta: Kementerian Pendidikan dan Kebudayaan.
- Kementerian Pendidikan dan Kebudayaan Republik Indonesia, (2017). *Tata Bahasa Baku Bahasa Indonesia*. Badan Pengembangan dan Pembinaan Bahasa Kementerian Pendidikan dan Kebudayaan. ISBN 978-979-069-263-3.
- Kholifah, U., & Sabardila, A. (2020). Analisis Kesalahan Gaya Berbahasa Pada Sosial Media Instagram dalam Caption dan Komentar. *Jurnal Nusa*, 15(3), 352-364.
- L Bin & L Xiaoxi & L You (2022). 有关零基础细分级的高职英语词汇教学改革的实验研究 (An Experimental Study on the Reform of Vocabulary Teaching in Higher Vocational Colleges with Zero-Based Subdivision.). *English on Campus*. 36, 6-8. <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XXXX202236002&DbName=CJFN2022>
- L Haitao. (2020). *An Introduction to Quantitative Linguistics*. The Commercial Press: Beijing. ISBN 978-7-100-15021-7
- L Hua. (2020). *语料库语言学——理论、工具与案例* (Corpus Linguistics- Theory, Tools, and Cases. Foreign Language Teaching and Research Press. Beijing. ISBN 978-7-5213-2283-5
- L Mingfu, Y Liqin, & L Jing (2011), *The Compilation of Level — based College English Word List for Level — based Teaching*. *Journal of Higher Education Research*. Vol. 34, No.3, Sep, 2011
- L Qiudong. (2023). Research on the Lexical List Distribution of Word Families in Business English Listening and Speaking Textbooks based on Range BNC. *Journal of Heilongjiang Vocational Institute of Ecological Engineering*. 36 (2) DOI: 10.3969/j.issn.1674-6341.2023.02.29

- Listiani, N. M. (2017). Pengaruh Kreativitas Dan Motivasi Terhadap Hasil Belajar Mata Pelajaran Produktif Pemasaran Pada Siswa Kelas XI SMK Negeri 2 Tuban. *Jurnal Ekonomi Pendidikan Dan Kewirausahaan*, 2(2), 263. <https://doi.org/10.26740/jepk.v2n2.p263-275>
- Marlina, E. (2020). Pengembangan Model Pembelajaran Blended Learning Berbantuan Aplikasi Sevima Edlink. *Jurnal Padagogik*, 3(2), 104–110. DOI <https://doi.org/10.35974/jpd.v3i2.2339>
- ML Roca-Varela. (2013). High-frequency English Words in Spoken Learner Language: Actual, career and pretend as a Case in Point. *Procedia - Social and Behavioral Sciences*. 95(25), 557-562. DOI: <https://doi.org/10.1016/j.sbspro.2013.10.682>
- M Toha (2013). Derivasi dan Infleksi Bahasa Simeulue. *Kekelpot*: 9 (1), 2013 ISSN 1858-3032
- Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N., & McCarthy, M. (Eds.), *Vocabulary: Description, acquisition, and pedagogy* 6-19, Cambridge: Cambridge University Press. ISBN 0-521-58484-i
- NH Wicaksono, I Nurhayani, I Khasanah. (2022). Makna Derivasional dan Infleksional Verba Nasal Bahasa Jawa Dialek Arek. *SeBaSa: Jurnal Pendidikan Bahasa dan Sastra Indonesia*. 5(2), November 2022. E-ISSN: 2621-0851
- P Nation & D Crabbe. (1991). A Survival Language Learning Syllabus for Foreign Travel. *System*: 19 (3), 191-201. ISSN: ISSN-0346-251X
- Pratiwi, D. F. (2018). Kesulitan yang Dihadapi oleh Para Pemelajar dalam Mempelajari Bahasa Indonesia di Universitas Nottingham, Inggris. Long, M. (1988). *Understanding second language acquisition*. Rod Ellis. Oxford: Oxford University Press, 1985. Pp. 327. *Studies in Second Language Acquisition*, 10(1), 79–82. <https://doi.org/10.1017/S0272263100007038>
- Renhoat, S. F., Faitah, I., Yunita, N., & Fadly, A. (2020). Pemelajaran kosakata bahasa indonesia bagi penutur asing: adaptasi pesantren. *Jurnal Genre (Bahasa, Sastra, Dan Pembelajarannya)*, 2(1), 1-4.
- S Fangwen. (2020). 近二十年国内外生词注释对词汇附带习得的影响研究综述(A Review of Research on The Impact of New-word Annotations on Incidental Vocabulary Acquisition in The Past Two Decades). *Think Tank Era*. 10, 260-261. 2096-4609 (2020) 10-0260-002. <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZKSD202010125&DbName=CJFQ2020>
- Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge University Press, Cambridge. ISBN 0-521-66048-3 (hb) – ISBN 0-521-66938-3 (pb)
- Sulistiyawati & Wahyudi & Trinuryono.(2022). Analisis (Deskriptif Kuantitatif) Moticasi Belajar Siswa dengan Model Blended Learning di Masa Pandemi Covid 19. *Kadikma* 13(1) 68-73, ISSN 2686-3243. doi: <https://doi.org/10.19184/kdma.v13i1.31327>.
- T N Y Dang. (2020). High-frequency words in academic spoken English: corpora and learners. *ELT Journal*, 74(2), 146–155, <https://doi.org/10.1093/elt/ccz057>
- Vilkaitė-Lozdienė L. ir Schmitt N. (2020). Frequency as a guide for vocabulary usefulness: High-, mid- and low-frequency words. Webb S. (ed.) *The Routledge Handbook of Vocabulary Studies*
- Winarsunu T. (2017). *STATISTIK Dalam Penelitian Psikologi dan Pendidikan*. (n.p.): UMMPress.
- Y Qiuyu & X Yuequan. (2020). An Empirical Study on Several Methods to Determine the Threshold of High Frequency Words in Co-word Analysis. 38(9), DOI: 10.13833/j.issn.1007-7634.2020.09.014
- Yusoff, Z., A. A. Gurmani, M., Sanif, S., & Mohd Noor, S. (2022). The Effect of Mobile-Assisted CEFR English Vocabulary Profile Word Lists on L2 Students' Vocabulary Knowledge. *Asian Journal Of University Education*, 18(2), 526-543. doi:10.24191/ajue.v18i2.18159
- Yudha, R., Hilal, I., & Suyanto, E. (2019). Hubungan Memahami Kosakata dengan Menulis Teks Pidato Siswa SMA Plus Muhammadiyah Natar. *Jurnal Kata: Bahasa, Sastra, dan Pembelajarannya*, 7(1 Apr).
- Y Shengran. (2014). 政府文件语篇分析综述. *Oriental Enterprise Culture* (09),247-248. doi:CNKI:SUN:DFQY.0.2014-09-232.
- Y Jingan & H Jianing (2016). A Comparative Study of General Academic Word Lists: Criteria of Word Selection, List Evaluation and Rank Ordering. *Shandong Foreign Language Teaching*. Dec. 2016, Vol.37, No.6 DOI: 10. 16482 /j. sdwy37 — 1026. 2016 — 06 — 007
- Y Yani, NS Idris, Y Mulyati. (2019). Kajian Bahan Ajar BIPA Guna Pemenuhan Kebutuhan bagi Pembelajaran BIPA di Tiongkok. *Seminar Internasional Riksa Bahasa*. Retrieved from <http://proceedings.upi.edu/index.php/riksabahasa/article/view/863>
- Z Xianjun. (2014). Corpora-based Assessment of Difficulty of Reading Text in National Entrance Test of English for MA/MS Candidates. *Enrollment and Examination in Hubei*, 10, 44-49. 1008-004X(2014)10-044-06 <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=HJKS201405012&DbName=CJFN2014>

- Z Xiaobin & L Yali. (2012). Research into Vocabularies of Chinese Elementary Intensive Reading Textbooks. *Language Teaching and Linguistic Studies*, 5, 26-33
<https://kns.cnki.net/kcms/detail/detail.aspx?FileName=YYJX201205006&DbName=CJFQ2012>
- Z Gangqin & Z Wei (2010). *Bahasa Indonesia Tingkat Dasar untuk Kuliah Intensif*. World Publishing Corporation: Guangzhou, ISBN 978-7-5192-9766-4

About the author

Lu Yingxuan is a doctoral student at the Faculty of Language and Literature Education in the Indonesian Education University. Her primary research focus is on teaching Indonesian as a foreign language. Email: luyingxuan.9321@gmail.com

Vismaia S. Damayanti is a professor at the Indonesian Education University, with her primary research focus on teaching Indonesian as a foreign language. Email: vismaia@upi.edu

Andoyo Sastromiharjo is the head of the Indonesian Education Department at the Faculty of Language and Literature Education, Indonesian Education University. His primary research focus is on Indonesian language teaching. Email: andoyo@upi.edu

Yeti Mulyati is a professor at the Indonesian Education University, with her primary research focus on teaching Indonesian as a foreign language. Email: yetimulyati@upi.edu