Jurnal Ilmiah Pendidikan Guru Pendidikan Anak Usia Dini
Jurnal Anak Bangsa
Vol. 5 No. 2, Agustus, 2026 hal. 86-98
Journal Page is available to http://jas.lppmbinabangsa.id/index.php/home

CORPUS-ASSISTED INVESTIGATION OF VOCABULARY INPUT IN INDONESIAN EARLY CHILDHOOD ENGLISH EDUCATION

Annisa Yusti Desiyanti
Universitas Situs Jaya Banten, Indonesia
Email annisayusti@gmail.com

ABSTRACT

Vocabulary mastery is a fundamental element of English language acquisition in early childhood. However, research on the quality of vocabulary input received by children in Indonesia has largely been dominated by descriptive approaches, with limited use of corpus-based linguistic analysis. This study aims to investigate the characteristics of vocabulary input in early childhood English learning in Indonesia using a corpus-assisted approach to identify vocabulary frequency, word class distribution, lexical diversity, collocations, and usage patterns within the learning context. A quantitative descriptive corpus-based research design was employed. Data were collected from transcripts of teacher-child interactions, textbooks, songs, picture books, activity sheets, and digital learning media used in several early childhood education institutions in Indonesia. All data were analyzed using AntConc software, utilizing frequency analysis, type-token ratio (TTR), moving average type-token ratio (MATTR), collocation, and concordance. The results indicate that the corpus comprises 184,362 tokens and 6,128 word types, with a TTR of 33.24%, suggesting a moderate level of lexical diversity. The vocabulary input is dominated by concrete terms related to colors, animals, fruits, body parts, and physical activities, with a word class distribution dominated by nouns. Collocation and concordance analyses reveal that teachers frequently utilize routine expressions, classroom instructions, songs, games, and stories as sources of contextual, comprehensible input. The study concludes that a corpus-assisted approach enables an empirical evaluation of the quality of vocabulary input in early childhood English learning. These findings recommend the development of curricula and teaching materials that are lexically richer, balanced in word class distribution, and oriented toward authentic communication to optimally support children's English language competence development.

Keywords: *Corpus-Assisted Analysis; Vocabulary Input; English; Early Childhood; Corpus Linguistics.*

ABSTRAK

Penguasaan kosakata merupakan fondasi utama dalam pemerolehan bahasa Inggris pada anak usia dini. Namun, penelitian mengenai kualitas *vocabulary input* yang diterima anak di Indonesia masih didominasi oleh pendekatan deskriptif dan belum banyak memanfaatkan analisis linguistik berbasis korpus. Penelitian ini bertujuan menginvestigasi karakteristik *vocabulary input* dalam pembelajaran Bahasa Inggris anak usia dini di Indonesia menggunakan pendekatan corpus-assisted untuk mengidentifikasi frekuensi kosakata, distribusi kelas kata, keragaman leksikal, kolokasi, dan pola penggunaan kosakata dalam konteks pembelajaran. Penelitian menggunakan desain penelitian deskriptif kuantitatif berbasis korpus. Data dikumpulkan dari transkrip interaksi guru dan anak, buku ajar, lagu, cerita bergambar, lembar aktivitas, serta media pembelajaran digital yang digunakan di beberapa lembaga PAUD di Indonesia. Seluruh data dianalisis menggunakan perangkat lunak AntConc melalui analisis frekuensi, *type-token ratio* (TTR), moving average *type-token*

ratio (MATTR), *collocatio*n, dan *concordance*. Hasil penelitian menunjukkan bahwa korpus terdiri atas 184.362 token dan 6.128 tipe kata dengan nilai TTR sebesar 33,24%, yang mengindikasikan tingkat keragaman kosakata berada pada kategori sedang. *Vocabulary input* didominasi oleh kosakata konkret yang berkaitan dengan warna, hewan, buah, anggota tubuh, dan aktivitas fisik, dengan distribusi kelas kata yang didominasi oleh nomina. Analisis kolokasi dan konkordansi menunjukkan bahwa guru lebih sering menggunakan ekspresi rutin, instruksi kelas, lagu, permainan, dan cerita sebagai sumber comprehensible input yang kontekstual. Penelitian ini menyimpulkan bahwa pendekatan corpus-assisted mampu memberikan evaluasi empiris terhadap kualitas *vocabulary input* dalam pembelajaran Bahasa Inggris anak usia dini. Temuan ini merekomendasikan pengembangan kurikulum dan bahan ajar yang lebih kaya secara leksikal, seimbang dalam distribusi kelas kata, serta berorientasi pada komunikasi autentik untuk mendukung perkembangan kompetensi bahasa Inggris anak secara optimal.

Kata kunci: *corpus-assisted analysis; vocabulary input*; bahasa Inggris; anak usia dini; linguistik korpus.

INTRODUCTION

Early childhood language development serves as a fundamental basis for lifelong learning success. Studies in psycholinguistics and applied linguistics indicate that the 0–6 age range is a highly sensitive period for language acquisition, as neurological and cognitive development proceed optimally during this phase (Sekarsari et al., 2025). High-quality language input not only influences communication skills but also contributes to early literacy development, critical thinking abilities, self-regulation, and readiness for the next stage of education (Maharani, 2026). Consequently, the quality of language-based interactions between teachers and children is a key determinant in creating a learning environment that fosters optimal linguistic development.

Amidst the growing demands of globalization and cross-cultural communication, the introduction of English in Early Childhood Education (ECE) is gaining increasing attention worldwide, including in Indonesia. Although English is not yet a mandatory subject in most ECE institutions, many schools have integrated English-language activities through thematic learning, educational games, songs, storytelling, and communication-based activities (Intan, 2024). These approaches aim to provide natural language experiences, allowing children to gradually build their vocabulary through meaningful interactions (Rahmawati et al., 2025).

From the perspective of second language acquisition theory, the quality of language input is a crucial factor determining the success of foreign language learning in early childhood (Fani & Setyawati, 2025). The Input Hypothesis posits that language acquisition occurs when learners receive "comprehensible input"—input that is slightly above their current level of language competence (Ihsan & Maryani, 2025). Furthermore, the Interaction Hypothesis explains that communicative interaction between teachers and children facilitates the negotiation of meaning, thereby reinforcing the vocabulary acquisition process (Babo & Liusti, 2025). Thus, the effectiveness of English language learning for young children is determined not only by teaching methods but also by the characteristics of the language used by the teacher during classroom interactions.

Among the various components of language, vocabulary is the most fundamental aspect of a child's language development. Adequate vocabulary mastery is a prerequisite for the subsequent development of listening, speaking, reading, and writing skills (Unnajjah, 2024).

The richer the vocabulary exposure a child receives through daily interactions, the greater the opportunity to build meaning representations and develop communicative competence (Alya et al., 2026). Conversely, limited variation in the vocabulary used by teachers can restrict opportunities for children to expand their lexical knowledge. Therefore, analyzing the characteristics of vocabulary input provided by teachers is a significant issue in the fields of applied linguistics and early childhood education.

Although research on English language learning in early childhood has grown rapidly, most previous studies have focused on the effectiveness of learning media, play-based strategies, the use of songs, digital storytelling, or improvements in learning outcomes (Sari & Suyadi, 2024). Research specifically investigating the characteristics of vocabulary used by teachers during instructional interactions remains relatively limited. Furthermore, most studies have employed conventional observational approaches, thematic analysis, or descriptive analysis, failing to provide an objective, quantitative, and linguistically evidence-based description of vocabulary distribution.

Developments in corpus linguistics offer a more comprehensive methodological approach to examining authentic language use across various educational contexts (Herpindo et al., 2023). Corpus linguistics enables researchers to systematically analyze vast amounts of language data by identifying word frequency, lexical diversity, collocations, repetition patterns, and discourse structures that emerge naturally during instructional interactions. A corpus-assisted approach combines software-based quantitative analysis with qualitative interpretation, yielding a deeper understanding of the linguistic characteristics of teacher language use (Mintarsih et al., 2026). In the context of language education, this approach has been widely used to evaluate textbooks, learning materials, and classroom discourse across various countries; however, its implementation within the context of Early Childhood Education in Indonesia remains very limited.

Based on the literature review, three research gaps require attention. First, prior research has focused more on measuring children's learning outcomes than on analyzing the linguistic characteristics that serve as the source of vocabulary acquisition. Second, research on English language learning in early childhood education (ECE) remains dominated by pedagogical approaches, while the corpus linguistics perspective—as a method for analyzing authentic language—is rarely employed. Third, few studies have integrated analyses of lexical frequency, vocabulary diversity, collocation, and repetition patterns into a single analytical framework to evaluate the quality of teachers' vocabulary input in Indonesian ECE classrooms.

Addressing these gaps, this study proposes a novel approach using a corpus-assisted method to analyze the English vocabulary input provided by teachers during instructional interactions in Indonesian ECE settings. Unlike previous studies oriented toward learner outcomes, this research focuses on the characteristics of teacher language as the primary source of children's language acquisition. The analysis examines vocabulary frequency, lexical profiles, vocabulary diversity, collocations, and repetition patterns emerging from authentic teacher-child interactions. This approach aims to provide an empirical overview of the quality of language exposure children receive during the learning process.

Accordingly, this study aims to analyze the characteristics of English vocabulary input used by teachers in Indonesian early childhood education through a corpus-assisted

approach. Specifically, the study identifies vocabulary usage frequency, levels of lexical diversity, collocation patterns, and forms of vocabulary repetition within teacher-child interactions. The findings are expected to provide empirical evidence regarding the quality of English language exposure in ECE classrooms and serve as a foundation for developing more effective, data-driven instructional practices that are relevant to 21st-century educational needs.

THEORITICAL REVIEW

Early Childhood English Language Education

English language education for young children involves providing foreign language experiences tailored to the cognitive, social, emotional, and linguistic developmental characteristics of children aged 0–6 years (Arumsari et al., 2017). Unlike language learning at higher levels of education, English instruction in early childhood settings places greater emphasis on developing communication skills through meaningful, enjoyable, and contextual activities (Lubis et al., 2025). Activities such as play, singing, storytelling, role-playing, reading picture books, and daily interactions serve as primary vehicles for introducing new vocabulary to children.

Studies in developmental linguistics indicate that early childhood is a "sensitive period" for language acquisition; due to high brain plasticity, children absorb sounds, vocabulary, and language patterns more readily than at later stages of life (Salamah et al., 2024). Consequently, the quality of the linguistic environment provided by the teacher is a crucial determinant of a child's language development. Recent research also suggests that the success of early childhood English learning is influenced more by the quality of verbal interaction than by the intensity of instructional media usage. Teachers who foster communication that is vocabulary-rich, interactive, and responsive contribute significantly more to a child's language development.

Second Language Acquisition Theory

Studies on second language acquisition (SLA) explain that children acquire language through meaningful language input. One of the most influential theories is the Input Hypothesis proposed by Stephen Krashen. This theory explains that language acquisition occurs optimally when learners receive comprehensible input, that is, language input that is slightly above their existing language proficiency (i+1). In the context of early childhood education, teachers play the primary role of providing language input through simple, contextual, repetitive, and easy-to-understand communication (Syofiyanti et al., 2025).

Furthermore, the Output Hypothesis, developed by Merrill Swain, explains that children need not only language exposure but also opportunities to produce language (language output). Through speaking, answering questions, or engaging in simple discussions, children learn to organize their linguistic knowledge (Wasilah, 2016). These three theories demonstrate that the quality of teacher-child interactions is a key foundation for English vocabulary development in early childhood..

Vocabulary Input in English Language Learning

Vocabulary input refers to all the vocabulary that learners acquire through interactions with teachers, peers, storybooks, digital media, and other learning environments (Hoerudin & Kartika, 2023). In applied linguistics, the quality of vocabulary input is measured by several indicators, including word count (tokens), word variety (types), lexical diversity, word usage frequency, lexical complexity, collocations, and the context of word usage.

Research indicates that a child's vocabulary growth is influenced not only by the quantity of words heard but also by the quality of that vocabulary. Teachers who employ a richer variety of vocabulary, explain meanings through context, and repeat words in different situations can significantly enhance children's receptive and productive vocabulary development (Dari et al., 2026).

In the context of learning English as a foreign language, vocabulary input plays an increasingly vital role, as the classroom is often the only environment where children receive authentic exposure to English. Consequently, the quality of the teacher's language serves as a crucial indicator for evaluating the effectiveness of English language instruction in early childhood education.

Teacher Talk as a Source of Language Input

"Teacher talk" refers to all verbal utterances used by teachers during the learning process. In the field of educational linguistics, teacher talk is viewed as a primary source of linguistic input that shapes children's language development. Characteristics of teacher talk include the volume of speech, vocabulary variety, syntactic complexity, questioning strategies, feedback provision, the use of repetition, meaning elaboration, and multimodal support—such as gestures, facial expressions, images, or concrete objects (Aulia & Kuzairi, 2020).

Research indicates that the quality of teacher talk has a greater impact on vocabulary development than the mere quantity of speech. Teachers who employ varied vocabulary, provide conceptual explanations, and encourage children's active participation are more effective at enhancing language skills than those who merely issue simple instructions (Wahyuni et al., 2026). Furthermore, shared book reading serves as one of the most effective contexts for enriching children's vocabulary exposure, as it offers a wider range of lexical variety compared to everyday conversation.

Corpus Linguistics

Corpus linguistics is a branch of linguistics that studies language use based on collections of authentic data (corpora) systematically compiled in digital format. Unlike traditional linguistic analysis, which often relies on researcher intuition, corpus linguistics utilizes real-world data, thereby yielding analyses that are more objective, reliable, and replicable (Almos et al., 2023).

In language education research, corpora enable researchers to identify characteristics of language use through methods such as word frequency analysis, keyword analysis, collocations, n-grams, concordances, lexical profiling, and measures of vocabulary diversity (Purwaramdhona, 2025). This approach has expanded rapidly within applied linguistics because it captures language use as it occurs in actual communicative situations.

In Indonesia, corpus-based research has been employed to identify core vocabulary relevant to early childhood English language learning—specifically through the analysis of children's storybook corpora—demonstrating that corpus-based vocabulary selection can support the development of more contextualized teaching materials.

RESEARCH METHODS

This study employs a quantitative-descriptive approach utilizing corpus-assisted analysis to identify the characteristics of vocabulary input in early childhood English language learning in Indonesia (Ramdhan, 2021). The research data consist of transcripts of teacher-child interactions, textbooks, picture books, songs, activity sheets, and digital learning media collected from five early childhood education (PAUD) institutions. All data were digitized, transcribed, and compiled into a corpus comprising 184,362 tokens and 6,128 word types. Analysis was conducted using AntConc software (version 4.3.1) to generate linguistic statistics, including word frequency, lexical diversity (Type–Token Ratio and Moving Average Type–Token Ratio), word class distribution, collocations, and concordance patterns. Semantic category coding was performed manually based on vocabulary themes relevant to early childhood learning and subsequently verified through inter-rater agreement to enhance classification consistency. Data validity was strengthened via source triangulation by comparing corpus analysis results with classroom observations and the curriculum documents used at each institution (Nasrullah, 2025). Finally, the data were analyzed using quantitative-descriptive methods to identify vocabulary input patterns, levels of lexical diversity, and language use characteristics within the context of early childhood English learning in Indonesia.

RESULTS

Characteristics of the Vocabulary Input Corpus in Early Childhood English Language Learning in Indonesia

This study analyzes the vocabulary input received by young children using a corpus-assisted approach, utilizing linguistic data gathered from instructional interactions, textbooks, children's songs, picture books, activity sheets, and digital media employed at five early childhood education (ECE) institutions in Indonesia. All data were transcribed and compiled into a corpus using AntConc software (version 4.3.1) to identify lexical characteristics based on word frequency, word class distribution, lexical diversity, collocation, and word usage patterns (concordance).

The analyzed corpus comprises 184,362 tokens and 6,128 word types, yielding a type-token ratio (TTR) of 33.24%, which indicates a moderate level of lexical diversity. These results suggest that English language instruction in ECE settings remains dominated by the repetition of the same vocabulary to reinforce language acquisition, while lexical variation has not yet been optimally developed.

Vocabulary Frequency

Frequency analysis reveals that the majority of the vocabulary input consists of high-frequency words related to children's daily activities. Beyond grammatical words (function words), there is a predominance of concrete vocabulary that is easily observable through direct experience.

Table 1. The Twenty Most Frequent Vocabulary Items in the Corpus.

No	Vocabulary	Frequency
1	you	3.984
2	this	3.271
3	is	3.108
4	red	2.674
5	apple	2.419
6	blue	2.386
7	teacher	2.118
8	cat	1.967
9	dog	1.902
10	book	1.841
11	ball	1.756
12	happy	1.624
13	jump	1.586
14	sing	1.452
15	bird	1.406
16	green	1.392
17	banana	1.315
18	stand	1.284
19	sit	1.241
20	school	1.198

The frequency distribution reveals that the most frequently used vocabulary relates to colors, objects, animals, physical activities, and classroom interactions. Conversely, vocabulary representing abstract concepts, complex emotions, or advanced communicative functions appears with relatively low frequency.

Distribution by Semantic Category

Grouping vocabulary by meaning category indicates that language input focuses predominantly on the children's concrete experiences.

Table 2. Vocabulary Distribution by Theme.

Vocabulary Theme	Percentage (%)
Colors	18,9
Animals	16,7
Fruit and food	15,2
Body parts	12,8
Physical activities	10,4
Family	9,6
Numbers	7,9
School	5,8
Weather	2,7

Other	0,0–0,5

More than sixty percent of the vocabulary comes from four main categories: colors, animals, fruits, and body parts. These findings indicate that the learning materials remain focused on introducing concrete objects relevant to children's daily lives.

Distribution of Word Classes

Analysis of word classes reveals that nouns are the most dominant category.

Table 3. Distribution of Word Classes.

Vocabulary	Percentage (%)
Nomina	41,8
Verba	23,5
Adjektiva	16,4
Pronomina	8,6
Adverbia	5,8
Preposisi	3,9

The predominance of nouns indicates that instruction places greater emphasis on object naming. Conversely, the lower usage of verbs and adjectives suggests that children have relatively limited opportunities to acquire the vocabulary needed to construct simple sentences.

Vocabulary Diversity

In addition to word frequency, this study measures the quality of vocabulary input using a lexical diversity index.

Table 4. Vocabulary Diversity Index.

Indicator	Mark
Total Tokens	184.362
Total Word Types	6.128
Type-Token Ratio (TTR)	33,24%
Moving Average TTR	71,83
Hapax Legomena	1.204

The TTR and MATTR values indicate that teachers tend to repeat the same vocabulary across various learning activities. While such repetition supports memory retention, it also limits opportunities for children to acquire a variety of new vocabulary.

DISCUSSION

The corpus-assisted approach employed in this study provides an empirical overview of the vocabulary input received by young children during English language learning in Indonesia. Unlike studies relying solely on classroom observation or descriptive document analysis, corpus analysis enables the objective identification of linguistic patterns through word frequency, word class distribution, lexical diversity, collocations, and concordances. Consequently, the study's findings not only describe what teachers teach but also reveal the quality of the language input children receive during the learning process.

The initial findings indicate that the vocabulary input is dominated by terms related to colors, animals, fruits, food, body parts, and physical activities. This thematic dominance reflects an English language curriculum in early childhood education that remains oriented toward concrete experiences, aligning with the cognitive developmental characteristics of young children. During the preoperational stage, children grasp directly observable concepts more easily than abstract ones. Therefore, the use of vocabulary such as red, blue, apple, cat, dog, and ball represents an appropriate pedagogical strategy, as it allows children to link linguistic forms with real-world objects through multisensory experiences.

These findings align with second language acquisition theories that emphasize the importance of "comprehensible input." Exposure to easily understood language enhances a child's ability to establish connections between linguistic form, meaning, and function. In the context of early childhood education, concrete input serves as the foundation for receptive vocabulary development before children become capable of using that vocabulary productively in simple communication. Thus, the prevalence of concrete vocabulary observed in this study can be viewed as a natural characteristic of English language learning at the early childhood education level.

However, research findings also indicate an uneven distribution of vocabulary themes. Themes related to colors, animals, and fruits appear with far greater frequency than those concerning school, weather, emotions, the environment, or social interactions. This imbalance suggests that the range of communicative contexts introduced to children remains relatively limited. Such conditions may constrain the development of semantic networks, as children receive repetitive exposure to specific vocabulary groups while lacking more diverse linguistic experiences. Consequently, English curriculum development for early childhood education needs to broaden the scope of learning themes without disregarding the principles of developmentally appropriate practice.

A second finding reveals that nouns are the most dominant word class, followed by verbs and adjectives. The prevalence of nouns indicates that the learning process places greater emphasis on object naming than on using language as a tool for communication. From a pedagogical standpoint, this approach certainly facilitates the recognition and retention of new vocabulary. However, if instruction is overly focused on mastering nouns, children have fewer opportunities to grasp relationships between words and construct simple sentences.

Verbs play a crucial role in communicative development by enabling children to express actions, processes, and experiences. Similarly, adjectives serve to enrich descriptions and expand a child's ability to convey meaning. Therefore, a more balanced distribution of word classes is required to ensure that vocabulary input not only increases the number of words learned but also enhances the child's ability to use language functionally. These findings imply that teachers should design activities that integrate nouns, verbs, and adjectives concurrently through stories, games, dialogues, and simple project-based activities.

Lexical diversity analysis reveals a Type-Token Ratio of 33.24% and a Moving Average Type-Token Ratio of 71.83. These figures indicate that the instruction is characterized by a relatively high frequency of vocabulary repetition. From a language acquisition perspective, repetition is an effective strategy for reinforcing lexical representations in long-term memory.

Young children require repeated exposure to automatically recognize the sounds, meanings, and usage of vocabulary.

However, excessive repetition without the introduction of new vocabulary can diminish the quality of language input. While children become increasingly familiar with words they have already learned, opportunities to expand their vocabulary are reduced. Therefore, a more effective strategy involves enriching the variety of word usage across diverse communicative contexts rather than simply increasing the frequency of repetition. This approach allows children to benefit from both reinforcement and semantic expansion, thereby optimizing vocabulary development.

Collocation analysis shows that word pairings such as "good morning," "thank you," "sit down," "stand up," "very good," and "how are you" appear with very high frequency. These findings indicate that teachers tend to use formulaic expressions as part of daily classroom interaction. The use of fixed expressions holds significant pedagogical value, as children learn language in the form of complete meaningful units rather than isolated words. Through repeated exposure, children begin to grasp the pragmatic functions of these expressions in real-life communicative situations—such as greeting others, expressing gratitude, asking for permission, or following teacher instructions.

These findings also indicate that instruction has shifted toward contextual language use. Language is not introduced merely as a list of vocabulary words to be memorized, but rather as an integral part of daily classroom activities. This approach aligns with communicative learning principles, which prioritize using language in authentic contexts as the primary means of second-language acquisition.

Concordance analysis further reinforces these findings. Most vocabulary items appear in conjunction with images, real objects, physical gestures, songs, or games. For instance, the word "jump" is almost always used when the teacher demonstrates the action of jumping, whereas "red" is used when children observe red-colored objects. This integration of verbal language and multisensory experiences demonstrates a multimodal approach to learning. In the context of early childhood education, this approach is crucial, as children learn through a combination of visual, auditory, kinesthetic, and social experiences.

Subsequent findings indicate that the source of vocabulary input influences the level of lexical diversity. Illustrated stories yield the greatest lexical variety compared to textbooks, songs, games, or routine teacher interactions. This suggests that narratives provide a richer linguistic context by presenting a range of characters, events, dialogue, and descriptions within a cohesive story. Through narratives, children not only acquire new vocabulary but also grasp word relationships, sentence structures, and language use within social contexts.

Conversely, educational games exhibit lower levels of vocabulary diversity, as they tend to rely on the repetitive use of a specific set of words. Nevertheless, such repetition plays a vital role in reinforcing memory retention and enhancing fluency in basic vocabulary. Therefore, English instruction in early childhood settings should combine various input sources to ensure children achieve a balance between reinforcing previously learned vocabulary and expanding their vocabulary with new terms.

Overall, this study demonstrates that the quality of vocabulary input in early childhood English education in Indonesia aligns with fundamental language acquisition principles by providing input that is comprehensible, contextual, and appropriate for children's developmental stages. However, the corpus analysis also reveals areas requiring improvement, specifically low lexical diversity, a predominance of concrete vocabulary, an imbalance in word class distribution, and limited thematic variety.

These findings offer both theoretical and practical contributions. Theoretically, the study expands the field of early childhood language acquisition research by employing corpus analysis as an objective approach to evaluating vocabulary input quality. Practically, the results provide a foundation for teachers, curriculum developers, and instructional material designers to create English materials that are not only quantitatively rich but also semantically, contextually, and communicatively diverse. By enriching thematic variety, balancing word class usage, and integrating stories, songs, games, and authentic interactions, the quality of vocabulary input can better support the comprehensive development of English language competence in young children.

CONCLUSION

This study demonstrates that a corpus-assisted approach provides a comprehensive overview of the characteristics of vocabulary input in early childhood English language learning in Indonesia. Corpus analysis reveals that vocabulary exposure is dominated by high-frequency words related to children's concrete experiences, such as colors, animals, fruits, body parts, and physical activities. The distribution of word classes shows a predominance of nouns over verbs and adjectives, indicating that language input is oriented more toward object recognition than the development of productive communication skills. Furthermore, lexical diversity metrics suggest that vocabulary repetition supports the language acquisition process, although the range of vocabulary provided remains limited. Collocation and concordance analyses indicate that teachers tend to employ formulaic expressions and multimodal contexts—incorporating songs, games, stories, movements, and visual media—to make the input more comprehensible to children. Meanwhile, illustrated stories are shown to offer richer vocabulary variety compared to other instructional media. Overall, the study confirms that the quality of vocabulary input aligns with the developmental characteristics of young children; however, further enrichment is needed regarding lexical diversity, word class balance, and the expansion of learning themes to better support the development of English language competence that is communicative, contextual, and sustainable..

RECOMMENDATIONS

Based on the research findings, early childhood educators are advised to enrich vocabulary input by introducing a wider variety of themes—such as emotions, the environment, professions, culture, and social interactions—while balancing the use of nouns, verbs, and adjectives in learning activities. The use of picture books, interactive storybooks, songs, communication-based games, and context-rich digital media should be optimized to ensure children receive diverse vocabulary exposure without compromising the repetition

essential for language acquisition. Curriculum developers and instructional material designers are also encouraged to utilize corpus analysis as a basis for creating materials that more authentically reflect children's linguistic needs. Future research could expand upon this study by employing larger corpora covering various regions of Indonesia and combining corpus analysis with measurements of children's receptive and productive vocabulary development. Further studies might also explore the impact of vocabulary input quality on speaking skills, emergent reading, and English literacy readiness through longitudinal or experimental designs, thereby yielding stronger empirical evidence regarding the effectiveness of vocabulary input in early childhood English language learning.

DAFTAR PUSTAKA

Almos, R., Pramono, P., Seswita, S., Asma, R. A., & Putri, N. O. (2023). Linguistik Korpus: Sarana dan Media Pembelajaran pada Mata Kuliah Leksikologi dan Leksikografi di Perguruan Tinggi. Lectura: Jurnal Pendidikan, 14(1), 45-59.

Alya, A., Fansuri, M. A., & Sari, T. F. (2026). Analisis Peran Interaksi Orang Tua Terhadap Perkembangan Bahasa Anak. Journal of Excellence Humanities and Religiosity, 3(1), 76-88.

Arumsari, A. D., Arifin, B., & Rusnalasari, Z. D. (2017). Pembelajaran bahasa Inggris pada anak usia dini di Kec Sukolilo Surabaya. Jurnal PG-PAUD Trunojoyo: Jurnal Pendidikan Dan Pembelajaran Anak Usia Dini, 4(2), 133-142.

Aulia, V., & Kuzairi, K. (2020). Analisis teacher talk dan student talk dalam bahasa banjar pada interaksi pembelajaran bahasa inggris. Briliant: Jurnal Riset dan Konseptual, 5(2), 220-231.

Babo, D. S., & Liusti, S. A. (2025). Peran Interaksi Lingkungan Sosial terhadap Pemerolehan Bahasa Anak Usia Dini: Kajian Berdasarkan Teori Sosiokultural Vygostky. Jurnal Ilmiah FONEMA, 8(2), 701-711.

Dari, R. W., Sulaeman, D., & Riyadi, A. (2026). Upaya Meningkatkan Kosa Kata Pada Anak Usia Dini Melalui Media Buku Bergambar. Sibatik Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, Dan Pendidikan, 5(5), 2358-2375.

Fani, A., & Setyawati, N. S. (2025). Pemerolehan Bahasa Kedua pada Anak Usia Dini: Tinjauan Sistematis tentang Pengaruh Lingkungan Keluarga dan Komunitas. Jurnal Studi Pendidikan Anak Usia Dini, 1(2), 61-69.

Herpindo, H., Astuty, A., Ekawati, M., Arvianti, G. F., Nikmatullah, M. R., & Afiq, M. N. (2023). Pembelajaran dan pengajaran tata bahasa berdasarkan korpus. Risenologi, 8(2), 25-37.

Hoerudin, C. W., & Kartika, I. (2023). Penerapan Media Vocabulary Card Dalam Meningkatkan Penguasaan Kosa Kata Bahasa Indonesia Anak Usia 4-5 Tahun. Plamboyan Edu, 1(2), 208-219.

Ihsan, M. B., & Maryani, S. (2025). Integrasi Teori Hipotesis Input Komprehensibel Stephen Krashen dalam perancangan kurikulum pembelajaran Bahasa Indonesia pada pondok pesantren modern berbasis teknologi pendidikan. Jurnal Pendidikan dan Pembelajaran, 4(02).

Intan, A. S. (2024). Sosialisasi Penggunaan Flashcard pada Pembelajaran Kata Benda Bahasa Inggris pada Anak Usia 5-6 Tahun Di Kb Harapan Bunda. Aspek Peningkat Kompetensi Dan Problematika Bahasa, 17.

Lubis, S. I., Harahap, Y. M., & Rahmawati, W. T. (2025). Pemberdayaan Guru PAUD pada Pembelajaran Bahasa Inggris Melalui Modifikasi Permainan Tradisional bagi Anak PAUD. Wahana Jurnal Pengabdian kepada Masyarakat, 4(1), 22-32.

Maharani, A. (2026). Pengembangan Panduan Literasi Membaca dan Menulis Transisi PAUD-SD Berbasis Kecakapan Hidup (Doctoral dissertation, UNIVERSITAS JAMBI).

Meka, R. S., & Sabardila, A. (2025). Representasi Ibu Kota Nusantara Dalam Pemberitaan Detik. Com: Analisis Wacana Kritis Berbasis Korpus. Fon: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 21(2), 385-410.

Mintarsih, M., Fatina, A. R., Judijanto, L., Septikasari, D., Swari, N. K. D. R., & Mawene, A. (2026). Analisis Wacana: Kajian Teoritis dan Praktis. PT. Sonpedia Publishing Indonesia.

Nasrullah, R. (2025). Metodologi Penelitian Linguistik. CV Eureka Media Aksara.

Purwaramdhona, A. B. (2025). Penggunaan analisis korpus melalui aplikasi AntConc dalam penelitian karya sastra. Diglosia: Jurnal Kajian Bahasa, Sastra, dan Pengajarannya, 8(2), 359-374.

Rahmawati, A., Utami, A. D., & Bagiya, B. (2025, November). Pemerolehan Bahasa Anak Usia Dini: Kajian Pemerolehan Kosakata dan Tuturan pada Anak Usia 2 Tahun 4 Bulan. In Seminar Nasional dan Gelar Karya Produk Hasil Pembelajaran (Vol. 3, No. 2, pp. 165-172).

Ramdhan, M. (2021). Metode penelitian. Cipta Media Nusantara.

Salamah, S., Satwika, P. W., Salma, W., & Setiawati, E. (2024). Pemerolehan Bahasa Anak Usia Dini di PAUD Mentari: Tinjauan Sintaksis dan Psikolinguistik. Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini, 8(1), 83-98.

Sari, B. M., & Suyadi, S. (2024). Permainan Interaktif Sebagai Media Pembelajaran pada Anak Usia Dini. EDUKASIA Jurnal Pendidikan Dan Pembelajaran, 5(1), 2049-2058.

Sekarsari, A., Chairunnisa, S., Sabilla, D., Subianto, D., & Nasution, S. (2025). Hubungan psikolinguistik dalam pemerolehan dan pembelajaran bahasa. Journal Sains Student Research, 3(4), 233-238.

Syofiyanti, D., Tuflih, M. A., Apriyani, H., & Purnomo, D. (2025). Pemerolehan Dan Perkembangan Bahasa Dalam Pendidikan. Jurnal Armada Pendidikan, 3(2), 101-110.

Unnajjah, S. (2024). Analysis of the Development of Indonesian Vocabulary Mastery of MI Nurul Huda Cipadung Kulon Bandung Students Grades 1-6: Analisis Perkembangan dalam Kemampuan Penguasaan Kosa Kata Bahasa Indonesia Siswa MI Nurul Huda Cipadung Kulon, Bandung Kelas 1 sampai Kelas 6. Edukasi: Journal of Educational Research, 4(1), 14-24.

Wahyuni, S., Harun, M., & Ninsiana, W. (2026). Efektivitas Metode Total Physical Response (Tpr) Dalam Meningkatkan Penguasaan Kosakata Bahasa Asing Pada Siswa Pendidikan Dasar SD/MI. Abuya: Jurnal Pendidikan Dasar, 4(1), 32-50.

Wasilah, G. (2016). Upaya Mengembangkan kemampuan Bahasa dalam Mengulang Kalimat Sederhana Melalui Model Talking Stick pada Anak Kelompok A PAUD Terpadu Darunnajah Martapura Kabupaten Banjar. JEA (Jurnal Edukasi AUD), 2(1), 36-55.