ENGLISH EDUCATION JOURNAL
(EEJ)
Journal Homepage: https://jurnal.usk.ac.id/EEJ/

Corpus Linguistics in Teaching Vocabulary for EFL Learners: A Study of Foreign Speakers

Dina Irmayanti Harahap[1*]
Yenita Uswar[2]
Nova Andriani[3]
Nining Sudarwita[3]
Enni Maysaroh[3]

[1]*English Education Department, Faculty of Social Science and Education, University of Potensi Utama, Medan, Indonesia*
[2]*Informatic Department, Faculty of Engineering and Computer Science, University of Potensi Utama, Medan, Indonesia*
[3]*Management of Islamic Education Study Program, Faculty of Education, STIT Hamzah Fansuri (HAFAS), Subussalam, Aceh, Indonesia*

ABSTRACT

The objective of this research was to investigate the vocabulary, terminology, and grammar utilized by EFL learners through the corpus linguistics tool, Sketch Engine. This study was conducted using a qualitative design at a private university in North Sumatera, involving 50 EFL learners. A mixed-methods approach was employed, combining quantitative methods for data collection and qualitative methods for in-depth analysis. The corpus-based approach was specifically implemented with first-semester students from the academic year 2022/2023. The research findings indicated a dominant mastery of vocabulary among EFL learners, primarily comprising nouns, verbs, adverbs, pronouns, and adjectives. Notably, 11 vocabulary items emerged as having the highest frequency of usage, highlighting learners' engagement with these terms. The analysis underscored the potential for effective and efficient vocabulary learning, emphasizing the integration of technology in EFL education to expose learners to diverse word types and variations in word formation within the Indonesian context. Furthermore, this study seeks to promote a greater emphasis on grammatical aspects in language pedagogy. It suggests that heightened focus on these components could enhance language acquisition and fluency. Ultimately, the research aspires to provide valuable insights for educators and curriculum designers, contributing to the development of more comprehensive and effective teaching methodologies in EFL settings, thus benefiting both learners and instructors alike.

Keywords: *lecturer-student interaction, online speaking class, perception on online learning.*

*Corresponding author, email: dinairmayanti4@gmail.com

Citation: Harahap, D. I., Uswar, Y., Andriani, N., Sudarwita, N., & Maysaroh, E. (2025). Corpus linguistics in teaching vocabulary for EFL learners: A study of foreign speakers. *English Education Journal, 16*(1), 24-41.

https://doi.org/10.24815/eej.v16i1.41762
Received 8 October 2024; Revised 30 December 2024; Accepted 18 January 2025
Available online 31 January 2025

1. INTRODUCTION

Teaching English to foreign language students has been a major challenge that has evolved throughout history in terms of pedagogical approaches, techniques, application methods and teaching aids. The development of technology has brought different and new knowledge to language teaching and learning. The use of technology to enhance language use has become a central issue in language pedagogy. In the field of EFL, educators have accepted that they are a key player, especially in preparing their students for success.

An Indonesian language (BI) study program geared towards foreigners is called Indonesian Language for Foreign Speakers (BIPA). Indonesian functions as a foreign language in this regard. The government can also use the BIPA program as a diplomatic instrument to improve Indonesia's standing internationally. Other nations can quickly learn about Indonesia and build bilateral or multilateral ties through the BIPA initiative. BIPA also accomplishes its goal of shedding light on Indonesia's many and rich cultures. At the moment, BIPA is being developed quickly. In 2021, 279 BIPA teachers from 204 educational institutions joined 10,730 BIPA students from 38 countries. It demonstrated the huge demand for Indonesian culture and language abroad. Additionally, it is evident in certain BIPA programs that are set up in. It can also be seen in some BIPA programs organized in Indonesian universities (Junpaitoon, 2017; Ningsih et al., 2021; Saddhono, 2015).

A teacher must be aware of planning, process, and evaluation phases as well as instructional strategies, media, and materials used in BIPA learning. The ability to speak, listen, read, and write in Indonesian with accuracy is one of the most crucial things to educate students. Language proficiency needs the mastery of a sufficient vocabulary to ensure the success of any message that is delivered. The level of someone's language depends on how well they know their vocabulary. A person cannot effectively employ whatever structure or function they have acquired to speak in a second language if they do not have a sufficient vocabulary, to be useful, vocabulary learning must be done in an efficient and relevant manner (Naji Meidani et al., 2022; Saddhono & Wahyono, 2019).

In order to support language skills, vocabulary is a crucial component of language development. Implicit learning is used to teach BIPA vocabulary. The four language skills—speaking, listening, reading, writing, and grammar—all contribute to vocabulary integration. Teaching vocabulary is crucial to improving BIPA students' proficiency in the Indonesian language. Students in the BIPA program have a good correlation between their vocabulary and their ability to learn Indonesian. A person with a larger vocabulary can definitely use Indonesian better to listen, speak, read and write (Presbitero, 2020).

Effective vocabulary acquisition is inherently communicative, serving as a tool for communication that addresses the speaking needs of students in the BIPA program (Uchihara & Clenton, 2020; Uchihara et al., 2019). Consequently, vocabulary is essential for teaching methodologies. As students learn vocabulary, they are also expected to remember and not lose it. Therefore, effective vocabulary teaching techniques are essential. This article explores vocabulary teaching principles and strategies aimed at helping students retain vocabulary and achieve their communication objectives within the BIPA program (Zou & Xie, 2018).

The purpose of this research is to examine the vocabulary, terminology, and grammar used by students enrolled in the BIPA program. The results will guide the enhancement of vocabulary instruction in the Indonesian language within the BIPA program, improving students' understanding of both the language and its texts. This

approach will not only assist BIPA students in successfully finishing their courses but will also support them in their daily communications in Indonesian.

The innovative aspect of this research lies in its approach to developing Indonesian language vocabulary instruction based on corpus linguistics data analysis derived from the BIPA program, utilizing the SKELL (Sketch Engine for Language Learning) application.

SKELL (Sketch Engine for Language Learning) application is a free search tool based on large text data (corpus). Users can quickly check their doubts about languages by using three different functions: (1) searching for words and phrases in their contexts, (2) generating collocational candidates and grammatical behaviour for a specific word, (3) using thesaurus providing a list of similar and related words to a particular word. SKELL (Sketch Engine for Language Learning) functions are available in 6 languages: Czech, English, Estonian, German, Italian, and Russian.

Researchers, instructors, lecturers, and students can all utilise SKELL (Sketch Engine for Language Learning), a web-based digital application on data corpora, to learn and take courses in processing the language corpora in the Indonesian archipelago. The vocabulary, terminology, and grammar used by BIPA program students at Potensi Utama University were obtained from the results of this application's analysis of Indonesian language data corpora. The application's results can be examined from a number of angles to characterise the proficiency of Potensi Utama University BIPA program students in Indonesian.

As a result of such integration, CL can help undergraduate EFL students to achieve better learning outcomes and become independent English learners. Therefore, this research targets EFL students to use their knowledge of English in academic classes where CL can be used and applied in multiple-tasks. Focusing on the experiences of individual students in the learning environment enriches understanding and deepens the research analysis of CL integration in EFL classrooms and especially in higher education. Therefore, the developed research questions are as follows:

1. How does EFL students implement CL in their L2?
2. How does CL integration affect EFL students in terms of L2 language learning?
3. What are the contextual factors that affect CL upon the EFL students?

2. LITERATURE REVIEW

2.1 Corpus Linguistics

Formally and functionally, a "corpus" is a collection of natural language examples that includes a few lines from written texts or recordings gathered for linguistic research. The texts are then arranged in a systematic manner (Pishghadam & Zabihi, 2012; Stefanowitsch, 2020; Zeldes, 2018). The corpus is deemed "natural" since the texts gathered are those that are created and utilised naturally, whether or not they are utilised as such. Textbooks, novels, magazines, newspapers, talk programs, interview transcripts, and more are examples of texts. A thorough set of guidelines and procedures for using corpora in language study, instruction, and learning is known as corpus linguistics (Callies, 2019; Reppen & Simpson-Vlach, 2019).

Corpus linguistics is also a field that focuses on procedures or methods of language learning (Gholaminejad & Sarab, 2020). Based on these definitions, it can be concluded

that corpus linguistics is the study of language using real-world or natural language examples recorded in a corpus. The inaugural speeches given by 50 BIPA students at Potensi Utama University in 2022 serve as the study's corpus. Because corpus linguistics possesses the qualities needed to accomplish the research goal, it is appropriate for use in this study.

Linguistic features in corpus linguistics are occurrences that refer to the occurrence of a word in a corpus or text (Crossley, 2020). It can be used to count grammatical, semantic, or other categorisation frequencies in addition to the frequency of a single word. Additionally, frequency might guide the researcher to more comprehensive observations. Frequency in corpus linguistics refers to the number of times a term appears in a corpus. Finding terms that appear frequently in a given corpus and then comparing and differentiating them from other words is possible with frequency analysis (Oakes, 2019). The quantity and diversity of commonly used language that students believe will help them comprehend Indonesian are examined in this study using frequency. Students will be able to pronounce and comprehend a list of important vocabulary words in this way. In addition, knowing frequently occurring words automatically indicates which language features are of primary importance to students in BIPA program.

Concordances are lists or sequences of occurrences of words, word parts, or word combinations found within a specific context, derived from a text corpus (Abbasi & Beltiukov, 2018). In corpus linguistics, a keyword is referred to as a research item, and these keywords can be presented in various formats. One of the most widely used techniques in corpus linguistics is the use of keywords in context (KWIC) (Safii et al., 2021). Concordance plays a vital role in corpus linguistics by enabling qualitative analysis of corpus data, allowing researchers to closely examine specific instances. Before modifying language or frequency, a compatibility assessment is typically necessary. To display concordances, specialized software is used to analyze a particular linguistic item within its context, considering the surrounding words that may appear to the left or right of the target word (Dunder & Pavlovski, 2018; Rădescu, 2021). By employing the concordance method, researchers can conduct qualitative analyses that involve studying the text's themes alongside the linguistic elements. This study adopts concordance to facilitate the qualitative analysis of existing data. The data analysis focuses on exploring the linguistic characteristics of a word by examining both the target word and its surrounding context. The concordance technique can also help identify a word's grammatical category, analyze different types of active and passive sentences, explore morphological processes affecting certain words, examine variations in word positioning within sentences, and more.

2.2 The Relationship between Corpus Linguistics and Language Learning

Numerous studies have explored the relationship between corpus linguistics and language learning from various perspectives. Al-Sulaiti and Atwell (2006) and Hizbullah et al. (2016), for instance, conducted a study on Arabic corpus linguistics that focused on Arabic language learning in Indonesia. Depicting the diverse contents of corpus linguistics, as well as its evolution and dynamics in Arabic studies globally, was its main objective. The researcher employed a corpus-linguistic technique in this study, which was thought to be the most thorough and representative. Basicly, it was capable of processing text in any language because it was outfitted with system tips. The file needs to be changed in this way before processing. For instance, documents created with Microsoft

Word typically end in.doc or.docx. These file formats need to be saved with a.txt extension after being converted using the UTF-8 encoding technique. Lastly, Wordsmith can be used to analyse the data. The creation of a corpus in Arabic and associated tools may be utilised for appropriate processing and analysis. It can also serve as the foundation for scholarly conversations and debates about the potential for creating an Arabic corpus model in Indonesia and advancing the study and acquisition of Arabic in higher education.

Isam and Abd Mutalib (2019) carried out another investigation on the use of corpus analysis as a substitute for grammar instruction and learning. They concentrated on the selection of corpus databases and dictionary software, material conversion procedures, linguistic data, dictionary software usage techniques, and language analysis instances. The data for the managed corpus language method, developed using AntConc software, was sourced from the corpus database of the Malaysian Institute of Language and Literature (Dewan Bahasa dan Pustaka). To complete the process, the researchers undertook several steps: First, students were instructed to download the AntConc software and access the data from the DBP (Institute of Language and Literature) corpus database (Anthony, 2024). Next, they were tasked with entering specific data (prepositions) into the AntConc application. Following that, the teacher encouraged the students to discuss interesting findings (dynamic characteristics) based on the morphological and syntactic behaviors of the analyzed prepositions. A key objective of corpus linguistics in this context was to foster student engagement and active participation in the teaching and learning process. Moreover, the technology implemented in this study was carefully designed for optimal use. Consequently, employing teaching and learning strategies rooted in corpus data analysis has the potential to enhance the study of linguistics and the Malay language in Malaysia, positively impacting student education development.

The development of the "Az-Ziro'ah" lexicon as an Arabic learning resource was another important aspect of optimizing corpus linguistics (Suryadarma & Fakhiroh, 2020). In this context, the researchers detailed how a corpus linguistics approach was utilized to create a bilingual dictionary focused on the agricultural sector. This process employed a qualitative descriptive method for lexicography and incorporated techniques for collecting corpus data using AntConc and Sketch Engine software. The workflow began with the gathering of agricultural vocabulary (including terminology related to technology and industry) from articles in both print and online media. Initial data collection was performed in .doc format, which was subsequently converted to plain text (.txt). The language data was then analyzed using Sketch Engine, with researchers conducting data cleaning while accounting for basic, derived, and compound words through the N-Grams tool in AntConc. This highlighted that corpus linguistics can effectively underpin the compilation of a bilingual dictionary in specialized fields. Furthermore, this methodology likely produced significant results in systematically organizing specific vocabulary. Key considerations in the dictionary's compilation included the selection of data sources, the processing mechanisms of the corpus data platform, and the vocabularies or lemmas included in the final draft. These components were crucial for the successful development of a dictionary based on corpus linguistics (Suryadarma & Zakaria, 2022).

A subsequent noteworthy study examined the application of corpus linguistics to identify high-frequency words in "Sahabatku Indonesia (BIPA 1)" (Wahyuningtyas & Kesuma, 2021). Utilizing a corpus linguistics approach, the researchers compiled a ranked list of words based on their frequency of use in the text. A qualitative methodology

guided the selection of data, analysis, and design of research materials. The process began with selecting data in .pdf format from the e-book, which was converted to .txt format using file conversion software available in the Apple Store. The converted .txt file was then uploaded to AntConc, where the word list function was employed to ascertain word frequencies. Findings revealed that the most frequently occurring word classes were not necessarily the most densely packed. For instance, function words such as “and,” “in,” “that,” and “so” significantly contributed to the list of high-frequency words. As a result, it was concluded that a list of high-frequency words may not always consist of technical terms directly related to the research topic. The identification of high-frequency words using corpus linguistics can be applied to various materials, including books, newspapers, and literature, allowing researchers to create a mini dictionary of specific words or terms for educational purposes.

Numerous studies examining corpus linguistics and education have affirmed their close interrelations and mutual advantages for analysis and research. Corpus linguistics can enhance and refine analyses by leveraging extensive data, significantly aiding educational planning that is informed by corpus linguistic analysis. The crucial component of such research is the corpus linguistic data analysis outcomes in the form of vocabulary, which can help students design a vocabulary learning process that is accessible, communicative, and practical for both daily life and academic pursuits.

3. RESEARCH METHODS

3.1 Participants

This study was conducted at Potensi Utama University in North Sumatera Province, Indonesia, with 50 BIPA students participating in 2022. They were students from the first semester of the BIPA course with different linguistic and cultural backgrounds.

3.2 Instruments

The vocabulary of students in the BIPA program at Potensi Utama University has been converted to plain text for processing with the Sketch Engine application. Utilizing the frequency characteristics of this tool, the words employed in the introductory discourse were compiled in a comprehensive list, organized by frequency from the most common to those occurring only once or twice. The software-generated list underwent a reduction process to differentiate complete words from other entries like abbreviations, syllable fragments, numbers, etc.

3.3 Procedures

A mixed-method approach was adopted during the initial phase of the study. A quantitative method was employed to gather research material, while a qualitative method was applied during the analytical process. Initially, quantitative data was acquired using the qualitative approach with SKELL (Sketch Engine for Language Learning). This application allows students and teachers of English to easily verify how actual English speakers use certain phrases or words without any need for registration or payment; users simply type in a word and click a button.

3.4 Data Collection

The data collection for this study was an introductory discourse written by 50 students about students in the BIPA program. The research data were words, which were then categorized according to the purpose of the analysis. Words that had a complete meaning were then categorized according to their function in the sentence.

3.5 Data Analysis

After the data were obtained, the words were qualitatively analyzed to conclude the linguistic features specific to each word class. After the analysis, the ability level of the students of the BIPA program was found, after which principles and techniques of vocabulary learning were prepared to improve the ability of the students.

4. RESULTS AND DISCUSSION

4.1 Results

In the initial stage, the introductory discourse of students in the BIPA program at Potensi Utama University was digitized into a standard Word file to facilitate processing using a word processor. The software analyzed the document utilizing a feature dictionary, which generated a categorized list of words from the conversation, organized by word type in Indonesian. This categorized information is summarized in Table 1.

Table 1. Word types category.

No.	*Word Types*	*Frequencies*
1.	Noun	741
2.	Verb	469
3.	Adverb	351
4.	Pronoun	237
5.	Adjective	156
6.	Number	98
7.	Others	61
8.	Preposition	48
9.	Interjection	37
10.	Conjunction	34
11.	Article	12
	Total	**2.234**

#Word Types: 2,234
#Word Tokens: 9,664

The results of the corpus linguistics analysis using the Sketch Engine indicate that nouns are the most frequently used words among BIPA students, with over 741 occurrences. This is followed by verbs at 469 occurrences, adverbs at 351, pronouns at 237, and adjectives at 156. These findings suggest that BIPA students demonstrate a stronger mastery of nouns compared to other parts of speech, highlighting the importance of nouns in their language acquisition and usage. The predominance of nouns in their

discourse indicates not only familiarity with concrete subjects but may also reflect the types of activities and discussions they engage in during their studies.

A noun is the name of all things and of everything that can be objectified, and it is divided according to form into (1) a concrete noun, which is the name of things perceived by the five senses, and (2) an abstract noun, which is the name of things that the five senses cannot endure. Noun properties are some words that can be clarified or expanded by adding the adjective yang or yang sangat after the word. Below are examples of two nouns.

Concrete Noun:
[1] rumah [house]: rumah yang besar [a big house]
[2] batu [stone]: batu yang kecil [a small stone] Abstract Noun
[3] keagungan [great]: ke-an + agung [great]
[4] kekuatan [strength]: ke-an + kuat [strong]

In a sentence, a noun can be a subject (S) and an object (O). For example, "Ku-harus membeli kopi". The phrases "Ku-harus" and "kopi" are nouns. According to the formation, Indonesian nouns are divided into basic and derived nouns. A base noun is an original noun (both concrete and abstract) that is not suffixed. A derivative is a noun that is given a specific suffix.

Basic Noun:
[5] Adik [little brother/sister]
[6] Mobil [car] Derivative Noun
[7] makanan [food]: makan [eat] + an
[8] perumahan [housing]: pe-an + rumah [house]

The above examples are the classification of corpus linguistic data by word type in Indonesian. The results of the SKELL application also identify words that are mostly used by students of the BIPA program. The frequency of occurrence according to corpus linguistics (number of occurrences) is given in Table 2.

Table 2. Words list of frequency.

No.	*Word Data*	*Frequencies*
1.	saya	530
	Saya	406
	saya,	147
	(saya	1
	saya!	1
	Total	**1085**
2.	di	583
	Di	128
	Total	**711**
3.	dan	430
	Total	**430**
4.	suka	154
	Suka	1
	Total	**155**

No.	*Word Data*	*Frequencies*
5.	adalah	144
	adalah:	1
	Total	**145**
6.	kota	105
	Kota	19
	kota,	1
	kota-kota	1
	kota.	2
	Total	**121**
7.	dekat	119
	dekat	1
	Dekat	1
	Total	**121**
8.	belajar	112
	belajar.	4
	Belajar	1
	belajar,	2
	Total	**119**
9.	dari	114
	Total	**114**
10.	ke	106
	Ke	1
	Total	**107**
11.	Yang	104
	yang	2
	Total	**106**

In the introductory discourse of students in the Potensi Utama University BIPA program, a total of 2,234 word types and 9,664 word tokens were identified. Word types refer to the different kinds of words used, while word tokens reflect the total number of words, including repetitions. Notably, the most frequently occurring word in this discourse is "saya," which appears in five variations: "saya," "Saya," "saya," "saya," and "(saya!" The only context where the word "Saya" does not appear is at the end of a phrase; however, even in such cases it is present as an interjection ("saya!").

The prevalence of the word "saya" in the students' discourse is quite striking, particularly compared to the lone occurrence of "aku," which is synonymous with "saya." This linguistic pattern suggests that BIPA program instructors effectively conveyed the distinctions between "saya" and "aku" to their students, emphasizing their appropriate contexts and meanings in the Indonesian language. The emphasis on "saya" over "aku" may also reflect an inclination towards politeness and respect in formal discourse, which aligns with the cultural norms of communication in Indonesian. Overall, the data indicates a strong grasp of vocabulary usage among BIPA students, particularly in understanding the nuances of first-person pronouns in various contexts.

According to *Kamus Besar Bahasa Indonesia* or The Great Dictionary of Indonesian Language, the term "aku" means a first-person pronoun who speaks or writes (in a familiar/friendly manner); to say to feel; Saya and the term "saya" mean a pronoun or one who speaks or writes (formally or casually); battery Based on the explanation, we

can explain that the terms “aku” and “saya” are both first person pronouns (Badan Pengembangan dan Pembinaan Bahasa, n.d.). Examples are given in the following sentences:

> [10] Ilham berkata, “Saya belajar sejarah di Potensi Utama University”. [Danny said, “I study history at Potensi Utama University.”]
> [11] Danny berkata, “Aku belajar sejarah di Potensi Utama University”. [Danny said, “I study history at Potensi Utama University”.]

The term “saya” in sentence [10] has exactly the same meaning as the term “Aku” in sentence [11]. The terms “saya” and “aku” in both sentences are equally related to the speaker, namely: Danny.

When we understand the similarity between the terms “saya” and “aku,” we can also recognize the differences in their usage. Some may argue that the term “saya” is more accurate and preferable to “aku.” However, this statement is misleading. Neither term is necessarily more accurate, preferable, or incorrect; both have their appropriate contexts. This distinction applies not just to Indonesian, but to languages in general, as each word carries its own nuances depending on the context of use.

If the term “aku” were considered inappropriate, it is unlikely that it would still exist in the language; if a word is unnecessary, it will naturally fall out of usage. Both “saya” and “aku” are correct and polite when used appropriately. The term “saya” is a more polite first-person pronoun, used to express self-respect and respect for the interlocutor. It may be employed by someone who sees themselves as lower in status than the listener, such as a subordinate addressing a boss. Conversely, a superior may also choose to use “saya” to demonstrate respect towards their subordinate.

On the other hand, “aku” is used to express a sense of intimacy between speakers. Thus, close friends, regardless of age differences, can comfortably refer to themselves as “aku.” The term “saya” is generally more polite and respectful to both oneself and the interlocutor; therefore, using “saya” is often preferable in most contexts. In a formal setting, “saya” is the standard, while “aku” is more commonly used in casual conversation. The clear distinction between “aku” and “saya” lies in their connotation—where “saya” tends to convey humility, “aku” may suggest self-assertion, depending on the context. The term “saya” is etymologically linked to “hamba sahaya,” which translates to “subordinate.”

Effective vocabulary teaching principles must be adhered to so that vocabulary learning is integrated with other elements of the language (Barclay & Schmitt, 2019; Charkazova, 2018). Integrating vocabulary with various language aspects adds significance to the vocabulary, enhancing its communicative potential and improving retention among BIPA program students. Key principles of vocabulary teaching include ensuring that vocabulary lessons are comprehensive, including the teaching of meanings, and incorporating repetition rather than presenting vocabulary only once. Vocabulary should be embedded in relevant and authentic contexts, employing diverse teaching techniques (Khasanova & Safarova, 2022).

Vocabulary learning should be approached holistically, taking different aspects into account, including phonetic elements such as intonation, pronunciation, accent, and the positioning of expressive instruments (Akramovna, 2022; Al-Faris & Jasim, 2021). It is essential to consider the backgrounds of BIPA program students, as well as the competencies of each individual. For instance, a foreign student in the United States might

struggle to differentiate between sounds like [a] and [e] or [θ] and [g]. The students in this BIPA program study are expected to distinguish between correct and incorrect vocabulary usage.

In corpus linguistics, analyses reveal different aspects of vocabulary learning, such as identifying vocabulary with different letters and meanings, extra letters, similar letters, identical letters, and spelling errors. This diversity in vocabulary types arises from the detailed information available to BIPA students for vocabulary identification. Teaching vocabulary should align with conveying its meanings; it is crucial for BIPA students to grasp how the language operates. At this stage, conveying the nuances and cultural implications of vocabulary is also important (Heidari et al., 2020; Hilte et al., 2018; Martinez-Vazquez, 2017).

According to vocabulary teaching principles, repetition is necessary; teaching vocabulary just once is insufficient (Dakhi & Fitria, 2019). Thus, the organization of learning materials should ensure that the vocabulary introduced in one lesson is revisited in subsequent lessons. Below is an example of how to teach vocabulary to students in the BIPA program.

Vocabulary should not be taught in isolation; rather, it should be contextualized in real-life situations. This means that vocabulary should be incorporated into authentic sentences that BIPA students are likely to use. Examples of sentences should align with the learning objectives of BIPA program students. Understandably, BIPA students learn Indonesian for various reasons—some want to communicate with Indonesian friends, while others may aim to study and earn a degree in an Indonesian institution, or even conduct research. Such differences must be carefully considered when designing teaching materials, as they will influence the vocabulary taught. It is essential to avoid creating artificial sentences that are disconnected from real-life usage, as seen in examples like "Danny memukul kucing" or "kucing mengejar tikus," which are inappropriate for BIPA students. Instead, we should offer examples of authentic sentences that reflect the goals of BIPA program students.

Several techniques have been proposed for teaching vocabulary in ways that help it stay firmly in the memories of BIPA program students while maintaining communicative elements. Vocabulary learning can occur through direct demonstration, using object replicas, and employing images. When teaching vocabulary, especially challenging or new words, teachers are encouraged to point to relevant items. This technique can also evolve into grouping words, where BIPA students categorize vocabulary into specific classes, such as nouns, verbs, adjectives, etc., as illustrated in the examples below.

Noun:
[12] negara dan kota [country and city] → these common nouns are used to refer to an object in general.
[13] kampus dan tanah [campus and land] → Base nouns, also known as original nouns, can be identified by referring to the form of the object without an affix.

Verb:
[14] makan dan minum [food and drink] → Basic verbs, verbs without suffixes or stems
[15] membaca dan bertemu [reading and meeting] → Derivative verbs, verbs with suffix(es) or plural forms Adverb

[16] banyak, sedikit, cukup, dan kira-kira [many, a few, adequate, and approximately] → Quantitative adverbs describe meaning related to number/quantity
[17] hanya, saja, dan sekadar [only, merely, and just] → Restrictive adverbs describe meaning related to limitation.

Vocabulary learning can be effectively enhanced through specific activities designed to engage students. When teachers introduce challenging vocabulary to BIPA program students, it is beneficial to provide illustrations by performing actions or activities that represent the vocabulary words. This interactive approach is particularly useful for clarifying understanding of abstract nouns, verbs, and adjectives. This teaching method is commonly referred to as the "Tarzan language technique," which emphasizes learning through action and demonstration. Below are examples of vocabulary to be taught.

Adjective:
[18] bersih dan nyaman [clean and comfortable] → A quality that gives a physical or mental quality and intensity to an adjectival state.
[19] ringan dan banyak [light and many] → The adjective "quantitative" expresses quality measured by quantitative measurement.

Vocabulary can also be taught by clearly and slowly repeating the pronunciation of words. If BIPA students struggle to understand how the teacher pronounces certain vocabulary, the teacher should continue to pronounce the same sentence without replacing the word the student finds difficult with another word. Instead, the teacher should explain the pronunciation by slowing it down. For example, in the sentence "Mau berangkat dengan siapa?" if BIPA students do not understand the word "berangkat," they should not simply replace it with the word "pergi." Instead, they should repeat the entire sentence containing the word "berangkat." This repetition is intended to help BIPA students learn the word "berangkat" rather than substituting it with "pergi."

As in the previous example, the teacher should not only verbalize the difficult vocabulary but also write it down. This practice is essential for helping BIPA students learn to pronounce and write challenging words. The goal is to engage multiple senses in the learning process. Involving various senses in vocabulary learning strengthens the retention of the vocabulary in BIPA students' memories, ensuring that the words are well-preserved and do not quickly fade from their memory.

If BIPA students encounter difficulties with certain vocabulary, the teacher can clarify the meaning by providing the antonym of the challenging word. Understanding vocabulary alongside its antonyms can enhance BIPA students' memory retention. Often, students in the BIPA program may only remember certain vocabulary, but when they learn the antonyms, they are likely to retain more words. Examples of antonyms include:

[20] berdiri [standing]>< duduk [sitting]
[21] besar [big] >< kecil [small]

If BIPA students have difficulty with certain vocabulary, the teacher can explain that vocabulary by providing synonyms for the difficult vocabulary. Because of the vocabulary of certain things, BIPA students are expected to have a better understanding

of different vocabularies at the same time. The students can say the synonym of this vocabulary.

[22] Siswa [student]→ Murid [student]
[23] Susah [troubled]→ sedih, tidak senang [sad, unpleasant]

If BIPA learners find it difficult to remember vocabulary, teachers should not simply hand out vocabulary lists. Instead, they can provide hints using memory prompts, such as with the word "berjalan." The teacher can give clues by indicating that the word has eight letters, the first letter is /b/, the third letter is /j/, and the last letter is /n/. This keyword approach, using one or more related vocabulary words, is intended to help students accurately guess the target vocabulary.

Writing complex and unfamiliar vocabulary on flashcards is an effective technique for helping students remember words. There are two options for using flashcards. First, vocabulary can be written on one side, leaving the other side blank. Second, difficult vocabulary can be written on one side, with the Indonesian spelling and translation on the opposite side. These cards can serve as reminders, and students can divide them into two groups: one for words whose meanings they recall and another for forgotten words. The second group, made up of forgotten vocabulary, will need to be memorized again. After this process, BIPA students can categorize vocabulary into memorized and forgotten groups, continuing until all cards move to the remembered group.

Vocabulary review can also occur without writing if there are at least two students present in the class. This technique requires students to memorize a list of words, with one student mentioning several vocabulary words they have memorized, while another student provides instructions. For example, if four random words are presented, students may be asked to sequence them correctly. Another technique involves vocabulary building through the song "What Are You Doing," where students must listen carefully to identify whether the vocabulary words mentioned match those they have memorized. Techniques may include (a) one student sharing vocabulary with another, and (b) students exchanging words, particularly with peers who may not remember certain vocabulary. When multiple students or groups are present in the class, the vocabulary memorization technique can be particularly effective. If only one student is present, the teacher can take on the role of the second student to facilitate the activity.

Vocabulary learning techniques can also involve the Order Words command. Once students understand the meaning of vocabulary as a learning tool, they should organize vocabulary into sentences. This word order technique allows students the freedom to complete the task independently or with guidance. An example might entail that students are provided with four random words and are tasked with arranging them correctly. Another approach could involve matching words while singing the "What Are You Doing" song. This technique can also be combined with word-guessing exercises designed to assist IPA program students in recalling specific vocabulary. Additional methods that can be utilized include free word guessing (close model), guessing puzzle words, and playing hangman. Another effective vocabulary teaching technique involves word-finding in texts, where BIPA students are asked to locate specific vocabulary, such as adjectives or verbs with affixes like me-kan, di-kan, ter, ber, etc. Students may also be asked to write specific vocabulary or engage in dialogues about technical language related to areas such as economics, politics, and culture.

BIPA students can also utilize vocabulary instructional technology, such as matching words with corresponding vocabulary items. This vocabulary-finding technique can be applied when students are presented with vocabulary words, and a student or teacher reads a specific term. Students then match the term to related vocabulary. This technique is particularly effective for listening exercises. BIPA students can learn vocabulary by developing root words into derived forms and vice versa, dividing words into their root components and constructing suffixes. Mastery of this technique is critical, as the Indonesian language features a unique system of affixation.

The explanation above outlines various vocabulary learning principles and techniques for BIPA students. Given the complexity of the diverse vocabularies in the Indonesian language, it is crucial to develop a comprehensive vocabulary learning method. A well-structured curriculum model can provide guidance for the direction of the curriculum and outline specific objectives to be achieved in each lesson. With careful planning, teachers can tailor curriculums to effectively meet educational goals.

Teaching units can be organized based on language materials that consider functions, situational contexts, grammatical rules, and classroom activities (Forey & Cheung, 2019; Kalinowski et al., 2019; Pishghadam et al., 2021). The curriculum design model, which integrates aspects of validity, language function, and contextual use, is particularly comprehensive and well-suited for teaching Indonesian to BIPA students (Saddhono et al., 2019; Suyitno et al., 2019).

To make these activities more engaging, one effective method could involve role-playing. For example, some students can take on the roles of waiters, while others act as restaurant guests. When a guest places an order, the waiter can respond with phrases such as “Baiklah” (Alright) or “Mohon maaf, makanan/minuman itu tidak tersedia” (Sorry, that food/drink is not available). Students can then switch roles to gain a fuller understanding of the vocabulary and its application in different contexts.

After completing these role-playing exercises, students can participate in a discussion comparing Indonesian lunch menu preferences—including types and quantities of food—with customs from their home countries (Lin, 2013; Parvaresh & Dabghi, 2013). This approach not only helps to highlight cultural differences but also contributes to a more immediate and deep understanding of Indonesian vocabulary, thereby enriching the students’ learning experience.

This research, which employed corpus linguistics, is valuable for identifying and analyzing corpus data. This corpus can originate from any language. The dataset derived from vocabulary created by BIPA students can be utilized to enhance Indonesian vocabulary learning for non-native speakers. Other researchers may also use this study’s results to explore different aspects and connections of interest, such as linguistic errors, morphological processes, syntactic studies, discourse analysis, or specialized vocabulary usage. The research context of these elements spans various languages, not just Indonesian. This study serves as an example of corpus linguistics applied to Indonesian research. Manual research and traditional linguistic data analysis need to transition to digital technology, utilizing corpus applications like the one applied in this research. The advantage of this corpus application, compared to many others worldwide, is its ability to process targeted groups according to the research corpus’s object, drawing from a variety of languages and contexts.

5. CONCLUSION AND SUGGESTION

Based on an analysis of the language used in introductory discourse by BIPA program students at Potensi Utama University, employing the Sketch Engine application within corpus linguistics, it was found that students possess a dominant command of vocabulary beyond just nouns. Among nine types of Indonesian words, the eleven most frequently used include saya, di, dan, suka, adalah, kota, dekat, belajar, dari, ke, and yang. The word "saya" is undoubtedly the most frequently occurring term because the speech context is introductory; thus, it plays a significant role in self-presentation. Although BIPA students produced a variety of vocabulary, there were errors in their writing due to a lack of meticulousness and carefulness in the editing process. The analysis of the Sketch Engine application, which included 2,334 vocabulary items from 50 students, allows for the organization of principles and techniques for effective and efficient vocabulary learning tailored to BIPA students, introducing them to the types of Indonesian words and variations in word forms. A rich vocabulary is essential for meeting students' needs as they master language. Therefore, the results of the SKELL corpus linguistics application are highly valuable in the BIPA learning process, providing guidance based on students' language skills to achieve optimal outcomes in vocabulary acquisition.

ACKNOWLEDGMENT

The researcher would like to express gratitude to the University of Potensi Utama for providing financial support to publish this article.

REFERENCES

Abbasi, M. M., & Beltiukov, A. P. (2018). Analyzing emotions from text corpus using word space. *Industry 4.0, 3*(4), 161-164.

Akramovna, A. I. (2022). The importance of teaching vocabulary. *Journal of Pedagogical Inventions and Practices, 6*, 23-27.

Al-Faris, S., & Jasim, B. Y. (2021). Memory strategies and vocabulary learning strategies: Implications on teaching and learning vocabulary. *Journal of Humanities and Social Sciences Studies, 3*(10), 11-21. https://doi.org/10.32996/jhsss.2021.3.10.2

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics, 11*(2), 135-171. https://doi.org/10.1075/ijcl.11.2.02als

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. https://www.laurenceanthony.net/software/antconc/

Badan Pengembangan dan Pembinaan Bahasa. (n.d.). Aku. In *KBBI Daring.* https://kbbi.kemdikbud.go.id/entri/aku

Barclay, S., & Schmitt, N. (2019). Current perspectives on vocabulary teaching and learning. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 799-819). Springer. https://doi.org/10.1007/978-3-030-02899-2_42

Callies, M. (2019). Integrating corpus literacy into language teacher education. *Learner Corpora and Language Teaching, 92*, 245-263. https://doi.org/10.1075/scl.92.12cal

Charkazova, S. S. (2018). Principles of learning and teaching vocabulary. *International Scientific and Practical Conference World Science, 5*(5), 27-29. https://elibrary.ru/item.asp?id=34924699

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.01

Dakhi, S., & Fitria, T. N. (2019). The principles and the teaching of English vocabulary: A review. *Journal of English Teaching, 5*(1), 15-25. http://dx.doi.org/10.33541/jet.v5i1.956

Dunder, I., & Pavlovski, M. (2018). Computational concordance analysis of fictional literary work. In K. Skala (Ed.), *Proceedings of The 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 644-648). IEEE. https://doi.org/10.23919/MIPRO.2018.8400121

Forey, G., & Cheung, L. M. E. (2019). The benefits of explicit teaching of language for curriculum learning in the physical education classroom. *English for Specific Purposes, 54*, 91-109. https://doi.org/10.1016/j.esp.2019.01.001

Gholaminejad, R., & Sarab, M. R. A. (2020). Academic vocabulary and collocations used in language teaching and applied linguistics textbooks: A corpus-based approach. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, 26*(1), 82-107. https://doi.org/10.1075/term.00043.gho

Heidari, F., Jalilifar, A., & Salimi, A. (2020). Developing a corpus-based word list in pharmacy research articles: A focus on academic culture. *International Journal of Society, Culture & Language, 8*(1), 1-15.

Hilte, L., Vandekerckhove, R., & Daelemans, W. (2018). Social media writing and social class: A correlational analysis of adolescent CMC and social background. *International Journal of Society, Culture & Language, 6*(2), 73-89.

Hizbullah, N., Fazlurrahman, F., & Fauziah, F. (2016). Linguistik korpus dalam kajian dan pembelajaran bahasa Arab di Indonesia [Corpus linguistics in the study and learning of Arabic in Indonesia]. *Prosiding Konferensi Nasional Bahasa Arab, 1*(2), 11-20. https://prosiding.arab-um.com/index.php/konasbara/article/view/84

Isam, H., & Abd Mutalib, M. (2019). Pemanfaatan analisis korpus sebagai teknik alternatif pengajaran dan pembelajaran tatabahasa [Utilization of corpus analysis as an alternative technique of teaching and learning grammar]. *International Journal of Language Education and Applied Linguistics, 9*(1), 13-31. https://doi.org/10.15282/ijleal.v9.594

Junpaitoon, P. (2017). Enrichment of vocabulary in BIPA learning for beginner Thai students. *Journal of Innovative Studies on Character and Education, 1*(1), 88-103.

Kalinowski, E., Gronostaj, A., & Vock, M. (2019). Effective professional development for teachers to foster students' academic language proficiency across the curriculum: A systematic review. *AERA Open, 5*(1), 1-23. https://doi.org/10.1177/2332858419828691

Khasanova, K. B., & Safarova, D. A. (2022). Teaching vocabulary: Methods and approaches. *Global Scientific Review, 3*, 15-16.

Lin, Y. (2013). Vague language and interpersonal communication: An analysis of adolescent intercultural conversation. *International Journal of Society, Culture & Language, 1*(2), 69-81.

Martinez-Vazquez, M. (2017). Cultural influence on the expression of cathartic conceptualization in English and Spanish: A corpus-based analysis. *International Journal of Society, Culture & Language, 5*(2), 1-14.

Naji Meidani, E., Makiabadi, H., Zabetipour, M., Abbasnejad, H., Firoozian Pooresfehani, A., & Shayesteh, S. (2022). Emo-Sensory communication, emosensory intelligence and gender. *Journal of Business, Communication & Technology, 1*(2), 54-66. https://doi.org/10.56632/bct.2022.1206

Ningsih, R. Y., Rafli, Z., & Boeriswati, E. (2021). Linguistic creativity in BIPA students (Indonesian for foreign speakers). *Lingua Cultura, 15*(2), 199-206. https://doi.org/10.21512/lc.v15i2.7613

Oakes, M. (2019). *Statistics for corpus linguistics*. Edinburgh University Press.

Parvaresh, V., & Dabghi, A. (2013). Language and the socio-cultural worlds of those who use it: A case of vague expressions. *International Journal of Society, Culture & Language, 1*(1), 74-88

Pishghadam, R., Derakhshan, A., Jajarmi, H., Tabatabaee Farani, S., & Shayesteh, S. (2021). Examining the role of teachers' stroking behaviors in EFL learners' active/passive motivation and teacher success. *Frontiers in Psychology, 12*, 707314. https://doi.org/10.3389/fpsyg.2021.707314

Pishghadam, R., & Zabihi, R. (2012). Crossing the threshold of Iranian TEFL. *Applied Research on English Language, 1*(1), 57-71. https://doi.org/10.22108/are.2012.15446

Presbitero, A. (2020). Foreign language skill, anxiety, cultural intelligence and individual task performance in global virtual teams: A cognitive perspective. *Journal of International Management, 26*(2), 100729. https://doi.org/10.1016/j.intman.2019.100729

Rădescu, R. (2021). Concordance techniques in lossless data compression of text files. In *Proceedings of The 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (pp. 1-4). IEEE. https://doi.org/10.1109/ATEE52255.2021.9425067

Reppen, R., & Simpson-Vlach, R. (2019). Corpus linguistics. In N. Schmit & M. P. H. Rodgers (Eds.), *An introduction to applied linguistics* (pp. 91-108). Routledge.

Saddhono, K. (2015). Integrating culture in Indonesian language learning for foreign speakers at Indonesian universities. *Journal of Language and Literature, 6*(2), 349-353.

Saddhono, K., Hasibuan, A., & Bakhtiar, M. I. (2019). Facebook as a learning media in TISOL (Teaching Indonesian to speakers of other languages) learning to support the independency of foreign students in Indonesia. *Journal of Physics: Conference Series, 1254*, 012061. https://doi.org/10.1088/1742-6596/1254/1/012061

Saddhono, K., & Wahyono, H. (2019). Learning vocabulary using multimedia based teaching Indonesian to speakers of other languages (TISOL). *Journal of Physics: Conference Series, 1339*, 012108. https://doi.org/10.1088/1742-6596/1339/1/012108

Safii, M., Harsiati, T., Kurniawan, T., Asari, A., Rahmania, L. A., Prasetyawan, A., & Arraja, Y. G. (2021). Index development with the keyword in context method in the online library catalog. *Journal of Physics: Conference Series, 1839*, 012001. https://doi.org/10.1088/1742-6596/1839/1/012001

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Suryadarma, Y., & Fakhiroh, A. Z. (2020). Optimalisasi penggunaan corpus linguistics dalam penyusunan Kamus Azziro'ah sebagai media pembelajaran Bahasa Arab [Optimizing the use of corpus linguistics in compiling the Azziro'ah Dictionary as a medium for learning Arabic]. *ISoLEC Proceedings, 4*(1), 123-128. http://isolec.um.ac.id/proceeding/index.php/issn/article/view/59

Suryadarma, Y., & Zakaria, G. A. N. (2022). Korpus Arab pesantren: Digitizing the work of Arabic non-Arabic speakers at modern Islamic institution Darussalam Gontor. *At-Ta'dib, 17*(1), 52-66. http://doi.org/10.21111/at-tadib.v17i1.7067

Suyitno, I., Pratiwi, Y., Roekhan, R., & Martutik, M. (2019). How prior knowledge, prospect, and learning behavior determine learning outcomes of BIPA students? *Journal Cakrawala Pendidikan, 38*(3), 499-510. http://dx.doi.org/10.21831/cp.v38i3.27045

Uchihara, T., & Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research, 24*(4), 540-556. https://doi.org/10.1177/1362168818799371

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning, 69*(3), 559-599. https://doi.org/10.1111/lang.12343

Wahyuningtyas, D., & Kesuma, T. M. J. (2021). Pemanfaatan linguistik korpus dalam menentukan kata berfrekuensi tinggi pada buku 'Sahabatku Indonesia' BIPA 1 [Utilization of corpus linguistics in determining high-frequency words in the book 'Sahabatku Indonesia' BIPA 1]. *Journal Bahasa Indonesia bagi Penutur Asing, 3*(2), 60-69.

Zeldes, A. (2018). *Multilayer corpus studies.* Routledge.

Zou, D., & Xie, H. (2018). Personalized word-learning based on technique feature analysis and learning analytics. *Journal of Educational Technology & Society, 21*(2), 233-244.