International
Journal of Society, Culture & Language
IJSCL

Journal homepage: www.ijscl.net
ISSN 2329-2210 (online)

Corpus Linguistics Use in Vocabulary Teaching Principle and Technique Application: A Study of Indonesian Language for Foreign Speakers

Kundharu Saddhono[1a], Muhammad Rohmadi[2a], Budhi Setiawan[3a], Raheni Suhita[4a], Ani Rakhmawati[5a], Sri Hastuti[6a], Islahuddin Islahuddin[7b]

ARTICLE HISTORY:

Received August 2022
Received in Revised form October 2022
Accepted November 2022
Available online November 2022

KEYWORDS:

Corpus linguistics
Vocabulary
Teaching principle
Indonesian language for foreign speakers
Thailand

Abstract

Indonesian Language for Foreign Speakers (BIPA) is Indonesian language learning intended for foreigners. The aim of this research was to examine the vocabulary, terminologies, and grammar used by BIPA students with the corpus linguistics application, *Kortara*. This research was conducted at Fatoni University and Yale University with 51 BIPA students. This research used a mixed-methods approach, comprising the quantitative method that was used in the early stage of research to obtain the research data and the qualitative method for the analysis process. The research results showed dominant mastery of vocabulary by BIPA students, mostly nouns, verbs, adverbs, pronouns, and adjectives. There were 11 vocabularies with the highest frequency. Based on the results of the analysis, effective and efficient vocabulary learning principles and techniques were structured for BIPA students as an introduction to word types and variations of word formation in the Indonesian language.

[1] Associate Professor, Email: kundharu_s@staff.uns.ac.id (Corresponding Author)
Tel: +62-815-6610804
[2] Associate Professor, Email: mamad_r76@staff.uns.ac.id
[3] Associate Professor, Email: kaprodipbi@staff.uns.ac.id
[4] Associate Professor, Email: rahenisuhita@staff.uns.ac.id
[5] Associate Professor, Email: anirakhmawati@staff.uns.ac.id
[6] Lecturer, Email: srihastuti@staff.uns.ac.id
[7] Lecturer, Email: islahuddin@ftu.ac.th
[a] Universitas Sebelas Maret, Indonesia
[b] Fatoni University, Thailand
http://dx.doi.org/10.22034/ijscl.2022.1971972.2823

1. Introduction

Indonesian Language for Foreign Speakers (BIPA) is an Indonesian language (BI) learning program with foreigners as the subjects. In this regard, the Indonesian language serves as a foreign language for foreign speakers. In addition, the BIPA program is empowered as a means of diplomacy that the government may use to strengthen Indonesia's position in the world. Through the BIPA program, other countries can easily get to know Indonesia and establish bilateral or multilateral relations. BIPA also carries out the mission to introduce insight into Indonesia's rich and varied cultures. BIPA is currently developing rapidly (Anonim, 2021). In 2021, 10,730 BIPA learners in 38 countries were affiliated through 279 assigned BIPA teachers in 204 institutions. This shows that the Indonesian language and culture are in great demand in other countries. This is also reflected in the BIPA programs organized in some universities in Indonesia (Junpaitoon, 2017; Ningsih, 2021; Saddhono, 2015).

In BIPA learning, a teacher needs to pay attention to the planning, process, and evaluation stages and also to the teaching materials, media, and methods used. One of the important things to teach is Indonesian language skills covering good and correct listening, speaking, reading, and writing skills. Language skills require mastery of sufficient vocabulary so that any idea to deliver will be well received. Good mastery of vocabulary determines a person's language quality. Sufficient vocabulary acquisition is important in second language mastery since, without sufficient vocabulary, a person cannot use any structure and function that has been learned to communicate well (Naji Median et al., 2022; Saddhono & Wahyono, 2019). That is why vocabulary learning should be performed effectively and contextually in order to be applied and achieve the purpose of daily communication and serve as significant learning, and be done actively, effectively, and in a fun way.

Vocabulary is an important part of language learning in support of language skills. Vocabulary in the BIPA program is taught through implicit learning. Vocabulary is integrated through four language skills: listening, speaking, reading and writing skills, and grammar material. Vocabulary teaching plays an important role in increasing the success of Indonesian language teaching to BIPA program students. The number of BIPA program students' vocabulary is positively correlated with the success of Indonesian language mastery. A person with more vocabulary will surely be more skillful in using the Indonesian language at listening, speaking, reading, and writing levels (Presbitero, 2020).

Good vocabulary learning is performed communicatively for it to serve as the means of communication and to support BIPA program students' need for speech act (Uchihara et al., 2019, 2020). Therefore, there is a need for vocabulary teaching principles. It is also expected in vocabulary learning that students retain them and will not forget them. In this regard, good vocabulary teaching techniques are needed. It is the vocabulary teaching principles and techniques for students to retain and not forget vocabulary and to meet the purpose of BIPA program students' communication that will be discussed in this paper (Zou & Xie, 2018).

The aim of this research is to examine the vocabulary, terminologies, and grammar used by BIPA program students. The research results will be used in preparing Indonesian language vocabulary learning in the BIPA Program so as to facilitate students' comprehension of the Indonesian language, and also to help them understand Indonesian language texts. This will thus not only facilitate BIPA program students completing their course well but also be beneficial for them in their daily activities using the Indonesian language.

The novelty of this research is that the basis of Indonesian language vocabulary learning development is based on the analysis of corpus linguistics data produced from BIPA program students' writing using the *Kortara* (Korpus Nusantara) application. *Kortara* is a web-based digital application on data corpus that can be used by researchers, teachers, lecturers, and students for learning and course in processing the corpora of language in the archipelago (Indonesia). *Kortara* was officially registered with the Ministry of Law and Human Rights in protecting creation in the field of science, arts, and literature under law Number 28 of 2014 on Copyright. The developed and redesigned application was re-launched on 3 September 2022 and can be accessed at https://korpusnusa

ntara.fbs.unp.ac.id/. It is from this application that the results of Indonesian language data corpora produced in the form of vocabulary, terminologies, and grammar of BIPA program students at Fatoni University and Yale University were obtained. The results of the *Kortara* application can be analyzed from various aspects so as to describe the BIPA program students' ability at Fatoni University and Yale University to produce writing in the Indonesian language.

2. Theoretical Framework

By form and purpose, 'corpus' as a collection of natural language examples consists of some sentences in a set of written texts or records collected for linguistic study, and the texts are then arranged systematically (Pishghadam & Zabihi, 2011; Stefanowitsch, 2020; Zeldes, 2018). Corpus is declared "natural" since the texts collected are those produced and used naturally and as is or not made up. The texts include textbook journals, novels, newspapers, magazines, and records of conversation broadcasts, interview results, and many more. Corpus linguistics is a complete system containing methods and principles to apply corpus in linguistic research and teaching or learning (Callies, 2019; Reppen & Simpson-Vlach, 2019). Corpus linguistics is also a field focusing on a set of procedures or methods for learning a language (Gholaminejad & Sarab, 2020). Based on these definitions, we may conclude that corpus linguistics is linguistic research that uses examples of daily or natural language stored in a corpus. The corpus in this research is the introductory discourse of 51 BIPA program students at Fatoni University and Yale University year 2022. Corpus linguistics is suitable for application in this research since corpus linguistics has the features needed to achieve the research aim.

The linguistic features include frequency in corpus linguistics, referring to the appearance of a word in a corpus or text (Crossley, 2020). Not only used to count the appearances of a single word, the frequency also allows the counting of grammatical, semantic, or other categorical frequencies. Frequency can also guide the researcher to wider findings. Frequency in corpus linguistics shows how many times a word appears in a corpus. Frequency analysis allows the researcher to recognize words often appearing in a certain corpus and then compare and distinguish them from other words (Oakes, 2019). In this research, frequency is used to examine the number and variations of vocabulary often appearing that students need to know in order to comprehend the Indonesian language more. This way, students will know the list of vocabularies of priority to pronounce and comprehend. Besides, the fact that frequently appearing words are known will automatically show what linguistic features are the priority to teach BIPA program students.

The other linguistic feature is concordance which is a list or order of examples of words, part of a word, or a combination of words existing in the context taken from text corpus (Abbasi & Beltiukov, 2018). The main word as the target researching in the corpus is called keyword. There are many methods to display keywords. One of the mostly-used methods in corpus linguistics is the use of keywords in a context, known as KWIC (keyword in context) (Safii et al., 2021). Concordance is an important aspect of corpus linguistics, allowing qualitative analysis of corpus data. This allows the researcher to explore individual cases in detail. Concordance analysis is usually important to do before claiming language variation or change by frequency. Displaying concordance requires software to investigate a certain linguistic item in the context in consideration of surrounding words which may start from a word leftward or rightward from the item in the whole context if necessary (Dunder & Pavlovski, 2018; Rădescu, 2021). The concordance technique also allows the researcher to do a qualitative analysis by learning the item in the text along with it. In this research, concordance is applied to help the qualitative analysis process on the available data. The data analysis is conducted to observe the linguistic features inherent in a word by viewing the word itself and also by viewing the surrounding words. Through the concordance technique, it will find the word class of a word, types of active and active sentences, words subject to morphological process, variation of word position in the sentence, etc.

Various studies have explored the connection between corpus linguistics and learning by engaging different perspectives. For example, research on Arabic corpus linguistics was widely carried out, one of which was in the context of Arabic language learning conducted

in Indonesia (Al-Sulaiti & Atwell, 2006; Hizbullah et al., 2016). It aimed to describe several contents of corpus linguistics and its development and dynamics in Arabic studies internationally. Researchers employed a corpus linguistic approach using the Wordsmith software, which was considered the most comprehensive and representative. Primarily, it could adequately process text in any language since it was equipped with a feature in the form of system adjustments. In this regard, the file must be altered before being processed. For instance, the texts typed in Microsoft Word software usually have a .doc or .docx extension. Such file formats must be converted by the encoding method using UTF-8 and then saved in the .txt extension. Eventually, the data can be analyzed using the Wordsmith. The development of the Arabic corpus and related software could be utilized for relevant processing and analysis. It might also be used as a basis for academic discourse and discussion regarding the possibility of developing a model for the Arabic language corpus in Indonesia and the involvement in studying and learning Arabic in higher education.

Another research on applying corpus analysis as an alternative to grammar teaching and learning was conducted by Isam and Mutalib (2019). They focused on the selection of word list software and corpus databases, material variant processes, language data, procedures for using word list software, and examples of linguistic analysis. The corpus linguistic approach being administered was based on the AntConc software, and it employed the data from a corpus database of the Institute of Language and Literature (Dewan Bahasa dan Pustaka). Researchers completed the procedure with the following stages: 1) students were asked to download data from the corpus database of the DBP (Institute of Language and Literature) (2012); 2) students were instructed to download the AntConc software; 3) students input the data (prepositions downloaded from the DBP database) into the AntConc; 4) the teacher asked students to convey something interesting (dynamic features) that could be explained based on the morphological and syntactic behavior of the analyzed prepositions. Likewise, corpus linguistics was intended to facilitate learners' interest and active involvement during teaching and learning. This study also revealed that the technique being examined was designed systematically; thereby, using teaching and learning strategies based on corpus data analysis could positively impact the development of student education and support the studies in linguistics and the Malay language in Malaysia.

The optimization of corpus linguistics was also engaged in preparing the "Az-Ziro'ah" dictionary as an Arabic language learning media (Suryadarma & Fakhiroh, 2020). In this regard, researchers described the involvement of the corpus linguistic approach in designing a bilingual dictionary in Agro-industry or Agricultural Industrial Technology. They incorporated a qualitative description approach in creating a dictionary (lexicography) and corpus linguistic data collection techniques based on the AntConc and Sketch Engine software. The process began by collecting vocabulary related to agro-industry (technology, industry, and agriculture) through print and online media articles. Data input and selection were carried out in .doc format; the file was subsequently converted to plain text (.txt); the corpus linguistic-based data were then analyzed using the Sketch Engine software by which researchers performed data cleaning by considering the basic, derived, and affixed words using the N-Grams tool of the AntConc software. Hence, corpus linguistics could be used as an approach or basis for compiling a bilingual dictionary in a particular field. Moreover, it was believed to generate significant results in mapping and grouping specific vocabularies efficiently and systematically. The crucial aspects in composing a corpus linguistic-based bilingual dictionary were the selection of data sources, the mechanism in the corpus data processing platform, and the ability to choose vocabularies or lemmas to be included in the dictionary's draft. These three were the primary requirements for successfully preparing a dictionary based on corpus linguistics (Suryadarma & Zakaria, 2022).

The subsequent relevant study investigated the use of corpus linguistics in determining high-frequency words in the book entitled "Sahabatku Indonesia (BIPA 1)" (Wahyuningtyas & Kesuma, 2021). Using corpus linguistics, researchers described a list of words based on their frequency of use in the book. A qualitative approach was involved in obtaining the research data, analyzing it, and drawing

conclusions. The processes started with selecting data in the form of a .pdf electronic book, which was converted into .txt using a file converter software. In this research, the "PDF to Text" software was downloaded from the Apple Store and then utilized to convert a .pdf file to .txt. Besides, researchers also used the AntConc 3.5.9. After the software was installed, a book file in the extension of .txt was uploaded into the AntConc (tool-based corpus), and then researchers accessed the Word List feature to specify the frequency of words. The results indicated that the word classes with the highest occurrence rate did not necessarily make them high-frequency words. Anchored on the findings, particles such as and, in, that, with, and so were the most significant contributors to high-frequency words. Therefore, it was concluded that the list of high-frequency words did not always include the technical terms related to the investigated topic. Depending on the researcher's needs, the determination of high-frequency words employing corpus linguistics could be applied to various materials (books, newspapers, literary works, and others). This kind of study might generate a list of specific words or terms packaged in the form of a mini dictionary as a reference for learners.

Various research on corpus linguistics and learning corroborated that they were closely interconnected and mutually helpful in terms of analysis and study. Corpus linguistics could help deepen and sharpen the analysis, considering the availability of comprehensive data, which would be very helpful in designing learning based on the results of corpus linguistic analysis. The most fundamental element in this type of investigation was the results of corpus linguistic data analysis in the form of vocabularies to be used by learners as a basis for designing a vocabulary learning process that was easy to master, communicative, and applicable to language skills for daily use or academic activities.

3. Methodology

3.1. Participants

This research was conducted at Fatoni University, Thailand, and Yale University, USA, with 51 BIPA students year 2022. They were students in the early semester of the BIPA program class with various language and cultural backgrounds.

3.2. Instruments

The introductory discourse of BIPA Program students at Fatoni University and Yale University in pdf form was converted into plain text so that it could be processed using the *Kortara* application. Through the feature frequency in this instrument, the whole words used in the introductory discourse were produced based on the frequency of appearance, from the most frequently appearing to only one or two appearances. The list produced by the software was sorted in the reduction process to separate words with full meaning and other words such as abbreviations, syllable pieces, numbers, etc.

3.3. Procedure

The method used was a mixed method in the early stage of research. A quantitative method was used in order to obtain the research data, and a qualitative method was used for the analysis process. In the early stage, the qualitative method was used to obtain quantitative data through the software *Kortara* which was a corpus linguistics application (Durrant et al., 2021; Egbert & Baker, 2019; Yadav et al., 2020).

3.3.1. Data Collection

The data collection in this study was an introductory discourse on BIPA program students written by 51 students. Research data was a word that was then classified according to the purpose of analysis. Words with full meaning were then classified into their respective class as per function in a sentence.

3.3.2. Data Analysis

After the data were obtained, the words were then analyzed qualitatively for linguistic features inherent in each class of words for a conclusion. After the analysis, the BIPA program students' ability level was found, and vocabulary learning principles and techniques were then designed for the students' improved ability.

4. Results

In the early stage, the introductory discourse of BIPA Program students at Fatoni University and Yale University in word form was converted into a plain file so that it could be processed in the software *Kortara*. Through the

feature wordlist, the document was processed by the software, producing a list of words existing in the discourse by type of words in the Indonesian language. The list of word types is presented in Table 1.

Table 1
Category by Word Type

No	Word Type	Freq.
1	Noun	871
2	Verb	398
3	Adverb	215
4	Pronoun	130
5	Adjective	121
6	Number	92
7	Others	31
8	Preposition	28
9	Interjection	26
10	Conjunction	25
11	Article	10
	Total	1.947
#Word Types: 1,947		
#Word Tokens: 9,664		

The result of corpus linguistics using the *Kortara* application can be explained that noun is the most dominantly used word with over 871 appearances, followed by the verb with 398 appearances, adverb with 215 appearances, pronoun with 130 appearances, and adjective with 121 appearances. This result shows that the BIPA program students master nouns the most compared to other word types.

Noun is the name of all objects and anything objectified, and by form, divided into (1) Concrete noun, which is the name of objects that can be caught by the five senses, and (2) Abstract noun, which is the name for objects that cannot be caught by the five senses. The characteristics of a noun are all words that can be explained or extended by adding ***yang + adjective*** or ***yang sangat + adjective*** behind the word. Examples of the two nouns are given below.

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, the noun can take a position as Subject (S) and Object (O). For example, "*Ku-Ares membeli kopi*". In the sentence, "*Ku-Ares*" and "*kopi*" are nouns. By formation process, the noun in the Indonesian language is divided into a basic noun and a derivative noun. The basic noun is an original noun (both concrete and abstract) that is not given any affix. The derivative noun is a noun that is given with a certain affix.

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 examples above are the classification of corpus linguistics data by word type in the Indonesian language. Through the result of *Kortara* application, the words mostly used by BIPA program students will also be identified. Frequency (number of appearances), according to corpus linguistics, is presented in Table 2.

Table 2
List of Words by Frequency

No	Word Data	Freq.	No	Word Data	Freq.
1	*saya*	530		*kota-kota*	1
	Saya	406		*kota.*	2

	saya,	147		*Total*	121
	(saya	1	7	*dekat*	119
	saya!	1		*dekat.*	1
	Total	1085		*Dekat*	1
2	*di*	583		*Total*	121
	Di	128	8	*belajar*	112
	Total	711		*belajar.*	4
3	*dan*	430		*Belajar*	1
	Total	430		*belajar,*	2
4	*suka*	154		*Total*	119
	Suka	1	9	*dari*	114
	Total	155		*Total*	114
5	*adalah*	144	10	*ke*	106
	adalah:	1		*Ke*	1
	Total	145		*Total*	107
6	*kota*	105	11	*Yang*	104
	Kota	19		*yang*	2
	kota,	1		*Total*	106

In the introductory discourse of BIPA program students at Fatoni University and Yale University, there are 1,947 Word Types and 9,664 Word Tokens. Word Type shows word type individually used by the students in the program in the introduction, and Word Token means the number of whole words existing in the introductory discourse of BIPA program students, including repeated words. The application results show that the word with the highest frequency is *"saya"* with five variations of *saya, Saya, saya, (saya, dan saya!*. The only position where the word *saya* is not found in the discourse is at the end of the sentence, but even if it is found at the end of a sentence, it is an interjection (*saya!*). It is interesting in the discourse that with so many words, *saya* is found. There is only one-word *aku* which is the synonym of *saya*. This shows that the BIPA program teachers have well-explained to their students the difference in the meaning between *saya* and *aku* in Indonesian speech.

Based on Kamus Besar Bahasa Indonesia (2021), the term '*aku'* means first person pronoun who is talking or writing (in a familiar/friendly way); *diri sendiri*; *saya* and the term '*saya'* means pronoun or one that is talking or writing (in an official or casual way); *aku*. Based on the explanation, we may explain that the term '*aku'* and '*saya'* are both first-person pronouns. Examples are given in the sentences below.

[10] *Ilham berkata, "Saya belajar sejarah di Fatoni University".*
[Ilham said, "I study history at Fatoni University."]

[11] *Ilham berkata, "Aku belajar sejarah di Yale University".*
[Ilham said, "I study history at Yale University".]

The term '*saya*' in sentence [10] bears an exactly similar meaning to the term '*aku'* in sentence [11]. The terms '*saya'* and '*aku'* in the sentences equally refer to the one speaking, namely: Ilham.

Having identified the similarity of the terms '*saya'* and '*aku'*, we can then understand the difference in their use. Some say that the term '*saya'* is more precise and recommended than '*aku'*. This is a false statement. Neither is more precise and recommended or false between the terms '*saya'* and '*aku'*. The same also applies to a statement that the term '*saya'* is more polite than '*aku'*. This is not completely correct. Each word in the Indonesian language, also in other languages, has a different dimension and uses context appropriate to the word.

If the term '*aku'* is deemed not good, it is unlikely necessary for the word to exist, and if that word is not needed, it will surely be not used or be removed from Indonesian vocabulary. It is correct and polite when it is used as per the context. The term '*saya'* is also correct and polite when it is used as per the context. The term '*saya'* is a way to express oneself (first person pronoun) more politely on the premise of respecting oneself as well as the interlocutor. The term '*saya'* is used by a person who considers himself lower than the interlocutor with the intention to respect the interlocutor. For example, a subordinate towards a superior. Not limited to that, however, the term '*saya'* can also be used by a superior towards a subordinate with the intention to respect the interlocutor despite being subordinate (Murtisari et al., 2019).

The term '*aku'* is correctly used to convey intimacy with the interlocutor. Thus, when people are good friends despite the age gap, they can use the term '*aku'* to refer to themselves. The important thing is that the term '*saya'* is more polite and respects oneself and the interlocutor more. In order to respect oneself as well as the interlocutor, it would be better to use the term '*saya'* instead of '*aku'*. In a good and correct sentence context, '*saya'* is considered more standard than '*aku'*. The term 'saya' is used in an official way, while '*aku'* is used in conversation. The obvious difference between '*aku'* and '*saya'* is that '*saya'* is to humble oneself and '*aku'* is used to exalt oneself (in a certain context). The term '*saya'* is a derivative form of root word '*hamba sahaya*' which means '*orang rendahan*'.

5. Discussion

The vocabulary teaching principles need to be observed for the vocabulary learning process to be integrated with other linguistic elements (Charkazova, 2018; Barclay & Schmitt, 2019). A vocabulary teaching that is integrated with other linguistic aspects will make the vocabulary more meaningful in that they have communicative power and are retained in BIPA program students' memory. Some vocabulary teaching principles include the following: vocabulary must be intact, along with teaching the word's meaning, with repetition instead of only once. Vocabulary should be included in actual and scientific contexts, using varied vocabulary teaching techniques (Khasanova & Safarova, 2022).

Vocabulary learning must be performed as a whole from various aspects covering sound elements such as how to pronounce intonation, utterance, utterance accent, and position of utterance instruments (Akramovna, 2022; Al-Faris & Jasim, 2021). In this vocabulary learning, it is necessary to pay attention to the BIPA program students' origin. This is also related to the competence of each individual or the BIPA program student. For example, a foreign student from the United States will find it difficult to distinguish the sound [a] and [e], [] and [g]. The targets in this research, BIPA program students, can distinguish correct utterances from the false utterance of a vocabulary taught.

The corpus linguistics analysis finds some vocabulary learning from identification, that is, vocabulary with different letters and meanings, Vocabulary with Additional Letters, Vocabulary with Almost the Same Letter, Vocabulary with the Same Letter, and Vocabulary Typo. This division of types of vocabulary is possible since the data used by the BIPA students can be identified in detail for vocabulary appearance. Teaching vocabulary should also be in line with teaching the meaning that the vocabularies bear. This is important for the BIPA program students to be aware of the exact meaning to apply in making language act. It is also necessary to convey in this stage the nuance of the meaning of the vocabulary, including the cultural value contained therein (Martinez-Vazquez, 2017; Hilte, Vandekerckhove, & Daelemans, 2018; Heidari & Salimi, 2020).

The vocabulary teaching principle states it should be repeated; thus, it is insufficient to teach vocabulary only once (Dakhi & Fitria, 2019). Consequently, learning material arrangement should be planned in such a way so that vocabularies in the first learning are to be brought in the next learning and so on. Below is an illustration of vocabulary teaching for BIPA program students.

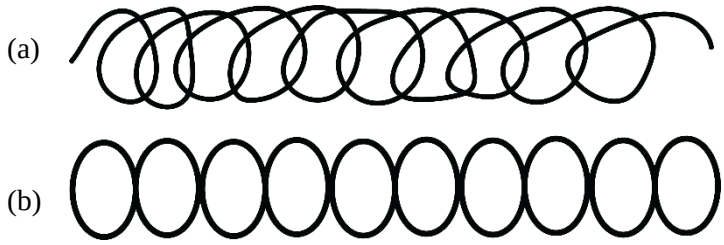

Figure 1
Vocabulary Teaching Principle that is Recommended (a) and Avoided (b)

Vocabulary taught to BIPA program students should not be taught separately but be included in the context of actual and natural word use. This means that the vocabularies are to be used in natural sentences that are actually used by BIPA program students. Examples of sentences are those adjusted to BIPA program students' learning objectives. We may understand here that BIPA program students learn the Indonesian language with different objectives. Some simply want to talk to Indonesian friends; some want to take study and title at an Indonesian school; and some intend to do research, etc. These differences should be responded to carefully with regard to the learning material presented, which will, in turn, influence the vocabulary to be taught. Making artificial sentences, far from the context of use as BIPA program students need, is to be avoided. For example, sentences *Ihsan memukul anjing*, *anjing mengejar kucing*, and the like (for old primary school students) should be avoided since BIPA program students will never use such sentences. Below are examples of actual and natural sentences as per the BIPA program students' objective.

Some techniques are suggested for vocabulary teaching so that the vocabulary will be well retained in BIPA program students' memory and have communicative elements. Vocabulary learning is performed by showing items directly, item replicas, and pictures of the vocabulary. Teaching vocabulary, especially difficult or new vocabulary, within teacher's and student's reach, teachers are suggested to show the concerned items or goods by showing the items or goods directly. This vocabulary learning technique can also proceed to Words Grouping, where BIPA program students group vocabulary into certain categories, such as noun group, verb group, adjective group, etc., as given in the examples below.

Noun

[12] *negara dan kota* [country and city] → these general nouns are used to refer to an item in general.

[13] *kampus dan tanah* [campus and land] → Base nouns, also known as original nouns, that can be recognized by referring to the form of an item without any affix.

Verb

[14] *makan dan minum* [food and drink] → Base verbs, verbs without any affix or root words

[15] *membaca dan bertemu* [reading and meeting] → Derivative verbs, verbs with an affix(es) or subject to pluralization

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] → Limitative adverbs describe meaning related to limitation.

Vocabulary learning can give an illustration of vocabulary with certain actions or activities. When teaching or explaining a difficult vocabulary, the teacher is suggested to give illustrations to BIPA program students by doing an act or activity that describes the vocabulary. This technique is suggested to explain the vocabularies of abstract nouns, verbs, and adjectives. This vocabulary teaching technique is often called the Tarzan language technique. Below are examples of vocabulary to teach.

Adjective

[18] *bersih dan nyaman* [clean and comfortable] → Property giving adjective state physical or mental quality and intensity.

[19] *ringan dan banyak* [light and many] →

Size adjective states measurable quality through the quantitative measure.

Vocabulary can also be taught to repeat vocabulary pronunciation clearly and slowly. When BIPA program students find it difficult to understand how a teacher pronounces certain vocabulary, the teacher should keep pronouncing the same sentence, not replacing vocabulary that is not understood by the student with another vocabulary. What the teacher should do is clarify the pronunciation by slowing it down. Example: *Mau berangkat dengan siapa?* When BIPA program students do not understand the word *berangkat*, the word *berangkat* should not be directly replaced with *pergi*, but the sentence containing the word *berangkat* is to be repeated. The repetition is intended for the BIPA program students to learn the word *berangkat* instead of replacing it with *pergi*.

Just like the example above, the teacher should not simply say the difficult vocabulary but also write the vocabulary. It is necessary to do so that the BIPA program students would know how to pronounce and write the difficult vocabulary. This is intended to involve some senses in a learning activity. Involving some senses in learning vocabulary will strengthen the vocabulary learned in the BIPA program students' memory; thus, the vocabulary will be well retained and not quickly lost from the BIPA program students' memory.

If the BIPA program students find difficulty in certain vocabulary, the teacher can explain the vocabulary by giving the difficult vocabulary antonym. Understanding vocabulary and its antonym can sharpen the BIPA program students' memory. Sometimes BIPA program students only remember certain vocabulary, and when it is understood along wing its antonym, more vocabularies are to be expectedly to be retained by them. Examples of antonyms are as follows:

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

In case BIPA program students find difficulty in certain vocabulary, the teacher can explain the vocabulary by giving synonyms to the difficult vocabulary. With some vocabulary for certain things, BIPA program students are expected to understand various vocabulary better simultaneously, for that matter. Forgetting to eat, the students can say a synonym of the concerned vocabulary.

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

In case BIPA program students find it difficult to remember a vocabulary, the teacher should not directly give the vocabulary but give them bait for them to remember, such as the word '*berjalan'*. The teacher gives bait by informing that there are eight letters, the first letter is /b/ and the third letter /j/, and the last letter /n/. The mention of bait with one or several letters from the vocabulary is intended so that students can guess the vocabulary correctly.

The technique of writing difficult and new vocabulary on cards is very good for helping students remember the vocabulary. There are two methods in this case. First, simply write vocabulary on a card while the opposite side is left empty. Second, writing the difficult vocabulary on a card, and the opposite side is written with how to write in the Indonesian language and their translation. These cards may serve as a reminder, such as dividing into two groups. The first group consists of a vocabulary group whose meaning is remembered, and the second group consists of a forgotten vocabulary group. The second group, the forgotten vocabulary, is to be remembered again. After this process, BIPA program students divide the vocabulary into two, remembered vocabulary and forgotten vocabulary. This goes on until the second group consisting of forgotten vocabulary gets empty, and all cards move to the remembered vocabulary group.

The technique of remembering vocabulary without writing them can be performed when there are at least two students in a class. In this technique, a list of vocabulary is given for the students to memorize. Student one mentions a number of vocabularies that he has memorized while the other student gives guidance. For an example of giving guidance, suppose there are four random words, and students are asked to put them in order. Another technique is to construct words by singing 'what are you doing' song 'what are you doing', listening carefully to mark whether the vocabularies mentioned are the same or different from their memorized vocabulary. The technique to use: (a) student one gives a number of vocabulary to

the other student, (b) the student exchanges vocabulary with the other, especially vocabulary that is not remembered by the other student. The technique of remembering vocabulary is suggested in case there is more than one student or a group of students in class. In the case of only one student, the teacher is suggested to play the role of the other student as a co-learner.

Vocabulary learning techniques can also be performed by *Arrange the Words*. If students are aware of the meaning of vocabulary as the learning target, they are to arrange the vocabulary into a sentence. This word-arranging technique can be performed freely, and students are asked to make sentences by themselves or by giving guidance. For an example of a guidance-giving technique, suppose there are four random words, and students are to put them in order. The other technique to use is to arrange words by singing a song 'what are you doing'. This technique can also be combined with the Guess the Word aiming to remind IPA program students of certain vocabulary. The techniques that can be chosen and used include: guessing the free word (cloze model), guessing the mysterious word, and playing *hangman*. A vocabulary teaching technique can also be Find the Word. In a text, BIPA students are asked to find certain vocabulary, such as an adjective or verb with an affix(es) *me-kan, di-kan, ter, ber*, etc. Students can also be asked to write certain vocabulary from a dialogue for technical vocabulary in the fields of economy, politics, culture, etc.

Vocabulary teaching technique or BIPA program students can also use the *Match the Words* with similar vocabularies. This word-matching technique is performed when there is some vocabulary in front of students, another student or teacher reads a certain vocabulary, then the students match it by choosing which vocabulary is relevant to what is talked about. This technique is recommended for listening. BIPA students can learn vocabulary by developing root words into formed words, and vice versa with existing formed words, and students are asked to break the word down into root words and construct its affix(es). Vocabulary learning with this technique is very important since the Indonesian language has uniqueness in affix(es) formation.

The explanation above is the description of vocabulary learning principles and techniques for BIPA program students. With the complicatedness of various vocabularies in the Indonesian language, it is necessary to design a comprehensive vocabulary learning method. A certain curriculum model determines the direction of the teaching program and describes the goals to be achieved in each learning activity. Syllabus design allows teachers to arrange teaching plans in line with the teaching goals to be achieved. A curriculum unit arranges language materials in line with language function, situation, grammar rule, and type of class activity (Kalinowski et al., 2019; Forey & Cheung, 2019; Pishghadam et al., 2021). The curriculum design model, which combines validity, language function, and language use situation aspects, is quite comprehensive that can be adopted to teach the Indonesian language to BIPA program students (Suyitno et al., 2019; Saddhono et al., 2019).

These activities can be made varied by asking some students to play as a waiter and others as the guest of a restaurant. When the guest orders food, the waiter responds with expressions such as *baiklah*, *mohon maaf*, *makanan/minuman itu tidak tersedia*, etc. The BIPA program students then exchange their roles and continue the activities. After performing these activities, the BIPA program students can be asked to compare which menus are preferred by Indonesians for lunch (types and amount of food) from the habit in their own country (Lin, 2013; Parvaresh & Dabghi, 2013). This way, cultural differences between countries can be identified. This activity accelerates a direct, in-depth understanding of Indonesian vocabulary for BIPA program students.

This research using corpus linguistics is very helpful in identifying the corpus data to be analyzed. This data corpus can come from any language in the world. With this data corpus, researchers can choose what topics will be analyzed and studied. In this study, the corpus of data in the form of vocabulary produced by BIPA students can be used in the development of learning about Indonesian vocabulary by foreign speakers. Other researchers can also use the results of the corpus data for other aspects and what contexts they want. For example, it will examine language errors, morphological processes, syntactic studies, discourse analysis,

the specific use of words or terms, and others. The research context of these examples is from various languages, not only in Indonesian. This study is only an example of the use of corpus linguistics in the study of Indonesian. Manual language data study and research must be abandoned and must switch to using digital technology such as utilizing corpus applications such as this study. The advantage of this corpus application, which, when compared to several corpus applications in the world, is that this corpus application can process the desired target corpus according to the object of the research corpus from various languages in the world and from various contexts.

Based on the results of the analysis of language use in the introductory discourse of BIPA program students at Fatoni University and Yale University using the corpus linguistics application *Kortara*, it is found that there is dominant mastery of vocabularies by the students over nouns out of the nine types of Indonesian words. Out of the vocabularies, 11 vocabularies have the highest frequency, including *saya*, *di*, *dan*, *suka*, *adalah*, *kota*, *dekat*, *belajar*, *dari*, *ke*, and *yang*. The word *saya* is certainly the vocabulary that appears the most since the discourse made by them is an introduction; thus, the *saya* is dominant in self-introduction. There are many vocabularies produced by the BIPA students, but mistakes are found in their writing due to a lack of strictness and carefulness in the editing process. Based on the result of the analysis of the *Kortara* application covering 1,947 vocabularies by the 51 students, we may arrange an effective and efficient vocabulary learning principle and technique for BIPA students to introduce types of words and variations of word formation in the Indonesian language. With mastery of abundant vocabulary, BIPA learning will run well as per students' needs based on their mastery of words. Therefore, the results of analysis on the corpus linguistics application Kortara are quite helpful in the BIPA learning process so that it is directed better that the students' linguistic ability is taken as the basis of vocabulary learning in BIPA for maximum outcome.

Acknowledgments

We would like to thank the Institute for Research and Community Service, Universitas Sebelas Maret, and the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for supporting this research.

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

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. *World Science*, *5*(5), 27-29.

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*, *52*(21), 261-274. https://doi.org/10.33541/jet.v5i1.956

Dunder, I., & Pavlovski, M. (2018). Computational concordance analysis of fictional literary work. In K. Skala, M. Koricic, T. G. Grbac, M. Cicin-Sain, V. Sruk, S. Ribaric, S. Gros, B. Vrdoljak, M.

Mauher, E. Tijan, P. Pale, & M. Janjic. (Eds.), *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

Durrant, P., Brenchley, M., & McCallum, L. (2021). *Understanding development and proficiency in writing: Quantitative corpus linguistic approaches*. Cambridge University Press.

Egbert, J., & Baker, P. (2019). *Using corpus methods to triangulate linguistic analysis*. Routledge.

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*, *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.

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.

Murtisari, E. T., Fabrian, D. D., Lolyta, R. D., Lukitasari, D. R., & Rahardjono, V. C. (2019). The use of Indonesian first-singular-pronouns by students interacting with teachers: Saya or Aku?. *Kajian Linguistik Dan Sastra*, *4*(1), 79-90. https://doi.org/10.23917/kls.v4i1.7811

Naji Meidani, E., Makiabadi, H., Zabetipour, M., Abbasnejad, H., Firoozian Pooresfehani, A., Shayesteh, S. (2022). Emo-Sensory communication, emo-sensory 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*, 1-17. 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), 1-13. https://doi.org/10.1016/j.intman.2019.100729

Rădescu, R. (2021). Concordance techniques in lossless data compression of text files. In M. Alexandru, M. Mihaela, O. P. Mihai, S. Alexandru, A. Mihaela, R. Vladimir, & T. Lucian (Eds.), *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., & Wahyono, H. (2019). Learning vocabulary using multimedia-based teaching Indonesian to speakers of other languages (TISOL). *Journal of Physics: Conference Series*, *1339*(1), 012108. https://doi.org/10.1088/1742-6596/1339/1/012108

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*(1), 012061. https://doi.org/10.1088/1742-6596/1254/1/012061

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*(1), 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 az-ziro'ah sebagai media pembelajaran bahasa Arab [Optimizing the use of corpus linguistics in compiling the az-ziro'ah dictionary as a medium for learning Arabic]. *ISoLEC Proceedings*, *4*(1), 123-128.

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. https://doi.org/10.21831/cp.v38i3.27045

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

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/10.1177/1362168818799

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. https://doi.org/10.26499/jbipa.v3i2.4125

Yadav, H., Vaidya, A., Shukla, V., & Husain, S. (2020). Word order typology interacts with linguistic complexity: A cross-linguistic corpus study. *Cognitive Science*, *44*(4), e12822. https://doi.org/10.1111/cogs.12822

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. https://doi.org/2018-18436-018