

Creating an Islamic boarding school English corpus: corpus metadata, frequently used words, and unique words

Yulia Agustina^{1,2}, Pratomo Widodo¹, Margana Margana¹

¹Department of Language Education Science, Faculty of Languages, Arts, and Cultures, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

²Department of Early Childhood Education, Faculty of Education, Universitas Hamzanwadi, Lombok Timur, Indonesia

Article Info

Article history:

Received Jul 7, 2023

Revised Sep 14, 2023

Accepted Oct 23, 2023

Keywords:

Corpus metadata

Frequently words

Islamic boarding school context

Leipzig corpora

Unique words

Voyant tool

ABSTRACT

In the current era, the use of corpora in language teaching is mainly explored in English classes as it has become a trend in education. Hence, this research aimed to identify the corpus metadata, frequently used words, and unique words related to the Islamic boarding school context to be used in the English instructional process. This research employed a mixed method combining quantitative and qualitative data analysis methods. Two English Islamic boarding school books, several articles covering the scope of Islamic boarding school, and students' speech texts were selected as the data. Then, they were analyzed using the Voyant tool. The finding showed total words of 49,970: 5,417 specific words, 0.108 vocabulary density, and a 12,980-readability index. The finding will be incorporated into instructional resources for developing Islamic boarding school students' general and/or specialized vocabulary. The words, in particular, will provide a foundation for students in constructing Islamic speech texts, delivering speeches, and using English in an Islamic boarding school environment.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Yulia Agustina

Department of Language Education Science, Faculty of Languages, Arts, and Cultures

Universitas Negeri Yogyakarta

Yogyakarta, 55281, Indonesia

Email: yulia0012pasca.2019@student.uny.ac.id

1. INTRODUCTION

Using corpora in language teaching has become a trend in education and is primarily discussed in English instruction nowadays. A corpus (sing.) is a collection of texts, spoken or written, stored electronically [1]. In addition, it is a set of texts, either written or spoken, stored on a computer using an application compiled for particular purposes [2], [3]. Another opinion comes from McEnery and Hardie [1] stating that a corpus is a sizable, ethical collection of writings that is naturally occurring and is meant to be representative of a particular language or language variant. Hence, a corpus is a systematic collection of texts representing a language or language variety. As a result of its careful design, it is a helpful tool for linguistic analyses, language samples, language instructions, and a variety of other language-related studies and uses.

As stated earlier, corpora (plural from corpus) can be formed by spoken or written texts. A spoken corpus takes considerably longer to build because speech has to be transcribed and possibly coded for some of its non-verbal features. On the other hand, written corpora can be made very quickly using the internet as a

source [2]. Building a corpus from written language requires some steps: creating a design rationale, input text, and database text [2]. Design rationale indicates activity in deciding what corpus is to be made and how many texts and sources are needed. Meanwhile, input text refers to re-type and scanned activity before uploading the tool. Then, database text refers to tracing the text used in making the corpus. Based on these considerations, this present study focused on the written language, particularly the ones made by several researchers who had discussed Islamic boarding school and Islamic values.

Furthermore, operating corpora requires the tools and an internet connection to download the software. Some corpora can be used freely while some others commercially, such as the corpus of contemporary American English (COCA), AntConc, British national corpus (BNC), WordSmith tools, TIME magazine corpus of American English and MonoConc. According to O'keeffe *et al.* [2], the overview of the basic analytical activities in corpora includes word frequency count, concordance, and collocation. A concordance is a list of the terms found in the collection of texts that includes information on where and how frequently each word appears in the text while collocation refers to a set of two or more words that typically go together and constitute a natural combination of words that are closely related to one another. Moreover, the basic corpus linguistics techniques contain concordance, wordlists or word frequency, keyword analysis, and cluster analysis: i) key word in context (KWIC) of concordance refers to the method of setting up and presenting textual information to highlight the context in which specific keywords occur, ii) wordlist is a compilation or grouping of words that are typically arranged alphabetically, iii) keyword analysis is to identify keyword in the text; and iv) cluster analysis is the process of grouping data into coherent groups or clusters to find significant patterns or structures within a dataset. Those corpus linguistics techniques can be used based on the researchers' needs in building the corpus.

Additionally, teaching English using corpora can be used to create instructional materials, language tests, grammar exercises, classroom activities, and syllabus designs [4]. Teachers usually integrate corpus into their teaching in three ways: first, they get information through corpus searches; second, they produce materials on a level basis; and third, they have students work with these materials. Moreover, teachers can generate specialized corpora from authentic texts or students' papers and then assign them to analyze the data [4] or use the online corpora that are already accessible to teach a particular language pattern. Leech [5] also elaborates on using direct corpora by mentioning teaching about, teaching to exploit, and exploiting to teach. According to McEnery and Xiao [6], teaching refers to the academic study of corpus linguistics and linguistics concepts like syntax and pragmatics. Meanwhile, teaching to exploit relates to the student's practical experiences and knowledge that enable them to utilize the corpora for their needs, which means that educational activity focuses on the students. Lastly, using the corpus-based technique to teach courses in sociolinguistics and discourse analysis belongs to exploiting to teach [6].

According to Hunston [7], a corpus benefits researchers and language learners since it effectively informs people what language is like. In the context of English teacher training, a corpus has great potential to help teachers design effective teaching activities as it provides authentic language data/examples and collocation learning [8]. Further, Fauzi [9] states that teaching and learning based on the corpus-based approach should be utilized to assist the beneficial effects of employing corpora in vocabulary instruction because it focuses on learning words and how to keep them in long-term memory, then use them in speaking. It aligns with the needs of Islamic boarding school students, who require vocabulary mastery to communicate and interact in English daily. Nonetheless, existing literature suggests that many students have difficulties in learning English vocabulary and studying the language in general [10]. As languages are built on words, teaching vocabulary is essential to language learning [10]. Therefore, corpus linguistics gives new insight into language learning, which provides help like a dictionary to find familiar, unfamiliar, and word arrangements. The primary argument favoring the employment of a corpus is that it is a more trustworthy guide to language use than native-speaker intuition [9].

Thus far, several researchers have created a remarkable corpus to produce specific vocabulary lists for particular purposes: Islamic religious studies textbooks vocabulary (IRSTV) corpus [3], nursing [11], English for young learners [12], medical English [13], and tourism [14], [15]. It can be concluded that the existing literature focused on the manufacture of the corpus. Although this present study has similar purposes and interests in creating a corpus, it differs from the previous ones as it concentrates on applying a unique word list in English language instruction in Islamic boarding school, an Islamic learning center providing faithful formal education and religious teaching [16]. Islamic boarding school is also defined as educational establishments that respect and preserve both the scientific tradition and morality of Muslims where the students spend most of their time; hence, they can live and learn about Islam from the cleric [16], [17]. By far, English instruction in Islamic boarding school had used general English similar to English instruction in other public schools, which did not fully consider Islamic boarding school students' needs. It is necessary that Islamic boarding school students learn vocabulary related to Islamic boarding school context. Relevant vocabulary items will be used for more specific purposes such as constructing speech texts, delivering speeches on designated days, and using English in Islamic boarding school environment. Some of the

context-relevant words include *makmum* (congregation), *wudlu* (ablution), *riya* (showing off), fasting, prayer, and recite. Therefore, the English corpus Islamic boarding school needs to be made to ascertain the uniqueness of vocabulary related to the Islamic boarding school context. The research questions can be found: i) what is the metadata for building the Islamic boarding school English corpus?, ii) what are the most frequently used and unique words in the Islamic boarding school English corpus?, and iii) how are lexical terms and their collocation used in the Islamic boarding school context compared to Leipzig corpora?.

2. METHOD

The method used in this study was a multiple-method design that combined both qualitative and quantitative [18]. Creswell and Creswell [19] states that mixed-method research is a method used for conducting research involving collecting, analyzing, and integrating quantitative and qualitative data. The qualitative method refers to the information in textual forms which were analyzed using qualitative data analysis techniques. In contrast, the quantitative method refers to the numerical forms analyzed using quantitative data analysis techniques. It was applied to complement one another to present a clear and complete description to provide comprehensive explanations.

2.1. Data collection

The researchers collected several English books and articles explaining Islamic boarding school's scope. The books were written by Solahudin [20], which described the Islamic creativity of Daarut Tauhid Islamic boarding school in Bandung, West Java, and Srimulyani [21], which described the women from traditional Islamic educational institutions in Indonesia. Then, some relevant articles were also used [22]–[24]. These sources were chosen because they explained the Islamic values in Islamic boarding school and were considered to already represent the data of the Islamic boarding school being sought. This study aimed to discover the frequent and unique words in the Islamic boarding school English corpus. The researchers used these data to establish a corpus-based English instructional model for Islamic boarding school students.

2.2. Procedures of the study

Corpus approach comprises three primary characteristics: i) practicality, significance, and morality, ii) extensive use of computer analysis, and iii) reliance on qualitative and quantitative analytical techniques [3], [14]. In addition, three factors must be considered while constructing a corpus: the collection must be ethical, contain authentic texts, and be electronically maintained [25]. To compose this corpus, the researchers took the steps of collecting two books and several articles related to the Islamic boarding school context, choosing representative texts that suit the needs, and then uploading them to the corpus tool, Voyant. The activities required extensive use of computers and were analyzed qualitatively and quantitatively.

2.3. Corpus tool

The use of corpus is always related to software analysis mediated by computers. There are many software analyses to examine the language data, such as Voyant, AntConc, Slect engine, MonoCon Pro, and WordSmith Tool. In this study, the Voyant application was chosen as the primary tool. This straightforward tool covered all the researchers' needs with one click. The only way to view word frequency lists, frequency distribution plots, KWIC displays, and unique words. is to upload the texts, which included all language data from the books and articles, to the device and then press enter. The initial results indicate that there are 49,970 total words associated with the word Islamic boarding school. To narrow them down, 5,417 unique words are found that can be used in the analysis of teaching materials. Meanwhile, the vocabulary density is 0.108, which can be useful as a measurement of vocabulary usage in comparison to the length of the texts. Finally, the readability index, with a number of 12,980, shows the estimation of how difficult the text is to read. The data of Islamic boarding school English corpus has been presented in Table 1.

Table 1. Islamic boarding school English corpus

Corpus name	Total words	Unique words	Vocabulary density	Readability index
Islamic boarding school English corpus	49,970	5,417	0.108	12,980

3. RESULTS AND DISCUSSION

The research results are presented in sub-sections along with the discussion. The results are presented in such a way as to answer the research questions. The data presented is the authentic data taken from the written sources described above.

3.1. Corpus metadata

Islamic boarding school English corpus was constructed from two English Islamic boarding school books and several articles covering the scope of Islamic boarding school. The books were chosen from international publishers, A New University (ANU) Press and Amsterdam University Press while the papers were selected from credible Indonesian journals. However, not all the texts were included in the analysis; only those represented the texts following the purpose of this study were selected. The first thing to do in building this corpus was to decide on a design rationale and how many texts and sources were needed for creating the corpus [2]. After collecting the required texts, the next steps were re-typing and scanning the written text from the sources. In converting the data that had been scanned, the researchers uploaded the document, and the results could be seen easily. The Table 2 consists of the corpus metadata in designing English corpus Islamic boarding school.

Table 2. Corpus metadata

No	Metadata	Descriptions
1	Corpus name	Islamic boarding school English corpus
2	Corpus language	English
3	Corpus size	49,970
4	Corpus type	Specialized corpus
5	Authors	Yulia Agustina, Pratomo Widodo, Margana Margana
6	Institution	Yogyakarta State University (UNY, Universitas Negeri Yogyakarta)
7	Aim	This corpus is a representation of the language used in the Islamic boarding school context. It consists of several writers and researchers in the field of Islamic boarding school, such as <i>pesantren</i> history, The existence of <i>Pondok Pesantren</i> (Islamic boarding school), Islamic values in <i>pesantren</i> , <i>pesantren</i> role, changes and future of <i>pesantren</i> , <i>I'tikaf and Lailatul Qodar</i> , the Al-hajj and the Umrah, <i>silaturrahmi</i> , <i>tausyiah</i> , moral decadence. This corpus created is part of the first author's dissertation. It is intended to get any unique words related to the Islamic boarding school context and then use them in teaching materials as part of the model developed.
8	Authors and Materials' titles	<ul style="list-style-type: none"> - Solahudin [20], the workshop for morality: the Islamic creativity of <i>Pesantren</i> Daarut Tauhid in Bandung, Java, 2008. - Srimulyani [21], women from traditional Islamic educational institutions in Indonesia negotiating public spaces, 2012. - Suhartini [22], the internalization of Islamic values in <i>Pesantren</i>, (December 2016). - Thahir [23], the role and function of Islamic boarding school: an Indonesian context, (April 2014). - Zakaria [24], <i>Pondok Pesantren</i>: changes and its future, (April 2010).
9	Publisher/website	Books: ANU E Press and Amsterdam University Press. Articles: Journal of Islamic and Arabic Education, TAWARIKH: International Journal for Historical Studies, Jurnal Pendidikan Islam (Islamic educational institutions concerning Islamic education).
10	Types	Books, articles, and English speech texts

3.2. Word list and unique words in Islamic boarding school English corpus

3.2.1. Most frequently used words

The most frequently used English words that are found in the Islamic boarding school English corpus are summarized in Table 3. It displays the estimated word calculation: 5,417 specific words, 0.108 vocabulary density, and 12,980 readability indexes. In total, there are 50 most frequently used words for Islamic boarding school English corpus. They comprise the words Islamic (616), *pesantren* (Islamic boarding school) (484), school (334), boarding (278), education (275), *pondok* (228), and Allah SWT (191).

3.2.2. Unique words

This subsection consists of the findings in regard to the unique words. Unique words are defined as words that are present in the target texts but uncommon or absent from the other texts in a corpus. Yet, they can represent a rather general phenomenon [26]. The number of unique words, 50, that are spread in Islamic boarding school English corpus has been summarized in Table 4 (in Appendix).

Table 4 shows that the word 'Islamic' is still the first unique word that occurred in the Islamic boarding school English corpus; this finding is similar to the one reported in Table 3. This means that the word *pesantren* can always be associated with the word 'Islamic'. Moreover, since the words in this corpus are more limited than the other corpora like COCA, NBC, Corpus Mate, or Leipzig corpora, it was then compared with one of them, the Leipzig corpora, particularly because it contains a collection of Indonesian texts. This comparison also aimed at providing additional information to readers or students at the time of their study in the classroom.

Nevertheless, there are four more powerful words in the English corpus Islamic boarding school than those in the Leipzig corpora, although the range is not much different, such as mosque (57:52), repentance (10:9), and preach (11:8). On the other hand, the range for the word proselytizing (20:2) is far

different which means that the word is equally powerful because it is more frequently used. However, when it comes to the reason for constructing a corpus, the quantity of the word is not a critical factor. The results of the most frequent words on the English corpus Islamic boarding school will be inserted later in the teaching materials.

Table 3. Fifty most frequently used words in Islamic boarding school English corpus

No	Words	Frequency in corpus	No	Words	Frequency in corpus
1	Islamic	616	26	Shalat	70
2	<i>Pesantren</i> (Islamic boarding school)	484	27	Muslim	69
3	School	334	28	Modern	68
4	Boarding	278	29	Muslims	64
5	Education	275	30	Followers	63
6	<i>Pondok</i>	228	31	Traditional	58
7	Allah SWT	191	32	Schools	57
8	Religious	187	33	Mosque	57
9	Educational	155	34	Development	56
10	<i>Santri</i> (students)	144	35	Society	53
11	Institution	131	36	Java	53
12	Kyai	130	37	Said	52
13	Values	127	38	Community	52
14	Islam	119	39	Technology	49
15	Knowledge	103	40	Study	49
16	Students	102	41	Social	49
17	Indonesia	87	42	Quran	48
18	Tauhid	85	43	Activities	46
19	Tradition	78	44	Teaching	45
20	Life	78	45	Institutions	45
21	Good	76	46	Value	44
22	Process	73	47	Muhammad	44
23	People	73	48	Human	44
24	World	70	49	Learning	43
25	Time	70	50	Prophet	42

3.3. Collocation

3.3.1. Collocation of some of most frequently used words

Collocation is a fixed or semi-fixed phrase formed by two or more words regularly occurring together in a particular order. O'keeffe *et al.* [2] state that collocation refers to three or more occurrences of words displayed in sentences. There are several examples of frequent word collocations, for example, Islamic, *pesantren*, and boarding. Each of these will be compared to the Leipzig corpora and elaborated.

a. Islamic

The word 'Islamic' is the most frequently used word in the Islamic boarding school English corpus. Some of the examples of the sentences consisting of the word 'Islamic' can be found in the following Figure 1. It can be inferred from the figure that the word 'Islamic' refers to nouns (N) or adjectives (Adj), which can be used in the beginning, middle, and end of a sentence. For example, in line 5, "...*Pesantren* is the oldest *Islamic* institution growing in this country..."; this means that the word 'Islamic' functions as an adjective in the sentence. Moreover, the word 'Islamic' is preceded by an article or conjunction and followed by adjectives or nouns. To conclude, this 'Islamic' word functions as a noun or adjective in the Islamic boarding school English corpus.

Meanwhile, the word 'Islamic' in the Leipzig Corpora is associated with Islamic centres, Islamic schools, Islamic villages, Islamic malls, Islamic preschools, and Islamic studies. The word 'Islamic' is closely related to the activities of the Islamic people. Slightly different from the English corpus Islamic boarding school, the word Islamic here only indicates an adjective (adj) followed by a noun (N) object. In detail, Figure 2 highlights how the word 'Islamic' is used in sentences in the Leipzig Corpora.

b. *Pesantren*

The word '*pesantren*' can be defined as Islamic boarding schools or places of recitation activities [27]. When it comes to a part of speech, Figure 3 explains that *pesantren* is the name of a place and is preceded by the preposition place and conjunction. For instance, at the end of the sentence in the 10th line, "conducted the study at various *pesantren*, this paper presented result...", it can be seen that the word '*pesantren*' is placed after a preposition of place, *at*. Moreover, the word '*pesantren*' is usually found in the beginning and at the end of sentences. Figure 3 consists of examples of the word '*pesantren*' in the developed corpus.

Voyant Tools			
Contexts			
Document	Left	Term ↑	Right
file untu...	in fostering morality of santri	islamic	boarding school students in Pesantren
file untu...	boarding school students) in Pesantren (islamic	Boarding School) Miftahul Muhajirin Cidadap
file untu...	recite Yasin after dawn. The	islamic	internalization process of Islamic values
file untu...	The Islamic internalization process of	islamic	values takes several steps namely
file untu...	school and through education in	islamic	boarding school. Pesantren is the
file untu...	school. Pesantren is the oldest	islamic	institution growing in this country
file untu...	under the guidance of kiai (islamic	teacher), and it has a
file untu...	results of studies related to	islamic	values internalized in fostering akhlak
file untu...	santri; the internalization process of	islamic	value in fostering akhlak santri
file untu...	and guided the akhlak of	islamic	boarding school students of Miftahul
file untu...	Miftahul Muhajirin Subang through the	islamic	values internalization. The data was
file untu...	and questionnaires. DISCUSSION Akhlak in	islamic	Point of View In Islam
file untu...	the actualization of a personal	islamic	and faith. A personal character
file untu...	or his definition toward the	islamic	faith and values attached to
file untu...	Islam is identical with the	islamic	religion implementation in all areas
file untu...	done through the implementation of	islamic	values internalization strategy and the
file untu...	syukur (Gratitude); h) sabar (patient).	islamic	values, the essence, captured from
file untu...	work. istiqamah. Ikhlas. and patience.	islamic	values can be internalized through

Figure 1. The word 'Islamic' in Islamic boarding school English corpus

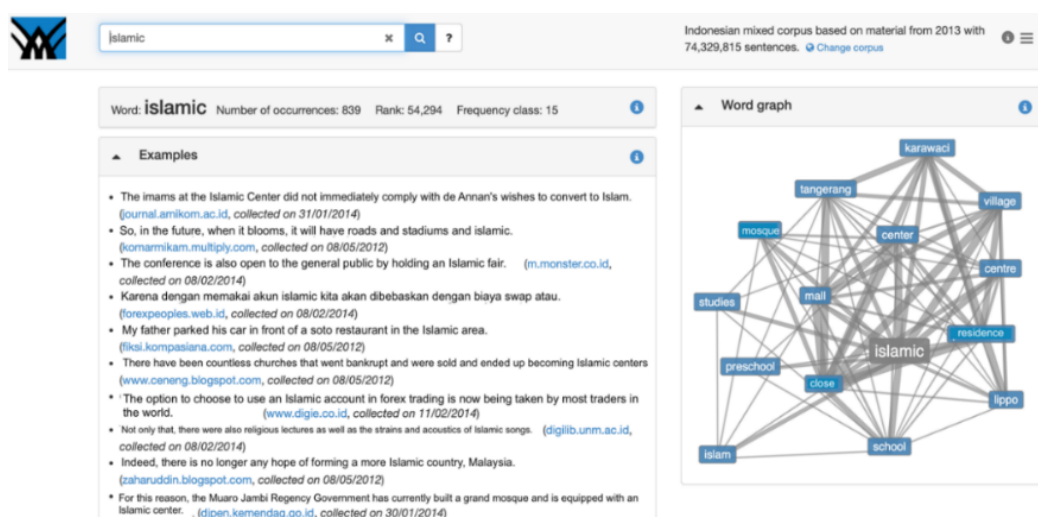


Figure 2. The word 'Islamic' in Leipzig corpora

Voyant Tools			
Contexts			
Document	Left	Term	Right
file untu...	INTERNALIZATION OF ISLAMIC VALUES IN	pesant...	ABSTRACT This article aims to
file untu...	Islamic boarding school students) in	pesant...	(Islamic Boarding School) Miftahul Muhajirin
file untu...	education in Islamic boarding school.	pesant...	is the oldest Islamic institution
file untu...	in Indonesia history in Indonesia,	pesant...	has been proven to be
file untu...	to create noble generations. In	pesant...	, the santri obtain a continuous
file untu...	facilitates the head of the	pesant...	, kiai and teachers to be
file untu...	the property owned by the	pesant...	. The values of togetherness with
file untu...	survive Development of akhlak in	pesant...	can be applied through internalization
file untu...	The internalization process in each	pesant...	has its model and advantage
file untu...	conducted the study at various	pesant...	, this paper presented the results
file untu...	seeing. Fostering Akhlak Santri in	pesant...	One of the akhlak santri
file untu...	the akhlak santri enhancement at	pesant...	is done through the implementation
file untu...	by them. Islamic values in	pesant...	Miftahul Muhajirin Subang From the
file untu...	interviews with the head of	pesant...	and the observations at Miftahul
file untu...	in Fostering Students' Akhlak in	pesant...	Miftahul Muhajirin Subang Internalization of
file untu...	kiai and santri in this	pesant...	was conducted with three patterns
file untu...	attitude and habit as in	pesant...	. As well as in muwajjahah

Figure 3. The word 'pesantren' in Islamic boarding school English corpus

At the Leipzig corpora, the word ‘*pesantren*’ is always accompanied by *pesantren Nahdlatul Ulama* (NU), an Islamic organization in Indonesia, *pesantren santri*, *pesantren boarding*, *pesantren kiai* (an expert in Islam), *pesantren school (madrasah)*, and Islamic *pesantren*. In this case, the word ‘*pesantren*’ can be written before or after another word. In grammar, this structure is called the noun phrase. In a sentence, it serves as the subject, the object, or the complement. A noun phrase is a collection of words that identifies or labels a person, place, thing, or idea [28]. For more clarity, look at Figure 4 for more sentences with the word *pesantren* according to Leipzig corpora.

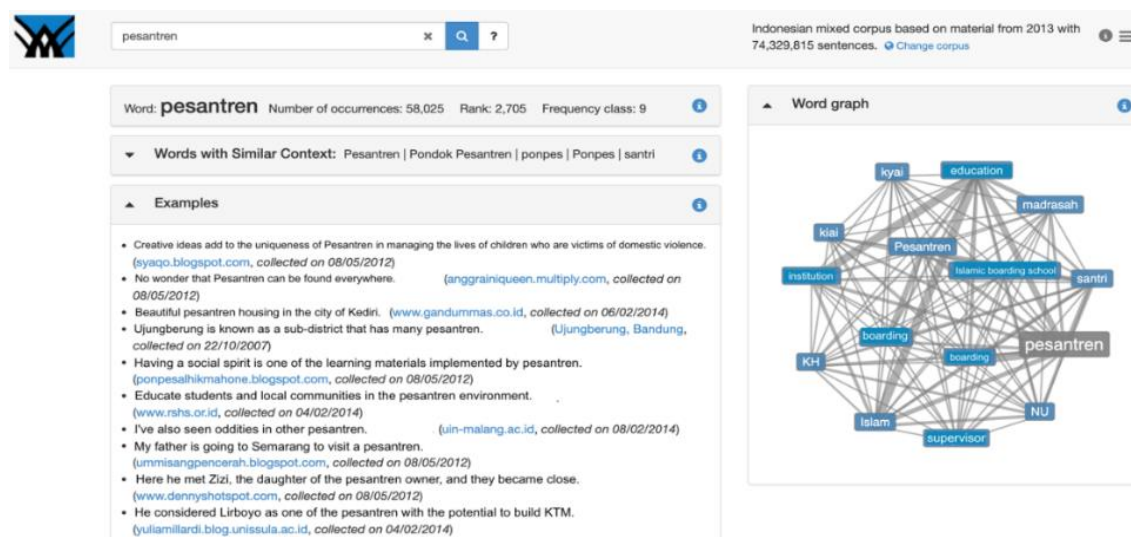


Figure 4. The word ‘*pesantren*’ in Leipzig corpora

c. Boarding

The word ‘boarding’ refers to a place or a dormitory where students learn to live and socialize in various aspects of social life [29], [30]. In simple terms, it is where students live in the school neighborhood. In more detail, the word ‘boarding’ can be preceded by an adjective, article, or link clause. To be precise, the examples can be found in Figure 5 in line 1, line 5, and line 6, respectively. Another finding from the English corpus Islamic boarding school is that the word ‘boarding’ is always followed by “school”. The examples of the sentences consisting of the word ‘boarding’ can be located in Figure 5.

Voyant Tools				
Contexts				
Document	Left	Term	Right	
file untu...	fostering morality of santri (Islamic	boarding	school students) in Pesantren (Islamic	
file untu...	school students) in Pesantren (Islamic	boarding	School) Miftahul Muhajirin Cidadak, Pagaden	
file untu...	and through education in Islamic	boarding	school. Pesantren is the oldest	
file untu...	supported by the fact that	boarding	schools are the traditional educational	
file untu...	within the scope of the	boarding	school community can learn together	
file untu...	only with the conviction that	boarding	schools are local agencies that	
file untu...	the history of the trip	boarding	schools. Communities and schools like	
file untu...	guided the akhlak of Islamic	boarding	school students of Miftahul Muhajirin	
file untu...	is simply understood as a	boarding	school which carries on the	
file untu...	educational system and system seed	boarding	school education, relevant done as	
file untu...	with cultural local is Islamic	boarding	school, yet in general, the	
file untu...	knowledge in the future, Islamic	boarding	school is more strengthening or	
file untu...	be a "reflection" that Islamic	boarding	school is urgent to revival	
file untu...	competitive (Mastuhu, 1999: 276), Islamic	boarding	school is faced by the	
file untu...	constructing the religious society. Islamic	boarding	school in the future is	
file untu...	demand of collaboration of Islamic	boarding	school with favorite school is	
file untu...	weakness (Muhaimin, 2009: 105). Islamic	boarding	school is assessed as the	

Figure 5. The word ‘boarding’ in Islamic boarding school English corpus

In contrast, the word ‘boarding’ in Leipzig corpora is mostly associated with airport activity. It is one of the operational processes by which passengers get into the aircraft after completing check-in procedures [31]. It is further proven by some possible word constructions such as boarding pass, boarding gate, boarding room, boarding system, boarding lounge, and boarding check-in. That being said, some sentences associated with Islamic boarding schools can still be found in examples 1 and 7 in the following Figure 6. From the examples, the word tends to be followed by a noun, and all sentences with the word ‘boarding’ indicate to have a noun phrase structure. The examples of the word ‘boarding’ in the Leipzig corpora can be identified in Figure 6. Thus, it can be concluded that ‘boarding’ in Islamic boarding school English corpus relates to boarding activity in Islamic boarding school. Meanwhile, although it still has a similar meaning, such a word in Leipzig corpora is dominant for boarding activity in the airport.

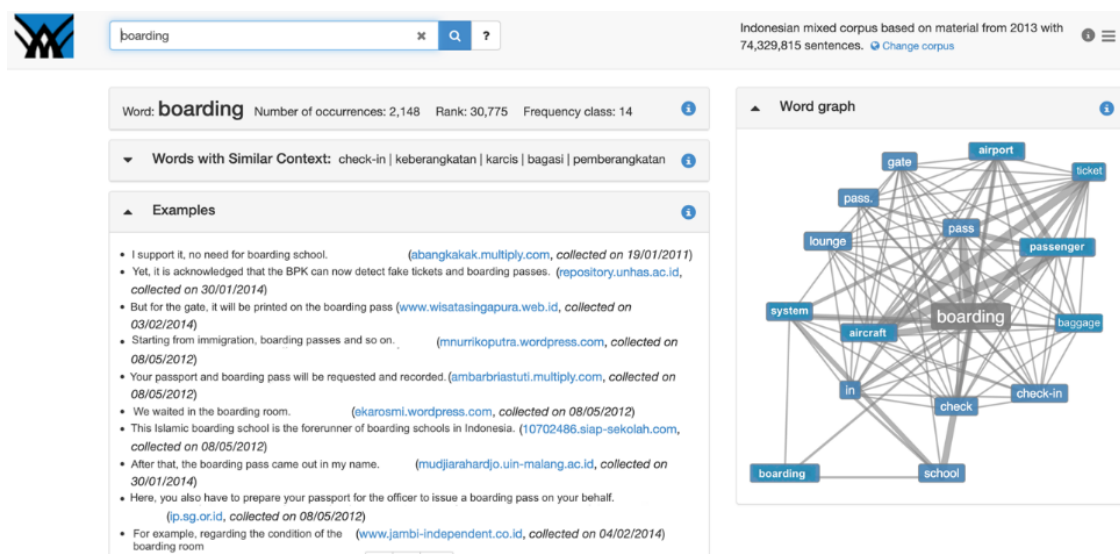


Figure 6. The word ‘boarding’ in Leipzig Corpora

3.3.2. Collocation of some unique words

There were some unique word collocations produced by the Islamic boarding school English corpus. Three examples of collocation of unique words that belong to Islamic boarding school English corpus and Leipzig corpora will be elaborated. The words include resignation, Allah Swt., and recitation.

a. Resignation

The word ‘resignation’ or known as *tawakal* is the attitude of surrendering everything to God, after an effort has been made, and believing that whatever happens in the world will never happen without God's intervention [32]. In the sentences in Figure 7, most of the word ‘resignation’ indicates a noun such as in line 3, faith, Islam, charity, piety, sincerity, resignation, gratitude, patience, honesty, fairness, and responsibility. Meanwhile, in line 5, ‘resignation’ can function as a noun phrase. Despite the part of speech they represent, this word is not mentioned much in this corpus as can be seen in Figure 7.

Unlike the Islamic boarding school English corpus, the word ‘resignation’ is mentioned 55 times in the Leipzig corpora. This word always comes with the word syndrome, letters, and addresses. More interestingly, the word ‘resignation’ here has three meanings according to the context: i) refers to people's guilt in the crime; ii) indicates the option for the minister; and iii) shows an official letter written by an employee to inform his intention to resign from a position or job in place. For more details, the use of the word ‘resignation’ can be seen in the following Figure 8.

b. Allah

“Allah” is one word that often appears in Islamic boarding school English corpus. In the doctrine of Islamic teaching, Allah is the creator of the universe, including humans [33] so he is known as the God for Muslims. In the sentence structure, He refers to a noun that a preposition or connection clause can precede. If the word ‘Allah’ is in the initial sentence, it will automatically be followed by a verb. The instance can be seen in line 14, “On this night, Allah determines the future fate of...”. This sentence reveals that the word ‘Allah’ is followed by the verb ‘determines’. The remaining examples can be identified in Figure 9.

On the other hand, the word ‘Allah’ in the Leipzig Corpora still appears more than it does in the English corpus Islamic boarding school because of the data used, 32,196,275 sentences. The word ‘Allah’

concordance lines with the words: Miyetti, Almighty, Insyaa Allah, Messenger of Allah, Ansar Allah, and in the name of Allah. To conclude, both the English corpus Islamic boarding school and the Leipzig corpora describe Allah as the Lord of humanity, almighty, and the universe's creator. The words associated with Allah in Leipzig corpora can be seen in Figure 10.

Voyant Tools			
Contexts			
Document	Left	Term	Right
file untu...	ihsan; d) taqwa; f) Tawakal (resignation	; g) syukur (Gratitude); h) sabar
file untu...	faith, Islam, charity, piety, sincerity,	resignation	, gratitude, patience, honesty, fairness, responsibility
file untu...	faith, islam, ihsan, taqwa, ikhlas,	resignation	, gratitude, patience, honesty, fairness, responsibility
file untu...	Islam value described as a	resignation	and obedience to the rule
file untu...	is the Creator; with their	resignation	as a slave who should

Figure 7. The word 'resignation' in Islamic boarding school English corpus

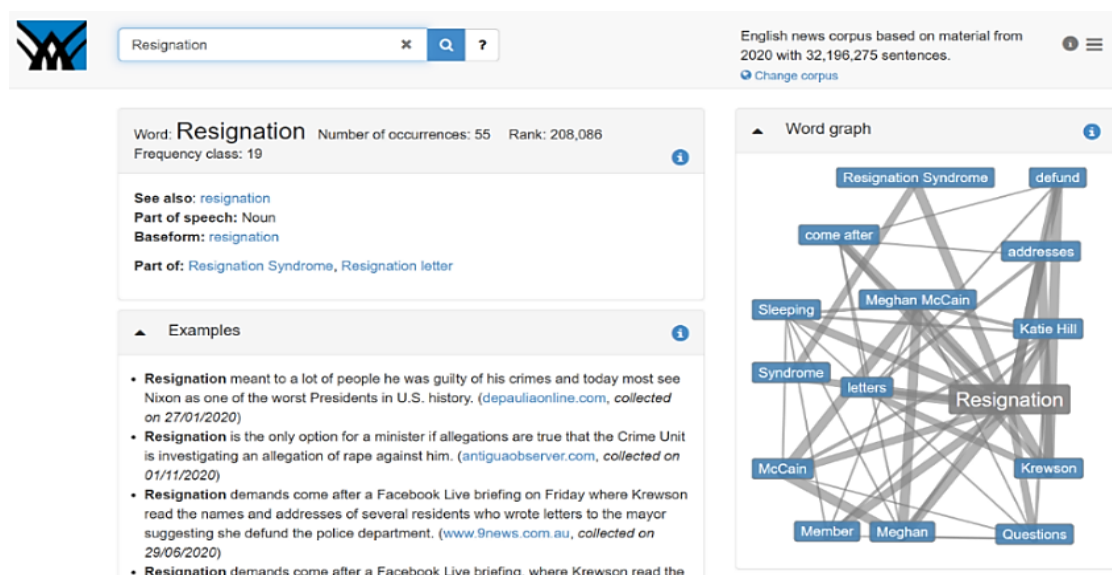


Figure 8. The word 'resignation' in Leipzig corpora

Voyant Tools			
Contexts			
Document	Left	Term	Right
file untu...	except in the name of	allah	While the act is defined
file untu...	akhlak explains relationship pattern with	allah	. Horizontally, akhlak lead lifestyles with
file untu...	have akhlak identified disobedience to	allah	and threatened with a painful
file untu...	the vertical dimension (relationship with	allah), horizontal (fellow creatures) and internal
file untu...	of religion and belief that	allah	is the All-seeing. Fostering
file untu...	inner attitude of trust to	allah	. Islam value described as a
file untu...	abandoned, because besides forbidden by	allah	these, also, will bring madlarat
file untu...	awareness and ikhlas because of	allah	. In sunnah fasting every Monday
file untu...	prayer with their belief that	allah	is the Creator; with their
file untu...	with being watched feeling by	allah	and doing all things because
file untu...	doing all things because of	allah	. They were also conditioned to
file untu...	really confidence and believe that	allah	is exist and Allah is
file untu...	that Allah is exist and	allah	is the Watcher. Akhlak lived
file untu...	lailatul qadar. On this night,	allah	determines the future fate of
file untu...	that it is considered by	allah	as better than a thousand
file untu...	one he needs to approach	allah	. One such way is the
file untu...	is to be blessed by	allah	a thousand fold. It was

Figure 9. The word 'Allah' in Islamic boarding school English corpus

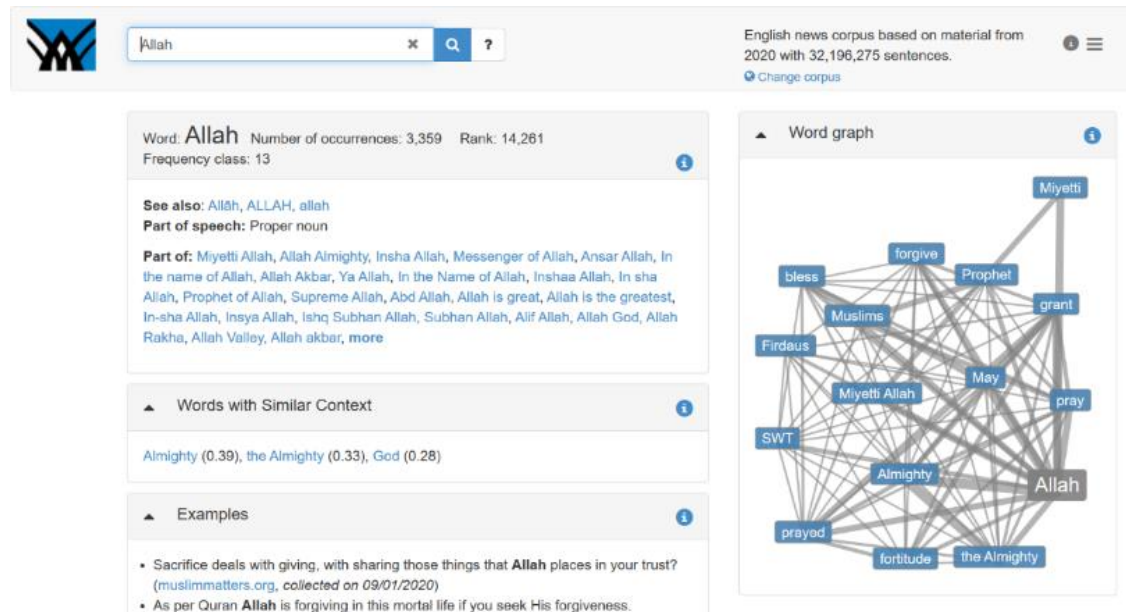


Figure 10. The word 'Allah' in Leipzig corpora

c. Recitation

Similar to the word 'resignation', the word 'recitation' in Islamic boarding school English corpus also appears for the limited amount of time, 6 times. Its uses in sentences can be located in Figure 11. It can be inferred that the word 'recitation' is seldom used in Islamic boarding school English corpus. It is striking that this word is rarely used in Islamic boarding school activities. Instead, most people prefer to use the word reading the Qur'an rather than reciting the Qur'an. Unfortunately, this word of choice is not appropriate as reading is the practice of extracting and constructing meaning from written texts [34]. Meanwhile, the purpose of reciting is to acquire the necessary skills to read the Holy Quran through Arabic calligraphy by ensuring the correct pronunciation, compliance with the rules of recitation science, the ability to stop at appropriate moments, as well as the ability to absorb sounds and tone [35].

Moreover, the word 'recitation' in the Leipzig corpora is a noun with a similar meaning to the one in Islamic boarding school English corpus. The use of the word 'recitation' in Leipzig corpora is presented in the Figure 12. The Figure 12 shows that the word 'recitation' is usually related to the recitation in religious activities. In Islam, it is often associated with Quran recitation, Quranic recitation, letter recitation, verses recitation, and poem recitation. Meanwhile, in other religions, it is often used with rosary recitation and hanuman chalisa recitation. Therefore, for the purpose of religious activities, the word 'recitation' is more suitable when it is used to recite the Qur'an, letter, verses, or other religious books.

Voyant Tools				
Contexts				
Document	Left	Term	Right	
file untu...	because in Islam the mere	recitation	of the Qur'an is a	
file untu...	Gym believes that the mere	recitation	of the Qur'an itself is	
file untu...	shalawat as well as the	recitation	of the holy Qur'an. He	
file untu...	to pengajian or the melodious	recitation	of Holy Qur'an. Men's eyes	
file untu...	tajwid (proper pronunciation for correct	recitation	of the Al- Qur'an), mantiq	
file untu...	sad voice in the shalat	recitations	. Aa Gym's sad voice is	

Figure 11. The word 'recitation' in Islamic boarding school English corpus

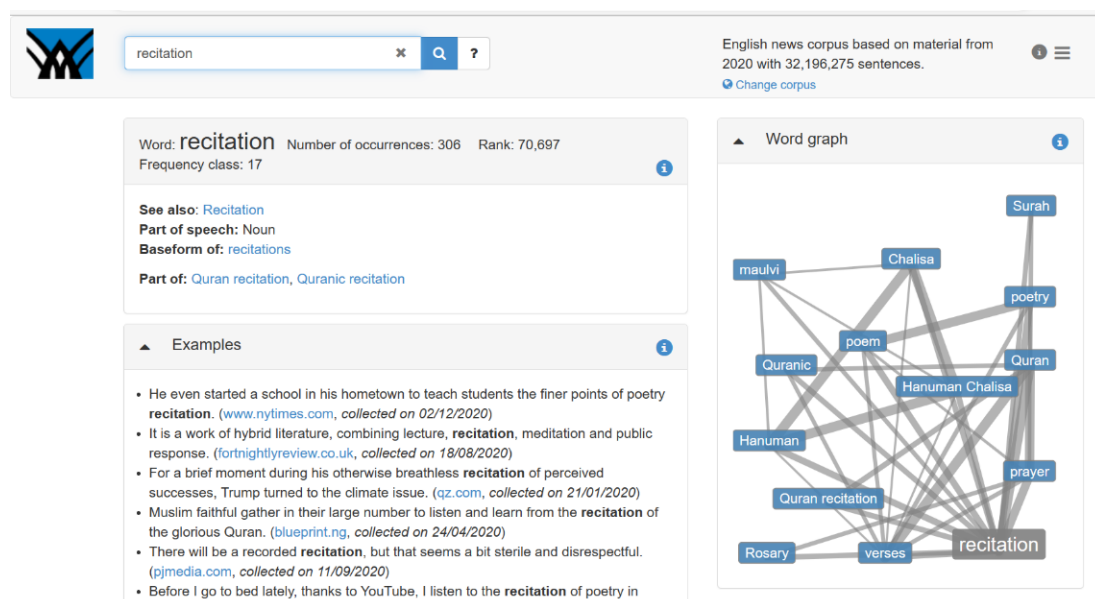


Figure 12. The word 'recitation' in the Leipzig corpora

4. CONCLUSION

Islamic boarding school English corpus is built from the work of several authors and researchers in the field of Islamic boarding school. It is built on 49,970 words comprising 5,417 specific words, 0.108 vocabulary density, and a 12,980-readability index. The output of this corpus will be incorporated into instructional resources for developing Islamic boarding school students' general and/or specialized vocabulary. They can be learned and applied by students in the learning process and the daily activities in the Islamic boarding school environment. Furthermore, the researchers will use the data from this corpus as a reference in developing an English language teaching model for Islamic boarding school students. Then, the data will be included in the development of teaching materials. Thus, it is recommended for future researchers who are interested in creating a corpus to know the purpose. Researchers can use various tools to analyze language data and then adjust to the needs of their research.

APPENDIX

Table 4. Fifty unique words in Islamic boarding school English corpus

No	Words	Frequency in Islamic boarding school English corpus	Frequency in Leipzig corpora	No	Words	Frequency in Islamic boarding school English corpus	Frequency in Leipzig corpora
1	Islamic	616	13,393	26	Saint (wali Allah)	8	259
2	<i>Pesantren</i>	484	5,025	27	Proselytizing (<i>taushiyah</i>)	20	2
3	Allah Swt.	191	69,464	28	Miraculous (<i>Ma'unah</i>)	10	23
4	Religious	187	653	29	Miracle (<i>mu'jizat</i>)	15	498
5	<i>Santri</i> (students)	144	3,958	30	Haughty (<i>takabbur</i>)	15	378
6	Monotheism	85	6,005	31	Modesty (<i>tawadhu'</i>)	10	654
7	Prayer (shalat)	70	121,937	32	Permissible (halal)	10	8,162
8	Muslim	69	126,379	33	Forbidden (haram)	7	297
9	Mosque	57	52	34	Recitation (<i>pengajian</i>)	20	14
10	Quran	48	25,050	35	Recite (<i>tadarrus</i>)	35	227
11	Leader (kiai)	130	25,739	36	Gossip (<i>ghibah</i>)	7	316
12	Congregation (<i>makmum</i>)	14	18	37	Steadfastness/consistency (<i>istiqamah</i>)	10	412

Table 4. Fifty unique words in Islamic boarding school English corpus (continue)

No	Words	Frequency in Islamic boarding school English corpus	Frequency in Leipzig corpora	No	Words	Frequency in Islamic boarding school English corpus	Frequency in Leipzig corpora
13	Character (<i>akhlaq</i>)	78	2,537	38	Piety (<i>taqwa</i>)	20	14
14	Morality (<i>akhlaq</i>)	22	63	39	Trustworthiness (<i>amanah</i>)	5	33
15	Prophet	42	58	40	Verse (ayat Quran)	40	82
16	Faith (<i>iman</i>)	27	57	41	Ablution (<i>wudlu</i>)	5	3
17	Resignation (<i>tawakal</i>)	5	57	42	Intention (<i>niat</i>)	15	133
18	Sincere (<i>ikhlas</i>)	20	56	43	Repentance (<i>taubat</i>)	10	9
19	Courtesy (<i>ihsan</i>)	11	53	44	Reward/merit (<i>pahala</i>)	12	1,227
20	Gratitude (<i>rasa syukur</i>)	8	53	45	Introspection (<i>muhasabah</i>)	3	2
21	Showing off (<i>riya</i>)	7	1,504	46	Benefit (<i>maslahat</i>)	15	931
22	Immorality (<i>maksiat</i>)	3	2	47	Teacher (<i>ustadz/ustadzah</i>)	41	1,158
23	Emergency (<i>madllarat</i>)	2	84	48	Preach (<i>khotbah</i>)	11	8
24	Obligation (<i>fardlu</i>)	18	829	49	Charity (<i>shodaqoh</i>)	5	995
25	Initiative (<i>ikhtiar</i>)	5	35	50	Fasting (<i>puasa</i>)	13	45




REFERENCES

- [1] T. McEnery and A. Hardie, *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.
- [2] A. O'keeffe, M. McCarthy, and R. Carter, *From corpus to classroom: Language use and language teaching*. Cambridge University Press, 2007.
- [3] S. Simbuka, F. Abdul Hamied, W. Sundayana, and D. A. Kwary, "A corpus-based study on the technical vocabulary of Islamic religious studies," *TEFLIN Journal - A publication on the teaching and learning of English*, vol. 30, no. 1, p. 47, Jul. 2019, doi: 10.15639/teflinjournal.v30i1/47-71.
- [4] R. Reppen, "Using corpora in the language classroom," *Materials development in language teaching*, vol. 2, pp. 37–50, 2011.
- [5] G. Leech, "Teaching and language corpora: A convergence," in *Teaching and language corpora*, 2014, pp. 1–24.
- [6] T. McEnery and R. Xiao, "What corpora can offer in language teaching and learning," in *Handbook of research in second language teaching and learning*, Routledge, 2011, pp. 364–380.
- [7] S. Hunston, *Introduction to a corpus in use*. Cambridge: Cambridge University Press, 2018.
- [8] N. Schmidt, "Unpacking second language writing teacher knowledge through corpus-based pedagogy training," *ReCALL*, vol. 35, no. 1, pp. 40–57, Jan. 2023, doi: 10.1017/S0958344022000106.
- [9] A. R. Fauzi, "Designing an English vocabulary workbook based on corpus-based approach: What actual learning task to incorporate target vocabularies into speaking," 2020.
- [10] A. R. M. Altkhaineh, M. Alaghawat, and A. Younes, "Challenges with online teaching and learning of English vocabulary," *International Journal of Information and Education Technology*, vol. 13, no. 3, pp. 577–586, 2023, doi: 10.18178/ijiet.2023.13.3.1841.
- [11] M.-C. Toriida, "Steps for creating a specialized corpus and developing an annotated frequency-based vocabulary list," *TESL Canada Journal*, vol. 34, no. 1, pp. 87–105, 2016.
- [12] K. Chujo, K. Oghigian, M. Utiyama, and C. Nishigaki, "Creating a corpus-based daily life vocabulary for TEYL," *Asian EFL Journal*, vol. 49, pp. 30–59, 2011.
- [13] R. Panocová, *The vocabulary of medical English: A corpus-based study*. Cambridge Scholars Publishing, 2017.
- [14] A. R. Fauzi and S. Suradi, "Building the students' English vocabulary for tourism through computer-based corpus approach," *Indonesian Journal of Integrated English Language Teaching*, vol. 4, no. 2, pp. 133–148, 2018.
- [15] J. Kaceti and B. Klímová, "English vocabulary for tourism – a corpus-based approach," 2015, pp. 489–494.
- [16] T. Taufikin, "Hermeneutic of pesantren with the "fusion of horizons" gadamer's theory," *Southeast Asian Journal of Islamic Education*, vol. 1, no. 1, pp. 37–58, Dec. 2018, doi: 10.21093/sajie.v1i1.1335.
- [17] T. H. Aprilia, A. H. Masrof, and A. Humaidi, "Pesantren in social construction perspective (the educational orientation of the sidogiri pesantren)," *Tarlim: Jurnal Pendidikan Agama Islam*, vol. 5, no. 2, Nov. 2022, doi: 10.32528/tarlim.v5i2.7903.
- [18] S. C. Curral and A. J. Towler, "Research methods in management and organizational research: Toward integration of qualitative and quantitative techniques," Sage Publications, 2003.
- [19] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.
- [20] D. Solahudin, *The workshop for morality: The Islamic creativity of Pesantren Daarut Tauhid in Bandung, Java*. ANU Press, 2008, doi: 10.22459/WM.08.2008.
- [21] E. Srimulyani, *Women from traditional Islamic educational institutions in Indonesia: Negotiating public spaces*. Amsterdam University Press, 2012.
- [22] A. Suhartini, "The internalization of Islamic values in pesantren," *Jurnal Pendidikan Islam (Islamic educational institutions concerning Islamic education)*, vol. 2, no. 3, p. 429, Dec. 2016, doi: 10.15575/jpi.v2i3.827.
- [23] M. Thahir, "The role and function of Islamic boarding school: An Indonesian context," *TAWARIKH*, vol. 5, no. 2, 2014.
- [24] G. A. N. Zakaria, "Pondok pesantren: changes and its future," *Journal of Islamic and Arabic Education*, vol. 2, no. 2, pp. 45–52, 2010.




- [25] G. Bennet, *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press ELT, 2010.
- [26] O. Kushnir, V. Yaremiv, I. Dovhan, and A. Kashuba, "Influence of unique words on the performance of corpus-based keyword detection methods," *Proceedings of X International Scientific and Practical Conference "Electronics and Information Technologies"*, p. 22, 2018, doi: 10.30970/elit2018.A22.
- [27] C. Rois, M. S. Dewi, and N. Robaniyah, "The historicity of pesantren: An overview of civilization discourse and the religion moderation of Islamic boarding school members," *Progresiva : Jurnal Pemikiran dan Pendidikan Islam*, vol. 12, no. 01, pp. 115–130, Jun. 2023, doi: 10.22219/progresiva.v12i01.24473.
- [28] M. Malakhovskaya, L. Beliaeva, and O. Kamshilova, "Teaching noun-phrase composition in EAP/ESP context: a corpus-assisted approach to overcome a didactic gap," *Journal of Teaching English for Specific and Academic Purposes*, p. 257, Mar. 2021, doi: 10.22190/JTESAP2102257M.
- [29] M. Yamin, H. Basri, and A. Suhartini, "Learning management in Salaf Islamic boarding schools," *At-tadzkir: Islamic Education Journal*, vol. 2, no. 1, pp. 25–36, Feb. 2023, doi: 10.59373/attadzkir.v2i1.10.
- [30] Y. Bakhtiar, D. Yanuarmawan, A. Tri Andari, and B. S. Jannah, "Critical review of income accounting on Islamic boarding school accounting guidelines," *Journal of Applied Business and Technology*, vol. 4, no. 2, pp. 130–133, May 2023, doi: 10.35145/jabt.v4i2.124.
- [31] T. Memika and T. K. Polat, "Internet of things supported airport boarding system and evaluation with fuzzy," *Intelligent Automation and Soft Computing*, vol. 35, no. 3, pp. 2687–2702, 2023, doi: 10.32604/iasc.2023.026955.
- [32] L. N. Amalia and A. Saifuddin, "Tawakal and academic stress in assignment completion of university students," *Gadjah Mada Journal of Psychology (GamaJoP)*, vol. 8, no. 2, p. 203, Oct. 2022, doi: 10.22146/gamajop.75621.
- [33] H. M. Hibatillah, "The concept of akhlaq in Islamic educational curriculum," *Educational Review: International Journal*, vol. 19, no. 3, pp. 7–17, 2016.
- [34] T. Hanghoj, K. Kabel, and S. H. Jensen, "Digital games, literacy and language learning in L1 and L2," *L1-Educational Studies in Language and Literature*, pp. 1–44, Jul. 2022, doi: 10.21248/11esll.2022.22.2.363.
- [35] A. Rezai, A. Professor, E. Namaziandost, and A. Amraei, "Exploring the effects of dynamic assessment on improving Iranian Quran learners' recitation performance," *Critical Literary Studies*, vol. 5, no. 1, pp. 159–176, 2022.

BIOGRAPHIES OF AUTHORS






Yulia Agustina    is a doctoral student from Universitas Negeri Yogyakarta. She received her master's degree in English Language Education from Universitas Negeri Sebelas Maret Surakarta (UNS) Indonesia and her bachelor's degree from Universitas Negeri Siliwangi (UNSIL), Tasikmalaya, Indonesia. In 2014, she joined the Department of English Language Education at the Faculty of Teacher Training and Education of Universitas Hamzanwadi, Lombok Indonesia as an English lecturer. She has written several papers in the field of English education. Her research interests also include developing a model of English learning for Islamic boarding school students. She can be contacted at email: yulia0012pasca.2019@student.uny.ac.id.



Pratomo Widodo    is a professor in the field of Germanistic Studies at Universitas Negeri Yogyakarta. He teaches at the Department of German Language Education (S1), the Department of Applied Linguistics (S2), and the Department of Language Education Doctoral (S3) at the Faculty of Languages, Arts, and Cultures. He also serves as the Chair of the Indonesian Germanistic Association (*Indonesischer Germanistenverband*). Several works in the form of research, scientific journal articles, seminar papers and books have been published, especially those related to the fields of Linguistics, Germanistics, and Language Teaching. He can be contacted at email: pratomo@uny.ac.id.



Margana Margana    is the vice rector for research, cooperation, information systems, and business from Universitas Negeri Yogyakarta. He was confirmed as a professor in September, 2017. His expertise in: linguistics, English curriculum and material development, educational research, TEFL methodology, literature, and has produced many scientific publications related to his expertise. He can be contacted at email: margana@uny.ac.id.