Jurnal Ilmiah Iqra'
2541-2108 [Online] 1693-5705 [Print]
Available online at: http://journal.iain-manado.ac.id/index.php/JII

Students' High-Frequency Vocabulary Profile: A Corpus Study at Manado State Institute of Islamic Studies

Muhammad Raihan Praba Tahir[1], Srifani Simbuka[2], Lies Kryati[3]*

[1,2,3]IAIN Manado, Manado, Indonesia

*Corresponding E-mail: raihan.tahir@iain-manado.ac.id

Abstract

This research examines the high-frequency vocabulary knowledge of students at Manado State Institute of Islamic Studies utilizing a corpus-based approach. By applying a quantitative research design and using the AntWordProfiler software to generate statistical data on the vocabulary, the research identifies and analyzes the students' lexical proficiency within the first and second General Service Lists (GSL). The results indicate that students demonstrated limited mastery of high-frequency words, covering only 18% of GSL 1 and 10% of GSL 2, with an overall inclusion of merely 15% across both lists. These findings highlight a critical gap in students' foundational English vocabulary, which has significant implications for English language teaching and curriculum development at the institution. To improve language acquisition outcomes, tailored vocabulary instruction strategies and curriculum adjustments are suggested to enhance students' mastery of high-frequency vocabulary.

Article Info

Article History

Submitted / Received: 19-03-2025
First Revised: 20-04-2025
Accepted: 30-05-2025
First Available online: 24-06-2025
Publication Date: 25-06-2025

Keywords:
vocabulary,
corpus linguistics,
students'

How to Cite:
Tahir, M. R. P., Simbuka, S., & Kryati, L. (2025). Students' high-frequency vocabulary profile: A corpus study at Manado State Institute of Islamic Studies. *Jurnal Ilmiah Iqra*, *19*(1), 86–96.

Introduction

Vocabulary acquisition is central to second language learning as the foundation for reading, writing, speaking, and listening skills. Without sufficient vocabulary knowledge, language learners struggle to comprehend texts, engage in conversation, or express complex ideas (Read, 2004). Furthermore, Richards and Renandya (2002) suggested that vocabulary is central to communication in a second language, and inadequate vocabulary knowledge is often the primary obstacle for learners attempting to read, write, speak, or understand English. Similarly, Alqahtani (2015) emphasizes that vocabulary proficiency is crucial for language learners, particularly because words are building blocks for comprehension and communication.

Vocabulary is categorized into four main types: high-frequency words or words that occur very frequently in any (general) texts, academic words or words that are highly frequent in any academic texts, and technical words or words that occur very frequently in specific texts (Nation, 2001). High-frequency vocabulary is critical as it constitutes many words encountered in everyday communication (Nation, 2001). These words are vital for academic success and practical language use, making them a priority in English language teaching (ELT).

Despite extensive English education in Indonesia, students' performance using essential vocabulary remains understudied, particularly in non-native English-speaking environments like IAIN Manado. The current study investigates the high-frequency vocabulary proficiency of first-year students at Manado State Institute of Islamic Studies (IAIN Manado). The research question posed in this study is "How is the students' vocabulary profile of high-frequency words at Manado State Institute of Islamic Studies?" Given that language proficiency heavily depends on a learner's mastery of high-frequency words, this research aims to evaluate students' knowledge using corpus linguistics tools.

Theoretical Review

The underlying theory that bases this present study is Corpus linguistics, which involves the analysis of extensive, authentic collections of spoken or written texts. Although it initially emerged as one of the methods in discourse analysis, Corpus Linguistics has become a valuable tool in researching topics in language education, especially in foreign language vocabulary teaching and learning (Nation, 2016). It allows researchers to quantify vocabulary usage, identify patterns, and determine learners' language proficiency (Biber, 2015). Unlike many conventional

methods of analyzing discourses, Corpus Linguistics uses specifically built corpus software as its analysis tool. Antconc 4.3.1 (Anthony, 2024) and AntWordProfile 2.2.1 are some of these tools. Developed by Lawrence Anthony (Anthony, 2024), Range based on Heatley and Nation (1994) and adapted to the web-based corpus tools (Cobb, 2023), LancsBox X developed by Brezina and Platt (2024), and many others. As one of the useful mediums in corpus studies, AntWordProfiler provides detailed analyses of vocabulary frequency and coverage, making it an ideal instrument for evaluating students' lexical knowledge (Anthony, 2013).

Vocabulary in Language Learning

Vocabulary is central to communication in a second language, and inadequate vocabulary knowledge is often the primary obstacle for learners attempting to read, write, speak, or understand English (Richards & Renandya, 2002). Alqahtani (2015) emphasizes that vocabulary proficiency is crucial for language learners, particularly because words serve as building blocks for comprehension and communication.

Nation (2001, 2016) proposes that at least three categories of vocabulary exist in any text in the English language. Learners of ELT should master this vocabulary. The three categories suggested by Nation (2001) are high-frequency vocabulary, which is defined by Nation (2001) as words that appear frequently across various contexts and constitute the core of language knowledge. High-frequency words are compiled into a word list named The General Service List (GSL), developed by Michael West (1953). This list is widely used to identify these essential words, listing approximately 2,000 word families most useful for English learners.

Corpus Linguistics and Vocabulary Profiling

Corpus linguistics, which involves the analysis of extensive, authentic collections of spoken or written texts, has become a valuable tool in language education. It allows researchers to quantify vocabulary usage, identify patterns, and determine learners' language proficiency (Biber, 2015). For instance, as one of the useful mediums in corpus studies, AntWordProfiler provides detailed analyses of vocabulary frequency and coverage, making it an ideal instrument for evaluating students' lexical knowledge (Anthony, 2013).

Relevant Research of Corpus Linguistics

Corpus Linguistics researches involve analyzing extensive collections of texts to explore language aspects such as vocabulary usage, composition, and profiling.

As the research purpose, the studies in this field aim to inform language teaching and learning practices, including materials development and curriculum design, and to better understand learners' vocabulary needs.

Simbuka (2019) developed the Technical Vocabulary of Islamic Religious Studies (IRSTV) to meet the needs of first-year English language learners in Indonesian Islamic higher education. Her study aimed to compile a comprehensive vocabulary list from five major sub-disciplines of Islamic Religious Studies (IRS): the Science of Qur'an, the Science of Hadiths, Islamic Law and Jurisprudence, Islamic Philosophy and Theology, and Islamic Mysticism. These are commonly taught at Indonesian Islamic universities and colleges. Using the Corpus of Islamic Religious Studies Textbooks, which included 18,058 word types and 305,701 tokens, Simbuka identified 262 word types (239 lemmas) in the IRSTV list that are crucial for English as a Foreign Language (EFL) students at these institutions.

Furthermore, Rozaq (2019) found that students had a high usage of academic vocabulary, with about 11.92% of the 66,419 running words being academic tokens. This indicates that the student's work is categorized as scholarly writing and contributes to understanding their proficiency with academic vocabulary, which could impact language instruction and curriculum development. In addition, Novi (2021) used a quantitative corpus analysis to examine the blogs of five international students for analyzing and categorizing various word classes, including nouns, adjectives, adverbs, verbs, interjections, conjunctions, pronouns, prepositions, auxiliaries, and the verb "be."

Moreover, Khasanah (2021) investigated the vocabulary in five textbooks created by the English teacher forum of Junior High School in Tuban. Using AntWordProfiler, the study identified 82,918 tokens and 6,669 types in the textbooks. High-frequency words such as nouns, verbs, and determiners were most common. The study also identified 282 academic word families, representing 2.51% of the tokens. This research gives teachers insights into the vocabulary that should be taught according to the textbooks to meet curriculum expectations.

Lastly, Asnidar (2021) conducted a study to evaluate vocabulary coverage and word types based on the Nation's typology in three English coursebooks used at the Manado State Institute of Islamic Studies. The study used a quantitative descriptive design and AntWordProfiler to collect statistical and frequency data on the vocabulary in these coursebooks. The study used a quantitative descriptive design and AntWordProfiler to collect statistical and frequency data on the

vocabulary in these coursebooks. The results showed that the coursebooks contained 63,130 tokens and 5,290 word types. High-frequency words comprised 87.76% of the tokens and 55.16% of the word types, highlighting their importance. The study also found that 6.17% of the words were Academic, and 6.07% were Low-Frequency. These findings have implications for curriculum design and English language teaching at Manado State Institute of Islamic Studies, suggesting that understanding the vocabulary composition of coursebooks can help educators make better decisions about vocabulary teaching.

Method

This study employed a quantitative research design using a corpus-based approach to profile students' high-frequency vocabulary knowledge. The data were collected from pre-tests administered to first-year students at IAIN Manado in the 2022/2023 academic year, focusing on their lexical mastery within the GSL 1 and GSL 2 word families.**3.2**

A total of 209 pre-test results were obtained from the Language Centre of IAIN Manado. The pre-tests were conducted as part of the English language matriculation program for first-semester students. Due to missing or incomplete data, only the results of 209 students were included in the analysis. The tests were converted into digital format and analyzed using AntWordProfiler software to measure the students' proficiency in high-frequency vocabulary.

The vocabulary tests were analyzed using AntWordProfiler, which compared the students' responses to the GSL 1 and GSL 2 word lists. The software generated statistical data on the students' coverage of high-frequency words, including the number of word types (distinct words) and tokens (total word occurrences) they produced.

Results

The analysis revealed that IAIN Manado students demonstrated low proficiency in high-frequency vocabulary. Out of the 2,000-word families in GSL 1 and GSL 2, the students covered only 15% of the total word types, with significantly lower coverage in GSL 2 than in GSL 1. In addition, students generated non-GSL words. The following table presents the result of the vocabulary analysis.

Category	Tokens	% of Total Tokens	Types	% of Total Types
GSL 1	5870	60.63%	731	47.19%
GSL 2	2557	26.41%	362	23.37%

Non-GSL	1254	12.95%	456	29.44%
Total	9681		1584	

Vocabulary Coverage in GSL 1

The student's performance in GSL 1 showed moderate proficiency, with 731 word types (18%) covered out of the 4,114-word kinds in this list. Regarding word tokens, the students produced 5,870 tokens, representing 60.63% of the total words in GSL 1.

Vocabulary Coverage in GSL 2

The student's performance in GSL 2 was notably lower than in GSL 1, with only 362 word types (10%) covered out of the 3,708 word types in GSL 2. The students produced 2,557 word tokens, accounting for 26.41% of the total tokens in this list. The top 25 words used in GSL 2 reflect a narrower vocabulary range, with common nouns such as "bag," "chair," and "pen" dominating the list (Table 2).

Non-GSL Vocabulary Coverage

Furthermore, students also produced 1,254 tokens and 456 types that did not belong to either GSL 1 or GSL 2. This suggests that students are familiar with some less frequent or specialized vocabulary, although this knowledge does not compensate for their limited grasp of high-frequency words

Discussion

The findings indicate that students' vocabulary coverage of high-frequency words is notably low. Examining the coverage for GSL 1, students generated 731 out of 4,114 word types, representing roughly 18% of the list. For GSL 2, they produced 362 out of 3,708 word types, which is about 10%. Overall, students demonstrate low proficiency in high-frequency words with only 15% coverage across both GSLs, leading to several outcomes referencing theories and previous studies of vocabulary learning.

Read's theory emphasizes the essential role of vocabulary as the building blocks of language, and the findings highlight the importance of vocabulary proficiency. The observed low vocabulary coverage among students may hinder their understanding of concepts and texts. As Read's theory suggests, insufficient vocabulary can lead to challenges in grasping larger linguistic structures like sentences and paragraphs. According to Nation's theory, high-frequency words

comprise a large portion of spoken and written language, appearing in diverse contexts and uses. Therefore, their importance is considerable. The findings reveal that students' low coverage of these high-frequency words may indicate a limitation in their overall language proficiency.

Furthermore, the findings of this study align with previous research, such as that by Simbuka (2019). In the field of Islamic Religious Studies (IRS) at Indonesian Islamic State Institutes (IISI), high-frequency words from General Service Lists (GSL) 1 and 2 account for 234,922 tokens, making up 76.85% of the total running words and 4,520 word types, or 25.03% of the total word types. These words are spread across five key IRS sub-disciplines: Islamic Philosophy and Theology, the Science of Qur'an, Islamic Law and Jurisprudence, the Science of Hadith, and Islamic Mysticism Theology. However, students have only covered about 15% of these GSLs. As a result, they may struggle with understanding complex religious texts and articulating sophisticated theological ideas, which could hinder their ability to engage thoroughly with the broader academic discourse in Islamic Religious Studies.

Similarly, research by Asnidar (2021) and Simbuka & Nagauleng (2021) indicates that four English for Specific Purposes (ESP) textbooks used in various schools at Manado State Institute of Islamic Studies include a substantial amount of high-frequency vocabulary, covering 84.14% of the words listed in the General Service List (GSL). Given the current study's finding of low high-frequency word coverage among students, this alignment raises essential concerns. The gap between the vocabulary emphasized in the ESP textbooks and the actual vocabulary proficiency of the students suggests a potential disadvantage. With only 15% vocabulary coverage, students may struggle to understand and effectively use the high-frequency words prevalent in the ESP textbooks. This highlights the need for a targeted approach in English Language Teaching (ELT). For example, language instructors at the language center should prioritize enhancing students' GSL vocabulary before introducing specialized vocabulary such as Islamic Religious Studies Technical Vocabulary (IRSTV), as Simbuka (2019) suggested.

To add to the previous one, utilizing corpus research findings, as seen in the current study, to inform learning materials is supported by various studies. For example, Tosun and Sofu (2023) explored the significant potential of pedagogical corpora for improving vocabulary acquisition among beginner learners. Their research showed that students who worked with corpora performed better in post-test and delayed post-test assessments than those who received traditional

vocabulary instruction. Additionally, these students demonstrated a heightened awareness of different aspects of words, including their parts of speech, multiple meanings and usages, lexical-grammatical structures, and collocations.

In addition to the advantages of using corpus research findings, another study by Paker and Ozcan (2017) evaluated the impact of corpus-based materials on vocabulary learning. Their results revealed a statistically significant difference in post-test scores between the experimental and control groups, with the experimental group achieving better outcomes. This suggests that corpus-based vocabulary materials are more effective than traditional textbooks or dictionaries.

Furthermore, a study by Elsherbini and Ali (2017) investigated the benefits of corpus-based activities in improving students' grammar and vocabulary. The study involved two groups of 104 freshmen from a business English course at Sadat Academy for Management Sciences in Egypt, with 54 students in the experimental group and 50 in the control group. Over 11 weeks, the experimental group received training in corpus usage and participated in corpus-based activities, while the control group used only the course book for instruction. The results showed that the experimental group significantly improved, with higher post-test scores than their pre-test scores. Moreover, the experimental group surpassed the control group in post-test scores for both grammar and vocabulary.

As a final implication, the study suggests a pattern of inadequate language learning in schools in North Sulawesi. Despite English being a mandatory subject at the elementary, junior high, and senior high school levels, students' vocabulary proficiency remains low. This highlights a need for further research to investigate the reasons behind this deficiency and underscores the necessity for university language instructors to intensify their efforts to address students' shortcomings.

Conclusion

The students' low proficiency in high-frequency vocabulary has several implications. First, the gap between the vocabulary they should know and their actual mastery indicates potential challenges in fully understanding texts. Second, these findings provide valuable insights for language tutors at language centers regarding students' coverage of high-frequency vocabulary. Third, the results highlight a trend of inadequate language learning in schools in North Sulawesi, suggesting the need for further research to explore the underlying causes and for tutors to increase their efforts to address students' deficiencies.

References

Alqahtani, M. (2015). The importance of vocabulary in language learning and how it can be taught. *International Journal of Teaching and Education, 3*(3), 21–34.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research, 30*(2), 141–161.

Anthony, L. (2023, July 7). *AntWordProfiler (Windows, Macintosh OS X, and Linux)* [Help documentation]. https://www.laurenceanthony.net/software/antwordprofiler/releases/AntWordProfiler141/help.pdf

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

Anthony, L. (2024b). *AntWordProfiler* (Version 2.2.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/AntWordProfiler

Ardila, E. M. G., & Meara, P. (2021). EFL learners' lexical availability: Exploring frequency, exposure, and vocabulary level. *System*, *99*, Article 102481. https://doi.org/10.1016/j.system.2020.102481

Asnidar. (2021). *Profiling the vocabulary of English coursebooks used in ELT for non-English majors at Manado State Islamic Institute: Corpus-based study* (Undergraduate thesis). Fakultas Tarbiyah dan Ilmu Keguruan, IAIN Manado.

Biber, D. (2015). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (2nd ed.). Oxford University Press.

Brezina, V., & Platt, W. (2024). *LancsBox X* [Computer software]. Lancaster University. http://lancsbox.lancs.ac.uk

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, *36*, 135–141. https://doi.org/10.1016/j.jeap.2018.08.005

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 26(3), 367–387. https://doi.org/10.1177/1362168820911189

Elsherbini, S., & Ali, A. (2017). The effects of corpus-based activities on EFL university students' grammar, vocabulary, and attitudes toward the corpus. *Journal of Research in Curriculum, Instruction and Education Technology, 3*(1), 133–161.

Farihah, I., & Nurani, I. (2017). Internalisasi nilai-nilai keislaman dalam skema hidden curriculum di MTs Nurul Huda Medini Demak. *Edukasia: Jurnal Penelitian Pendidikan Islam, 12*(1), 213–234. https://doi.org/10.21043/edukasia.v12i1.2347

Ha, P. T., et al. (2023). Developing and validating a mid-frequency word list for chemistry: A corpus-based approach using big data. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*, Article 12. https://doi.org/10.1186/s40862-023-00205-5

Indrajit, R. E. (2016). *E-learning dan sistem informasi pendidikan: Modul pembelajaran berbasis standar kompetensi dan kualifikasi kerja* (2nd ed.). Preinexus.

Khasanah, S. (2021). *A corpus study of English vocabulary in textbooks constructed by the English teacher forum of junior high school in Tuban* (Undergraduate thesis). Faculty of Tarbiyah and Teacher Training, UIN Sunan Ampel Surabaya.

Lei, S., & Yang, R. (2020). Lexical richness in research articles: Corpus-based comparative study among advanced Chinese learners of English, English native beginner students, and experts. *Journal of English for Academic Purposes*, *47*, Article 100894. https://doi.org/10.1016/j.jeap.2020.100894

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Novi, A. (2021). *Learners' corpora in students' blogs of English writing* (Undergraduate thesis). Faculty of Tarbiyah and Teacher Training, UIN Sunan Ampel Surabaya.

Paker, T., & Ozcan, Y. (2017). The effectiveness of using corpus-based materials in vocabulary teaching. *International Journal of Language Academy, 5*(1), 62–88.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146–161.

Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching*. Cambridge University Press.

Rozaq, M. (2019). *Corpus analysis of academic vocabulary in students' thesis proposals at the English teacher education department of UIN Sunan Ampel Surabaya*

(Undergraduate thesis). Faculty of Tarbiyah and Teacher Training, UIN Sunan Ampel Surabaya.

Simbuka, S. (2019). A corpus-based study on the technical vocabulary of Islamic religious studies. *TEFLIN Journal, 30*(1), 47–66.

Simbuka, S., & Nagauleng, A. (2021). A corpus-informed materials evaluation of EFL textbooks and teachers' generated materials in Indonesian Islamic universities. *Al-Ishlah: Jurnal Pendidikan, 13*(3), 2489–2500.

Tosun, S., & Sofu, H. (2023). The effectiveness of data-driven vocabulary learning: Hands-on concordance through a pedagogical corpus. *Journal of Language and Education, 9*(3), 176–190.