

**LEMBAR
HASIL PENILAIAN SEJAWAT SEBIDANG ATAU PEER REVIEW
KARYA ILMIAH : JURNAL ILMIAH**

Judul artikel : Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era
 Nama penulis : Prihantoro
 Jumlah penulis : 5
 Status pengusul : ~~Penulis pertama~~/penulis anggota/~~penulis korespondensi~~
 Identitas Jurnal
 a. Nama jurnal : World Journal of English Language
 b. Nomor ISSN : 1755-1676
 c. Vol, no, tahun : Vol 13 No 3 (2023)
 d. Penerbit : SCIEDU
 e. DOI : <https://doi.org/10.5430/wjel.v13n3p181>
 f. Alamat web jurnal : <https://www.sciedupress.com/journal/index.php/wjel>
 g. Alamat artikel : <https://www.sciedupress.com/journal/index.php/wjel/article/view/23445>
 h. Terindeks : SCOPUS (Q3) SJR 0.10

Kategori publikasi jurnal ilmiah v Jurnal ilmiah internasional (bereputasi, terindeks, faktor dampak)
 Jurnal ilmiah nasional terakreditasi
 Jurnal ilmiah nasional tidak terakreditasi

Hasil penilaian peer review 1

| Komponen yang dinilai | | Nilai maksimal jurnal ilmiah | | | Nilai akhir yang diperoleh |
|-------------------------------------|--|------------------------------|------------------------|------------------------------|----------------------------|
| | | Internasional | Nasional terakreditasi | Nasional tidak terakreditasi | |
| | | [40] | [] | [] | |
| a | Kelengkapan unsur isi jurnal (10%) | 4.00 | | | 10% x 40 = 4 |
| b | Ruang lingkup dan kedalaman pembahasan (30%) | 12.00 | | | 30% x 40 = 12 |
| c | Kecukupan dan kemutakhiran data / informasi dan metodologi (30%) | 12.00 | | | 30% x 40 = 12 |
| d | Kelengkapan unsur dan kualitas terbitan/jurnal (30%) | 12.00 | | | 30% x 40 = 12 |
| Total 100% | | 40.00 | | | |
| Nilai pengusul: 10% x 40 = 4 | | 40% x 40 : 4 = 4 | | | |

Catatan penilaian paper oleh reviewer 1

1. Kelengkapan unsur isi jurnal:

Kelengkapan komponen IMRAD dalam isi jurnal

2. Ruang lingkup dan kedalaman pembahasan:

Ruang lingkup dan membahas secara mendalam mumpuni sesuai bidang ilmu para penulis tingkat internasional

3. Kecukupan dan kemutakhiran data/informasi dan metodologi:

Memiliki kecukupan bahasan dan kebaruan data dengan metode yang membuktikan analisis dan pembahasan secara ilmiah

4. Kelengkapan unsur dan kualitas terbitan:

Kualitas terbitan jurnal ini internasional bereputasi terindeks scopus scimago quartile 3 dan SJR 0.10

Medan, 4 Mei 2023
Reviewer 1



Nama : Prof. T. Silvana Sinar, M.A., Ph.D.
NIP/NIDN : 195409161980032003
Unit kerja : Fakultas Ilmu Budaya Universitas Sumatera Utara

**LEMBAR
HASIL PENILAIAN SEJAWAT SEBIDANG ATAU PEER REVIEW
KARYA ILMIAH : JURNAL ILMIAH**

Judul artikel : Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era
 Nama penulis : Prihantoro
 Jumlah penulis : 5
 Status pengusul : ~~Penulis pertama~~/penulis anggota/~~penulis korespondensi~~
 Identitas Jurnal
 a. Nama jurnal : World Journal of English Language
 b. Nomor ISSN : 1755-1676
 c. Vol, no, tahun : Vol 13 No 3 (2023)
 d. Penerbit : SCIEDU
 e. DOI : <https://doi.org/10.5430/wjel.v13n3p181>
 f. Alamat web jurnal : <https://www.sciedupress.com/journal/index.php/wjel>
 g. Alamat artikel : <https://www.sciedupress.com/journal/index.php/wjel/article/view/23445>
 h. Terindeks : SCOPUS (Q3) SJR 0.10

Kategori publikasi jurnal ilmiah v Jurnal ilmiah internasional (bereputasi, terindeks, faktor dampak)
 Jurnal ilmiah nasional terakreditasi
 Jurnal ilmiah nasional tidak terakreditasi

Hasil penilaian peer review 2

| Komponen yang dinilai | | Nilai maksimal jurnal ilmiah | | | Nilai akhir yang diperoleh |
|--|---|------------------------------|------------------------|------------------------------|----------------------------|
| | | Internasional | Nasional terakreditasi | Nasional tidak terakreditasi | |
| | | [40] | [] | [] | |
| a | Kelengkapan unsur isi jurnal (10%) | 4.00 | | | 4.00 |
| b | Ruang lingkup dan kedalaman pembahasan (30%) | 12.00 | | | 12.00 |
| c | Kecukupan dan kemuakhiran data / informasi dan metodologi (30%) | 12.00 | | | 12.00 |
| d | Kelengkapan unsur dan kualitas terbitan/jurnal (20%) | 12.00 | | | 11.00 |
| Total 100% | | 40.00 | | | 39 |
| Nilai pengusul: 40% x 39 /4 = 3,9 | | | | | |

Catatan penilaian paper oleh reviewer 2

1. Kelengkapan unsur isi buku

Unsur yang ada dalam nasakah indexk kosakata cukup memadai dan lengkap

2. Ruang lingkup dan kedalaman pembahasan

Ruang lingkup terkait index kosakata sudah cukup mewakili materi yang penting untuk dikaji

3. Kecukupan dan kemuakhiran data/informasi dan metodologi: data dan informasi yang terpapar cukup lengkap dan komprehensif

4. Kelengkapan unsur dan kualitas penerbit: Badan Penerbit World Journal of English Language sudah cukup dikenal dalam bidang publikasi pembelajaran Bahasa Inggris

Bandar Lampung, 4 Mei 2023

Reviewer 2



Nama : Prof Dr. Cucu Sutaryah, Dip.TESL., MA
NIP/NIDN : 195704061986031002/0006045704
Unit kerja : FKIP Universitas Lampung

LEMBAR PERNYATAAN BEBAS PELANGGARAN KARYA ILMIAH

Yang bertanda tangan di bawah ini

Nama : Prihantoro
NIP : 198306292006041002
NIDN : 3374102906830004
Pangkat (golongan ruang) : Pembina/IV A
Jabatan Akademik : Lektor Kepala
Program Studi : Magister Linguistik
Fakultas/Sekolah : Fakultas Ilmu Budaya/Universitas Diponegoro

menyatakan bahwa karya ilmiah dengan judul “Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era” yang dipublikasikan pada (World Journal of English Language, vol 13 (3), 181–192); di mana saya sebagai (salah satu) penulis, bebas dari atau tidak mengandung pelanggaran kode etik ilmiah.

Demikian surat pernyataan ini kami buat untuk dipergunakan sebagaimana mestinya.

Semarang, 1 April 2023

Yang Menyatakan



Prihantoro
NIP. 198306292006041002

Scopus Impact Factor SJR Q3

| | |
|--|------------------------------|
| World Journal of English Language Scopus coverage years: from 2020 to Present Publisher: Sciedu Press ISSN: 1925-0703 E-ISSN: 1925-0711 Subject area: Arts and Humanities: Literature and Literary Theory Arts and Humanities: Language and Linguistics Social Sciences: Linguistics and Language Social Sciences: Education Source type: Journal | CiteScore 2021 0.2 |
| | SJR 2021 0.102 |
| | SNIP 2021 0.000 |



World Journal of English Language

HOME ABOUT LOGIN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS RECRUITMENT SUBMISSION EDITORIAL
BOARD INDEXES ETHICAL GUIDELINES SPECIAL ISSUES CONTACT

Journal Help

USER
Username
Password
 Remember me

World Journal of English Language
Literature and Literary Theory
best quartile
Q3
SJR 2021 0.1
powered by scimagojr.com


World Journal of English Language (ISSN 1925-0703 E-ISSN 1925-0711) is a peer-review journal, published by Sciedu Press. It is devoted to publishing articles in various aspects, fields and scopes of the English Language, such as but not limited to English literature, English linguistics, teaching and learning English as a Second Language (ESL), as an Additional Language (EAL) or as a Foreign Language (TEFL). It is published **Bimonthly** (January, March, May, July, September and November) in both online and printed versions.

The journal accepts article submissions [online](#) or by [e-mail \(wjel@sciedupress.com\)](mailto:wjel@sciedupress.com).

Abstracting and Indexing:

- AE Global Index
- CNKI Scholar
- DHET Accredited Journals 2022
- Elektronische Zeitschriftenbibliothek EZB
- EuroPub
- Google Scholar
- Harvard Library E-Journals





World Journal of English Language

HOME ABOUT LOGIN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS RECRUITMENT SUBMISSION EDITORIAL
BOARD INDEXES ETHICAL GUIDELINES SPECIAL ISSUES CONTACT

Journal Help

USER
Username
Password
 Remember me

World Journal of English Language
Literature and Literary Theory
best quartile
Q3
SJR 2021 0.1
powered by scimagojr.com

Home > Vol 13, No 3 (2023) > Lun

Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era

Wong Wei Lun, Mazura Mastura Muhammad, Warid Mihat, Muhammad Syafiq Ya Shak, Mairas Abdul Rahman, Prihantoro Prihantoro

Abstract

In the wake of the COVID-19 pandemic, Malaysian English teachers identified a pressing need to support upper primary school pupils, particularly those in the upper levels, in the effective composition of extended writing. Additionally, these educators required more innovative methodologies for teaching vocabulary in this context. Consequently, the current study aimed to develop a vocabulary index as a suggested resource for Malaysian English teachers instructing upper primary school pupils on extended writing. To achieve this, a quantitative computational research strategy and corpus-driven research design were employed. A purposive sampling technique was used to select 560 advanced upper primary school pupils from 28 schools, each with high English performance in the capital of each state and the federal territory of Malaysia, who produced a total of 152,187 words in extended writing for analysis. LncsBox, a primary computational linguistics application, was used for data processing. Given that the vocabulary index for extended writing necessitates a more comprehensive coverage of vocabulary, functional and content words were included, and keywords, raw and normalised frequencies were analysed and reported. Through the vocabulary index built in this study, the researchers found English teachers in Malaysia should utilise local issues in writing prompts, emphasise the use of both positive and negative adjectives, introduce complex sentence structures to enhance pupils' writing abilities and also train pupils to organise the ideas in their writing. Future linguistic studies could replicate the present investigation, so that it can respond to their classroom needs.

Full Text:
[PDF](#)

JOURNAL CONTENT
Search
Search Scope
All

Editorial boards minimal 4 negara

Editorial Team

Editor-in-Chief

[Andres Canga Alonso](#), University of La Rioja, Spain

Associate Editors

[Amelia Maria Cava](#), Università di Napoli Suor Orsola Benincasa, Italy

[Cheryl Caesar](#), Michigan State University, United States

[Leila A Lomashvili](#), Shawnee State University, United States

Editorial Assistant

[Mr. Joe Nelson](#), Sciedu Press, Canada

Editorial Board Members

[Abdulfattah Omar](#), Prince Sattam Bin Abdulaziz University, Saudi Arabia

[Acep Unang Rahayu](#), Poltekpar NHI Bandung, Indonesia

[Mr Aissa Hanifi](#), Chlef University, Algeria

[Ali Hussein Hazem](#), University of Patras, Greece

[AMER M TH AHMED](#), Dhofar University, Oman

[Ana Maria Costa Lopes](#), Higher School of Education of the Polytechnic Institute of Viseu, Portugal

[Anna Kuzio](#), University of Zielona Gora, Poland

[Antonio Piga](#), University of Cagliari, Italy, Italy

[Ayman Khafaga](#), Suez Canal University, Egypt

[Ayman Rashad Yasin](#), Princess Sumaya University for Technology, Jordan

[Bahram Kazemian](#), Islamic Azad University, Tabriz, Iran, Islamic Republic of

[Bhuvaneshwari G](#), Vellore Institute of Technology, Chennai, India

[Daniel Ginting](#), Universitas Ma Chung

[Chunlin Yao](#), Tianjin Chengjian University, China

[Deena Elshazly](#), Arab Academy for Science, Technology and Maritime Transport, Egypt

[Dr Don Anton Robles Balida](#), International College of Engineering and Management, Oman

[Elena Alcalde Alcalde Peñalver](#), University of Alcalá, Spain

[Fatma Abu-sweel](#), The university of Trinoli, Libya

Minimal 2 negara yang berbeda

[Saudi EFL Students' Responses to Written Corrective Feedback on Writing](#)

Abdulrahman Nasser Alqefari

[WhatsApp as a Supporter Tool in Language Learning: A Study of Saudi EFL Learners' Perceptions](#)

Mohammad Yahya Ali Bani Salameh

[Sylvia Plath's The Bell Jar: A Feministic Reading](#)

Naeemah Alrasheedi

[Saudi Translation Agencies and Translation Centers: A Study of Post-Editing Practices](#)

Bodour Ali Alshehri, Noha Abdullah Alowedi

[Improving Writing Constructs and Performance Through Vlog-Assisted Language Learning \(VALL\)](#)

Jupeth T. Pentang, Sanny S. Maglente, Ma. Estela A. Sescon, Francia Formalejo Murao, Minsoware S. Bacolod, Cheryl J. Juancho, Leonilo B. Capulso, Michael Bhubet B. Baluyot, Jaypee R. Lopres, Hajdari Hazir, Hajdari Besnik

[Studying English in the USA: A Study of Saudi Learners' Perceptions](#)

Sultan Ayed Alanazi

[My Self-Perspective as Future English Language Teacher Analysis of the Predictive Power of Mentoring Process](#)

Sanny S. Maglente, Merlyn N. Luza, Leonilo B. Capulso, Jaypee R. Lopres, Cinder Dianne L. Tabiolo, Eduardo C. Mira, Anshu Mathur, Pratimam Saxena, Jayashree Premkumar Shet, Hajdari Besnik, Hajdari Hazir

[The Impact of Teacher Quality Management on Student Performance in the Education Sector: Literature Review](#)

Guo Qingyan, Ali Sorayyaei Azar, Albattat Ahmad

[Reading Comprehension and Behavior in Children Using E-books vs. Printed Books](#)

Georgios A. Moutsinas, Juan Carlos Orosco Gavilán, Cesar Emmanuel Cubas Ramirez, Bernardo Cespedes Panduro, Oblitas Gonzales Anibal, Dang Lam Ngoc Dieu, Sadia Naz

[Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era](#)

Wong Wei Lun, Mazura Mastura Muhammad, Warid Mihat, Muhammad Syafiq Ya Shak, Mairas Abdul Rahman, Prihantoro Prihantoro

Vocabulary Index as a Sustainable Resource for Teaching Extended Writing in the Post-Pandemic Era

Wong Wei Lun¹, Mazura Mastura Muhammad¹, Warid Mihat², Muhammad Syafiq Ya Shak³, Mairas Abdul Rahman⁴ & Prihantoro⁵

¹ Department of English Language and Literature, Sultan Idris Education University, Malaysia

² Academy of Language Studies, University Technology MARA Kelantan Branch, Malaysia

³ Academy of Language Studies, University Technology MARA Perak Branch, Malaysia

⁴ Faculty of Languages and Communication, University Sultan Zainal Abidin, Malaysia

⁵ Universitas Diponegoro, Indonesia

Correspondence: Mazura Mastura Muhammad, Department of English Language and Literature, Sultan Idris Education University, Malaysia.

Received: January 5, 2023 Accepted: February 15, 2023 Online Published: March 17, 2023 doi:10.5430/wjel.v13n3p181

URL: <https://doi.org/10.5430/wjel.v13n3p181>

Abstract

In the wake of the COVID-19 pandemic, Malaysian English teachers identified a pressing need to support upper primary school pupils, particularly those in the upper levels, in the effective composition of extended writing. Additionally, these educators required more innovative methodologies for teaching vocabulary in this context. Consequently, the current study aimed to develop a vocabulary index as a suggested resource for Malaysian English teachers instructing upper primary school pupils on extended writing. To achieve this, a quantitative computational research strategy and corpus-driven research design were employed. A purposive sampling technique was used to select 560 advanced upper primary school pupils from 28 schools, each with high English performance in the capital of each state and the federal territory of Malaysia, who produced a total of 152,187 words in extended writing for analysis. LancsBox, a primary computational linguistics application, was used for data processing. Given that the vocabulary index for extended writing necessitates a more comprehensive coverage of vocabulary, functional and content words were included, and keywords, raw and normalised frequencies were analysed and reported. Through the vocabulary index built in this study, the researchers found English teachers in Malaysia should utilise local issues in writing prompts, emphasise the use of both positive and negative adjectives, introduce complex sentence structures to enhance pupils' writing abilities and also train pupils to organise the ideas in their writing. Future linguistic studies could replicate the present investigation, so that it can respond to their classroom needs.

Keywords: vocabulary index, teaching extended writing, Malaysian primary school learners, COVID-19

1. Introduction

In 2023, Malaysian primary school pupils returned to in-person instruction after a prolonged period of online learning (Jidin, 2020), as it became evident that the quality of education had been severely impacted (Thang et al., 2022). Scholars noted that pupils were less engaged in online lectures and often failed to submit assignments on time, resulting in lower academic performance and a need for increased revision (Thang et al., 2022). Such challenges were particularly pronounced in extended writing, which had been identified as a concern even prior to the pandemic (Cambridge Assessment, 2013). Given the adverse effects of online learning on writing skills, it is possible that only fewer pupils may now be at the CEFR-A2 writing level, as compared to before the pandemic. Nonetheless, researchers noted that pupils struggled to meet learning expectations and produce high-quality work due to the challenges they faced during online learning, including poor internet access, lack of personal coaching, and inadequate academic support from parents (Thang et al., 2022; Jaafar et al., 2022). These challenges were especially pronounced in rural areas where the lack of infrastructure compounded the difficulties associated with online learning. As schools reopened, many teachers faced significant challenges in teaching writing due to the gaps created by these phenomena.

In response to these issues, a study was proposed to produce a vocabulary index for Malaysian English teachers instructing extended writing to upper primary school pupils. This index aimed to provide a practical and relevant resource for teachers and pupils to use in the writing session. While some researchers suggested using textbooks and dictionaries, the study focused on identifying authentic and relevant vocabulary for Malaysian upper primary school pupils to use in their extended writing. By addressing the issues caused by online learning, this study aimed to help pupils produce high-quality writing that met the expectations of their teachers and achieved success in their assessments through this question: 1. What is the most salient vocabulary discovered from the learner corpora/Advanced Malaysian Upper Primary School Learners Corpus (henceforth, AMUPSLC)?

2. Literature Review

Corpus-driven research has developed as a crucial method for examining language use and development, especially in the fields of second language acquisition and language instruction (Wong et al., 2022). Corpus-driven research entails the collection and analysis of corpora of large language data to inform language education and language learning (McEnery & Wilson, 2001). The term "corpus" is derived from the Latin word "corpus," which means "body." It refers to a collection of authentic linguistic data that can be used to research language use, structure, and variation. In the context of corpus-driven research, corpora can be compiled from several sources, such as written texts, spoken language, and learner data (Biber et al., 1994).

It is impossible to overestimate the significance of corpus-driven research in second language acquisition and language instruction. This methodology offers researchers and language educators access to vast quantities of actual language data, which can enrich our understanding of language usage and development. By analysing patterns and structures within the language data, researchers can uncover crucial elements of second language acquisition and usage, which can inform the creation of more successful language teaching materials and pedagogical procedures (Tognini-Bonelli, 2001).

This literature review examines the significance of corpus-driven research in second language acquisition and language instruction, with an emphasis on its impact on the field of language education. This study seeks to provide a clear and short overview of the fundamental concepts, techniques, and applications of corpus-driven research, as well as to highlight its potential for boosting language teaching in both theory and practice.

2.1 The Significance of Corpus-Driven Research

2.1.1. Evolution and Advancement of Corpus-Driven Research Over Time

Over the years, corpus-driven research has seen substantial development and evolution. In the 1960s and 1970s, corpus-driven research arose as a methodology for researching language use by collecting and analysing vast corpora of linguistic data (Baker, 2006). Early corpora focused primarily on analysing language use in circumstances involving native speakers. With the advent of the computer age and the increasing availability of language data, however, there was a shift towards researching language use in second language environments. This trend has led to a greater emphasis on learner corpora as a research tool for second language acquisition and language instruction.

2.1.2 Effect of Corpus-Driven Research on Second Language Acquisition and Instruction

Corpus-driven research has had a substantial effect on our knowledge of second language acquisition and language instruction (Smith, 2018). With the collection and analysis of vast quantities of language data, corpus-driven research has provided academics with insights into the patterns of language usage among second language pupils, which has contributed to our knowledge of the development of second language competence. For instance, corpus-driven research has indicated that the use of formulaic language by second language pupils is a significant part of their capacity to effectively communicate in the target language. In addition, corpus-driven research has yielded important insights into the lexical and grammatical characteristics of second language learners, which can guide language instruction (Poole, 2021).

2.1.3 Learner Corpora's Function in Corpus-Driven Research

Learner corpora are now an integral part of corpus-driven research in second language acquisition and language instruction (Biber et al., 1994). These corpora are vast collections of language data generated by second language learners, and they provide academics and teachers with significant insights on the linguistic characteristics of second language learners' language use. Learner corpora have been used to explore a wide range of themes, including second language writing development, the acquisition of grammar and vocabulary, and second language pupils' use of formulaic language (Wong et al., 2022). In addition, learner corpora have been utilised to inform the creation of language teaching materials and pedagogical approaches that take into account the demands of second language pupils.

2.2 Data Collection in Corpus Studies

Collecting and analysing enormous amounts of language data from multiple sources is required for corpus-driven research. This section will focus on the methodologies and approaches utilised in corpus-driven research.

2.2.1 Collecting and Annotating Corpus Information.

Typically, corpus data is obtained from numerous sources, including written texts, spoken language, and social media platforms (McEnery & Wilson, 2001). The data is subsequently annotated with linguistic metadata in order to simplify analysis. Annotation can be performed manually or mechanically and involves tagging data with information such as part of speech, grammatical structure, and semantic data.

2.2.1 Analyzing Corpus Data Using Quantitative and Qualitative Approaches

In corpus-driven research, enormous amounts of language data are analysed to find patterns and tendencies. This analysis can be conducted both quantitatively and qualitatively (Rayson, 2008). Quantitative analysis uses statistical approaches to detect patterns in the data, whereas qualitative analysis requires a detailed examination of the data to identify themes and patterns.

2.2.3 Software Instruments for Corpus-Driven Research

For Corpus-driven research, a variety of software tools are available, including concordancers, which enable researchers to swiftly search and analyse vast amounts of text data (Sinclair, 1991). Additional software tools include annotation tools that ease the annotation of corpus data and statistical tools that enable academics to conduct statistical analysis on the data.

2.3 Challenges and Opportunities for Corpus-Driven Research in Bilingual Settings

Corpus-driven research has made a significant impact on second language acquisition and language teaching, as well as the development of language teaching materials and pedagogical approaches (Tognini-Bonelli, 2001). This part will focus on the application of corpus-driven research in bilingual settings, examining the challenges and opportunities, the role of corpus-driven research in understanding language use, and the use of corpus-driven research in developing language teaching materials for bilingual learners.

2.3.1 Obstacles and Prospects for Corpus-driven Research in Bilingual Contexts

Bilingual settings present unique challenges and opportunities for corpus-driven research (Goyak et al., 2021). Bilingualism involves the use of two or more languages in everyday communication, and it is often accompanied by language contact and code-switching. These factors can complicate the collection and analysis of corpus data, making it challenging to identify language-specific patterns of use. However, bilingual settings also offer valuable insights into the interaction between languages and the development of bilingual competence. Corpus-driven research can be used to investigate the acquisition and use of multiple languages in bilingual settings, providing a more comprehensive understanding of bilingualism.

2.3.2 Function of Corpus-driven Understanding Language Usage in Bilingual Environments Research

Research based on corpora has been essential in expanding our understanding of language use in bilingual situations (Baker, 2006). By analysing vast quantities of language data, corpus-driven research can detect patterns of use, such as code-switching and borrowings, and analyse how these vary across various bilingual settings. For instance, Li (2019) investigated the use of English and Cantonese in a multilingual community in Hong Kong using corpus data. The study indicated that code-switching was ubiquitous in this community's communication and was utilised to communicate a variety of communicative roles. In addition, the study discovered variations in code-switching patterns across various social and situational circumstances. These results demonstrate the significance of corpus-driven research for comprehending the intricacies of language use in bilingual contexts.

2.3.3 Use of Corpus-Driven Research in the Creation of Instructional Materials for Bilingual Students

Research based on corpora has also been utilised to create language training resources that are specifically targeted to the needs of bilingual learners (Andrushenko, 2021). By examining the language use of bilingual speakers, corpus-driven research can discover language elements that are particularly difficult for bilingual learners and build instructional resources to target these difficulties. For instance, Liu (2018) investigated the use of connectives in English writing by bilingual Chinese students using a corpus-driven approach. The study discovered frequent errors in the use of connectives and produced a pedagogical intervention centred on these problem areas. It was determined that the intervention improved the students' use of connectives in their writing.

In conclusion, corpus-driven research has proven to be a powerful tool for examining language use in multilingual settings, expanding our understanding of the intricacies of bilingualism, and producing language teaching resources to meet the requirements of bilingual learners (Stewart et al., 2018). Even though there are problems in collecting and evaluating corpus data in multilingual contexts, the opportunity to gain insights into the interaction between languages and the development of bilingual competence make this area of research very relevant. As a result, corpus-driven research should remain a central focus of research in bilingual contexts, leading to the creation of more effective language teaching resources and pedagogical techniques for bilingual learners.

2.3.4 Corpus Studies in Malaysia

Corpus studies have become increasingly important in language education, offering valuable insights into language teaching and learning. In Malaysia, corpus studies have the potential to contribute to the creation of pedagogical interventions, analysis of current achievement, and policy improvement. This literature review explores the potential benefits of corpus studies in Malaysia.

To begin with, corpus studies can be used to create effective pedagogical interventions. For example, Mustafa and Abdullah (2017) used corpus data to identify common writing errors made by Malaysian ESL learners, which informed the development of a targeted intervention that significantly improved their writing skills. Similarly, Ibrahim and Latif (2020) developed a vocabulary learning intervention for Malaysian students using corpus data, resulting in improved vocabulary acquisition.

Corpus studies can also provide insights into current levels of language proficiency and achievement. Azhar et al. (2018) used corpus data to analyse the language proficiency of Malaysian undergraduates in their academic writing, revealing areas of weakness that could be addressed through targeted interventions. Furthermore, corpus studies can contribute to policy improvement by providing evidence-based insights into the effectiveness of current language policies. Mukundan and Nimehchisalem (2018) used corpus data to evaluate the Malaysian English Language Teaching Curriculum for secondary schools, identifying areas for improvement such as the need to focus more on grammar and vocabulary.

In addition to these potential benefits, corpus studies in Malaysia can contribute to the development of a corpus-based syllabus. This approach uses corpus data to identify the most relevant language items for teaching, ensuring that language teaching focuses on the language learners are most likely to encounter and use in real-life situations (Mohamed, 2016). Corpus studies can also be used to explore language variation in Malaysia, given its multilingual and multicultural context. They can help to investigate patterns of language use and variation across different linguistic and social contexts, which can inform language policy and planning (Mukundan & Krishnasamy, 2018). Finally,

corpus-based assessments can provide a more accurate representation of language proficiency, helping to prepare learners for language exams and real-life language use (Abdullah et al., 2019).

Overall, corpus studies have significant potential to contribute to language education in Malaysia, including effective pedagogical interventions, improved language assessment, and policy improvement. Further research is recommended to explore the full potential of corpus studies in the Malaysian context. In conclusion, corpus studies offer a promising approach for improving language teaching and learning in Malaysia. Hence, this study addressed the research gap by utilizing a corpus-driven methodology for the construct of the vocabulary index.

3. Research Design of the Study

3.1 Method

This study was of corpus driven research aiming at finding salient vocabulary used in extended writing among advanced Malaysian upper primary school pupils. Corpus research design is a rigorous methodology that employs large electronic databases of texts, commonly referred to as corpora, to investigate patterns and use of language. The process involves compiling a representative corpus of a particular language or genre and using computational tools and techniques to analyze the data within it. This method is versatile, enabling researchers to study a broad range of linguistic phenomena such as vocabulary use, grammatical structures, discourse patterns, and stylistic features. It is a valuable tool for both descriptive and applied linguistics, providing insights into language use and variation that can inform language teaching, language policy, and other language-related fields.

3.2 Participants

Table 1. Demographic data of AMUPSLC

| East/West Malaysia | Region | Capital | No. of Learners/Extended Writing | No. of Tokens | No of Tokens per Region |
|-----------------------|--------------------|-----------------|-------------------------------------|---------------|----------------------------|
| West | Northern Region | Perlis | 40 | 9,855 | 46,185 |
| | | Alor Setar | 40 | 8,260 | |
| | | Georgetown | 40 | 10,392 | |
| | | Ipoh | 40 | 17,678 | |
| | Central Region | Kuala Lumpur | 40 | 17,519 | 26,969 |
| | | Shah Alam | 40 | 11,264 | |
| | | Malacca City | 40 | 12,951 | |
| | | Johor Bahru | 40 | 6,981 | |
| | Southern Region | Negeri Sembilan | 40 | | 31,017 |
| | | | | | |
| | East Coast | Kuantan | 40 | 8,977 | 28,443 |
| | | Kuala | 40 | 10,946 | |
| | | Terengganu | 40 | | |
| | | Kota Bahru | 40 | 8,520 | |
| Sabah | Kota Kinabalu | 40 | 8,692 | 19,573 | |
| | | | | | |
| East | Sarawak | Kuching | 40 | 10,881 | |
| Total | 6 | 14 | 560 | 152,187 | 152,187 |

Table 1 displays the demographic characteristics of the participants who took part in the study. To ensure representativeness, the participants were drawn from six distinct regions. The selection process was carefully designed to ensure that each region was equally represented in the study. It is worth noting that the study focused on a very specific subset of the participants: only twenty advanced primary school pupils from each school were selected to complete the extended writing assignment. Hence, there were forty advanced primary school pupils from each state. This allowed for a more focused and in-depth analysis of the writing abilities of this group of students. In order to form a vocabulary index as a reference for Malaysian upper primary school pupils and English teachers in extended writing, the vocabulary employed by advanced upper primary school pupils appeared to be a suitable reference for intermediate and low English proficiency pupils. As a consequence of this targeted approach, the number of advanced primary school pupils and the number of extended writing assignments were determined concurrently. This allowed for a more precise assessment of the writing abilities of this specific group of students. It is important to clarify that the number of tokens identified during the analysis process and the overall number of tokens present in each region are two separate metrics. The former refers to the individual units of language that were identified and analysed during the study, while the latter refers to the total number of words in the region as a whole.

3.3 Data Collection Flow

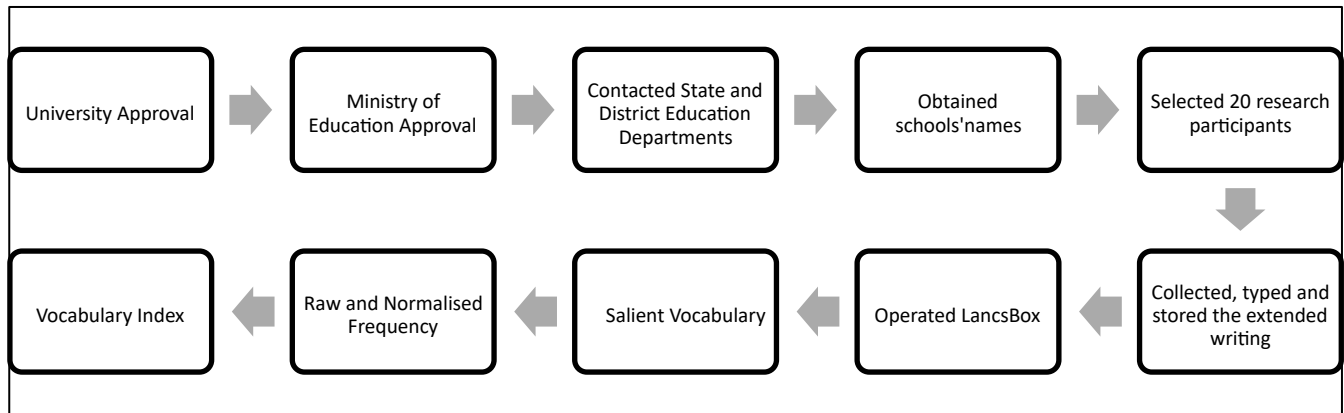


Figure 1. Research Flow

The researchers started by obtaining consent from the institution (Ethics referral code: 2021-0454-01) to conduct the study, and subsequently secured the approval of the Ministry of Education Malaysia (Ethics referral code: KPM.600-3/2/3-eras(11456)). The next phase involved reaching out to state and district education authorities to compile a list of 28 schools that had demonstrated high levels of English proficiency. Subsequently, a sample of 20 advanced upper primary pupils from each school was selected based on their ability to produce extended writing. The researchers allowed the 20 advanced upper primary pupils from each school in the study to choose the topic of their writing, without any restrictions on the scope or subject matter. This approach has several benefits. Firstly, it increases the authenticity of the writing produced and can lead to higher quality output. Secondly, it allows for a wider range of writing topics and styles to be captured and thirdly, it promotes ownership and autonomy among the participants, which can have positive effects on motivation and self-esteem. To this end, a total of 560 extended writing samples were collected, transcribed, and compiled in a laptop for further analysis using the software program LancsBox. The analysis focused on identifying salient vocabulary, both functional and content-related. The top 20 most frequently occurring vocabulary items, based on raw and normalised frequency, were selected for inclusion in the resulting vocabulary index. Overall, the study yielded a valuable resource in the form of the vocabulary index, which has potential applications for enhancing the teaching and learning of extended writing skills among primary school learners in Malaysia

3.4 Data Analysis

All 560 extended writing were divided into the following categories: Northern, Central, Southern Regions, East Coast, Sabah, and Sarawak. The type-token-ration was stated first. The type-token ratio (henceforth, TTR) is a metric for determining the degree of vocabulary variety within a written text or in spoken language. Two real-world examples are used to calculate and analyse TTR. The TTR is demonstrated to be a useful indicator of textual lexical variation. It can be used to track changes in children and adults who have difficulty with their language. TTR is a ratio calculated by dividing the types (the total number of distinct words) contained in a text or utterance by the tokens (the total number of words). A high TTR reflects a great degree of lexical diversity, whereas a low TTR reflects the inverse. The range is theoretically between 0 (infinite repetition of a single kind) and 1 (the complete non-repetition found in a concordance). Researchers have occasionally reported this TTR as a percentage by multiplying the ratio by 100. This is an unnecessary calculation, as the ratios are sufficiently illustrative.

Additionally, some studies (Kliefgen, 1985) adopt a “token-type” ratio rather than the more prevalent “type-token” ratio. The number of tokens is divided by the number of kinds in these experiments. The results are reported in a range, with a TTR of 100 indicating the maximum degree of variance conceivable and higher ratios indicating minor variation. TTRs were used to investigate the lexical complexity (or, more precisely, the lexical variation revealed by TTRs) of AMUPSLC in this study. The formula used is as follows:

$$TTR = (number\ of\ types/number\ of\ tokens)*100$$

4. Findings

What is the most salient vocabulary discovered from the learner corpora/Advanced Malaysian Upper Primary School Learners Corpus?

Table 2 and Figure 2 below illustrate the TTR for West Malaysia (Northern Region, Central Region, Southern Region, and East Coast) and East Malaysia (Sabah and Sarawak).

Table 2. TTR of West and East Malaysia

| Malaysia | Region/State | No. of Types/No. of Tokens | TTR (%) |
|----------|-----------------|----------------------------|---------|
| West | Northern Region | 5,675/46,185 | 12.28 |
| | Central Region | 4,263/26,969 | 15.81 |
| | Southern Region | 4,511/31,017 | 14.54 |
| | East Coast | 3,589/28,443 | 12.62 |

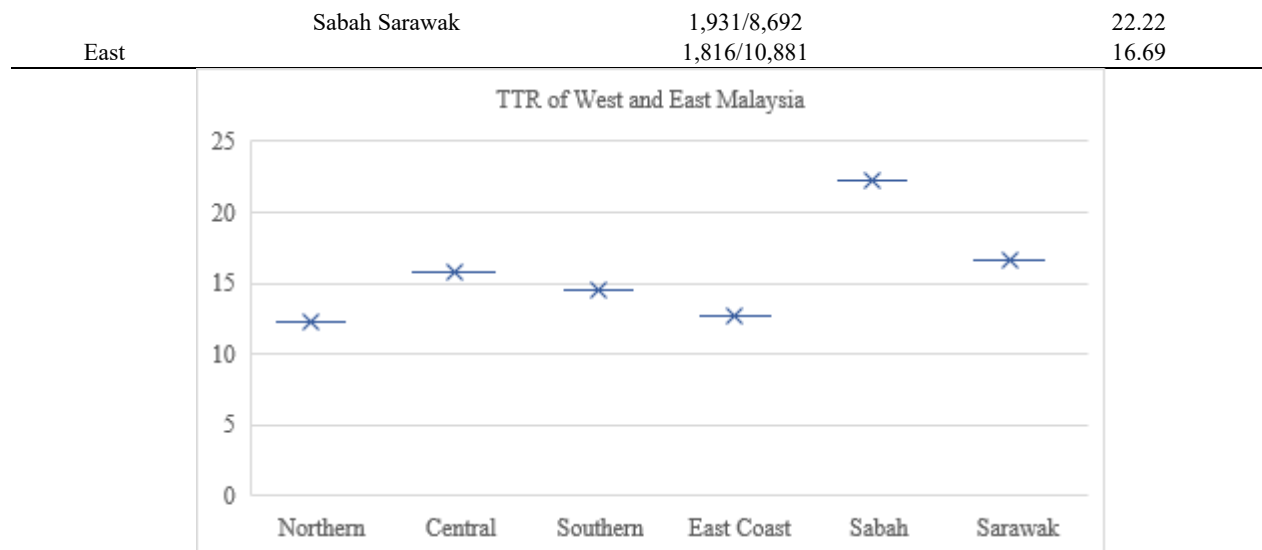


Figure 2. TTR of West and East Malaysia

As depicted in both Table 2 and Figure 2, the Type-Token Ratio (TTR) is not impacted by the number of tokens and types. For instance, despite the Northern Region having the most tokens and types, its TTR is the lowest among the six regions, standing at 12.28%. This low TTR suggests that the sub corpus in question displays limited lexical variance. Notably, the extended writing of 160 advanced Malaysian upper primary school pupils from the Northern Region revealed a dearth of vocabulary. Conversely, Sabah exhibits the highest TTR rate of 22.22%, indicating considerable lexical variance, despite having the fewest tokens and the second-fewest types.

Examining Figure 2, the TTR of the East Coast is comparable to that of the Northern Region, both standing at 12.62% and 12.28%, respectively. Allegedly, 120 advanced Malaysian upper primary school pupils from the East Coast employed a limited vocabulary in their extended writing. It is possible that the vocabulary utilized between the Northern Region and East Coast sub corpora is similar.

On the other hand, the Southern Region, the Central Region, and Sarawak have nearly identical TTRs of 14.54%, 15.81%, and 16.69%, respectively. While these TTR values are lower than Sabah's, they are higher than those of the Northern Region, implying moderate lexical variance within these three sub corpora. Consequently, it is reasonable to infer that 240 advanced Malaysian primary school pupils from these regions may exhibit similar English proficiency levels and vocabulary banks for extended writing. Tables 3 – 7 further describe the salient words observed in this study according to the categories and regions. Table 3. Summary of Regional Salient Functional Words

| Rank | NR | CR | SR | EC | Sabah | Sarawak |
|------|------|------|------|------|-------|---------|
| 1 | the | the | the | the | the | the |
| 2 | to | to | and | to | I | to |
| 3 | and | and | to | and | and | and |
| 4 | a | of | a | a | to | a |
| 5 | I | a | I | is | my | of |
| 6 | of | is | of | I | a | they |
| 7 | we | I | my | in | was | I |
| 8 | in | that | is | of | of | we |
| 9 | is | in | was | my | it | in |
| 10 | was | it | we | that | is | was |
| 11 | my | he | in | we | we | it |
| 12 | it | for | it | it | that | were |
| 13 | that | are | that | are | in | my |
| 14 | for | this | for | for | me | for |
| 15 | as | my | they | with | she | their |
| 16 | you | his | the | you | he | with |
| 17 | are | be | and | was | on | that |
| 18 | they | we | to | can | at | some |
| 19 | with | was | a | they | for | at |
| 20 | be | with | I | on | but | on |

Table 4. Summary of Regional Salient Nouns

| Rank | NR | CR | SR | EC | Sabah | Sarawak |
|------|----------|----------|-------------|--------------|---------|----------|
| 1 | people | people | time | person | day | Covid-19 |
| 2 | day | time | day | beauty | family | school |
| 3 | time | games | people | people | time | scouts |
| 4 | school | food | food | day | mother | bus |
| 5 | friends | life | friends | tiger | sister | trip |
| 6 | family | video | family | time | father | lunch |
| 7 | fox | monster | person | family | dad | time |
| 8 | food | things | school | children | water | camping |
| 9 | life | school | life | man | name | virus |
| 10 | Kaeya | Covid-19 | competition | universities | school | day |
| 11 | world | Gatsby | home | life | house | friends |
| 12 | way | family | father | school | things | campsite |
| 13 | grapes | world | place | friends | home | pandemic |
| 14 | person | day | friend | others | brother | teacher |
| 15 | home | movie | pollution | Covid-19 | night | home |
| 16 | students | Ralph | Malaysia | university | friends | twigs |
| 17 | things | friends | mother | mother | people | tents |
| 18 | car | society | world | vaccine | games | disease |
| 19 | house | pandemic | house | conclusion | fruits | life |
| 20 | mother | way | years | future | Alice | area |

Table 5. Summary of Regional Salient Verbs

| Rank | NR | CR | SR | EC | Sabah | Sarawak |
|------|---------|---------|---------|---------|---------|-----------|
| 1 | went | go | went | get | do | started |
| 2 | said | want | saw | go | went | went |
| 3 | go | get | make | do | get | go |
| 4 | got | going | help | make | told | got |
| 5 | get | love | go | makes | like | affected |
| 6 | make | hope | do | went | got | cleaned |
| 7 | do | make | got | think | heard | get |
| 8 | going | see | love | help | came | arrived |
| 9 | saw | went | took | give | saw | organised |
| 10 | asked | help | know | love | play | packed |
| 11 | need | think | get | need | see | made |
| 12 | know | take | going | play | go | make |
| 13 | take | become | started | spend | know | played |
| 14 | see | shows | came | take | woke | felt |
| 15 | took | find | told | look | called | came |
| 16 | eat | say | want | see | started | stay |
| 17 | called | started | decided | want | said | follow |
| 18 | started | need | take | receive | make | going |
| 19 | want | got | said | getting | spend | spread |
| 20 | made | live | felt | know | took | set |

Table 6. Summary of Regional Salient Adjectives

| Rank | NR | CR | SR | EC | Sabah | Sarawak |
|------|-----------|------------|-----------|-----------|-----------|-------------|
| 1 | good | good | best | beautiful | new | last |
| 2 | beautiful | new | beautiful | good | good | happy |
| 3 | happy | violent | good | favourite | big | enjoyable |
| 4 | new | online | happy | important | little | tired |
| 5 | best | better | new | kind | unique | excited |
| 6 | first | different | old | inner | excited | first |
| 7 | last | same | favourite | first | last | long |
| 8 | online | beautiful | first | best | old | new |
| 9 | long | long | last | proud | scared | online |
| 10 | hungry | best | great | physical | beautiful | hard |
| 11 | important | aggressive | different | little | best | fun |
| 12 | own | favourite | hard | hard | fun | respiratory |

| | | | | | | |
|----|-----------|-----------|-----------|--------|-------------|--------|
| 13 | big | first | online | new | different | sick |
| 14 | delicious | great | important | happy | traditional | early |
| 15 | better | free | same | public | amazing | small |
| 16 | old | local | better | better | happy | whole |
| 17 | same | healthy | mental | old | small | able |
| 18 | fun | big | delicious | last | favourite | known |
| 19 | favourite | social | long | shiny | wonderful | common |
| 20 | hard | important | true | own | louder | social |

Table 7. Summary of Regional Salient Adverbs

| Rank | NR | CR | SR | EC | Sabah | Sarawak |
|------|--------|--------|--------|----------|----------|----------|
| 1 | not | not | also | not | when | when |
| 2 | when | also | when | also | also | also |
| 3 | also | when | not | when | up | up |
| 4 | so | more | so | very | not | not |
| 5 | then | so | up | then | so | so |
| 6 | there | how | very | how | very | very |
| 7 | up | very | always | always | out | all |
| 8 | very | out | out | just | just | together |
| 9 | just | up | how | so | more | back |
| 10 | out | just | there | up | really | then |
| 11 | always | really | back | even | still | how |
| 12 | back | even | then | more | then | even |
| 13 | all | back | all | all | again | really |
| 14 | still | always | really | only | there | now |
| 15 | again | only | just | out | always | out |
| 16 | really | then | even | still | even | finally |
| 17 | even | there | down | together | back | later |
| 18 | now | most | as | too | off | more |
| 19 | how | too | only | however | now | still |
| 20 | well | again | well | most | together | again |

Tables 3-7 describe the 20 salient vocabulary (functional and content) used by the participants in this study in their extended writing products according to each category in each region. The raw frequencies have been normalised to as per ten thousand words for uniform representation. The formula applied is $(NRF \times 104) / T = NNF$ whereby, NR = value of the raw frequency of the salient vocabulary, T = total tokens of the subcorpus (46,185) and NNF = value of the normalized frequency. The figures for percentage of distribution are rounded off to the nearest two decimal digits. The formula applied is $(NNF / 104) \times 100\% = NPD\%$. whereby, NNF = value of the normalized frequency, NPD = value of the percentage of occurrence (%).

The findings presented in Tables 3 - 7 indicate that functional words and adverbs displayed similarities across different regions, while nouns, verbs, and adjectives exhibited considerable differences. These similarities and differences are critical as they reveal the relevance of specific vocabulary in the context of Malaysian primary school education. The fact that functional words and adverbs are commonly used across different regions implies that they are essential components of the English language that learners must acquire to communicate effectively. The disparities observed in the usage of nouns, verbs, and adjectives are also significant for English instructors as they highlight the areas where learners need to focus on to improve their writing skills.

Moreover, the collected nouns, verbs, and adjectives were all authentic, relevant, and important for every Malaysian primary school pupil for extended writing. This suggests that pupils who acquire these words will be better equipped to produce written works that are not only grammatically correct but also coherent, engaging, and informative. Additionally, the disparities observed in the usage of nouns, verbs, and adjectives provide English instructors with an opportunity to tailor their teaching methods to address the specific needs of their pupils. For instance, if pupils in a particular region struggle with the usage of adjectives, an English instructor can focus on providing more explicit instruction, examples, and exercises to improve their understanding and usage of this part of speech.

Overall, the findings in Tables 3 - 7 have significant implications for English language education in Malaysia. By identifying the similarities and differences in the usage of different parts of speech across different regions, English instructors can design more effective teaching strategies that are tailored to the needs of their pupils. Additionally, the importance of functional words and adverbs underscores the critical role that these components play in effective communication in extended writing activity. As such, English instructors should ensure that pupils have a solid grasp of these components before moving on to other parts of speech. By doing so, pupils will be better equipped to communicate effectively in written and spoken English, thereby improving their prospects in education and beyond.

5. Discussions

The primary research question concerns the salient vocabulary uncovered by AMUSLC. Upon reviewing the aforementioned data, it becomes apparent that the preferred language of advanced upper primary school pupils in each location shares similarities and variances.

The primary goal of corpus-driven research is to facilitate the development of genuine and authentic language components from research participants. Numerous scholars (Batchelor, 2023; Wong et al., 2022; Yang et al., 2022; Goyak et al., 2021) have employed corpus-driven studies for a variety of research purposes. Since different writers from various cultures and environments use different vocabularies in their extended writing, it is only possible for the results to be compared. However, the essence of the findings remains the same, as the language is allowed to develop naturally rather than a specific vocabulary being predetermined.

Sarjono et al. (2022) argued for the significance of function words such as *the*, *to*, and *because*, as they play a crucial role as articles, simple conjunctions, prepositions, and infinitives. Without these conjunctions, a statement may sound awkward and be grammatically incorrect. On the contrary, first-person pronouns such as *I* are frequently used in Sabah, as opposed to Struyk (2022), where most upper primary school pupils employ third-person pronouns in extended writing. Nonetheless, secondary school students frequently use *I* when required to provide their opinions and express their thoughts. The differences in the frequency of function words suggest that Sabah employs first-person pronouns frequently, as evidenced by the variances in the frequency of function words, where *me* is identified. Upper primary school pupils may be exposed to various types of essay writing and learning tools, resulting in the effective use of *I* in extended writing.

The word *person* is the most prominent noun on the East Coast. Using nouns, it can be said that upper primary school pupils in West Malaysia write about their environment and surroundings, such as their daily routines and the individuals they encounter regularly. This is similar to Siloka (2022), where the theme of the world of self, family, and friends is the primary theme introduced in primary school for learning English as a second language. *Day*, *time* (Sabah), and *school* (Sarawak) are the three most prominent nouns found in East Malaysia. Fischer et al. (2020) suggest a similar assumption that pupils find it easy to describe real-world issues. Interestingly, the differences in word frequency reveal even the new writing styles and essays. *Kaeya* (Northern), shows that sophisticated Malaysian upper primary school pupils might not like to use names like *Ali* and *Abu*, which were common in the 19th century because the learning resources are localised as opposed to globalised. They probably learned from the tale or film they watched. This result is comparable to the findings provided by Vanderplank (2019) because these names are uncommon in Malaysia. They were inevitably exposed to foreign substances. Because *fox* and *tiger* are generally introduced in linear and non-linear Malaysian English textbooks, advanced upper primary school pupils in these two locations are likely exposed to Malaysian folk or animal tales. It is supported by Nkomo (2022). Consequently, based on the identified foreign names and animal nouns, it would be possible to investigate the learning resources utilised by advanced Malaysian upper primary school pupils for extended writing.

The term *monster* suggests that virtual characters from anime or other media are a constant presence in the lives of Malaysian upper primary school pupils. These pupils may acquire knowledge through video games and online videos, as noted by Syafiq et al. (2021). The movement control order in 2019 led to the organisation of numerous online competitions, which may explain why pupils in this region are particularly interested in vocabulary *competitions*. It has been found that direct and genuine engagement fosters long-term memory retention, similar to the results proposed by Brady and Störmer (2022). Despite this, pollution remains a pressing concern for Malaysian school pupils, particularly in the Southern Region. Tay et al. (2021) suggest that this region is experiencing significant environmental challenges.

On the East Coast, there are different variations of the nouns *beauty*, *children*, *man*, *universities*, *vaccine*, *conclusion*, and *future*. The supplied vocabulary was highly factual and did not include any nonfiction works. Shakur et al. (2020) suggest that advanced upper primary school pupils in this region may have a greater sense of their future, especially concerning higher education and the significance of vaccines. Sarawak, on the other hand, included words such as *scouts*, *bus*, *trip*, *camping*, *campsite*, *twigs*, and *tents*, reflecting the culture of the school. Valdés (2018) has found that immersion in a culture can improve one's learning, which may account for the differences in vocabulary between regions. In terms of COVID-19 awareness, the Central Region and Sarawak demonstrate a similar level of awareness, with Sarawak having a higher number of COVID-19 related nouns.

In primary school, basic verbs such as *said*, *want*, *get*, *saw*, *make*, *do*, and *started* are fundamental to sentence formation. Suhaimi et al. (2019) have found this to be consistent with Malaysian upper primary school pupils. Northern Region pupils tend to favour the sentence structure "... (someone) said..." in their writing, primarily involving brief conversations. This suggests that pupils have their own preferences and linguistic skills beyond the fundamental writing pattern. Additionally, advanced Malaysian upper primary school pupils can write in multiple tenses, as demonstrated by Sahebkhair (2020). In Sarawak, verbs such as *affected* and *spread* are commonly used to describe COVID-19, indicating that verb usage reflects the contemporary challenges that pupils are facing. This aligns with Eggemeier et al. (2020). Looking at the adjectives they employ in their extended writing, the majority of advanced Malaysian upper primary school pupils have favourable moral beliefs. Their linguistics may reflect their moral principles (Białek et al., 2019) This is supported by investigation by Kubota (2020). In contrast, there is violence in the Central Region. Advanced upper primary school pupils in the Central Region are probably exposed to situations and issues such as online or domestic abuse. It is provided by Joharry and Turiman (2020). In addition, social media, such as Facebook and Tiktok, could spread violent content. Due to their substantial exposure to violence, pupils commonly use *violent* as an adjective in their essays. Under the influence of violence, people may suffer psychologically (Every-Palmer et al., 2020). Therefore, the scenario is highly perilous.

As demonstrated by Martins and Gresse Von Wangenheim (2022), pupils in the Northern Region of Malaysia begin their extended writing by describing their daily lives, with the words *hungry* and *fun* being particularly significant. Pupils find it simple to describe their daily lives since they are real. In the Central Region, the Southern Region, and the East Coast, pupils propose social science topics using factual adjectives, such as psychological issues (*mental*, *inner*), society issues (*local*, *public*), personalities (*kind*, *proud*), and health topics (*healthy*). It is possible that these pupils have been exposed to relevant environments or media, since technology is a vital tool for any family. The

Central Region is characterised by the pejorative adjective *aggressive*, which is intertwined with the topic of violence, bolstering the argument made in the preceding paragraph.

Sipatau et al. (2020) provide evidence that Sabah, a region in Malaysia, is rich in cultural elements, which can be characterised using unique adjectives such as *unique, old, beautiful, traditional, amazing, wonderful, and louder*. Sabah's culture is exemplified by its historical buildings and festivals. In contrast, Sarawak has a number of distinctive and unique words such as *tired, online, respiratory, and sick*, which may characterise their existence under COVID-19, with *respiratory* describing the effects of COVID-19 on people. Current events play an important role in extended writing, as noted by Eggemeier et al. (2020).

According to Appel and Szeib (2018), *when* and *also* play a crucial part in extended writing, with *when* supplying additional information by extending further as a clause and *also* demonstrating addition. Negation is a sentence form for writing in West Malaysia, with advanced upper primary school pupils preferring to use negation in extended writing, as shown in research by Goubault et al. (2020). This research demonstrates that primary school pupils have an advanced command of the English language, as they can form both affirmative and negative sentences as outlined in CEFR.

The frequency of adverbs in Malaysian learner corpora differs across regions, with the most prominent adverbs *now* in the Northern Region, *down* and *as* in the Southern Region, *together* and *however* in the East Coast, *just, there, always, and off* in Sabah, and *all, how, finally, and later* in Sarawak, as reported by Andrushenko (2021). The accessible types of adverbs include time (*now, later*), directions (*down, there*), linkers (*however, finally*), manners or degrees (*together, just, off, all, as, how*), and frequencies (*always*). Researchers such as Poole (2021), Andrushenko (2021), Stewart et al. (2018), and Li and Long (2022) have used adverbs for various research reasons, including stance adverbs, Middle English adverbs, adverbs related to fairness and justice, and temporal adverbial clauses for Chinese EFL learners, respectively. The significant adverbs revealed in this study contribute to the corpus research literature on Malaysian learner corpora, indicating that advanced Malaysian upper primary school students have the English proficiency to use adverbs for various purposes based on the categories listed above.

Based on the findings of the study, several recommendations related to vocabulary can be proposed. First, the topic for extended writing must be contextualised to pupils' social setting. Given the cultural diversity in Malaysia, English teachers can incorporate relevant local topics into their writing prompts, allowing pupils to apply their writing skills and experience in their writing product. Second, the analysis of regional adjectives highlights the significance of both positive and negative adjectives in extended writing. While it is important to instil moral values, it is equally important for pupils to develop proficiency in expressing both positive and negative aspects of a topic. Thus, English teachers should emphasise the usage of both types of adjectives to enhance pupils' writing abilities. The study also found that advanced upper primary school pupils are capable of using complex sentence structures in their writing. Therefore, English teachers should introduce compound and complex sentences as a training ground during writing activity in the classroom. Based on the data collected, these long sentences can be constructed from the use of *and, that, when* and *but*. Similarly, pupils should also get enough training in sequencing ideas especially using the word, *then*. In conclusion, the study recommends that English teachers in Malaysia should utilise local issues in writing prompts, emphasise the use of both positive and negative adjectives, introduce complex sentence structures to enhance pupils' writing abilities and also train pupils to organise the ideas in their writing. While these tips may appear trite, the vocabulary index constructed in this study identifies specific words that should be given priority during lessons, rather than from a vast repertoire.

6. Conclusion

The construction of a vocabulary index, in this case is identifying twenty significant vocabularies from diverse word categories in extended writing, presents a promising approach to enhance the teaching and learning of English, especially in the domain of extended writing among primary school pupils in Malaysia. The current discussion underscores the significance of fusing corpus-driven research into the production of instructional resources to bolster genuine English language acquisition. It is recommended that the Ministry of Education and education departments at state and district levels leverage the vocabulary index as a resource for English teachers from time to time, as it offers authentic and contextualized settings. Moreover, further research should be conducted to examine the effectiveness of the vocabulary index in elevating the quality of English language instruction in Malaysia. It is noteworthy that the vocabulary index approach is aligned with Malaysia's Sustainable Development Goals 2030, which accentuate the necessity of inclusive, equitable, and learner-centered quality education. The creation and application of pedagogical materials that amalgamate localised contexts and real language use can contribute towards achieving these objectives. To ameliorate English language teaching in Malaysia sustainably, policymakers and educators must scrutinise the potential of corpus-driven research and the generation of learner corpora.

Acknowledgments

The authors are grateful for the research funding provided by the Research Management and Innovation Centre (RMIC) UPSI for funding the journal publication.

References

- Abdullah, N. M., Mustafa, M., & Al-Shawawreh, F. A. (2019). Corpus-based assessment: Its advantages and disadvantages. *GEMA Online® Journal of Language Studies, 19*(4), 121-136.
- Andrushenko, O. (2021). Corpus-based studies of Middle English adverb largely: Syntax and information-structure. *X-Linguae, 14*(2), 60-75. <https://doi.org/10.18355/XL.2021.14.02.05>

- Appel, R., & Szeib, A. (2018). Linking adverbials in L2 English academic writing: L1-related differences. *System*, 78, 115-129. <https://doi.org/10.1016/j.system.2018.08.008>
- Azhar, N. A., Yusof, N. B., & Jaafar, N. (2018). Exploring language proficiency of Malaysian undergraduate students in academic writing using corpus-based analysis. *GEMA Online® Journal of Language Studies*, 18(3), 129-148.
- Baker, P. (2006). *Using corpora in discourse analysis*. Continuum International Publishing Group. <https://doi.org/10.5040/9781350933996>
- Batchelor, J. (2023). Just another clickbait title: A corpus-driven investigation of negative attitudes toward science on Reddit. *Public Understanding of Science*, 09636625221146453. <https://doi.org/10.1177/09636625221146453>
- Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, 85, 103855. <https://doi.org/10.1016/j.jesp.2019.103855>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511804489>
- Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(7), 942. <https://doi.org/10.1037/xlm0001014>
- Cambridge Assessment. (2013). Preparing for extended writing tasks: Challenges and supports. *Research Matters: A Cambridge Assessment Publication*, 15, 22-29.
- Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (2020). Workload assessment in multi-task environments. In *Multiple-task performance* (pp. 207-216). CRC Press. <https://doi.org/10.1201/9781003069447-12>
- Every-Palmer, S., Jenkins, M., Gendall, P., Hoek, J., Beaglehole, B., Bell, C., Williman, J., Rapsey, C. & Stanley, J. (2020). Psychological distress, anxiety, family violence, suicidality, and wellbeing in New Zealand during the COVID-19 lockdown: A cross-sectional study. *PLoS one*, 15(11), e0241658. <https://doi.org/10.1371/journal.pone.0241658>
- Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R. & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1), 130-160. <https://doi.org/10.3102/0091732X209093>
- Goubault, É., Lazić, M., Ledent, J., & Rajsbaum, S. (2020). A dynamic epistemic logic analysis of the equality negation task. In *International Workshop on Dynamic Logic* (pp. 53-70). Springer. https://doi.org/10.1007/978-3-030-38808-9_4
- Goyak, F., Muhammad, M. M., Mohd Khaja, F. N., Zaini, M. F., & Mohammad, G. (2021). Conversational mental verbs in English song lyrics: A corpus-driven analysis. *Asian Journal of University Education (AJUE)*, 7(1), 222-239. <https://doi.org/10.24191/ajue.v17i1.12619>
- Ibrahim, N. H., & Latif, A. A. (2020). A corpus-based vocabulary learning intervention for Malaysian ESL learners. *Indonesian Journal of Applied Linguistics*, 9(3), 728-738.
- Jaafar, N. M., Ng, L. S., Mahmud, N., Thang, S. M., & Mihat, W. (2022). An Investigation on the Online Learning Engagement of Malaysian Secondary School Students from Different School Types. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 12(4), 1-20. <https://doi.org/10.4018/IJCALLT.310079>
- Jidin, R. (2020). *Teks ucapan Sidang Pembukaan Semula Sekolah bagi Murid Buka Kelas Peperiksaan Awam*. Retrieved from <https://www.moe.gov.my/muat-turun/teks-ucapan-dan-slide/tu2020/3510-teks-ucapan-sidang-media-ybmk-pendidikan-berhubung-pe-mbukaan-semula-sekolah-bagi-murid-bukan-kelas-peperiksaan-awam-1-julai-2020/file>
- Joharry, S. A., & Turiman, S. (2020). Examining Malaysian public letters to editor on COVID-19 pandemic: A corpus-assisted discourse analysis. *GEMA Online Journal of Language Studies*, 20(3), 242-260. <http://doi.org/10.17576/gema-2020-2003-14>
- Kliefgen, J. A. (1985). „Skilled variation in a kindergarten teacher’s use of foreigner talk’. In Gass and Madden (1985). 89-114.
- Kubota, R. (2020). Confronting epistemological racism, decolonising scholarly knowledge: Race and gender in applied linguistics. *Applied Linguistics*, 41(5), 712-732. <https://doi.org/10.1093/applin/amz033>
- Li, W., & Long, Y. (2022). A Development Study on the Ordering Distribution of Temporal Adverbial Clauses by Chinese EFL Learners Based on Dependency Treebank. *Chinese Journal of Applied Linguistics*, 45(4), 551-565. <https://doi.org/10.1515/CJAL-2022-0404>
- Li, Y. (2019). Code-switching in a multilingual community in Hong Kong: A corpus-based investigation. *International Journal of Bilingualism*, 23(4), 914-934.
- Liu, S. (2018). Improving the use of connectives in English writing: A Corpus-driven pedagogical intervention for bilingual Chinese students. *Language Teaching Research*, 22(4), 449-469.
- Martins, R. M., & Gresse Von Wangenheim, C. (2022). Findings on Teaching Machine Learning in High School: A Ten-Year Systematic Literature Review. *Informatics in Education*. <https://doi.org/10.15388/infedu.2023.18>
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.

- Mohamed, N. (2016). The role of corpora in the development of a corpus-based syllabus for ESP writing. *Arab World English Journal*, 7(4), 296-307.
- Mukundan, J., & Krishnasamy, H. (2018). Variationist approaches to the study of Malaysian English: A review. *Journal of English as an International Language*, 13(2), 1-18.
- Mukundan, J., & Nimehchisalem, V. (2018). Corpus-based evaluation of the Malaysian English Language Teaching Curriculum for secondary schools. *The Asia-Pacific Education Researcher*, 27(4), 293-301.
- Mustafa, M., & Abdullah, N. (2017). Identifying common errors in Malaysian ESL learners' writing: A corpus-based study. *3L: Language, Linguistics, Literature*, 23(1), 1-12.
- Nkomo, S. (2022). Social networking sites in cultivating the reading habits of secondary school learners in Bulawayo, Zimbabwe. *South African Journal of Libraries and Information Science*, 88(1), 1-11. <http://dx.doi.org/10.7553/88-1-2114>
- Poole, R. (2021). A corpus-aided study of stance adverbs in judicial opinions and the implications for English for legal purposes instruction. *English for Specific Purposes*, 62, 117-127. <https://doi.org/10.1016/j.esp.2021.01.002>
- Rayson, P. (2008). From keywords to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549. <https://doi.org/10.1075/ijcl.13.4.06gray>
- Sahebkhair, F. (2020). Improving grammar achievement through using metatalk considering Iranian EFL learners with advanced language proficiency. *South Asian Research Journal of Arts, Language and Literature*, 2(4), 56-59. <https://doi.org/10.36346/sarjall.2020.v02i04.001>
- Sarjono, R. I. L., Heda, A. K., & Bram, B. (2022). Exploring collocations in Bahasa Inggris textbook: A corpus study. *Jurnal Penelitian Humaniora*, 23(2), 84-94.
- Shakur, E. S. A., Sa'at, N. H., Aziz, N., Abdullah, S. S., & Rasid, N. H. A. (2020). Determining unemployment factors among job seeking youth in the east coast of peninsular Malaysia. *The Journal of Asian Finance, Economics and Business*, 7(12), 565-576. <https://doi.org/10.13106/jafeb.2020.vol7.no12.565>
- Siloka, K. (2022). *A corpus linguistics study of English in written essays by third-year students in the Department of Wildlife Management and Ecotourism at the University of Namibia Katima Mulilo Campus* (Doctoral dissertation, Namibia University of Science and Technology). Retrieved from <http://ir.nust.na:8080/jspui/handle/10628/925>
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Sipatau, J. A., Mapjabil, J., & Imang, U. (2020). Diversity of rural tourism products and challenges of rural tourism development in Kota Marudu, Sabah. *Journal of Islamic*, 5(34), 51-61.
- Smith, J. (2018). The Use of Corpora in Second Language Acquisition Research. *Journal of Applied Linguistics*, 25(2), 45-62.
- Stewart, D., Biasi, P., & Binelli, A. (2018). *Using sketch engine: An analysis of five adverbs*. Academia.
- Struyk, M. W. (2022). *Exploring writing interventions for college students*. The University of Arizona.
- Suhaimi, N. D., Mohamad, M., & Yamat, H. (2019). The effects of WhatsApp in teaching narrative writing: A case study. *Humanities & Social Sciences Reviews*, 7(4), 590-602. <https://doi.org/10.18510/hssr.2019.7479>
- Syafiq, A. N., Rahmawati, A., Anwari, A., & Oktaviana, T. (2021). Increasing speaking skill through YouTube video as English learning material during online learning in pandemic covid-19. *Elsya: Journal of English Language Studies*, 3(1), 50-55. <https://doi.org/10.31849/elsya.v3i1.6206>
- Tan, K. E., & Ganakumaran, S. (2018). The Lexical Profile of Malaysian Upper Primary School Pupils' Writing. *Journal of Language and Communication*, 5(1), 1-14.
- Tay, S. I., Alipal, J., & Lee, T. C. (2021). Industry 4.0: Current practice and challenges in Malaysian manufacturing firms. *Technology in Society*, 67, 101749. <https://doi.org/10.1016/j.techsoc.2021.101749>
- Thang, S. M., Mahmud, N., Mohd Jaafar, N., Lay Shi Ng, L., & Noor Baizura, A. A. (2022). Online Learning Engagement Among Malaysian Primary School Students During the Covid-19 Pandemic. *International Journal of Innovation, Creativity and Change*. www.ijicc.net, 16(2), 2022. Retrieved from www.ijicc.net
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins. <https://doi.org/10.1075/scl.6>
- Vanderplank, R. (2019). „Gist watching can only take you so far”: Attitudes, strategies and changes in behaviour in watching films with captions. *The Language Learning Journal*, 47(4), 407-423. <https://doi.org/10.1080/09571736.2019.1610033>
- Wong, W. L., Muhammad, M. M., Zaini, M. F., Damit, A. R., Teoh-Ong, C., Singh, C. K. S., & Yusoff, N. (2022). Analysis of Covid-19 related phrases using corpus-based tools: Dualisms language & technology. *Journal of Positive School Psychology*, 6(3), 5034-5044.

Yang, Y., Yap, N. T., & Ali, A. M. (2022). A Corpus-based comparative study on syntactic complexity in university students' EFL writing in Southwestern China: A case of Pu'er University. *World Journal of English Language*, 12(8), 172-180. <https://doi.org/10.5430/wjel.v12n8p172>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).