Using Technologised Computational Corpus-Driven Linguistics Study on the Vocabulary Uses Among Advanced Malaysian Upper Primary School English as a Second Language Learners (ESL) in Northern Region

Wong Wei Lun1, Mazura Mastura Muhammad1,⁎, Warid Mihat2, Mairas Abdul Rahman3, Mohamad Syafiq Ya Shak4, Lee Mei Chew5

1 Department of English Language and Literature, Faculty of Languages and Communications, Sultan Idris Education University, 35900 Tanjong Malim, Malaysia
2 Academy of Language Studies, University Teknologi MARA Kelantan Branch, 18500 Machang, Malaysia
3 Faculty of Languages and Communication, University Sultan Zainal Abidin, 21300 Kuala Neru, Malaysia
4 Academy of Language Studies, University Teknologi MARA Perak Branch, 32610 Seri Iskandar, Malaysia
5 Department of Science and Mathematics, Faculty of Science and Mathematics, Sultan Idris Education University, 35900 Tanjong Malim, Malaysia

ARTICLE INFO

ABSTRACT

The vocabulary used by advanced Malaysian upper primary school learners is unknown after the Covid-19 pandemic, according to a corpus-driven study. This study primarily uses data from a learner corpus. This main purpose of this paper is to fill in this knowledge gap by examining the vocabulary usage in extended writing by advanced Malaysian upper primary school learners in the Northern Region. The study uses a corpus-driven research strategy and a quantitative methodology. Some 160 advanced upper primary school learners from the Northern Region, Malaysia were included in the study. They attended eight national primary schools with strong English programmes in the Northern Region. Purposive sampling technique was used to select the samples. This study’s main corpus-driven toolkit (LancsBox) was used to analyse the data. A total of 46,185 tokens from 160 extended writing tasks were analysed. The word frequencies were counted both raw and normalised. Results showed that 160 advanced Malaysian upper primary school learners in West Malaysia used the frequently and preferably in extended writing followed by the, and, vocabulary be is the twentieth salient vocabulary over other vocabulary such as to, and, a, and I. Depending on their perspectives, one may argue that they described their experiences in the first person. These findings imply that advanced Malaysian upper primary school learners use to, and, of, in, is, it, that, for, as and within their extended writing. The findings provide important knowledge and information for ESL teachers to improve the writing grammatical proficiency, especially for extended writing of ESL learners.

Keywords:
Corpus-driven; vocabulary; learner corpus; ESL learners

⁎ Corresponding author.
E-mail address: mazura@fbk.upsi.edu.my

https://doi.org/10.37934/araset.31.1.298314
1. Introduction

1.1 Research Background

Globally, corpus linguistics has grown into a major area of research [1]. Despite this, it has not been frequently subjected to rigorous exploration in Malaysia. Additionally, corpus-driven research focusing on upper primary school learners is still scarce. In other words, there is essentially a huge research gap for more local research on corpus linguistics [2]. In Malaysia, English is taught as a second language [3,4] and the English subject is compulsory for all students in national primary schools, regardless of their ethnicity or culture [5]. Furthermore, the Common European Framework of Reference (CEFR) serves as the key framework for assessment, teaching and learning English as a second language [6] and some of the success statements in CEFR are students can learn essay writing, vocabulary and phrases. The learning criteria are as follows: (a) 4.2.4 Use appropriate sentences to describe people, places, and objects; and (b) 4.3.2 Spell various high-frequency words accurately in autonomous writing. To meet these learning objectives, vocabulary and phrases are critical components of essay writing for language formation. When compared to other abilities such as listening, speaking, and reading, writing is considered the most challenging since it demands ESL students to be able to acquire both vocabulary and grammar concepts and use them to compose sentences correctly and effectively [7]. Teaching writing skills to ESL students is thus undoubtedly difficult [8-10].

As aforementioned, English is one of the compulsory subjects in Malaysian primary schools with the objectives of assisting the learners to master the language and the skills. The introduction of the CEFR emphasised the importance of students’ ability to write descriptive essays and to enable the students to compose clear and cohesive descriptive essays requires sufficient vocabulary and phrases. However, past research shows that many primary school learners struggled to employ suitable vocabulary and phrases to convey clear information in essays [11]. Moreover, research has shown that English teachers have encountered problems and are unsure of the appropriate phrases to teach to make a good essay [12,13] reiterates the lack of corpus research on the use of phrases for school learners and English teachers. [14] investigated the most prevalent errors made by Malaysian Chinese primary year 6 ESL students when writing in English as well as the degree to which the first language, Mandarin, affected ESL writing. English is written using alphabetic scripts, while Mandarin utilises a logographic writing system. Chinese ESL students find it difficult to master the English language because of the differences between the two language systems. A qualitative research methodology was employed which involved writing activities, with 15 ESL learners from primary year six selected as participants. The research found that the areas with the highest and most noticeable grammar mistakes were tenses, object pronouns, plurals, auxiliary verbs, prepositions, and articles. It was discovered that these inaccuracies were brought on by the influence of the mother tongue through direct translation from the Mandarin language. In conclusion, mistakes will inevitably occur, but they are essential for ESL students to advance in their English writing ability. Grammar is the key linguistic function of the English language that will be impacted by language transfer. For ESL learners, grammar is thought to be the most beneficial part of the language. Before the Malay language was replaced in public secondary schools in 1981, the language was widely utilised as a medium of teaching, hence it is essential for students in Malaysia to have a strong grasp of English grammar [7].

In another study carried out by [15] on Malaysian primary students’ problems with narrative writing, findings showed that the learners’ L1 had a significant impact on their L2. Other elements including the students' earlier experiences and helpful and constructive criticism on their writing could have contributed to their struggles with narrative writing. [16] agreed that writing is complex,
demanding, and requires organisation skills, writing mechanics, correctness in word choice, and appropriate knowledge of syntax and grammar; it is problematic for ESL learners. [17] concur that ESL learners struggle to create basic sentences, and the majority of them have difficulty creating compound and complicated sentences that contain dependent and independent clauses. Along with spelling, L1 interference, and language use (grammar), these are the main difficulties faced by ESL students [18]. However, previous studies have mostly focused on comprehending the difficulties that secondary and tertiary students have in learning. Very few studies explored vocabulary use among advanced learners in selected schools in Malaysia. Therefore, this study aims at collecting and analysing vocabulary use among advanced Malaysian upper primary school learners of English. The emphasis will be on writing ability, especially vocabulary components. The present study’s findings aimed at contributing to the current Malaysian expertise in CL research and the use of vocabulary in extended writing through examining advanced Malaysian upper primary school learners’ extended writing. This study will answer the following research question: What are the differences and/or similarities in the use of English vocabulary in the extended writing produced by advanced learners in upper primary schools in the northern region of Malaysia?

1.2 Literature Review and Previous Studies on Learner Corpora

Vocabulary is a critical component of language acquisition. School learners must improve their vocabulary to communicate their thoughts. According to [19], vocabulary is defined as school learners’ comprehension of oral and written words, including conceptual knowledge of the terms beyond their straightforward dictionary definitions. They emphasised that vocabulary acquisition is a continual process in which school learners make links to other words, study instances of related words, and eventually use the vocabulary correctly and appropriately within the context of the sentence. Furthermore, [20] defined vocabulary as a language’s words, including single words, sentences, or chunks that convey meaning. In this study, the term vocabulary refers to advanced Malaysian upper primary school learners’ words and phrases utilised in extended writing.

Based on the social constructivism perspective, humans view the world through eyes shaped and formed by experiences and interactions. Reality is unique to each individual and contextualised by the social context in which activities occur. If this is true, then learning is not universal, and learners will react differently to classroom surroundings, classes, and professors based on their prior experiences and previously constructed knowledge. Learners in a classroom bring a unique reality, or vision of reality, to their study, and together they will develop new understanding. With numerous alternate realities colliding in a common space, schools can be either rich or poor, depending on the atmosphere created. Teachers play a critical role in the design of a class and their responsiveness to their learners from a social constructivist perspective on learning. The school cannot serve as a preparation for social life unless it replicates the typical conditions of social life within itself (Dewey, 1909). This also means that if education’s goal is to prepare learners for life in the world, authentic learning is critical, ensuring that what we do in the classroom reflects real life.

Generally, a corpus-driven study employs empirical corpus data from which language features arise naturally through data analysis. The availability of prominent and representative corpora and the incorporation of computational software enables the investigation of linguistic variation from various angles. The prospective corpus is used to generate linguistic categories that have not been recognised by corpus linguistics or scholars, as the findings are intended to be exhaustive in terms of corpus evidence [21]. The linguistics categories are formed systematically from the recurrent patterns and frequency distributions when language is used in context [21]. According to Love [22], certain corpus-driven studies emphasised the importance of frequency evidence, particularly when
studying lexical bundles, while not prioritising frequency in analysing grammar patterns. Despite this contrast, it is assumed that the primary goal of a corpus-driven study is to discover novel language features inside a corpus inductively.

Before CL, linguists studied grammar based on the intuition of native speakers. The corpus has made a considerable contribution to linguists, researchers, teachers or students to use natural language data to show and understand how language works. They can analyse many aspects of language such as vocabulary, including frequency, collocation and grammar, grammatical rules in grammar books, describing language, and displaying qualitative and quantitative analysis. In addition, they can compile a corpus of a specific type of language and find commonly used words and expressions. They can even compare students’ written and spoken language to determine the types of mistakes they make [23].

Undeniably, the advantage of CL is providing the research to give quantitative manner to linguistic features through systematic statistical measures such as computer software, which is laborious if performed through manual steps. [24] elaborates on CL based on computer technology by stating that it is an advanced emerging methodology to analyse language and a new philosophical approach for language because it is a robust technological tool that has made CL possible. [25] added that to use appropriate software to measure language, tools for indirect observation such as collocates, query languages, aligners, concordances and parsers are required.

In recent years, CL has been employed in research on vocabulary. It includes using corpora in teaching and learning a language. For instance, [26] has shown corpora to teach a language. In 2020, the vocabulary programme he designed allowed language teachers to arrange their corpus by having 32 various texts with different lengths to identify words prevalent in them. [27] used learners corpora to investigate the use of boys’ and girls’ vocabulary in native and non-native contexts. Both of the researchers applied WordSmith Tools 5.0 to compare critical verbs that are key in domains related. In 2020, Love compared the vocabulary used by speakers of different genders, ages and social groups to study a spoken conventional component of BNC which consisted of 4.5 million words. From the gender aspect, 25 most essential words characteristic of both genders were found. In the findings, boys tend to use more cursing words while females prefer to use more feminine pronouns and first-person pronouns.

1.3 Corpus Linguistics (CL) in Malaysia

Regarding past studies conducted [28-32], Malaysians are becoming more interested in CL, especially in learner corpora. Other research areas of language are included, such as Malay linguistics by [33], and English for specific purposes by [34]. These studies employed a different set of methodological approaches to investigate various linguistic features, namely grammar and semantics. Interestingly, English was not the only language explored as the Malay Language was targeted for research. On the other hand, these studies have not examined vocabulary and phrases used by young learners in extended writing.

[35] conducted research similar to the present study, which investigated the types of discourse connectives used by 32 secondary students in narrative essays. In addition, interviews were carried out with ten selected students to explore how they perceived the use and meaning of some discourse connectives in their texts. The researcher examined 96 narrative essays and listed out the discourse connectives identified in them. Based on the preliminary findings, she found that the students tend to use different discourse connectives in their essays. However, they were alert that there were many other types of discourse connectives available in English. She added that students were innovative in applying discourse connectives in essays instead of strictly following the related frameworks.
Linguists were made aware that students were able to become accomplished English users despite imitating native speakers in using the discourse connectives for essay writing. Although the mentioned research examined student use of discourse connectives in narrative essays, it is meaningful to examine from the perspectives of advanced Malaysian upper primary school learners their use of vocabulary in extended writing. In this present study, learner corpora were the corpora chosen and collected since the data collected were authentic extended writing from advanced Malaysian upper primary school learners who learn ESL. It fulfils second language acquisition because they learn ESL in Malaysia [5]. The number of extended writings collected was around 160 essays. Hence, learner corpora were suitable as the extensive data were stored electronically and analysed through suitable software and applications.

This study was conducted using a corpus-driven analysis. The corpus obtained was learner corpora, meaning the results were based on actual data gathered during the investigation rather than on a reference corpus as in corpus-based analysis. The linguistic component examined was vocabulary used by 160 advanced Malaysian primary school learners in extended writing. Hence, the extended writing was used automatically in this study to gather learner corpora. To begin, LancsBox was utilised to extract salient vocabulary (functional & content words) from the extended writing. The research focused on functional and content words since functional words may not give meaningful results. Moreover, to illustrate, the log-likelihood calculator was utilised for the log-likelihood values while comparing the salient vocabulary (functional & content) to the L-O-B reference corpus.

![Learner Corpus -> Salient Vocabulary -> LancsBox](image)

**Fig. 1. Research conceptual framework**

2. Methodology

The study uses a corpus-driven research strategy and a quantitative methodology. The current study aimed at analysing, comparing and contrasting the top 20 salient vocabulary (functional & content) used among advanced Malaysian upper primary school learners of English. The study focused on vocabulary used by 160 advanced Malaysian primary school learners in their extended writings. To establish a comparable scenario for collecting written texts in the four national primary schools and contexts, it was determined that collection methods needed to be as similar as feasible.
The research site chosen was the northern states from Malaysia involving 4 national primary schools with high English performance located in Malaysia. The national primary schools chosen were Sekolah Kebangsaan (henceforth, SK) and Sekolah Kebangsaan Jenis Kebangsaan Cina (henceforth, SJKC) based on the data provided by officers from the District Education Office and State Education Department.

2.1 Participants of The Study

Based on Table 1, Peninsular Malaysia specifically the northern region was chosen as the research site. The study included only the capitals of each state because the target participants were advanced Malaysian upper primary school learners enrolled in national primary schools with high English proficiency, as determined by district education offices or state education departments. Only forty advanced primary school learners from each state were chosen to complete the extended writing assignment. As a result, the number of advanced primary school learners and extended writing were generated concurrently. Due to the nature of the extended writing, all advanced primary school learners were given the flexibility and opportunity to write on any subject they desired.

Table 1
Demographic data of AMUPSLC

<table>
<thead>
<tr>
<th>East/West Malaysia</th>
<th>Region</th>
<th>Capital</th>
<th>No. of Learners/Extended Writing</th>
<th>No. of Tokens</th>
<th>No of Tokens per Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peninsular Malaysia</td>
<td>Northern Region</td>
<td>Perlis</td>
<td>40</td>
<td>9,855</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alor Setar</td>
<td>40</td>
<td>8,260</td>
<td>46,185</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Georgetown</td>
<td>40</td>
<td>10,392</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ipoh</td>
<td>40</td>
<td>17,678</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>4</td>
<td>160</td>
<td>46,185</td>
<td>46,185</td>
</tr>
</tbody>
</table>

2.2 Data Collection

The following sections outline the main design concerns and criteria for the writing task:

i. each advanced upper primary school learner writes one extended writing of their chosen topic (without any prompts given by the teacher)

ii. face-to-face or online classroom environment

iii. no word limits

iv. 30 minutes length of time (flexible)

v. advanced upper primary school learners are not permitted to use dictionaries or other reference source.

When the writing task was assigned, the English teachers gave the following directions to the pupils: Today, you will write on whatever subject you want. Research participants were given a minimum of 30 minutes to complete a writing task in class or at home through online learning. It should be noted that time and word limitations were not strictly enforced since upper primary school learners were given flexibility if they needed more time or wrote over the word limit.

The handwritten writings of AMUPSLC were first typed and stored as doc/docx files. All files were separated by states into thirteen different folders. Subsequently, they were further separated by regions into five different folders. Spelling mistakes were rectified during transcription. Because the purpose of this current study was not to determine how much advanced Malaysian primary school...
learners’ extended writing varies from Standard English, the researcher did not analyse spelling, grammar, or punctuation mistakes. Rather, the researcher was interested in studying their choices of vocabulary. Therefore, some degree of standardisation needed to be included in the transcriptions to aid in identifying trends and patterns in phrase usage. For instance, the researcher wanted to be able to detect all instances of the term wake up, considering alternative spellings (wak up, wok up) as synonymous for the sake of this study. While converting the handwritten texts to computer format, the researcher encountered numerous typical difficulties when analysing the extended writing. Nonetheless, there is an inherent problem in transcribing their writing, based on unreadable texts and concerns about spelling and grammaticality mistakes. However, even if school learners sometimes made spelling or grammatical mistakes, it was usually simple to decipher what they meant to write.

To answer research question one which focused on emphasising the salient vocabulary (functional & content) the researcher chose to use a variety of automated techniques to establish significant variances and similarities. LancsBox is a suite of corpus software developed by Lancaster University. It includes tools for doing several types of linguistic analysis, including WordList, Keywords, concordances and log-likelihood calculator. The next step was to separate the frequency lists’ function and content terms. This was accomplished manually by scanning each list for the first twenty salient vocabulary (functional & content). The top twenty salient vocabulary (functional & content) were quantified in this study using frequency lists. What makes the usage of frequency lists so important is that they disclose information about the corpus. With reference to Morato et al. (2021), frequency lists constitute a fundamental kind of frequency information, but they also offer critical information to supplement the findings of thorough concordance studies. Frequency lists may be used to characterise the corpora from which they were derived. The researcher chose the top twenty salient vocabulary (functional & content). Typically, the top-ranking vocabulary (functional & content) on the list are functional words.

The keyword analysis employs two main corpuses namely AMUPSLC and L-O-B corpus, with the later serving as the reference corpus, which is compared using LancsBox. The frequency lists for each state were compared. The comparison findings include terms that are ‘key’ in one file but not in the other. After uploading the texts in the corpus to calculate the frequencies, these values were normalised to an incidence rate per 1,000 words of text. A review of concordance lines may aid in understanding the situations in which advanced Malaysian upper primary school learners employ this vocabulary (functional & content).

2.3 Data Analysis

To answer research question one which focused on emphasising the salient vocabulary (functional & content) the researcher chose to use a variety of automated techniques to establish significant variances and similarities. LancsBox is a suite of corpus software developed by Lancaster University. The process was outlined in the previous section. The top twenty salient vocabulary (functional & content) were quantified in this study using frequency lists. What makes the usage of frequency lists so important is that they disclose information about the corpus. With reference to [36], frequency lists constitute a fundamental kind of frequency information, but they also offer critical information to supplement the findings of thorough concordance studies. Frequency lists may be used to characterise the corpora from which they were derived. The researcher chose the top twenty salient vocabulary (functional & content). Typically, the top-ranking vocabulary (functional & content) on the list are functional words.
3. Results

3.1 Functional Words

The following part summarises the top twenty salient functional words in Peninsular Malaysia (Northern Region). Next, their log-likelihood values are analysed and the functional words are examined and analysed together with the concordances.

3.2 Salient Functional Words: Peninsular Malaysia

In the Peninsular Malaysia sub corpus, there are 480 extended writing with 132,614 tokens. First and foremost, a total of 160 extended writing was imported into LancsBox for data analysis from the sub corpus of the Northern Region. To determine the salient functional words, a total of 46,185 tokens were analysed. Through the acquisition of frequency lists, twenty salient functional words were found. Table 2 presents the twenty salient functional words in terms of raw frequency (henceforth, RF), normalised frequency (henceforth, NF) and percentage of occurrence (%) in each of the extended writing. The column N refers to the number of essays where these words appear, for example, the is found in 160 written texts in the frequency list.

Based on Table 2, 160 advanced Malaysian upper primary school learners in Peninsular Malaysia used the frequently and preferably in extended writing. The RF is 2417.00, which looks to be the highest of the remaining 19 functional words. Additionally, it is the sole vocabulary found in all 160 extended writing analysed. In contrast to the, be is the twentieth salient vocabulary, with a RF of 277.00. In extended writing, only 92 advanced Malaysian upper primary school learners utilised the vocabulary be. It might be deduced that be was not given priority over other vocabulary such as to, and, a, and I. Surprisingly, vocabulary I is the fifth salient functional word, appearing in just 95 of 160 extended writing. The percentage of N is 59.38. It reveals that 95 advanced Malaysian upper primary school learners frequently utilised the vocabulary I in their extended writing. Depending on their perspectives, one may argue that they described their experiences in the first person.

Table 2
Salient Functional Words of West Malaysia: Northern Region

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>RF</th>
<th>NF</th>
<th>%</th>
<th>N</th>
<th>N%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>2,417.00</td>
<td>523</td>
<td>5.23</td>
<td>160</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>to</td>
<td>1,658.00</td>
<td>359</td>
<td>3.59</td>
<td>159</td>
<td>99.38</td>
</tr>
<tr>
<td>3</td>
<td>and</td>
<td>1,400.00</td>
<td>303</td>
<td>3.03</td>
<td>157</td>
<td>98.13</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>1,050.00</td>
<td>227</td>
<td>2.27</td>
<td>155</td>
<td>96.88</td>
</tr>
<tr>
<td>5</td>
<td>I</td>
<td>888.00</td>
<td>192</td>
<td>1.92</td>
<td>95</td>
<td>59.38</td>
</tr>
<tr>
<td>6</td>
<td>of</td>
<td>752.00</td>
<td>163</td>
<td>1.63</td>
<td>138</td>
<td>86.25</td>
</tr>
<tr>
<td>7</td>
<td>we</td>
<td>699.00</td>
<td>151</td>
<td>1.51</td>
<td>90</td>
<td>56.25</td>
</tr>
<tr>
<td>8</td>
<td>in</td>
<td>681.00</td>
<td>147</td>
<td>1.47</td>
<td>142</td>
<td>88.75</td>
</tr>
<tr>
<td>9</td>
<td>is</td>
<td>607.00</td>
<td>131</td>
<td>1.31</td>
<td>116</td>
<td>72.50</td>
</tr>
<tr>
<td>10</td>
<td>was</td>
<td>603.00</td>
<td>131</td>
<td>1.31</td>
<td>93</td>
<td>58.13</td>
</tr>
<tr>
<td>11</td>
<td>my</td>
<td>590.00</td>
<td>128</td>
<td>1.28</td>
<td>93</td>
<td>58.13</td>
</tr>
<tr>
<td>12</td>
<td>it</td>
<td>541.00</td>
<td>117</td>
<td>1.17</td>
<td>124</td>
<td>77.50</td>
</tr>
<tr>
<td>13</td>
<td>that</td>
<td>476.00</td>
<td>103</td>
<td>1.03</td>
<td>123</td>
<td>76.88</td>
</tr>
<tr>
<td>14</td>
<td>for</td>
<td>367.00</td>
<td>79</td>
<td>0.79</td>
<td>123</td>
<td>76.88</td>
</tr>
<tr>
<td>15</td>
<td>as</td>
<td>296.00</td>
<td>64</td>
<td>0.64</td>
<td>103</td>
<td>64.38</td>
</tr>
<tr>
<td>16</td>
<td>you</td>
<td>290.00</td>
<td>63</td>
<td>0.63</td>
<td>46</td>
<td>28.75</td>
</tr>
<tr>
<td>17</td>
<td>are</td>
<td>290.00</td>
<td>63</td>
<td>0.63</td>
<td>98</td>
<td>61.25</td>
</tr>
<tr>
<td>18</td>
<td>they</td>
<td>286.00</td>
<td>62</td>
<td>0.62</td>
<td>82</td>
<td>51.25</td>
</tr>
<tr>
<td>19</td>
<td>with</td>
<td>278.00</td>
<td>60</td>
<td>0.60</td>
<td>102</td>
<td>63.75</td>
</tr>
<tr>
<td>20</td>
<td>be</td>
<td>277.00</td>
<td>60</td>
<td>0.60</td>
<td>92</td>
<td>57.50</td>
</tr>
</tbody>
</table>
Note: The raw frequencies have been normalized to as per ten thousand words for uniform representation. The formula applied is \( \frac{N_{RF} \times 10^4}{T} \) where \( N_{RF} \) = value of the raw frequency of the salient vocabulary, \( T \) = total tokens of the sub corpus (46,185) and \( N_{NF} \) = value of the normalised frequency. The figures for percentage distribution are rounded off to the nearest two decimal digits. The formula applied is \( \frac{N_{NF}}{10^4} \times 100\% = N_{PD}\% \) whereby, \( N_{NF} \) = value of the normalised frequency, \( N_{PD} \) = value of the percentage of occurrence (%).

Similarly, the vocabulary *my* and *was* were used in conjunction with *I*. *My* is the eleventh salient functional word, while *was* is the tenth salient vocabulary. Their RFs are 590.00 (*my*) and 603.00 (*was*) respectively. However, *my* and *was* are discovered in 93 extended writing. It is estimated that 93 advanced Malaysian upper primary school learners regularly used *my* and *was* in their essays. Furthermore, it is suggested that 93 extended writing were produced in the past tense in order to describe prior experiences. Logically, it is reasonable to suppose that advanced Malaysian upper primary school learners in the Northern Region prefer to write extended writing about previous experiences in the first person perspective using *I*, *my* and *was*.

Other vocabulary is discovered in numerous extended writings (≥ 100) due to the findings. They are the vocabulary from the second, third, fourth, sixth, eighth, ninth, twelve, thirteenth, fourteenth, fifteenth, and nineteenth ranks. *In* has a RF of 681.00, but it is found in 142 extended writing, which is more than the sixth functional word *of*, with just 138 extended writing examples. Following that, *is* has a RF of 607.00; advanced Malaysian upper primary school learners used it in total of 116 extended writing. Other functional words such as *it* (541.00), *that* (476.00), *for* (367.00) and *as* (296.00) with lesser RF as compared to *is* (607.00) seem to appear more in extended writing by advanced Malaysian upper primary school learners which are 124, 123, 123 and 103 extended writing respectively. These findings imply that advanced Malaysian upper primary school learners use *to*, *and*, *a*, *of*, *in*, *is*, *it*, *that*, *for*, *as* and *within* their extended writing to improve the organisation and explanation of essays through the use of pronouns, prepositions, linkers, and conjunctions.

On the other hand, the functional words which are not identified in most of the extended writing (≤ 100) are *I*, *we*, *was*, *my*, *you*, *are*, *they* and *be*. Their RFs are 888.00, 699.00, 603.00, 590.00, 290.00, 290.00, 286.00 and 277.00 respectively. Although they might have a relatively high frequency compared to other functional words, the number of extended writings using them is less. Only 46 to 98 extended writing have used these functional words. Hence, it is derived that most of the advanced Malaysian upper primary school learners in the Northern Region did not prefer to use these functional words and preferred to describe their essays in the third person instead of the first person.

With reference to Table 3, 14 salient functional words demonstrate significance due to their high G² value of ≥15.13 at \( p < 0.0001 \). The salient functional words are *the* (233.00), *to* (47.91), *a* (25.59), *I* (681.32), *of* (782.36), *we* (373.32), *in* (139.31), *was* (198.52), *my* (813.85), *it* (99.70), *for* (67.91), *you* (196.44), *they* (41.52), and *be* (42.50). The NF has little effect on the G², as certain functional words, such as *to*, have a G² of only 47.91. Simultaneously, *and* has an NF of 303 but only a G² of 2.41. Each of these 14 salient functional words is important compared to the L-O-B reference corpus.
Table 3
Log-Likelihood Values of Northern Region vs. L-O-B

<table>
<thead>
<tr>
<th>Subcorpus (Northern Region) vs. L-O-B</th>
<th>Rank</th>
<th>Word</th>
<th>NF</th>
<th>G²</th>
<th>Sig. Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>the</td>
<td>523</td>
<td>233.00</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>to</td>
<td>359</td>
<td>47.91</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>and</td>
<td>303</td>
<td>2.41</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>a</td>
<td>227</td>
<td>25.59</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>l</td>
<td>192</td>
<td>681.32</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>of</td>
<td>163</td>
<td>782.36</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>we</td>
<td>151</td>
<td>373.32</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>in</td>
<td>147</td>
<td>139.21</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>is</td>
<td>131</td>
<td>11.35</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>was</td>
<td>131</td>
<td>198.52</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>my</td>
<td>128</td>
<td>813.85</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>it</td>
<td>117</td>
<td>99.70</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>that</td>
<td>103</td>
<td>0.36</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>for</td>
<td>79</td>
<td>67.91</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>as</td>
<td>64</td>
<td>2.97</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>you</td>
<td>63</td>
<td>196.44</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>are</td>
<td>63</td>
<td>0.14</td>
<td>&lt; 0.1</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>they</td>
<td>62</td>
<td>41.52</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>with</td>
<td>60</td>
<td>9.26</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>be</td>
<td>60</td>
<td>42.50</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Note: $G^2 > 6.63$ at $p < 0.01$ or 1% level. Values of significance as given by McEnery et al. (2006, p. 55): Null hypothesis, $H_0 =$ There exists no significant association between salient vocabulary in the sub corpus of AMUPSLC with those in the reference corpora. Alternative hypothesis, $H_a =$ There exists a significant association between the occurrences of salient vocabulary in the sub corpus of AMUPSLC with those in the reference corpora ($G^2 \geq 10.83$ at $p < 0.001$, $G^2 \geq 15.13$ at $p < 0.0001$). Similar values of significance are applied for $G^2$ test in AMUPSLC.

Following that, just one salient functional word bears the $G^2 \geq 10.83$ at $p < 0.001$. *Is* has a $G^2$ value of 11.35. Compared to the reference corpus, however, L-O-B, it is still regarded as a relatively crucial functional word. The remaining five functional words are less or non-significant compared to the reference corpus, L-O-B, because their $G^2$ is less than 10.83, which complies with the $G^2 > 6.63$ at $p < 0.01$ or 1% level requirements. *With* (9.26) is the functional word with $p < 0.01$; the other four functional words with $p < 0.1$ are *and* (2.41), *that* (0.36), as (2.97), and *are* (0.14).

Overall, 15 salient functional words from the Northern Region's sub corpus are significant compared to the reference corpus, whereas the remaining five salient functional words are either less significant or insignificant. Additionally, one may argue that *my* is the most important functional word, with a $G^2$ of 813.85, while *are* is the least significant, with a $G^2$ of 0.14. Nonetheless, the five functional terms that are insignificant because their $G^2$ is less than 10.83 are moderately significant in the AMUPSLC learner corpus.

3.3 Concordances of Similar Functional Words: Peninsular Malaysia

The concordances of each similar salient functional word analysed above are presented below. As illustrated in Figure 2, the *is* employed as a determiner. It is used at the start of noun groups to refer to someone or something that has been stated or identified previously [37]. The use of the *identifies* a total of 6,768 lines of concordances (Northern: 2,417). It might be deduced that in all 2,417 lines of concordances the is used in the manner specified. Likewise, it is considered that
advanced Malaysian upper primary school learners in the Northern Region prefer to use the in their extended writing the most.

With regard to Figure 3, the word to can be employed in two ways. According to [37], the preposition to is used and it is also noticed before the basic form of a verb. Analysis showed that to is commonly employed before the base form of a verb despite of in place of a preposition. There are 1,658 concordance lines to count (Northern: 1,658). It demonstrates that advanced upper primary school learners in Peninsular Malaysia regularly used to preceding the basic form of a verb in their extended writing, instead of using it as a preposition.

---

**Fig. 2.** Concordances ‘the’: Northern Region (34 of 2,417 lines)

**Fig. 3.** Concordances ‘to’: Northern Region (34 of 1,658 lines)
The concordances of *and* in four sub corpora are depicted in Figure 4. It is obvious that *and* is utilised as a conjunction by every advanced Malaysian upper primary school learner in Peninsular Malaysia. It is used to establish a connection between two or more words, groupings, or phrases [37]. A total of 1,400 concordance lines are available (Northern: 1,400). The findings are comparable to those of *the* and *to*, as advanced Malaysian upper primary school learners in the Northern Region emphasised *and* in their extended writing. One could argue that they are more adept at structuring complicated sentences.

![Fig. 4. Concordances ‘and’: Northern Region (34 of 1,400 lines)](image)

[37] describes *a* as a determiner used when referring to someone or something for the first time or when people are unsure of who or what they are referring to. There are 1,050 concordance lines counted (Northern: 1,050). Through study, it was discovered that every advanced Malaysian upper primary school learner in Peninsular Malaysia mastered the correct and acceptable use of *a* as a determiner in their extended writing.
4. Discussions

The research question is: What are the differences and/or similarities in the use of English vocabulary in the extended writing produced by advanced learners in upper primary schools in the northern region of Malaysia?

Each difference and similarity of vocabulary (function & content) from the findings will now be discussed. As evidenced by [38] the significance of the, to, is that they play an essential function as articles, simple conjunctions, prepositions, and infinitives. Without the conjunctions the, to, and, a statement sounds awkward and may be grammatically wrong. Upper primary school learners prefer first-person pronouns rather than third-person ones in extended writing. It is in contrast to [39] because in his study, the majority of learners in upper primary school employed third-person pronouns for more extended writing. Despite this, I is frequently employed by secondary school students when they are required to provide their opinion and express their thoughts [40]. Possibly, learners in upper primary school are exposed to various types of essay writing and learning tools, resulting in the effective use of I in extended writing [41].

In reference to research question 1, it was reported that the learners in this study most recurrently used the infinitive ‘to’, conjunction ‘and’, and articles ‘a’ and ‘the’ that are fundamental for simple sentence formation in primary school. This aligns with findings obtained by [42] but on tertiary students who relied on using vocabulary including nouns and pronouns. Despite students producing a number of good vocabulary, students still tend to make mistakes in their writing because of carelessness. This study’s conclusion is consistent with those of [43] who discovered that Malaysian upper primary school students did not employ English language to its full potential. As a result of their poor vocabulary skills, they were unable to compose a strong essay. The writing skills of the students were low intermediate.

Malaysian English primary school teachers struggled to come up with language and phrases to teach their students for guided and extended writing [44]. English lessons are taught in ELT classrooms by Malaysian English primary school teachers utilising terminology from the Get Smart Plus 4 (Year 4) and English Plus 1 textbook (Year 5). There was no additional vocabulary or teaching
of phrases in the English lessons. Students studying ESL in upper primary who are not exposed to sufficient phrases in the teaching and learning of English struggle to develop phraseology competence [45]. In the past 10 years, several researchers have started to investigate the vocabulary used by primary school students who are learning a language other than their own. An example of this is the vocabulary of Spanish that students who have learned English as a foreign language. For instance, the Spanish vocabulary used by primary school students was the subject of research by [46]. Even while this kind of research is still in its infancy, it could undoubtedly be expanded to include additional students that study ESL or something similar. According to [47], more research should be done on the vocabulary use of primary school students so that language educators can select and create the best pedagogical procedures, including materials for teaching and learning, for second language acquisition.

Numerous past studies on language learning, specifically those looking at EFL learners, have made known that learners at practically all skill levels struggle to use common verbs such as “make.” Additionally, they suggest that activities using concordances can help learners become more conscious of the complexity of high-frequency verbs [48]. This study suggests that teachers must embed a detailed process of using corpora in language teaching merely for pedagogical goals as recommended by [49] in defining the “structure or language elements” for instruction. Teachers’ role is to provide students with corpus-based activities for practice. Teachers will be able to identify how proficient a learner is through the corpus-based activities. [50] agrees that learner corpora should be applied in facilitating second language learning and error analysis to promote data-driven learning that will assist learners in becoming more aware of their native language and accelerate their language acquisition.

5. Conclusions

Based on the findings, it is clear that functional words use in vocabulary including functional and content words are identified. The advanced upper primary school learners in different areas in the northern regions are exposed to varied learning resources and materials, causing them to use different vocabulary and phrases for extended writing. The theme of the world of self, family, and friends is fundamental and essential, as most of the identified vocabulary and phrases revolve around it. Hence, it could be said that the use of salient vocabulary is highly influenced by the theme introduced in the school. Furthermore, it could be assumed that this theme is authentic to school learners as they are encouraged to write based on authentic experiences which makes the theme of the world of self, family and friends become the first and significant theme introduced by the English teachers since Year 1. Finally, the AMUPSLC could serve as a writing reference for Malaysian English language teachers and upper primary school pupils. The salient vocabulary and phrases from the findings can be used for extended writing activities.

References


