Incorporating Corpus Linguistics in the Development of Vocabulary and Phraseological Index

Wong Wei Lun[1*], Warid Mihat[2], Mairas Abdul Rahman[3]

[1]Department of Educational Development and Innovation, School of Education, Sunway University,
47500 Petaling Jaya, Selangor, Malaysia
colinw@sunway.edu.my
[2] Academy of Language Studies, University Teknologi MARA Kelantan Branch,
18500 Machang, Malaysia
waridmihat@uitm.edu.my
[3]Faculty of Languages and Communication, Universiti Sultan Zainal Abidin,
Kuala Nerus 21300, Malaysia
mairas@unisza.edu.my
[*]Corresponding Author

https://doi.org/ *to be updated*

Received: 27 May 2024
Accepted: 30 October 2024
Date Published Online: 17 November 2024
Published: 17 November 2024

Abstract: This study highlighted the need for developing teaching and learning resources in this 21st century, especially in language learning. Hence, this study presented a contemporary method that combined corpus linguistics with developing a vocabulary and phraseological index. The aim was to improve the learning of English at all levels of education. Selecting a doctoral study as a reference, this study demonstrated the practical implementation of corpus linguistics integrating SkecthEngine. A substantial dataset involving 560 extended writing, amounting to 152,187 words, formed the basis for computational analysis. This analysis presented a list of twenty salient vocabulary and phrases, categorised by their literary genre and corresponding to the levels of the CEFR. Concordances were employed to display authentic example sentences. Finally, the vocabulary and phraseological index provided a convenient and valuable resource for language teachers, tailored to different educational settings.

Keywords: CEFR, Corpus linguistics, Extended writing, Vocabulary and phraseological index.

1. Introduction

With reference to Wong et al. (2023a), vocabulary and phraseological index was considered as a fundamental tool in language learning. It encompasses various vocabulary (functional and content) and phrases (bi-, tri- and qualgram). In addition, this vocabulary and phraseological index highlighted vocabulary that learners should learn since they were salient based on the corpus analysis (Wong et al., 2023a). The salient vocabulary and phrases included in the index were often identified in the English textbooks provided by the government and some learning resources selected by the English teachers. Eventually, the index could serve as an indicator for English teachers on what vocabulary and phrases learners should master for extended writing based on the salient words presented in the index. Usually, the vocabulary and phraseological index could be organised based on themes or learning units. However, in this study, the integration of CEFR in the policy introduced salient vocabulary and phrases to learners based on learning proficiency set by CEFR, from A1 to C2. Hence, this vocabulary and phraseological index could be used for English learning to provide

learners with authentic examples to understand and apply the salient vocabulary and phrases presented into their extended writing.

To elaborate, the importance of the vocabulary and phraseological index in learning English was obvious. It provided a reference to language learning, allowing learners to understand and use salient vocabulary and phrases authentically for extended writing. For instance, it helped them to comprehend the use and authentic content of various vocabulary and phrases and expanded their vocabulary bank (Wong et al., 2023b). On the other hand, it was crucial to mention that there must be a need to develop this vocabulary and phraseological index. It would be cliché if no learning needs were found for this development. Subsequently, it highlighted the need for English teachers to identify the learning needs among learners, perhaps those with weak and intermediate English proficiency prior to the development of vocabulary and phraseology index. By doing so, they could ensure the index was truly relevant and suitable for their learners, thereby improving the quality of teaching and learning of English for extended writing (Wong et al., 2023b).

Many past studies were conducted on the innovative teaching and learning resources for ESL including different models and strategies. To commence, ADDIE model (Hu, 2023; Jing & Eng, 2023; Khazali et al., 2023), Rasch model (Chen et al., 2024; Effatpanah et al., 2024; Iwamoto, 2023; Yan, 2022) and design and development research (Kudryavtseva et al., 2023; Wan Mahzan et al., 2020). They were used to develop some teaching and learning resources such as language modules or textbooks (Hašková et al., 2023), mobile augmented reality vocabulary reference (Jalaluddin et al., 2020), handbooks (Santosa et al., 2022), coursebooks, mobile-assisted language learning lesson (Adijaya et al., 2023), online course (Rajprasit, 2022) and evaluation checklist (Roberts et al., 2020). They aimed to improve the reading, writing, listening, speaking or grammar skills among learners.

Therefore, in this study, the introduction of corpus linguistics incorporation to develop vocabulary and phraseological index provided an option for researchers and educators to innovate their teaching and learning resources in ESL, aiming to improve learners' extended writing. Concurrently, it answered one research question of "How was the incorporation of corpus linguistics to develop a vocabulary and phraseological index?"

2. Literature Review

This section started by explaining corpus linguistics. Following that, corpus-based and corpus-driven approaches were introduced and discussed. Finally, a conceptual framework was provided for visualisation of this entire study.

2.1 Definitions of Corpus Linguistic

In this era of modern technology, corpus linguistics has emerged as a method for computational quantitative analysis, capable of processing a large size of text, commonly referred to as corpora (Katz, 2021). It could be understood as digital compilation of text analysis. It is different from a normal electronic archive because it allows the language patterns to emerge rather than preconceptions. Peñarroja (2020) highlighted the use of corpus linguistics software in facilitating non-linear examination of texts for both quantitative and qualitative analysis.

Following that, corpus linguistics offers a dual toolkit of software and theoretical frameworks, aiding researchers in detecting language patterns across a large volume of texts. The results often become focal points for in-depth qualitative analysis such as concordances and interviews. To illustrate, Camiciottoli (2020) conducted a study on the language used in investor relations communications of Japanese and Italian companies. By integrating mixed-method research approaches and employing corpus linguistics software, she extracted quantitative findings from a large textual dataset. Subsequent phases involved qualitative analysis of the results obtained.

However, Lu (2019) noted the different definitions attributed to corpus linguistics. Notably, there was a conceptual divergence. First of all, one viewed it as a linguistic subfield focused on evolving theoretical language models based on corpus data. Nevertheless, others regarded it as a domain aimed at forging new language theories from corpus-driven data. These viewpoints also informed corpus analysis approaches by addressing different linguistic challenges. Academic

definition occasionally distinguished between 'corpus-driven' and 'corpus-based' approaches (Tognini-Bonelli, 2002). Yet, as Hardie and Dorst (2020), and Collins et al. (2020) observed, these distinctions were often overstated, with people from both schools frequently collaborating across diverse academic disciplines.

On the other hand, corpus linguistics as a less discussed topic in the history contrasted with the frequently used text mining among researchers. While both using computational tools for text analysis, text mining was more user-friendly, due to the complex nature of data interpretation in corpus linguistics. This differences generated from their academic origins. To continue, corpus linguistics was a subset of linguistics, whereas text mining was rooted in natural language processing which was a branch of computer science. Although they have distinctions, they had developed alongside each other with intersections and overlaps (Nguyen, 2020).

Moreover, Chapelle (2019) explored the history of corpus linguistics claiming its nonexistence until the late 1950s. The focus then was on analysing authentic language, with the evolution of computer technology later enabling the improvement of corpus linguistics. The roots of corpus linguistics traced back to pre-computer times, with the design and development of manual concordances like Christian Bible (Brend, 2019). These efforts were seen as forerunners to the current corpus linguistic tools introduced in 2024 such as LancsBox and SketchEngine. Additionally, the use of frequency lists in the nineteenth century, as studied by Portelli (2021), played a significant role in corpus analysis

To summarise, corpus linguistics has computerised software for both quantitative and qualitative analysis of large text data. Its definition varied across different schools of thought such as the Neo-Firthian school focused on developing new language theories, while the methodologist focussed on its role in natural language processing. Despite these differences, there was a synergy between corpus linguistics, text mining, and computational linguistics. Yet, text mining was particularly notable for its qualitative focus in specific contexts.

2.2 Corpus-Based Approach

Undeniably, corpus-driven and corpus-based approaches are linked with their respective advocates, stemming from deep-seated convictions about the use of corpora in linguistic analysis. For corpus-based approach, the use of corpus as the main data for shaping preconceptual language descriptions is emphasised. Islentyeva (2020) stated that corpus-based approach was defined as an approach for illustrating current theories, which were not originally conceived with direct reference to a corpus. Hence, it involved using tagged corpora for speech analysis, examining the use of vocabulary and grammar frequencies, and conducting qualitative analyses in concordances. It demonstrated greater adaptability for researchers in various domains.

To illustrate, corpus-based approach was employed by many past studies. For instance, systemic functionalist, De Fina (2020) has utilized this approach. Research of diachronic language changes has been studied by researchers such as Bouzouita (2019) and Laske (2020). On the other hand, language variation studies have been conducted by Devereaux (2019) and Villena-Ponsoda (2019). All of them utilised corpus-based approach.

These past studies demonstrated corpus-based approach's efficacy as a methodology that helped researchers and linguists in substantiating and enhancing theoretical claims. However, it was worth noting that its reliance on predetermined methods for describing preconceptual language might constrain the discovery of novel emerge of language patterns. The nature of the findings could potentially be influenced by the researchers' preconceptions, as there was s a propensity to validate existing theories.

2.3 Corpus-Driven Approach

To initiate, Ramos and Guzmán (2021) offered a definition on corpus-driven approach, characterising it as an approach where findings were not forcibly aligned with pre-defined results established by researchers. This stance was evident in conventional grammar like clause components and parts of speech, which may not consistently align with authentic language dana used by learners

in this modern era. Instead of manipulating the results to conform to existing theoretical concepts, this approach allowed for the introduction of new language trends.

Next,researchers using corpus-driven approach gathered massive data devoid of overt preconceptions. They aimed to discover real, authentic, genuine and original language patterns emerged from the data. It ensured an untainted language description that parallelled corpus data (Keselman & Yakovleva, 2021). Practically, as proposed by Liu and Afzaal (2021), the results and language patterns derived from word-frequency lists, keywords and concordances in corpus-driven approach emphasised a reliance on frequency, a focus on vocabulary, the recognition of phrases as the primary unit of language analysis. In addition, it was a commitment to the descriptive potential of lexicogrammar in forming distinct meanings.

Therefore, corpus-driven approach is linked with methods of data analysis and extraction, linguistics, and a series of past studies utilising this approach. The past studies associated with corpus-driven approach were essential in shaping grammar patterns (Tichý, 2021), informing syllabi in language teaching and learning with a focus on vocabulary and phrases (Keselman & Yakovleva, 2021; He et al., 2021), and enhancing dictionaries (Dobrovoljc, 2020). Additionally, corpus-driven approach contributed significantly to the evolution of various domains in scientific and academic (Hussein et al., 2021; Cordeiro, 2019).

Finally, the corpus-driven approach is the chosen approach for this study, especifically for the incorporation of corpus linguistics in developing vocabulary and phraseological index. Its preference over the corpus-based approach was due to its concept-oriented nature, in contrast to the theoretical underpinnings of the latter. A conceptual framework, aimed at enhancing comprehension, was outlined in the subsequent part.

2.4 Conceptual Framework

In this part, the researchers described how one can incorporate corpus-driven approach for developing vocabulary and phraseological index, pivoting from theoretical to concept-oriented approaches. Consequently, a conceptual framework was necessary for a thorough understanding.

Based on Figure 1, at the onset of any corpus linguistics studies, essay collection was essential for developing vocabulary and phraseological index. This process required attention to several factors. The types of essays to be collected must be identified, varying significantly across educational levels. Primary schools may focus on guided or extended writing, whereas secondary schools often lean towards narrative or argumentative essays. At the tertiary level, the emphasis shifted to more complex forms of academic writing.

Next, the proficiency of learners was another vital aspect; selecting those with advanced English writing skills was preferable, as the goal is to establish a reference index for English teachers and learners. The volume of the learner corpus was also a key consideration. Setting a minimum word count for each educational level ensures a substantial corpus, which should aim for at least 100,000 words in total. This calculation was based on the number of learners and the minimum word count per essay.

Furthermore, thematic focus was crucial, with themes tailored to each educational stage. For instance, primary school essays might concentrate on the 'World of Knowledge' theme, while secondary schools could explore themes like 'Pollution' and 'Health'. In tertiary education, the themes may align with specific research fields, such as 'Education', 'Technology', or 'Linguistics'.

Following the successful collection of essays, the next step involved digitising these texts into DOCX format. This digital transformation was a prerequisite for their subsequent analysis using computational corpus linguistics software. Among the various tools available, such as LancsBox, SketchEngine, and WordSmith, SketchEngine has been selected for this project due to its exceptional suitability and capabilities in developing a vocabulary and phraseological index.

Once activated, SketchEngine facilitated the compilation of these essays into a learner corpus. The primary focus of this software was on functional and content words and phrases. This process began with a quantitative analysis, utilising frequency lists to pinpoint salient words and phrases that were crucial for the index. Subsequently, a qualitative analysis was conducted, involving

concordances to furnish real and authentic sample sentences. These sentences served as valuable references.

Upon completion of both analyses, the development of the vocabulary and phraseological index commenced. This index included several components: a cover page, a foreword, a user manual, and a comprehensive list of salient functional and content vocabulary and phrases. These were accompanied by concordances, providing sample sentences. This index was designed to aid English teachers and learners at primary, secondary, and tertiary levels in the teaching and learning of writing.

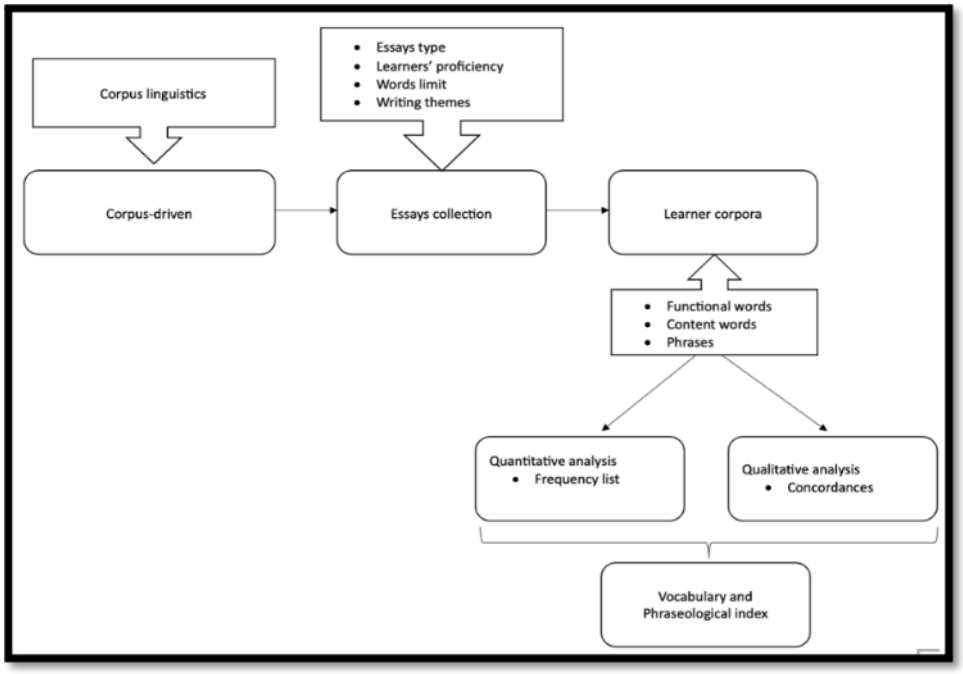


Fig. 1

2.5 Past Empirical Studies

While Google Scholar offers an extensive array of literature, including 7,480 articles related to 'corpus-driven' and 4,870 intersecting both 'corpus-driven' and 'teaching' since 2020, it was not selected as the primary research database. The focus on high-indexed journals to ensure academic rigour and quality led to the adoption of the SCOPUS database, known for its reputable collection of high-indexed journals. A targeted search in SCOPUS using 'corpus-driven' as a keyword yielded 284 articles from the period 2020 to 2024. Further refinement of the search terms to include 'corpus-driven and teaching' narrowed the results to 31 relevant articles. Out of these, 19 open-access articles were downloaded for analysis. A subsequent qualitative appraisal resulted in the exclusion of one review article, culminating in a curated list of 18 articles. These articles, which have been meticulously selected, are detailed in the table that follows, providing a comprehensive overview of the current research landscape in this field.

Table 1. Past studies reviewed

Author(s)	Title
Bocorny et al. (2021)	Writing scientific articles with the support of key lexical bundles: A corpus-driven study in the area of health sciences
Whitty et al. (2022)	Can and Could in academic writing: A corpus-driven comparison of English L1 and Vietnamese EFL students
Tian (2022)	Construction of a multimodal corpus of college students' spoken English based on semantic concepts
Rudneva (2020)	Corpus-driven ESP pedagogy: A preliminary case study
Gu (2022)	Corpus-driven resource recommendation algorithm for English online autonomous learning
Auziņa et al. (2020)	Creating target hypotheses in a learner corpus of Latvian
Salama & Altohami (2022)	Enhancing EFL students' COCA-Induced collocational usage of Coronavirus: A corpus-driven approach
Du et al. (2022)	Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study
Khedri & Basirat (2022)	Interactive metadiscourse in dentistry research articles: Iranian Vs Non-Iranian academic writers
Narkprom & Phoocharoensil (2022)	Lexical bundles in native English speakers' and Thai writers' dissertations
Fajri et al. (2020)	Lexical bundles of L1 and L2 English professional scholars: A contrastive corpus-driven study on applied linguistics research articles
Li (2022)	Meta-analytical approach to the impact of corpus-driven teaching on foreign language acquisition
Munalim et al. (2022)	Question-declaration coupling in a university meeting talk: Discourse of social inequality and collegiality
Keselman & Yakovleva (2021)	Short teacher responses in the EFL classroom: A corpus-approach assessment
Magalhães & Matos Rocha (2022)	The emergence of Fictive Interaction in the classroom as a teaching and learning strategy
Sanosi (2022)	The use and development of lexical bundles in Arab EFL writing: A corpus-driven study
Wu & Yang (2022)	Unpacking the functions of personal metadiscourse in teachers' classroom discourse
Wong et al. (2023)	Vocabulary index as a sustainable resource for teaching extended writing in the post-pandemic era

This review of corpus-driven research delineates the intersection between linguistic analytical methods and pedagogical approaches in Teaching English as a Second/Foreign Language (TESL/TOEFL). First and foremost, Bocorny et al. (2021) lay essential groundwork in exploring lexical bundles in health sciences, suggesting a pedagogical framework that shows promise but needs empirical substantiation across various fields for broader application. Concurrently, Whitty et al. (2022) contribute to our insight into the use of modal verbs in academic texts, yet their research could gain from a comparison with the language used in everyday contexts.

Next, Tian (2022) aims to enhance grammar teaching materials with semantic richness that mirrors authentic language usage, while also considering the necessity of integrating pragmatic language elements to avoid an overly theoretical concept of grammar. Following that, Rudneva (2020) progresses in teaching non-linguists, though the practicality of her detailed approach to semantics may be limited due to the varied backgrounds of learners. The algorithm introduced by Gu (2022) for

improving English education through technology shows potential for customized learning experiences, but its sustained effectiveness over time is yet to be fully determined.

Moreover, the studies by Auziņa et al. (2020) and Salama & Altohami (2022) focus on language learner output and collocational usage, respectively, with the latter centering on the timely term 'coronavirus'. While both studies are contemporary, the durability of their findings in the face of evolving language and collocational patterns requires ongoing verification. Du et al. (2022) present a thorough examination of collocational profiles varying by learner background, indicating a need for further investigation into the interplay between first language influence and proficiency levels.

On the other hand, the research by Khedri & Basirat (2022) on interactive metadiscourse markers in dentistry sets the stage for comparative studies across disciplines to better grasp academic writing standards. Lexical bundle analysis in dissertations and academic articles by Narkprom & Phoocharoensil (2022) and Fajri et al. (2020) respectively, reveals key structural and functional insights, underscoring the necessity for further dialogue about their role in academic communication effectiveness.

In addition, Li (2022) explores the impact of corpus-driven instruction on foreign language learning, a methodology whose influence on communicative competence needs critical assessment. Munalim et al. (2022) offer novel perspectives on discourse features in academic meetings, though their wider applicability requires validation. Studies by Keselman & Yakovleva (2021) and Wu & Yang (2022) enrich our understanding of teacher-student interactions through teacher response analysis and personal metadiscourse markers, yet the connection to student learning outcomes remains an area for further exploration.

Subsequently. Cruz Magalhães & Matos Rocha (2022) add a cognitive dimension to language analysis in medical education with their study on fictivity, presenting an intriguing prospect for educational integration. Sanosi (2022) examines lexical bundle usage among Arab EFL learners, prompting a call for a more thorough examination of its pedagogical impact. Wong et al. (2023) introduce a comprehensive vocabulary index for Malaysian English teachers, but its practical effectiveness in various educational settings is yet to be determined. Overall, these studies collectively enhance the body of knowledge in TESL and corpus linguistics, each offering unique insights into linguistic patterns and educational approaches.

Examining these past studies is crucial for their collective contributions to the expanding field of corpus linguistics and its application in language teaching. In the effort to incorporate corpus linguistic methods into developing a vocabulary and phraseology index in this article, these past studies provide an in-depth perspective on language learning and instructional methodologies. To recall, corpus linguistics lays the empirical groundwork for comprehending actual language usage, pivotal for developing pedagogical tools that accurately reflect real-life language patterns. The studies by Bocorny et al. (2021) and Whitty et al. (2022) underscore the relevance of real-world language usage in developing lexical bundles and modal verb instructions, mirroring our objective for the index to be grounded in genuine language use, as opposed to prescriptive norms.

To continue, Tian (2022) and Rudneva (2020) further reinforce the necessity for authenticity and comprehensive understanding in language education, advocating for materials that extend beyond superficial comprehension to embrace the full semantics of language. This aligns with the goal of developing an index that not only enumerates vocabulary and phrases but also illuminates their deeper semantic and pragmatic roles. On the other hand, the algorithmic method for self-directed learning proposed by Gu (2022) and the exploration of learner output deviations by Auziņa et al. (2020) are particularly pertinent to this article. These studies demonstrate how corpus-driven tools can facilitate personalized educational pathways, a principle guiding our index's development—to provide a resource adaptable to the unique needs of individual learners.

The pedagogical approach by Salama & Altohami (2022) for handling the collocational use of current terms such as 'coronavirus' underlines the necessity for the index to be dynamic and responsive to changing language trends. As language is inherently fluid, the index must be equipped to incorporate emerging phraseologies from ongoing linguistic discourses. Studies into metadiscourse and lexical bundles (Khedri & Basirat, 2022; Narkprom & Phoocharoensil, 2022; Fajri et al., 2020) enrich the understanding of the functioning and evolution of academic language. This insight is

crucial for the index to serve not just as a vocabulary source but also as a tool for constructing academically robust texts.

Additionally, the corpus-driven instructional approach investigated by Li (2022) is in harmony with the vision of the index as a facilitative tool for foreign language acquisition. The observations from Munalim et al. (2022) on discourse elements in academic meetings, and the communicative strategies outlined by Keselman & Yakovleva (2021) and Wu & Yang (2022), are key to our methodology in capturing the interactive dimensions of language. The study of language's cognitive aspects by Magalhães & Matos Rocha (2022), together with Sanosi's (2022) thorough examination of lexical bundle usage, further emphasize the complexity of language learning and the essential role of pedagogical tools in addressing these cognitive factors.

To conclude, the integration of these research findings into the development of a vocabulary and phraseology index for this article goes beyond mere academic pursuit. It is an essential step in forging a tool that truly represents the essence of corpus linguistics—showcasing language that is authentic, dynamic, and rich in functionality. The index is designed to be more than just a list of terms; it is intended to be a catalyst for effective language learning, bridging the gap between empirical linguistic studies and practical language instruction. This alignment with contemporary research ensures the index will be a vital asset in TESL, equipping both learners and educators with a resource that is not only authoritative but also pedagogically effective.

3. Proposed Incorporation

This article proposes the integration of corpus linguistics within the development of vocabulary and phraseology, particularly in an ESL/EFL context. The methodology is elucidated through a detailed flowchart, aligning with the conceptual framework outlined in the literature review. The authors illustrate this session using a doctoral study as a case example.

First of all, in this article, the focus is on upper primary schools in Malaysia, with the study framed within the realm of ESL education. The emphasis is placed on 'extended writing', a form of essay writing by ESL learners, typically exceeding 80 words. This study specifically targets advanced upper primary learners, characterized by their high proficiency in English. This proficiency is evidenced through formative assessments, with learners achieving either Band 5 or 6 in the CEFR-M (Malaysia’s version of CEFR) or scoring between 80 to 100 marks in summative assessments.

For the purpose of this article, 'advanced learners' were purposively selected from schools known for high English proficiency levels. This selection process involved 560 primary learners from the capital city of each Malaysian state and one federal territory. In each capital city, two primary schools with high English proficiency recommended by district education officers were chosen. From these schools, 20 upper primary learners, predominantly in Year 6 or Year 5, were selected for essay writing. Notably, exceptional Year 4 students were also included.

The essay writing stage is facilitated by English teachers who provide specific instructions and a time limit. The example instruction is as follows:

"Dear pupils, you are requested to compose an essay of at least 80 words. You may choose topics from areas such as 'World of Knowledge' (covering themes like occupations), 'World of Family and Friends' (including personal experiences), or 'World of Stories' (encompassing fictional narratives). A time frame of 60 minutes will be allotted for this task."

While illustrated in the context of upper primary education in Malaysia, this methodology is adaptable to secondary or tertiary education. The themes and time duration for essay writing can be modified to suit different educational levels and needs. It is essential that learners write legibly and neatly, as the essays are subsequently digitalized into DOCX format for analysis.

The legibility of essays is crucial as they need to be converted into DOCX format. This format compatibility is a prerequisite for the corpus linguistics software used in this study. Once the essays are digitalized, the authors employ SketchEngine as the recommended corpus analysis tool. A series of user manual snapshots are provided in the appendix to facilitate understanding.

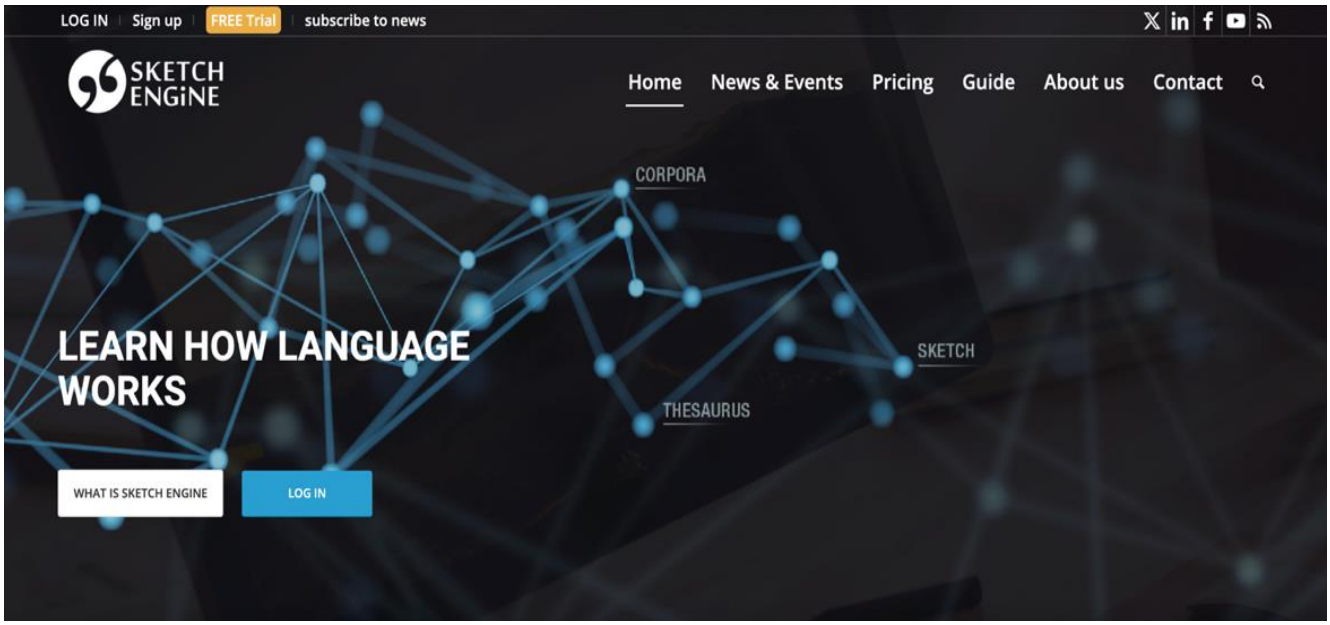


Fig. 2 SketchEngine

The first step in using SketchEngine involves accessing its web portal at https://www.sketchengine.eu/. New users must undergo a SIGN UP process, with a 30-day trial subscription available, requiring personal and professional details. Conversely, existing users can directly LOG IN. Given the academic nature of this research, selecting an academic account is advised. This requires information similar to that for the trial subscription, emphasizing SketchEngine's commitment to a personalized, academically-focused user experience.

Upon successful login, users are greeted with a dashboard, the central hub for corpus creation and analysis. The initial steps involve defining the corpus name and a brief description, followed by the crucial phase of uploading the essays. The interface displays the cumulative word count, offering a quantitative data perspective. Once the upload is complete, the researcher compiles these essays into a cohesive corpus. Using SketchEngine's WORDLIST feature, researchers can focus on specific linguistic elements, like prepositions in this study. This tool's versatility allows for targeted analysis of vocabulary and phrases, essential for this research.

The subsequent stage in this methodology involves identifying the top twenty salient functional and content words, including two-word to four-word phrases (bigrams, trigrams, qualgrams), marking the initiation of the quantitative analysis. This process thoroughly examines Keywords in Context (KWIC) or concordances to derive authentic sentence samples. For enhanced accuracy in this analysis, involving at least three language experts is highly recommended. Their collective expertise is pivotal in reducing human error, thereby strengthening the validity and reliability of the research outcomes. To aid in understanding, snapshots of the SketchEngine interface are provided in the appendix.

With the targeted learner corpora compiled and the salient functional and content words and phrases identified, the vocabulary and phraseological index development commences. This index comprises several components: a cover page, a foreword, a user manual, and a comprehensive list of identified vocabulary and phrases. The cover page features the index title and the names of the developers or researchers. The foreword introduces the index's purpose, while the user manual guides teachers and learners in utilizing the index for vocabulary and phrase learning in writing contexts.

Based on the doctoral study, two additional procedures are integrated. Firstly, despite predetermined writing themes, categorizing literacy genres is essential to inform users about the types of essays compiled in the index. These may include formal, informal, fiction, non-fiction, and other identified types. In the Malaysian context, where the CEFR framework is utilized, the index categorizes salient vocabulary and phrases from levels A1 to C2. This categorization ensures the index's relevance to the national educational system.

Secondly, sample sentences extracted from the concordances are included, providing learners with examples of authentic usage of the identified vocabulary and phrases. The index requires a final validation by three experts specializing in language, corpus linguistics, and learning materials. This validation may lead to adjustments in presenting technical aspects, enhancing the index's user-friendliness and effectiveness. Figure 3 below encapsulates the process involved in developing the indexes.

Fig. 3 Flowchart

4. Materials and Methods

The study under discussion, notably a doctoral research conducted by Wong et al. (2023a, 2023b), was selected to serve as a pertinent example in this study. Initially, the requisite human research ethics approval was secured, identified by the code 2021-0454-01 as authorised by the university. Additional approval was later obtained from the Ministry of Education, allowing the researchers to include 560 advanced upper primary school learners from the study. Selection of participants was planned: two schools from the capital city of each district, alongside a federal territory, were chosen. This encompassed 13 districts and one federal territory, targeting 28 primary schools noted for their higher proficiency in English as recommended by the respective district educational officers.

From this cohort, 20 advanced upper primary learners, typically in Year 5 or 6, who achieved band 5 or 6 in the classroom-based assessments were purposively selected by the English teachers. These learners were tasked with composing extended writing on themes such as 'world of family and friends', 'world of knowledge', and 'world of stories', with each piece consisting of no fewer than 100 words. The mode of submission allowed for both hard and soft copies, depending on the learners' preferences. Following the successful collection of data, hard copies were digitised for further analysis. All 560 extended writing was then compiled to form a learner corpora. This corpus was analysed to identify a frequency list of 20 salient functional and content vocabulary items and phrases, including bi-, tri-, and quadrigrams. Concordances for these elements were subsequently prepared, categorising them according to the Common European Framework of Reference for Languages (CEFR) levels ranging from A1 to C2. These concordances were utilised to illustrate authentic usage in sample sentences. Eventually, the final documentation included a cover page, a foreword, a user manual, and a comprehensive list of the identified salient functional and content vocabulary and phrases, accompanied by their respective concordances, thus providing examples of authentic sentences.

5. Results

This study aimed to elucidate the procedures for incorporating corpus linguistics to develop a vocabulary and phraseological index. Concurrently, certain results were presented as examples to enhance understanding of how to construct the index. These included a frequency list of functional and content words along with phrases. Sample concordances were displayed to demonstrate practical applications. Additionally, examples of categorising salient words according to the CEFR levels were provided.

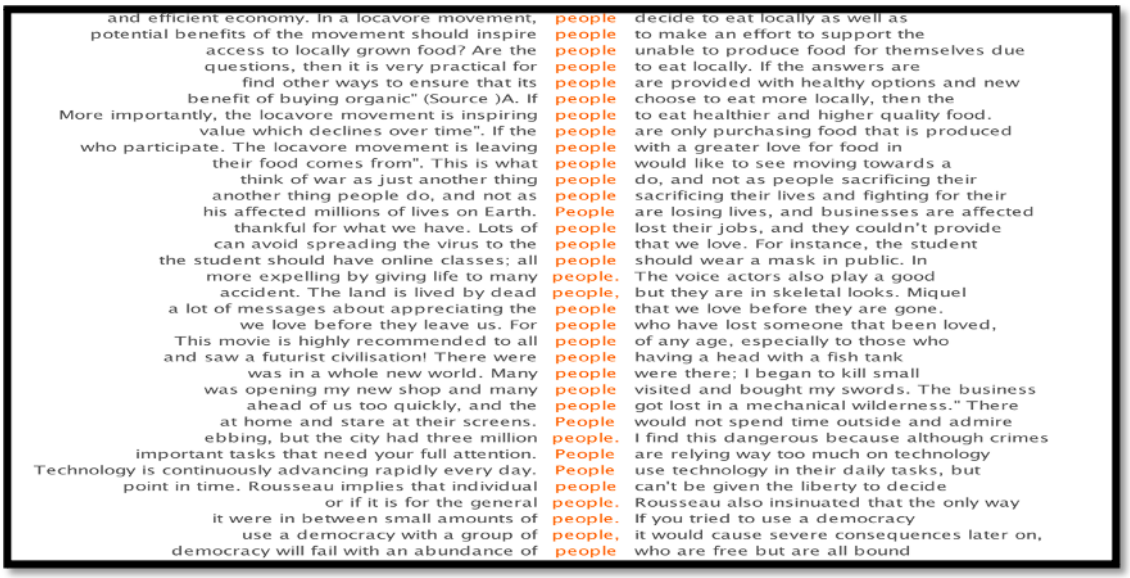
In corpus linguistic studies, representativeness was crucial for examining language patterns. In the doctoral study referenced, West and East Malaysia served as the primary areas for macro-level analysis of representativeness. However, micro-level analysis was conducted across various regions, including the Northern, Central, Southern, and East Coast regions. While Wong et al. (2023a) focused on findings from the Northern region, this study has chosen to illustrate the findings from the Central region. Below was a table presenting the salient vocabulary and phrases identified from the Central region.

Table 2. Salient vocabulary and phrases in Central region

Functional	Noun	Verb	Adjective	Adverb	Bigram	Trigram	Qualgram
the	people	go	good	not	of the	violent video games	I would like to
to	time	want	new	also	in the	one of the	the state of nature
and	games	get	violent	when	to the	would like to	eyes of T.J. Eckleburg
of	food	going	online	more	it is	I would like	is one of the
a	life	love	better	so	to be	the Covid-19 pandemic	one of the most
is	video	hope	different	how	the monster	I want to	the eyes of T.J.
I	monster	make	same	very	on the	there are many	of the movie is
that	things	see	beautiful	out	video games	looking forward to	of violent video games
in	school	went	long	up	and the	there is no	the locavore movement is
it	Covid-19	help	best	just	I would	the state of	am looking forward to
he	Gatsby	think	aggressive	really	there are	the movie is	I am looking forward
for	family	take	favourite	even	that he	state of nature	on the state of
are	world	become	first	back	it was	a lot of	there are many things
this	day	shows	great	always	that the	on the island	the past few years
my	movie	find	free	only	that I	the other boys	a very long time
his	Ralph	say	local	then	we can	because of the	video games as a
be	friends	started	healthy	there	for the	the end of	everyone has their own
we	society	need	big	most	is the	my dream school	from this Covid-19 pandemic
was	pandemic	got	social	too	with the	it was a	for me I want
with	way	live	important	again	in a	we need to	want my dream school

The table categorised various vocabulary including nouns, verbs, adjectives, adverbs, as well as phrases segmented into bigrams, trigrams, and qualgrams. First and foremost, nouns such as "people," "time," "games," and "food" primarily denoted tangible items, abstract concepts, or collectives, serving as focal points in sentences. Following that, verbs like "go," "want," "get," and "love" were essential for linking subjects to actions or states, suggesting dynamic sentence constructions. Then, adjectives including "good," "new," "violent," and "different" enriched extended writing by detailing noun characteristics. Moreover, adverbs such as "not," "also," "when," and "more" modified verbs, adjectives, or other adverbs, affecting the interpretation of actions or descriptors.

Subsequently, bigrams like "of the," "in the," and "to the" were crucial for phrase construction, often highlighting relationships within sentences. Finally, trigrams and qualgrams like "violent video games" and "I would like to" encapsulated specific expressions that may be culturally relevant or particularly useful in certain contexts. The results not only demonstrated how various vocabulary functioned within language but also underscored their utility in creating grammatically correct and contextually rich sentences. Such insights were invaluable for linguistics and language studies, helping to dissect how components contribute to overall meaning and interaction within texts. The focus on terms related to gaming, education, societal issues, and personal aspirations suggested a thematic concern with modern social issues and personal goals, reflecting prevalent discussions in society today. Therefore, three sample of concordances were presented below to illustrate some authentic sentences.



and efficient economy. In a locavore movement,	people	decide to eat locally as well as
potential benefits of the movement should inspire	people	to make an effort to support the
access to locally grown food? Are the	people	unable to produce food for themselves due
questions, then it is very practical for	people	to eat locally. If the answers are
find other ways to ensure that its	people	are provided with healthy options and new
benefit of buying organic" (Source)A. If	people	choose to eat more locally, then the
More importantly, the locavore movement is inspiring	people	to eat healthier and higher quality food.
value which declines over time". If the	people	are only purchasing food that is produced
who participate. The locavore movement is leaving	people	with a greater love for food in
their food comes from". This is what	people	would like to see moving towards a
think of war as just another thing	people	do, and not as people sacrificing their
another thing people do, and not as	people	sacrificing their lives and fighting for their
his affected millions of lives on Earth.	People	are losing lives, and businesses are affected
thankful for what we have. Lots of	people	lost their jobs, and they couldn't provide
can avoid spreading the virus to the	people	that we love. For instance, the student
the student should have online classes; all	people	should wear a mask in public. In
more expelling by giving life to many	people.	The voice actors also play a good
accident. The land is lived by dead	people,	but they are in skeletal looks. Miquel
a lot of messages about appreciating the	people	that we love before they are gone.
we love before they leave us. For	people	who have lost someone that been loved,
This movie is highly recommended to all	people	of any age, especially to those who
and saw a futurist civilisation! There were	people	having a head with a fish tank
was in a whole new world. Many	people	were there; I began to kill small
was opening my new shop and many	people	visited and bought my swords. The business
ahead of us too quickly, and the	people	got lost in a mechanical wilderness." There
at home and stare at their screens.	People	would not spend time outside and admire
ebbing, but the city had three million	people.	I find this dangerous because although crimes
important tasks that need your full attention.	People	are relying way too much on technology
Technology is continuously advancing rapidly every day.	People	use technology in their daily tasks, but
point in time. Rousseau implies that individual	people	can't be given the liberty to decide
or if it is for the general	people.	Rousseau also insinuated that the only way
it were in between small amounts of	people.	If you tried to use a democracy
use a democracy with a group of	people,	it would cause severe consequences later on,
democracy will fail with an abundance of	people	who are free but are all bound

Fig. 4 Concordances of 'people'

we can scroll on the internet. My	favourite	platform to buy anything that I want
In conclusion, I choose Zalora as my	favourite	platform for online shopping. I would prefer
My	favourite	movie this year that I watched is
exercise. I also can not join my	favourite	badminton and table, tennis classes. I can
Corona Spread, I was very sad. My	favourite	thing about travelling is memories! Like the
was extremely cold. But it was my	favourite	trip!
is playing badminton with my friends. My	favourite	place is City mall, and my favourite
favourite place is City mall, and my	favourite	shop is Popular. And my Happy Birthday
Birthday is on 27 May. And my	favourite	restaurant is Wajamama; my best friend is
best friend is Muhamad Adam Hayyqal. My	favourite	game is Call of Duty Modern Warfare,
Hi, I am Chee Shu Wei. My	favourite	food is Spaghetti because it is delicious.
Aglio e olio and Bolognese flavour. My	favourite	flavour is Carbonara because it has a
concert with my friends and play our	favourite	games, like badminton or swimming. Secondly, I
of Pink' situation, which is my personal	favourite.	It is in the category of flora
badminton, basketball, table tennis and hockey. My	favourite	sport is playing badminton. To play badminton,

Fig. 5 Concordances of 'favourite'

Fig. 6 Concordances of 'a lot of'

Given the abundance of authentic sentences for each salient vocabulary item, it was prudent to choose one or two authentic sentences as examples to include in the vocabulary and phraseological index. This facilitated a clear and practical demonstration of how each vocabulary item could be used in context. Subsequently, this section continued by categorising the salient vocabulary listed in Table 1 into the respective CEFR levels.

Table 3. Categorisation of Vocabulary

Vocabulary/ CEFR Level	A1	A2	B1	B2	C1	C2
Functional	the, my, a, that, this, his I, it, we, he, and, to, of, for, with, was, is, are, be	in	-	-	-	-
Noun	people, day, school, friends, family, food, world, things, games, movie	time, video	life, way, COVID-19, monster,	society	-	-
Verb	went, go, get, make, want, got, started, love, going, find, say, need, see	take, help, hope, become,	live	think, shows	-	-
Adjective	good, beautiful, new, best, first, important, better, favourite, long, same, big, different, great,	free, healthy,	local, social	violent, aggressive, social	-	-
Adverb	not, when, also, very, always, again, more, too,	up, just, even, how, there, back, really, only, most,	so, out	then	-	-
Bigram	in the, it is, it was, and the, is the, there are, in a, I would, we can	to the	on the, to be, the monster, with the	of the, for the, that I, that he, that the	-	-
Trigram	one of the, it was a,	violent video	would like to,	I want to, the state of,		a lot of

Vocabulary/ CEFR Level	A1	A2	B1	B2	C1	C2
	we need to, there are many, there is no, the movie is	games, on the island,	I would like, looking forward to, the other boys, because of the, the end of, my dream school,	state of nature		
Qualgram	is one of the, one of the most, there are many things	of the movie is, of violent video games, video games as a	I would like to, am looking forward to, I am looking forward, the past few years, everyone has their own, from this COVID- 19 pandemic, for me I want, want my dream school	the state of nature, eyes of T.K. Eckleburg, the eyes of T.J., the locavore movement is, on the state of	a very long time	-

The table organised salient vocabulary and phrases such as functional words, nouns, verbs, adjectives, adverbs, bigram, trigram and qualgrams into their respective CEFR levels. Some proper nouns such as 'Gatsby' and 'Ralph' were not included as proper nouns could not be given any level. Moreover, 'pandemic' was out of the range of A1 to C2. Hence, it was not identified above.

Firstly, the categorisation by parts of speech was pivotal for grammatical comprehension and application. This vocabulary and phraseological index aided English teachers in developing writing lessons focused on word usage and sentence construction, which were fundamental across all CEFR levels. Additionally, the inclusion of bigrams, trigrams, and qualgrams proved especially beneficial. Teaching these facilitated learners' understanding of context and usage, crucial for advancing from basic (A1-A2) to intermediate (B1-B2) CEFR levels. Such phrases exemplified how vocabulary can be naturally integrated into English, enhancing both comprehension and speaking fluency.

Moreover, the progression from single words to complex phrases reflected the linguistic complexity that increased as one advanced through the CEFR levels. English teachers can leverage this structured progression to introduce more sophisticated language as learners' proficiency developed, thereby making this index a dynamic tool for curriculum planning. The practical application of phrases like "violent video games" or "the Covid-19 pandemic" underscored contemporary relevance, rendering the learning process more engaging and applicable. Discussing current topics can boost learners' capabilities to employ English in real-world scenarios, thus augmenting their practical language skills across all CEFR levels.

For learners, the index served as a self-assessment tool, allowing them to identify their comfort levels with various parts of speech and their effectiveness in using certain phrases. This awareness can guide their personal study and practice. English teachers must, therefore, contextualise the use of this vocabulary based on their learners' proficiency levels and learning objectives.

In conclusion, the categorisation of vocabulary and phrases in the index was an effective educational tool for teaching English for writing. It clearly delineated the language learning journey from basic word recognition to complex phrase construction. By providing explicit linguistic targets, this categorisation not only facilitated teaching but also enabled learners to monitor their progress and tailor their studies accordingly. English teachers can utilise this index to customise lessons that were appropriate to the proficiency level and focus on practical language use.

6. Discussion

This study posited a novel integration of corpus linguistics in ESL education. It addressed the challenge of diverse learner proficiencies in ESL/EFL settings, proposing a vocabulary and phraseological index developed through corpus linguistics. This index, aimed at primary, secondary, and tertiary levels, utilised SketchEngine for analysis of a substantial dataset from learners' essays. Critically examining this method, its foundation laid in the corpus-driven approach, focusing on authentic language data without conforming to predefined theoretical constructs. This was aligned with contemporary pedagogical needs for authentic language usage and teaching materials that mirror real-world language patterns, as evidenced in studies like those by Bocorny et al. (2021) and Whitty et al. (2022). The corpus-driven approach was advantageous for its quantitative linguistic analysis, offering educators a time-efficient and effective resource for language teaching.

However, the practicality and adaptability of this index in varied educational contexts warranted consideration. Past studies, such as those by Tian (2022) and Rudneva (2020), emphasised the importance of integrating comprehensive language understanding in teaching materials, suggesting that the proposed index should extend beyond mere vocabulary listing to include deeper semantic and pragmatic roles of language. The emphasis on a robust, corpus-driven approach was commendable, but the index's effectiveness in fostering comprehensive language acquisition remained to be empirically validated across diverse educational settings.

This study also referenced the challenges and advancements in corpus-driven ESL/EFL teaching methods. Studies like those by Gu (2022) and Auziņa et al. (2020) highlighted the potential of corpus-driven tools in facilitating personalised educational pathways. This underscored the need for the proposed index to be dynamic and responsive to changing language trends and learner needs.

The notion of a corpus-driven vocabulary and phraseology index, as highlighted in this study, was innovative and aligned with current trends in applied linguistics. However, the effectiveness of such tools in enhancing language proficiency must be scrutinised through empirical studies. As noted by Crossley and Salsbury (2019), the integration of corpus data into teaching materials can significantly impact learners' language development. Yet, the question of how this integration translated into improved linguistic competence remained.

Moreover, the adaptability of the proposed index to different educational contexts, particularly those with limited technological resources, raised practical concerns. While this study references advanced tools like SketchEngine, the accessibility of such tools in under-resourced educational settings was a critical issue. Studies by Levitt et al. (2021) underscored the importance of considering technological equity in the implementation of digital language teaching tools. Additionally, this study suggested a need for ongoing updates to the index to reflect evolving

language usage. This dynamic nature of language, particularly in an era of rapid digital communication, necessitated continuous corpus analysis and updates to teaching materials. However, the feasibility of such ongoing updates in real-world educational settings was a pertinent concern.

In summary, while the proposed corpus-driven approach to vocabulary and phraseology index in ESL/EFL instruction was theoretically sound and innovative, its practical implementation, adaptability, and effectiveness in diverse educational settings require comprehensive empirical validation. The potential benefits of such an approach, if realised, could significantly advance the field of TESL/TOEFL and contribute to more effective language learning outcomes. However, the journey from theoretical innovation to practical application in diverse educational contexts was fraught with challenges that must be carefully navigated through ongoing research and collaborative efforts between linguists, educators, and technologists.

7. Potential and Challenges

This article presents a nuanced exploration of the potential and challenges inherent in integrating corpus linguistics into ESL/EFL education. The potential of this integration is multifaceted, primarily revolving around the enhancement of vocabulary and phraseological learning through corpus-driven insights. Corpus linguistics, with its empirical approach to language analysis, offers a robust framework for understanding and teaching real-world language use, as supported by scholars like Chen et al., (2023), Bocorny et al. (2021) and Whitty et al. (2022). This methodology allows for creating resources that are theoretically sound and empirically validated, reflecting actual language use in various contexts.

However, this article also acknowledges significant challenges in this endeavour. One primary concern is the complexity of corpus analysis tools, which may pose a barrier to educators and learners unfamiliar with computational linguistics, a sentiment echoed in Tian (2022) and Rudneva (2020) studies. Additionally, the diversity of learner needs and contexts necessitates the creation of adaptable and flexible resources, a challenge that is both pedagogical and logistical in nature. There is also the need to constantly update these resources to reflect the dynamic nature of language, as indicated by Li (2022).

Despite these challenges, this article underscores the transformative potential of corpus linguistics in ESL/EFL education, advocating for a more integrated and nuanced approach to language teaching and learning. The development of the vocabulary and phraseological index is a testament to the possibilities of bridging theoretical linguistic research with practical educational needs, paving the way for future innovations in the field.

Expanding on the potential and challenges in incorporating corpus linguistics into ESL/EFL education, this article illuminates the profound impact this integration can have on pedagogical practices. The potential lies in creating highly relevant and context-specific teaching materials, which are crucial for effective language acquisition. By utilizing corpus linguistics, educators can provide learners with authentic examples of language use, facilitating a deeper understanding of both common and nuanced language patterns.

However, challenges persist, particularly in the accessibility and usability of corpus tools for educators and learners. The technical nature of corpus analysis can be daunting for those without a background in computational linguistics. Additionally, the dynamic nature of language and the vast diversity of learner profiles call for continuous adaptation and customization of teaching materials. Overcoming these challenges requires ongoing professional development for educators, along with the development of more user-friendly corpus tools and resources. This approach will ensure that the benefits of corpus linguistics can be fully harnessed in ESL/EFL education, leading to more effective and engaging learning experiences.

8. Conclusion

The culmination of this study synthesised pivotal insights and implications stemming from the integration of corpus linguistics into ESL education. Anchored in a corpus-driven methodology, the study proposed the development of a vocabulary and phraseological index, utilising the SketchEngine.

This advancement strategically catered to the varied linguistic needs across educational levels, providing a bespoke and efficacious tool for language acquisition.

Central to the contribution was its practical application of corpus linguistics in generating authentic educational resources. This approach aligned with the findings of Bocorny et al. (2021) and Whitty et al. (2022), who emphasised the significance of real-world language application in teaching contexts. The article highlighted the crucial role of authentic, context-specific vocabulary and phrases for ESL learners, reflecting the fluid and ever-evolving nature of language. Drawing from diverse corpus-driven research, including works by Tian (2022) and Rudneva (2020), the study underlined the need for a deep and comprehensive approach to language education, advocating for materials that go beyond basic language comprehension.

The constructed vocabulary and phraseological index transcended a simple aggregation of terms. It embodied the core principles of corpus linguistics, offering language that was not only authentic and dynamic but also functionally relevant. This index specifically addressed the distinct requirements of individual learners, resonating with the corpus-driven instructional strategies explored by Li (2022). Additionally, it laid a foundational framework for subsequent research and applications in TESL, ensuring that both educators and learners have access to resources that are authoritative, instructive, and pedagogically sound.

In summation, this study marked a significant contribution to the fields of TESL/TOEFL and corpus linguistics. It effectively bridged the divide between empirical linguistic research and practical language instruction, representing a critical advancement in the creation of innovative, effective, and contextually appropriate teaching materials (in this case, vocabulary and phraseological index). The application of corpus linguistics in ESL education, as demonstrated in this study, reaffirmed the method's potential to revolutionise language learning and teaching methodologies.

9. Co-Author Contribution

The authors affirmed that there is no conflict of interest in this article. Wong carried out the field work, prepared the literature review and overlook the writeup of the whole article. Warid wrote the research methodology and data analysis. Mairas carried out the discussion and conclusion.

10. Acknowledgement

Artificial Intelligence (AI) was used ethically to improve the coherence of the sentence, check for grammatical errors and to reduce typos in the writing.

11. References

Adijaya, M. A., Armawan, I. K., & Kristiantari, M. G. (2023). Mobile-assisted language learning (MALL) innovation for vocational education. *International Journal of Language Education*, *7*(3). https://doi.org/10.26858/ijole.v7i3.52910

Al Abrawi, N. A. (2022). The positive outcomes and challenges experienced by modern foreign language teachers in Oman. *Education 3-13*, *51*(7), 1130-1142. https://doi.org/10.1080/03004279.2022.2051578

Auziņa, I., Levāne-Petrova, K., & Kaija, I. (2020). Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā. *Valoda: nozīme un forma / Language: Meaning and Form*, *11*, 7-26. https://doi.org/10.22364/vnf.11.01

Awajan, N. (2022). Increasing students' engagement: The use of new instructional designs in English literature online courses during COVID-19. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.1060872

Bocorny, A. E., Rebechi, R., Reppen, R., Delfino, M. C., & Lameira, V. M. (2021). A produção de artigos Da area das ciências Da saúde com o auxílio de key lexical bundles: Um estudo direcionado POR corpus. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *37*(1). https://doi.org/10.1590/1678-460x2021370101

Bouzouita, M. (2019). *Cycles in language change*. Oxford University Press.

Brend, R. (2019). *The summer institute of linguistics*. Walter de Gruyter GmbH & Co KG.

Chapelle, C. (2019). *The handbook of technology and second language teaching and learning*. John Wiley & Sons.

Chen, M., Wong, W. L., Yin, G., Swaran Singh, C. K., Mihat, W., & Yoong, S. M. (2023). Global intellectual trend of corpus linguistics studies among scholars in social sciences from September 2013 – September 2021. *Asian Journal of University Education, 19*(4), 613-631. https://doi.org/10.24191/ajue.v19i4.24615

Chen, X., Yang, Y., & Zhang, X. (2024). Validating a measure of growing pattern understanding in Chinese preschool children. *Early Childhood Research Quarterly*, *67*, 24-33. https://doi.org/10.1016/j.ecresq.2023.11.006

Chigbu, G. U., Emelogu, N. U., Egbe, C. I., Okoyeukwu, N. G., Eze, K. O., Nwafor, C. K., Patrick, C. P., Okon, O. E., Agbo, P. A., & Okwo, F. A. (2023). Enhancing ESL students' academic achievement in expository essay writing using digital graphic organizers: A mixed-methods research. *Heliyon*, *9*(5), e15589. https://doi.org/10.1016/j.heliyon.2023.e15589

Collins, L. C., Semino, E., Demjén, Z., Hardie, A., Moseley, P., Woods, A., & Alderson-Day, B. (2020). A linguistic approach to the psychosis continuum: (dis)similarities and (dis)continuities in how clinical and non-clinical voice-hearers talk about their voices. *Cognitive Neuropsychiatry*, *25*(6), 447-465. https://doi.org/10.1080/13546805.2020.1842727

Cordeiro, C. M. (2019). A corpus-based approach to understanding market access in fisheries and aquaculture international business research: A systematic literature review. *Aquaculture and Fisheries*, *4*(6), 219-230. https://doi.org/10.1016/j.aaf.2019.06.001

Crawford Camiciottoli, B. (2020). Using English as a Lingua Franca to engage with investors: An analysis of Italian and Japanese companies' investor relations communication policies. *English for Specific Purposes*, *58*, 90-101. https://doi.org/10.1016/j.esp.2020.01.003

De Fina, A. (2020). *The Cambridge handbook of discourse studies*. Cambridge University Press.

Devereaux, M. (2019). *Teaching language variation in the classroom*. Routledge.

Dobrovoljc, K. (2020). Identifying dictionary-relevant formulaic sequences in written and spoken corpora. *International Journal of Lexicography*, *33*(4), 417-442. https://doi.org/10.1093/ijl/ecaa008

Du, X., Afzaal, M., & Al Fadda, H. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.752134

Effatpanah, F., Baghaei, P., & Karimi, M. N. (2024). A mixed Rasch model analysis of multiple profiles in L2 writing. *Assessing Writing*, *59*, 100803. https://doi.org/10.1016/j.asw.2023.100803

Fajri, M. S., Kirana, A. W., & Putri, C. I. (2020). Lexical bundles of L1 and L2 English professional scholars: A contrastive corpus-driven study on applied linguistics research articles. *Journal of Language and Education*, *6*(4), 76-89. https://doi.org/10.17323/jle.2020.10719

Goldstone, R., McCarthy, R., Byrne, G., & Keen, D. (2023). Levelling the playing field for the international migration of nurses: The India English language programme. *BMC Nursing*, *22*(1). https://doi.org/10.1186/s12912-023-01308-7

Gu, L. (2022). Corpus-driven resource recommendation algorithm for English online autonomous learning. *Computational and Mathematical Methods in Medicine*, *2022*, 1-10. https://doi.org/10.1155/2022/9369258

Hardie, A., & Dorst, I. V. (2020). A survey of grammatical variability in Early Modern English drama. *Language and Literature: International Journal of Stylistics*, *29*(3), 275-301. https://doi.org/10.1177/0963947020949440

Hašková, A., Radulović, B., Mikla, Š., Stajić, S., & Zatkalík, D. (2023). Design and development of teaching materials aimed at mentor professional training. *Journal of Education Culture and Society*, *14*(2), 154-170. https://doi.org/10.15503/jecs2023.2.154.170

He, M., Ang, L. H., & Tan, K. H. (2021). A corpus-driven analysis of phrase frames in research articles on business management. *Southern African Linguistics and Applied Language Studies*, *39*(2), 139-151. https://doi.org/10.2989/16073614.2021.1920438

Hinkel, E. (2020). *Teaching academic L2 writing: Practical techniques in vocabulary and grammar*. Routledge.

Hu, T. (2023). Evaluation of the integration path of ideological and political elements in English major courses based on the ADDIE model. *Applied Mathematics and Nonlinear Sciences*, *8*(2), 2981-2992. https://doi.org/10.2478/amns.2023.2.00014

Hussein, R. F., Haider, A. S., & Al-Sayyed, S. W. (2021). A corpus-driven study of terms used to refer to articles and methods in research abstracts in the fields of economics, education, English literature, nursing, and political science. *Journal of Educational and Social Research*, *11*(3), 119. https://doi.org/10.36941/jesr-2021-0056

Islentyeva, A. (2020). *Corpus-based analysis of ideological bias*.

Iwamoto, N. (2023). Relationship between L2 proficiency and psychological traits with self-assessment bias among L2 speakers. *JALT Journal*, *45*(2), 185-210. https://doi.org/10.37546/jaltjj45.2-1

Jalaluddin, I., Ismail, L., & Darmi, R. (2020). Developing vocabulary knowledge among low achievers: Mobile augmented reality (MAR) practicality. *International Journal of Information and Education Technology*, *10*(11), 813-819. https://doi.org/10.18178/ijiet.2020.10.11.1463

Jing, G., & Eng, L. S. (2023). Comprehensive academic thesis writing module for English major undergraduates in a public University in China. *International Journal of Learning, Teaching and Educational Research*, *22*(10), 394-415. https://doi.org/10.26803/ijlter.22.10.22

Joannes, R., & Alsaqqaf, A. (2023). The effect of analytic text-based writing strategies on ESL argumentative writing among Malaysian form-six students in Sabah, Malaysia. *3L The Southeast Asian Journal of English Language Studies*, *29*(2), 169-185. https://doi.org/10.17576/3l-2023-2902-12

John, J. G., Gopal, R., Swaran Singh, C. K., Yesupatham, K. M., & Pratama, H. (2023). Reaching out to group contingencies in the ESL classroom. *Journal of Higher Education Theory and Practice*, *23*(1). https://doi.org/10.33423/jhetp.v23i1.5794

Katz, D. (2021). *Legal informatics*. Cambridge University Press.

Kazima, M., Jakobsen, A., Mwadzaangati, L., & Gobede, F. (2023). Teaching the concept of zero in a Malawi primary school: Illuminating the language and resource challenge. *ZDM – Mathematics Education*, *55*(3), 627-639. https://doi.org/10.1007/s11858-023-01473-8

Keselman, I., & Yakovleva, Y. (2021). Short teacher responses in the EFL classroom: A corpus-approach assessment. *Journal of Language and Education*, *7*(2), 175-188. https://doi.org/10.17323/jle.2021.9767

Khazali, N. A., Ismail, I., Sakamat, N., Mat Zain, N. H., Mohamed Noh, N. A., & Ishak, N. H. (2023). Smart pictorial dictionary via mobile augmented reality. *Bulletin of Electrical Engineering and Informatics*, *12*(2), 1019-1028. https://doi.org/10.11591/eei.v12i2.4009

Khedri, M., & Basirat, E. (2022). Interactive metadiscourse in dentistry research articles: Iranian vs non-Iranian academic writers. *Discourse and Interaction*, *15*(2), 77-100. https://doi.org/10.5817/di2022-2-77

Kubokawa, J. M. (2023). The multilingual poetry task: Innovating L2 writing pedagogy in the secondary classroom. *Journal of Second Language Writing*, *61*, 101039. https://doi.org/10.1016/j.jslw.2023.101039

Kudryavtseva, V., Barsuk, S., & Frolova, O. (2023). Promoting active online interaction with maritime English students. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *17*(3), 707-713. https://doi.org/10.12716/1001.17.03.23

Laske, C. (2020). *Law, language and change*. BRILL.

Lee, S., & Shin, S. (2021). Towards improved assessment of L2 collocation knowledge. *Language Assessment Quarterly*, *18*(4), 419-445. https://doi.org/10.1080/15434303.2021.1908295

Li, L. X. (2022). Meta-analytical approach to the impact of corpus-driven teaching on foreign language acquisition. *Mobile Information Systems*, *2022*, 1-11. https://doi.org/10.1155/2022/5049312

Liu, K., & Afzaal, M. (2021). Translator's style through lexical bundles: A corpus-driven analysis of two English translations of Hongloumeng. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.633422

Lu, X. (2019). *Computational and corpus approaches to Chinese language learning*. Springer.

Lyashevskaya, O., Vinogradova, O., & Scherbakova, A. (2022). Accuracy, syntactic complexity and task type at play in examination writing. *Studies in Corpus Linguistics*, 241-272. https://doi.org/10.1075/scl.104.10lya

Magalhães, L. C., & Matos Rocha, L. F. (2022). A emergência Da Interação Fictiva Em sala de aula Como estratégia de ensino E aprendizagem / The emergence of fictive interaction in the classroom as a teaching and learning strategy. *REVISTA DE ESTUDOS DA LINGUAGEM*, *30*(2), 496. https://doi.org/10.17851/2237-2083.30.2.496-518

Misesani, D., Janggo, W. O., & Wuwur, M. S. (2020). Need analysis in ADDIE model to develop academic speaking materials. *Ethical Lingua: Journal of Language Teaching and Literature*, *7*(2), 438-446. https://doi.org/10.30605/25409190.226

Mohd. Said, N. E., Thambirajah, V., Md Yunus, M., Tan, K. H., & Mohamed Sultan, F. M. (2022). Exploring English descriptive writing vocabulary acquisition through creative pedagogical strategies. *3L The Southeast Asian Journal of English Language Studies*, *28*(4), 212-228. https://doi.org/10.17576/3l-2022-2804-15

Munalim, L. O., Genuino, C. F., & Tuttle, B. E. (2022). Question-declaration coupling in a university meeting talk: Discourse of social inequality and collegiality. *Studies in English Language and Education*, *9*(1), 400-417. https://doi.org/10.24815/siele.v9i1.21293

Narkprom, N., & Phoocharoensil, S. (2022). Lexical bundles in native English speakers' and Thai writers' dissertations. *GEMA Online® Journal of Language Studies*, *22*(3), 43-62. https://doi.org/10.17576/gema-2022-2203-03

Nguyen, L. M. (2020). *Computational linguistics*. Springer Nature.

Peñarroja, M. R. (2021). Corpus pragmatics and Multimodality: Compiling an ad-hoc multimodal corpus for EFL pragmatics teaching. *International Journal of Instruction*, *14*(1), 927-946. https://doi.org/10.29333/iji.2021.14155a

Peng, N., & Chen, X. (2023). Model-based learning towards environment in cross-cultural communication: A mediating role of technology innovation acceptance in culture congruence and English language teaching for environmental education. *Economic Research-Ekonomska Istraživanja*, *36*(3). https://doi.org/10.1080/1331677x.2022.2162946

Portelli, S. (2018). The nineteenth-century Italian translators of Lord Byron's Marino Faliero. *Quaderni d'italianistica*, *38*(1), 153-171. https://doi.org/10.33137/q.i..v38i1.31155

Pratiwi, W. R., Kuswoyo, H., Puspitasari, M., Juhana, J., & Bachtiar, B. (2024). Driving to communicative approach: The innovative teaching speaking methods in Indonesian English immersion program. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(1), 626. https://doi.org/10.11591/ijere.v13i1.25420

Rajprasit, K. (2022). Design and development of an English as a global language MOOC to increase global Englishes awareness: Evaluation in a Thai University. *3L The Southeast Asian Journal of English Language Studies*, *28*(1), 121-138. https://doi.org/10.17576/3l-2022-2801-09

Ramos, F. P., & Guzmán, D. (2021). Examining institutional translation through a legal lens. *Target. International Journal of Translation Studies*, *33*(2), 254-281. https://doi.org/10.1075/target.21003.pri

Roberts, F., Aziz, A. A., & Matore, M. E. (2022). Establishing the validity and reliability of the Malaysian English language textbook evaluation checklist (MELTEC) using Rasch measurement model (RRM). *Journal of Language Teaching and Research*, *13*(1), 38-45. https://doi.org/10.17507/jltr.1301.05

Rudneva, M. (2021). Corpus-driven Esp pedagogy: A preliminary case study. *Journal of Teaching English for Specific and Academic Purposes*, 241. https://doi.org/10.22190/jtesap2003241r

Salama, A. H., & Altohami, W. M. (2022). Enhancing EFL students' COCA-induced Collocational usage of coronavirus: A corpus-driven approach. *International Journal of Advanced Computer Science and Applications*, *13*(2). https://doi.org/10.14569/ijacsa.2022.0130226

Sanosi, A. B. (2022). The use and development of lexical bundles in Arab EFL writing: A corpus-driven study. *Journal of Language and Education*, *8*(2), 106-121. https://doi.org/10.17323/jle.2022.10826

Santosa, M. H., Harismayanti, I., & Jaya Putra, I. N. (2022). Technology in action: Developing

Gamification handbook in English teaching and learning for the 21st century learners. *Teaching English as a Second or Foreign Language--TESL-EJ*, *26*(101). https://doi.org/10.55593/ej.26101a2

Sari, L. I., & Sari, R. H. (2020). Exploring English language needs of Indonesian marine pilots: A need analysis and its implications in ESP classrooms. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *14*(4), 909-917. https://doi.org/10.12716/1001.14.04.15

Shruthi, S., & Aravind, B. R. (2023). Cognitive load theory for ESL students: Mixed method to employ difficulty in using tenses while writing. *Journal of Higher Education Theory and Practice*, *23*(16), 168-179. https://doi.org/10.33423/jhetp.v23i16.6472

Slimi, Z., Al Alawai, F., Al Alyani, H., Al Abri, S., Al-Farsi, F. A., & Al Balushi, K. (2022). Writing issues in ESL and their potential solutions: Case study IMCO's Foundation students. *Journal of Educational and Social Research*, *12*(6), 81. https://doi.org/10.36941/jesr-2022-0146

Tarrayo, V. N., & Anudin, A. G. (2021). Materials development in flexible learning amid the pandemic: Perspectives from English language teachers in a Philippine State University. *Innovation in Language Learning and Teaching*, *17*(1), 102-113. https://doi.org/10.1080/17501229.2021.1939703

Tian, X. (2022). Construction of a multimodal corpus of college students' spoken English based on semantic concepts. *Mobile Information Systems*, *2022*, 1-10. https://doi.org/10.1155/2022/5270408

Tichý, O. (2021). Corpus driven identification of lexical bundle obsolescence in late Modern English. *Studies in Language Companion Series*, 101-129. https://doi.org/10.1075/slcs.218.04tic

Tognini-Bonelli, E. (2002). Corpus linguistics at work. *Computational Linguistics*, *28*(4), 583-583. https://doi.org/10.1162/coli.2002.28.4.583a

Villena-Ponsoda, J. A. (2019). *Language variation - European Perspectives VII*. John Benjamins Publishing Company.

Wan Mahzan, M. S., Alias, N. A., & Ismail, I. S. (2020). Unboxing the design of English as a second language (ESL) learning video game for Indigenous learners: An empathic Designbased approach. *Asia Pacific Journal of Educators and Education*, *35*(2), 39-56. https://doi.org/10.21315/apjee2020.35.2.3

Whitty, L., Parkinson, J., & Pham, H. T. (2022). Can and could in academic writing : A corpus-driven comparison of English L1 and Vietnamese EFL students. *The Journal of AsiaTEFL*, *19*(1), 93-108. https://doi.org/10.18823/asiatefl.2022.19.1.6.93

Wong, W. L., Muhammad, M. M., Mihat, W., Abdul Rahman, M., Ya Shak, M. S., & Lee, M. C. (2023a). Using Technologized 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. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *31*(1), 298-314. https://doi.org/10.37934/araset.31.1.298314

Wong, W. L., Muhammad, M. M., Mihat, W., Ya Shak, M. S., Rahman, M. A., & Prihantoro, P. (2023b). Vocabulary index as a sustainable resource for teaching extended writing in the post-pandemic era. *World Journal of English Language*, *13*(3), 181. https://doi.org/10.5430/wjel.v13n3p181

Wu, X., & Yang, H. (2022). Unpacking the functions of personal Metadiscourse in teachers' classroom discourse. *Sustainability*, *14*(20), 13502. https://doi.org/10.3390/su142013502

Yan, J. (2022). Deep integration of Rasch model and English classroom: Language teaching development under information technology. *Journal of Sensors*, *2022*, 1-11. https://doi.org/10.1155/2022/3744678

Zhang, J. (2020). The construction of college English online learning community under ADDIE model. *English Language Teaching*, *13*(7), 46. https://doi.org/10.5539/elt.v13n7p46