Global Intellectual Trend of Corpus Linguistics Studies among Scholars in Social Sciences from September 2013 – September 2021

Chen Minjie[1], Wong Wei Lun[2*], Yin Guojie[3], Charanjit Kaur Swaran Singh[4], Warid Mihat[5], Yoong Soo May[6]
[1 3]Mianyang Teachers' College, Sichuan Province, Mianyang City, P.R. China
chenminjie@mtc.edu.cn
yaleygj@mtc.edu.cn
[4]Faculty of Languages and Communications, Sultan Idris Education University,
35900 Tanjong Malim, Perak, Malaysia
colin_me-and-you@live.com.my
charanjit@fbk.upsi.edu.my
[5]Academy of Language Studies, University Teknologi MARA Kelantan Branch, 18500 Machang, Malaysia
waridmihat@uitm.edu.my
[6]Special Education Division, Ministry of Education Malaysia, 62604 Wilayah Persekutuan, Putrajaya, Malaysia
soomayyoong@gmail.com
[*]Corresponding Author

https://doi.org/10.24191/ajue.v19i4.24615

Received: 19 August 2023
Accepted: 23 October 2023
Date Published Online:6 November 2023
Published: 6 November 2023

Abstract: The primary objective of this article is to undertake a comprehensive and methodical examination of corpus-driven research endeavours carried out in the Asian region. A plethora of recent scholarly investigations have elucidated that the employment of corpus-driven methodologies across diverse fields and aims could derive advantageous outcomes through the utilisation of a diverse array of corpus linguistics software tools. The primary objective of corpus-driven analysis is to facilitate scholars in acquiring novel insights pertaining to a specific corpus, including the identification of pivotal keywords. The present investigation undertook a comprehensive examination of 21 scholarly research publications through the meticulous application of a systematic literature review methodology. The selection of Google Scholar and Scopus as the designated databases for the purpose of conducting a comprehensive search for pertinent scholarly resources has been made. A total of twelve research articles were procured from the Scopus database, whereas an additional eight articles were obtained from the esteemed Google Scholar platform. The empirical evidence presented in this study demonstrates the significance of corpus-driven research within the realm of Asian studies, primarily owing to the inherent ethnic diversity prevalent in this particular geographical region. The exploration of authentic language can be undertaken through a lens characterised by determination and focus. This statement elucidates the fact that the utilisation of corpus-driven research methodologies yields a greater abundance of relevant and genuine data, particularly within the realm of linguistics disciplines.

Keywords: Asia, Corpus-driven, Corpus linguistics, Literature Review

1. Introduction

History indicates that before the emergence of corpus linguistics (henceforth CL), most linguists studied English grammar using the native speaker's innate intuition (Meakins, 2018). Soon, they realised that the corpus made a major contribution to the work of linguists, researchers and educators in comprehending the natural use of language of certain communities. After several decades, it was noticed that many linguists and researchers were employing CL as a key methodology in social science and linguistics studies. However, young learners were not the main target group. Zhang and Yang (2021) claimed that the normal scenario was for mature learners to be chosen as research participants instead of those of other ages. This demonstrates the need for CL to be used with young learners.

Corpus-driven approach is one element of CL (Alsahlanee & Jaganathan, 2022; Friginal, 2018). The purpose of this approach was totally different from those used in other research as the analysed results the former produced were not intended to be adjusted to any levels to match the existed language categories proposed by other researchers (Hussein, 2017). The ultimate aim was to collect authentic and real data from a target group to determine the emergence of language patterns that were unique yet relevant to the particular target group. In this study, corpus-based approach (CDA) will be employed, corpus linguists accept the importance of language, multi-word units and collocation.

Corpus-based linguists employ tagged and annotated corpora as a principal resource to provide tangible evidence for existing language theories and models. This means, the application of data to investigate various levels of analysis, such as lexis, syntax, pragmatics, semantics or discourse in different study disciplines, including translation (Al-Wahy, 2021), stylistics (Gandón-Chapela, 2020), language teaching and learning (Qin & Pan, 2022; Roslim et al., 2021), forensic linguistic (Karlińska, 2021), and grammar theory contain systemic functionalism (Abney, 2021; Gabrielatos, 2019). The socio-pragmatic glossary scheme crafted by Athanasiadou (2020) was implemented into historical corpora (Campbell, 2021), and it defined data as sociolinguistic information, namely sex, status, age, and interaction roles. Based on these suggestions, corpus-based researchers cannot work in parallel with one language model, yet it is best explained as a methodology with which to explore a language from various usage-based perspectives.

Data analysis is critical in social sciences research (Lytras, 2020). Researchers have used various data analysis techniques based on their areas of specialisation and scientific objectives. A variety of analysis procedures have been published for different research purposes, including experimentation (Ge, 2018), introducing educational platforms (Romero-Hall, 2020), investigations (Ghavifekr & Wong, 2022), application (Sengupta, 2022), and evaluation (Hannis et al., 2022). Several studies have emphasised the importance of data analysis in terms of potential research solutions (Bibri, 2019), the link between developing trends (Satrovic, 2019), and the advantages of research (Hilton, 2020). Through critical systematic literature studies, several serious issues have been raised (Zainuddin et al., 2020). Other studies addressed critical issues such as suggesting future research (Chung et al., 2019), suitability, synthetisation, and framework development (Fetzer, 2018), addressing gaps (Hempel et al., 2019), and describing purpose (van der Meij et al., 2017). Nonetheless, it is undeniable that the source of data has a major impact on the way it is analysed.

Data types are the foundation of all study areas (Cohen, 2017). The relevance of a type of data frequently enables a researcher to reveal intriguing results when addressing their research questions. Types of data depend entirely on the interpretations made during the data analysis. Similarly, researchers discovered that the data types used in corpus-driven research indicated a growing trend in the application of the corpus-driven approach in different disciplines. Although the kind of data presented in corpus-driven research is somewhat relevant to learners' academic improvement, most researchers were targeting a certain type of linguistics as the main corpus, but not one that was related to education. Research can involve a variety of objectives, including efficacy (Singh et al. 2019), technique development (Ngulube, 2019), and experimentation (Dawson, 2019). The corpus-driven studies that used content and narrative analysis uncovered several themes; researchers used the data variously to address research gaps (Dehghanzadeh et al., 2019), develop a prototype (Singh et al., 2019), conduct investigations (Hatice et al., 2018), develop and design (Hung, 2018), and to define findings (Tan, 2018).

Since 1994, scholars have examined numerous different areas of corpus-driven approach using various approaches and formats (Soderland et al, 1994). These studies provided acceptable recommendations, as well as offering some insights and future directions for those involved in similar fields of study. The corpus-driven research that has been conducted thus far is satisfactory. The current study aims to illustrate the useful prior research undertaken in corpus-driven studies in Asia.

The purpose is to highlight prior efforts in corpus-driven research by mapping the research landscape into a coherent taxonomy and determining the relevant aspects and settings that have contributed to the research scope. This section is structured as follows: the first section discusses the corpus linguistics software employed in the data analysis. The second section describes the corpus used in the studies. Section three summarises the purposes of the reviewed studies. Section four focuses on the research design involved in each study. These existing studies were obtained from various journals relevant to corpus-driven research dating from 2013 to 2021.

2. Corpus Driven Analysis

Ramos and Guzmán (2021) provided a different opinion on the CDA by describing it as an approach with the findings that are not meant to be adjusted to any degree to match with the pre-assumed categories drawn by the researcher. It could be noticed through some traditional grammatical categories such as clause components and part of speech that might not always match with the authentic language data instead of adjusting the data just to fit with any concepts proposed by a theory. In order to solve the issue mentioned, new information and narration of language are demanded. Hence, linguists who practice a CDA put in efforts to collect data without any apparent preconceptions since the aim is to allow the emergence of authentic and original language patterns from the data without contamination, leading to a thorough description of the language in parallel with corpus data (Keselman & Yakovleva, 2021). Practically, with reference to Liu and Afzaal (2021), it could be said that the insights and data identified from word-frequency lists, keywords, concordances and clusters provide a reliable dependence on frequency, a focal point on lexis, recognition of lexical phrases as the fundamental unit of language description, and obligation to the potential of descriptive pattern in lexicogrammar to form specific significances. In other words, it suggests that the CDA has a connection with methods of data analysis and data extraction, corpus-driven linguists, and a set of observations that appeared from studies employing the CDA. Eventually, methods of CDA and observations conducted play a vital role in developing the patterns of grammar (Tichý, 2021), syllabuses of language teaching and learning to focus on lexis and phrases (Keselman & Yakovleva, 2021; He et al, 2021), finally, the dictionaries (Dobrovoljc, 2020). In addition, it also involves the development of different disciplines in scientific and academic (Hussein et al, 2021; Cordeiro, 2019).

3. Systematic Literature Review

Systematic literature reviews require a thorough explanation and a comprehensive evaluation of the selected themes, from which it may be possible to suggest some implications and insights for future researchers. The current study undertook a brief assessment to identify, examine, construe, and critically analyse the extant literature on the subject. This is the most common type of evidence employed by researchers to conduct many types of investigations. The systematic review method is well-known for its evidence-based approach, enabling researchers to identify research gaps and explain them scientifically. It was anticipated that this study might discover areas that require additional exploration. Systematic reviews include individual research (referred to as primary studies) and systematic reviews (referred to as secondary studies).

3.1 Search Strategy

The articles reviewed for this study were selected based on studies from different fields. The search strategy is illustrated in Figure 1. For this study, the researchers selected two digital databases to extract the relevant articles. Scopus offers global research outputs from various domains. Google Scholar includes journal articles from various domains, such as the social sciences, arts, and humanities.

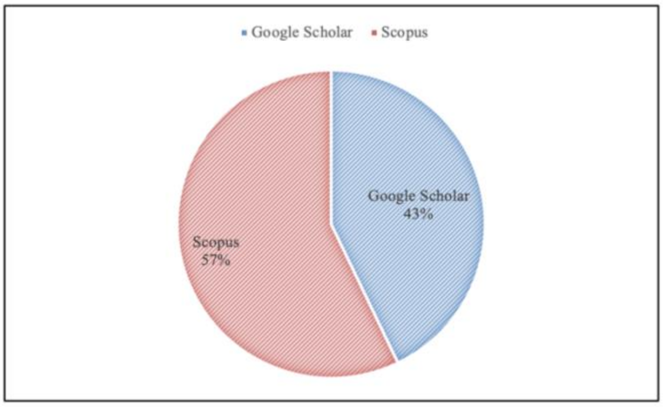

Fig. 1 Databased and percentage of articles collected

3.2 Search Queries

The researchers searched for the material at the end of September 2021 using Scopus and Google Scholar. The three focus areas were described using various word combinations and terms. The "AND" operator connected the focus areas, while the "OR" operator was used to group phrases and words with similar meanings in each focus area. The first area of interest was corpus-driven, which was followed by a search for alternative phrases for the data analysis searches identified in the literature. The researchers excluded book chapters, monographs, and any other types of reports that did not fall into the categories of conference papers/proceedings or journal articles. Only two selections were deemed eligible for inclusion: recent or updated scholarly papers relevant to the subject of the study.

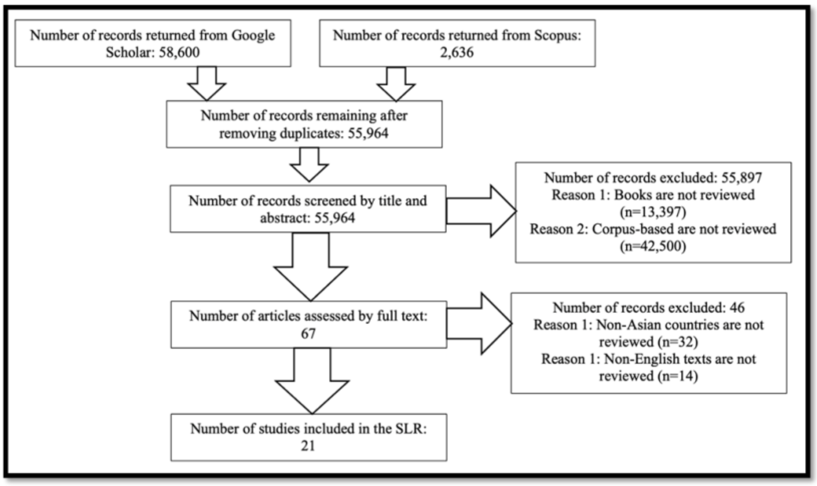
Number of records returned from Google Scholar: 58,600
Number of records returned from Scopus: 2,636
Number of records remaining after removing duplicates: 55,964
Number of records screened by title and abstract: 55,964
Number of records excluded: 55,897
Reason 1: Books are not reviewed (n=13,397)
Reason 2: Corpus-based are not reviewed (n=42,500)
Number of articles assessed by full text: 67
Number of records excluded: 46
Reason 1: Non-Asian countries are not reviewed (n=32)
Reason 1: Non-English texts are not reviewed (n=14)
Number of studies included in the SLR: 21

Fig. 2 Flow diagram of systematic literature review

3.3 Qualitative Checklist

A qualitative checklist was adopted from PRISMA checklist 2020 developed by Page et al. (2021). There were 27 items in the PRISMA checklist 2020. However, there were 14 selected for the use of corpus-driven systematic literature review.

Table 1. Qualitative checklist

Item	Remark
Title is corpus-driven related	
Full abstract	
Research rationale is identified	
Research objectives are identified	
Research purpose is mentioned	
Corpus-driven as the main research method	
Data collection is mentioned	
Corpus used is mentioned	
Software used is mentioned	
Research design is mentioned	
Results are reported	
Full access is available	
Asian countries as research site	
English as the medium	

3.4 Selection of article from various field of studies

The rationale for selection of articles from various fields of studies were to gain a comprehensive view of the of corpus-driven research conducted in Asia in various fields of studies and the solution was to comprise all studies to assist academics to obtain new insights into a given corpus, such as the most critical keywords.

3.4.1 Study Selection

The search queries were constructed by speculating on the section's primary phrases and their neighbouring meanings. The first segment addressed the synonyms and terms associated with corpus-driven research. The second stage sought the various data analysis methods. The subsequent section discussed the type of data used in the analysis, with the following search queries generated as a result of numerous search attempts. The queries were divided into two sets for use during both the title and abstract scanning, as well as the complete text reading phases.
(corpus-driven OR corpus driven OR corpus-assisted OR corpus assisted) AND (Asia OR Asian)
[[All: corpus-driven] OR [All: corpus driven] OR [All: corpus-assisted] OR [All: corpus assisted]] AND [[All: Asia] OR [All: Asian]]

3.4.2 Inclusion and Exclusion Criteria

During the first screening phase, a set of criteria was used to identify the research. Studies published between 2013 and September 2021 were included. An important aspect was that all papers in the databases had to be written and published using the English language. The Scopus and Google Scholar papers had to contain original theses, articles, reviews, or brief surveys. The first stage was to identify any corpus-driven approach employed in any research in the fields of corpus linguistics or linguistics, as well as other studies that employed a corpus-driven or corpus-assisted approach as the main design. Based on a variety of corpora, the corpus-driven approach comprises data-driven outcomes, such as language trends, the most primitive forms of collocation, and the factors involved in high-frequency terms. Next, the researchers examined the corpus selected, software utilised, procedures, data size and research aim. Then, the corpus-driven review was categorised into different novel emerging trends, before being further divided into the corpus linguistics software, corpus used, research purposes, and research design. A table is attached to illustrate which exclusion criteria were involved in this review.

Table 2. Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
-Title is in English -Scopus are theses, conference papers/proceedings, journal articles or review papers. -Google Scholar are theses, conference papers/proceedings, journal articles or review papers. -Corpus-driven approach is mandatory -Asia countries are compulsory.	-Corpus-assisted studies with corpus-based approach - Studies conducted out of Asia countries

3.4.3 Taxonomy

Taxonomy is a systematic approach to arranging and grouping similar items or concepts based on their attributes or characteristics. Within the purview of a systematic literature review (SLR), a taxonomy serves as a conceptual framework that facilitates the classification and arrangement of the research papers that have been identified and scrutinised during the review process. The utilisation of a structured and systematic approach is paramount in effectively managing the copious amount of information that is readily available in the literature. Prior to embarking on the development of a taxonomy, it is of paramount importance to unambiguously articulate the research inquiry or aim of the systematic literature review. The present inquiry shall serve as a guiding principle for the process of taxonomy development.

In accordance with the research inquiry, the initial measure entails discerning the principal categories that pertain to the subject matter being scrutinised. It is imperative that the categories employed accurately capture the primary facets or dimensions of the research domain.

It is noteworthy that every category has the potential to be subdivided into subcategories, thereby enabling the capture of more precise and nuanced concepts or ideas within the overarching category. The implementation of a hierarchical structure facilitates a comprehensive and meticulous examination of the literary works.

Upon the establishment of categories and subcategories, a coding scheme is formulated to allocate each research paper to its corresponding category. The proposed framework may encompass a series of key terms, expressions, or descriptors that aptly capture the essence or thematic orientation of each manuscript. The development of a taxonomy is a process that is characterised by iteration, whereby the categories and subcategories are refined and revised based on the findings of the literature review. As the review advances, novel insights and themes may surface, necessitating modifications to the taxonomy.

It is of utmost significance to ascertain the maintenance of consistency and reliability in the application of the taxonomy. It is recommended that the papers be categorised independently by multiple reviewers, and that measures of intercoder reliability be employed to gauge the degree of concurrence among the reviewers. The resolution of discrepancies can be achieved through the process of deliberation and the attainment of a shared agreement. Upon the successful categorisation of papers utilising the established taxonomy, the subsequent step entails the extraction of pertinent data from each paper. Subsequent to the collection of data, it can be subjected to a comprehensive analysis and synthesis process, with the ultimate aim of providing a response to the research inquiry or objectives of the systematic literature review.

Through adherence to a rigorously structured classification system, scholars are able to systematically arrange the body of literature, discern recurrent themes and tendencies, and furnish a thorough and all-encompassing survey of the extant scholarship within a given domain. The implementation of a rigorous and replicable review process serves to augment the scholarly inquiry and facilitates the derivation of meaningful conclusions and informed recommendations from the synthesised evidence.

4. Results

The following section presents the taxonomy, which summarises the results of the search process. The process began by searching, scanning, filtering, and full-text reading of all the selected articles. These were then classified into four major categories.

4.1 Corpus Linguistics Software

The primary category of the taxonomy comprises articles on corpus linguistics software (n= 20/21). This section explores corpus linguistics software further by dividing it into eleven types.

Table 3. Types of corpus linguistics software (2013-2021)

Corpus Linguistics Software	Author(s)/Year
AntConc	Awab & Norazit (2013); Pristian (2016); Chan (2017); Nor Fariza Mohd Nora & Adlyn Syahirah Zulcafli (2020); Muhammad Imran Shah & Singh (2021); Nor Fariza Mohd Nor et al. (2021) *f*=5
CANCODE	Tsuchiya & Handford (2014) *f*=1
ConcGram	Cheng & Lam (2013) *f*=1
Collocate 1.0	Ang & Tan (2018); Ang & Tan (2019) *f*=2
KfNgram	Ang & Tan (2019); He et al. (2021) *f*=2
LancsBox	Goyak et al. (2021) *f*=1
Langid	Mohammad Arshi Saloot (2018) *f*=1
Ngram	Muhammad Imran Shah & Singh (2021) *f*=1
Transana	Tsuchiya & Handford (2014) *f*=1
Wmatrix	Cheng & Lam (2013); Huang (2015) *f*=2
WordSmith	Siti Aeisha Joharry (2016); Leung (2016), Siti Aeisha Joharry & Hairani Wahab (2019); Norsimah Mat Awal et al. (2019); Nor Fariza Mohd Nor et al. (2019); Liu & Muhammad Afzaal (2021) *f*=6

With reference to the table 3, a pie chart in Figure 3 shows the use ratio of each corpus linguistics software program.

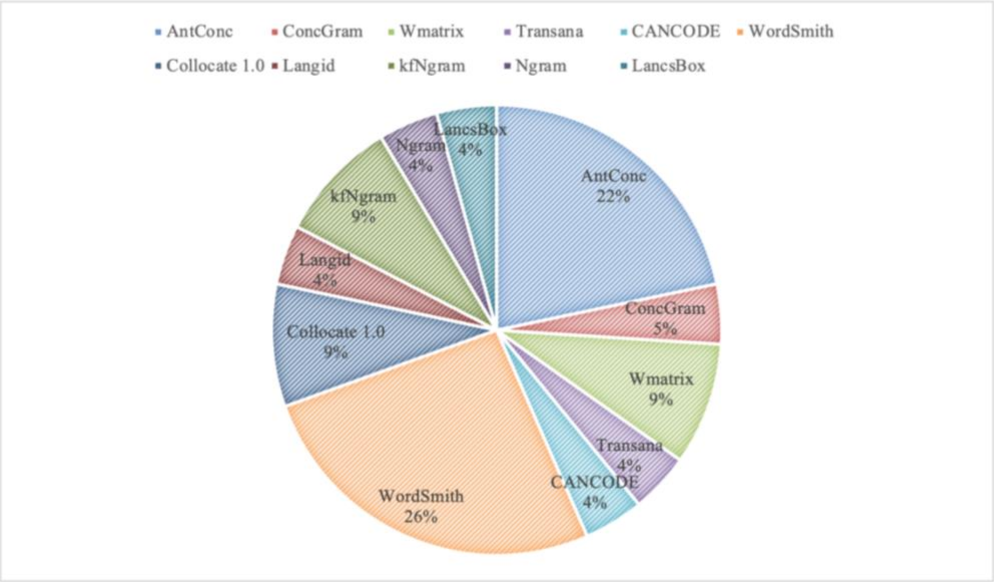

Fig. 3 Percentage of corpus linguistics software used (2013-2021)

Ten corpus linguistic software applications were discovered through an analysis of 21 studies published between 2013 and 2021. To begin with, WordSmith (26%) was well-known among corpus linguists doing research (Siti Aeisha Joharry, 2016; Leung, 2016; Siti Aeisha Joharry & Hairani Wahab, 2019; Norsimah Mat Awal et al., 2019; Nor Fariza Mohd Nor et al., 2019; Liu & Muhammad Afzaal, 2021). According to Scott (2012), the WordSmith program retrieves wordlists, keyword lists, concordances, and collocations from a corpus. Wordsmith's Wordlist function was frequently used to search for and retrieve three-word, four-word, five-word, and six-word amalgamations are identified empirically in a corpus of natural language. Thus, the salient bundles were analysed for the elements that occurred most frequently in the corpus.

AntConc (22%) appears to be the second most frequently utilised type of corpus analysis software since it was employed in five research studies (Awab & Norazit, 2013; Pristian, 2016; Chan, 2017; Nor Fariza Mohd Nora & Adlyn Syahirah Zulcafli, 2020; Muhammad Imran Shah & Singh, 2021; Nor Fariza Mohd Nor et al., 2021). AntConc's objective is to examine wordlists, keyword lists, concordances, and collocations from corpora that have been analysed (Zufferey, 2020). WordSmith and AntConc have fairly similar functions. Nonetheless, the latter is available for access while the former needs to be purchased. Furthermore, WordSmith is slightly advanced in terms of function compared to AntConc. Using AntConc, some researchers introduced different approaches. For example, the mutual information score was calculated to assess the collocational strength of two words using the Gaussier, Langé, and Maunier formulae, as reported by Oakes in 1998 (Muhammad Imran Shah & Singh, 2021).

Collocate 1.0 (9%), Wmatrix (9%), and kfNgram (9%) were used at a similar rate by numerous researchers (Ang & Tan, 2018; Ang & Tan, 2019; Cheng & Lam, 2013; Huang, 2015; Ang & Tan, 2019; He et al., 2021). First, Barlow (2004) defined Collocate 1.0 as a tool for automatically extracting lexical bundles via the span option. This corpus recognised plain text files ending in.txt. Collocate 1.0 uses two statistical measures to extract n-gram (lexical bundle) lists: frequency and mutual information. On the other hand, Wmatrix was used to search for patterns and frequencies in the analysed corpus (Rayson, 2001). It produced quantitative data on the most frequently occurring semantic categories in the corpora, followed by a key semantic tag analysis to determine statistically which semantic categories were substantially more frequent in one corpus than another. Meanwhile, kfNgram was used in conjunction with Collocate 1.0 in data analysis. Collocate 1.0 was used to identify lexical bundles. After identifying potentially useful lexical bundles, the kfNgram software (Fletcher, 2002) automatically extracted lexical frames from the inventory of lexical bundles.

Finally, based on the review, Ngram, Langid, CANCODE, Transana, and ConcGram all produced the same result, 6%, with one frequency (f=1). They were identified in research by Tsuchiya and Handford (2014), Cheng and Lam (2013), Mohammad Arshi Saloot (2018), and Muhammad Imran Shah and Singh (2021). First and foremost, the ngram was a feature of the AntConc program. It was utilised to ascertain the frequency of each corpus lexical pattern. Lui and Baldwin (2012) established Langid as an open-source Python probabilistic language detection library. It was used to find the most frequently used language in tweet messages. In 2014, the same investigation was conducted using CANCODE and Transana. Audio-recorded conversation data was transcribed and time-stamped in the research using the annotation software tool Transana (Fassnacht & Woods, 2002). Meanwhile, the transcriptions used in the study were annotated using the annotation symbols provided in CANCODE (Adolphs, 2006).

According to the review, WordSmith (26%) and AntConc (22%) were highly popular among corpus linguists between 2013 and 2021. However, additional types of corpus linguistics software were chosen to meet the aims of the various studies.

4.2 Corpus Used

The second category of the taxonomy contains articles about the corpus used (n= 21/21). This section explores the many features of the corpus used in greater depth by classifying them into ten distinct categories.

Table 4. Corpus used (2013-2021)

Author(s)/Year	Corpus Used	Corpus Size
Awab & Norazit (2013)	New Straits Times & Straits Times	267,804 words; 274,345 words
Cheng & Lam (2013)	1996–1998 & 2006– 2008 Newspaper Source and ProQuest (260 regional newspapers & 25 national newspapers and magazines)	2,427 texts; 1,686,424 words
Tsuchiya & Handford (2014)	conversations	34,675 words

Huang (2015)	ESL weblogs	200 weblogs;
		female: 41, 585 words;
		male: 39, 222 words

Leung (2016)	The Strait Times	889 articles: 615,827 words
Pristian (2016)	eight Muslim fashion shops from Indonesia, Malaysia, U.K., and U.S.A. websites	2,991 word types & 7,826 word tokens.
Siti Aeisha Joharry (2016)	The Malaysian Corpus of Students' Argumentative Writing Version 2; Louvain Corpus of Native English Essays (LOCNESS)	1,460 Malaysian essays:565,500 words; LOCNESS: 324,304 words
Chan (2017)	Hong Kong: SCMP Chinese: The standard & China Daily	SCMP:695 texts, 306,434 words Chinese: 661 texts, 294,534 words
Ang & Tan (2018)	138 research articles in International Business Management	1 million words
Mohammad Arshi Saloot (2018)	1 million Tweet messages	14,484,384 words; 646,807 unique vocabulary
Nor Fariza Mohd Nor et al. (2019)	Malaysian Hansard Corpus	157 million words; 3,511 parliamentary proceedings

Author(s)/Year	Corpus Used	Corpus Size
Norsimah Mat Awal et al. (2019)	Malaysian Hansard Corpus (MHC) from Parliament 1 (P1) to Parliament 13 (P13)	3,511 files: 157 million words
Siti Aeisha Joharry & Hairani Wahab (2019)	Code of Practice for Institutional Audit	34,287 words
Ang & Tan (2019)	138 research articles in International Business Management	1 million words
Nor Fariza Mohd Nor & Adlyn Syahirah Zulcafli (2020)	The Star online reports about Covid-19 from March 1st to March 31st, 2020	1018 news reports
Chareonkul & Wijitsopon (2020)	British English 2006 & American English 2006	500 texts; 1,147,097 words; 500 texts; 1,175,965 words
He, Ang & Tan (2021)	160 research articles (2013-2018)	1.5 million words
Goyak et al. (2021)	5000 song lyrics in the 1960s until 2000s comprising of 25 songs for every year of each genre	1 million words
Nor Fariza Mohd Nor et al. (2021)	585 news articles (January 2020 to December 2021) from the New Straits Times and The Star Online	23,856 word types; 320,509 words
Liu & Muhammad Afzaal (2021)	First 15 chapters of Hongloumeng Hawkes & Yangs	91,173 Chinese; 89,396 words 67,649 words + 761 footnotes
Muhammad Imran Shah & Singh (2021)	Students' editorials	1 million words

Based on Table 4, a graph and a bar chart are shown below to illustrate the minimum and maximum words for analysis, as well as the frequency of the number of words involved in the corpus chosen, from the review of the period 2013 to 2021.

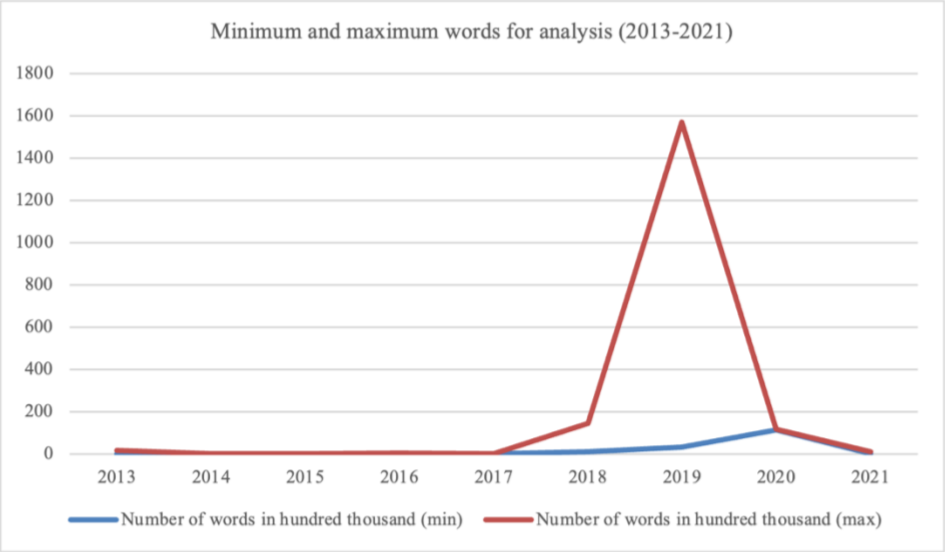

Fig 4 Minimum and maximum words for analysis (2013-2021)

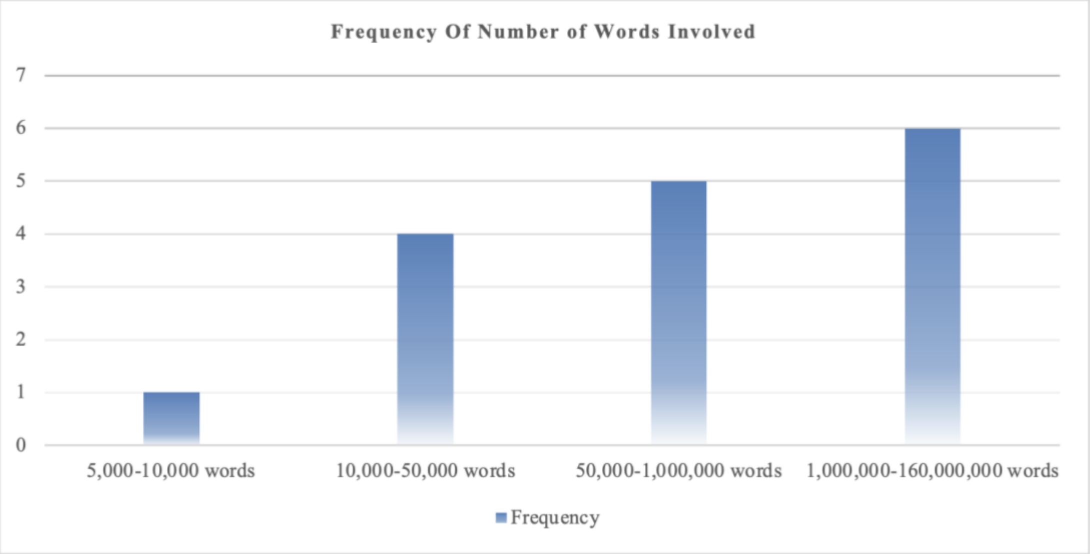

Fig 5 Frequency of number of words involved (2013-2021)

Table 4 illustrates the minimal number of words employed in the corpus analysis was 7,826 (Pristian, 2016). The researcher analysed the websites of eight Muslim fashion boutiques in different countries. As a result, a massive number of words may not have been included. Nonetheless, demonstrating high representation was sufficient because the websites might supply additional data for the researcher to analyse. Norsimah Mat Awal et al. (2019) undertook a corpus-driven analysis of the Malaysian Hansard Corpus (MHC) from Parliament 1 (P1) to Parliament 13 (P13) (P13). Numerous words were used, as this corpus is one of the primary political references and is amassed over many years. Each meeting or document was logged into the corpus. When the number of words studied ranges between 7,826 and 157 million, one could argue that a corpus-driven strategy is valid.

Following that, Figure 4 illustrates the minimum and maximum number of words required for analysis, as reported in the studies from 2013 to 2021. The maximum number rose between 2017 and 2020 (from 306,000 to 157 million words), indicating that corpus linguists and researchers began a pattern of utilising huge data sets containing a reasonably high number of words in a corpus. Simultaneously, the minimum number of words chosen for examination increased dramatically between 2017 and 2021 (from 295,000 to 1,1471,000 words). It was assumed that Asian countries began emphasising corpus-driven research by financing researchers to analyse massive amounts of data. Additionally, this represents a trend in which society is beginning to undertake large-scale corpus-driven analysis using multiple co-researchers to produce striking findings.

Besides, Figure 5 illustrates the frequency of the number of terms used from 2013 to 2021. It was evident that six researchers (Cheng & Lam, 2013; Mohammad Arshi Saloot, 2018; Nor Fariza Mohd Nor et al., 2019; Norsimah Mat Awal et al, 2019; Chareonkul & Wijitsopon, 2020; He et al., 2021) had analysed a corpus containing a large number of words, ranging from 1 million to 160 million. Merging the data from Figures 2 and 3 enables the deduction that beginning in 2021, a substantial number of corpus linguists and researchers began actively participating in corpus-driven research by analysing a large number of words from a corpus, perhaps because if a colossal corpus is used in a study, the results can be regarded as legitimate, dependable, and credible (Lange, 2019).

4.3 Research Purposes

The third category of the taxonomy contains articles discussing the purposes of research (n= 21/21). This section delves deeper into the various research purposes for which corpus-driven studies have been undertaken and identifies several recurring themes.

Table 5. Research purposes

Author(s)/Year	Research Purpose(s)	Recurring Theme(s)
Awab & Norazit (2013)	1. To compare the metaphors used in the two major English newspapers in Malaysia and Singapore.	Language patterns
	2. To show how the metaphors reflect current political ideology, the socio-cultural milieu, historical background, and even the physical and natural environment.	Language relationship
	3. To look at whether salient metaphors could have had any effect on mitigating the outcome of and/or shaping beliefs about attitudes toward or responses to the crisis.	Language impacts
Cheng & Lam (2013)	1. To compare the Western and Chinese media discourses related to Hong Kong's handover.	Language patterns
Tsuchiya & Handford (2014)	1. To examine turn-taking in a multiparty professional ELF meeting from a bridge-building project in South Asia.	Language assistance
Huang (2015)	1. To explore the content of Malaysian male and female English weblogs.	Language content
	2. To compare the male and female weblog sub-corpora.	Language patterns
	3. To identify key semantic domains and key parts of speech	Language patterns
Leung (2016)	1. To report on the discursive representation of gamblers in Singapore newspaper texts	Language patterns
Pristian (2016)	1. To identify the most frequently used words on the product descriptions.	Language patterns
	2. To investigate the differences and similarities using Multimodality theory	Language relationship
Siti Aeisha Joharry (2016)	1.To investigate the most salient linguistic items	Language patterns
	2. To check how are the items used similarly or differently	Language use
	3. To look for the most overused types of lexical bundles	Language patterns
	4. To explore how do these bundles function	Language impacts
Chan (2017)	1. To investigate for the most common words used to depict the Umbrella Movement event	Language patterns
	2. To check the extent of the discursive construction of the protest	Language patterns
Ang & Tan (2018)	1. To compare lists of phraseological sequences	Language patterns
	2. To identify a type of	Language patterns

Author(s)/Year	Research Purpose(s)	Recurring Theme(s)
	phraseological sequence	
Mohammad Arshi Saloot (2018)	1. To compile a corpus that represents the colloquial Malay language	Language patterns
	2. To analyse colloquial language and standard language corpora.	Language patterns
	3. To design a Malay Language Tweet normalisation architecture	Language design
	4.To evaluate the performance of the Malay language Tweet normalisation architecture	Language impacts
Nor Fariza Mohd Nor et al. (2019)	1. To report on a corpus- driven analysis around the word *ekonomi* (economy)	Language patterns
Norsimah Mat Awal et al. (2019)	1. To look at the issues that surround *air*/water in Malaysian Parliamentary debates	Language assistance
Siti Aeisha Joharry & Hairani Wahab (2019)	1. To investigate words that are used to describe the auditing process	Language patterns
Ang & Tan (2019)	1. To analyse the characteristics of a form of discontinuous phraseological sequence	Language patterns
Nor Fariza Mohd Nor & Adlyn Syahirah Zulcafli (2020)	1.To contribute to research in linguistics about Covid-19	Language patterns
Chareonkul & Wijitsopon (2020)	1. To investigate the use of present perfect in general corpora and samples presented in textbooks	Language use
He et al. (2021)	1. To compile lists of academic phraseological expressions	Language patterns
Goyak et al. (2021)	1. To identify the frequency distributions of lexical verbs, mental verbs.	Language patterns
	2. To analyze the language uses of mental verbs in the Diachronic Corpus of English Song Lyrics	Language use
Nor Fariza Mohd Nor et al. (2021)	1. To identify the adjective collocates for the node word ‘mental health’.	Language patterns
	2. To determine the semantic grouping of the collocates	Language patterns
	3. To investigate the portrayal of mental health issues in two newspapers	Language assistance
Liu & Muhammad Afzaal (2021)	1. To investigate the use of lexical bundles	Language patterns
Muhammad Imran Shah & Singh (2021)	1. To explore the effect of data-driven language learning (DDLL)	Language assistance

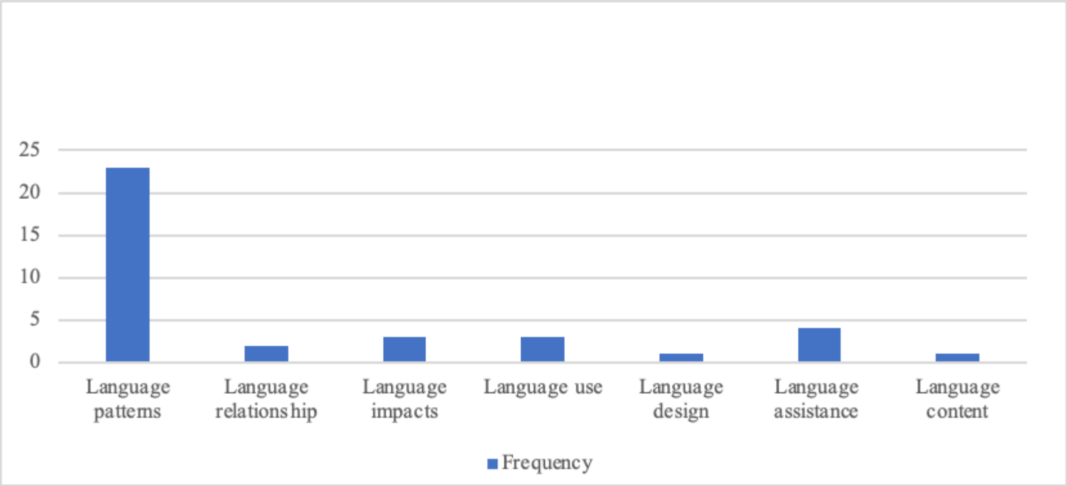

Fig. 6 Frequency of recurring themes

As illustrated in Table 5 and Figure 6, the theme of 'language patterns' emerged as the most frequently identified theme, based on the research purposes reviewed for the period 2013 to 2021. A total of 23 research objectives were associated with this theme. They were all connected with the study of language patterns such as metaphors (Awab & Norazit, 2013), media discourses (Cheng & Lam, 2013), semantic domains and key parts of speech (Huang, 2015), discursive representations of gamblers (Leung, 2016), linguistic items such as salient lexical bundles (Siti Aeisha Joharry, 2016), the most common words used and the discursive construction of a protest (Chan, 2017), types of phraseological sequences (Ang & Tan, 2018), the compilation of a corpus for language representation (Mohammad Arshi Saloot, 2018), word patterns associated with ‘economy’ (Nor Fariza Mohd Nor et al., 2019), words used for auditing (Siti Aeisha Joharry & Hairani Wahab, 2019), the characteristics of a form of discontinuous phraseological sequence (Ang & Tan, 2019), COVID-19 linguistics (Nor Fariza Mohd Nora & Adlyn Syahirah Zulcafli, 2020), a compilation of lists of academic phraseological expressions (He et al., 2021), the frequency distributions of lexical and mental verbs (Goyak et al., 2021), semantic groups of adjective collocations (Nor Fariza Mohd Nor et al., 2021), and lexical bundles used to trace stylistic features and differences (Liu & Muhammad Afzaal, 2021). As mentioned previously, a corpus-driven approach could be used to research a wide variety of various topics connected to language patterns. Between 2013 and 2021, a corpus-driven approach to language patterns appears to have been prevalent. Thus, one could argue that the primary purpose of a corpus-driven approach is to investigate or locate authentic and genuine language patterns across diverse corpora and disciplines.

Following that, the themes of 'language impacts' (*f*=3), 'language use' (*f*=3), and 'language assistance' (*f*=4) were discovered to have been used with a similar frequency. First, the term 'language impacts' was used to refer to language functions, as in the research conducted by Awab and Norazit (2013) to assess whether salient metaphors could have had any effect on mitigating the outcomes of and/or shaping beliefs about attitudes or responses to crises. Further examples were the research conducted by Siti Aeisha Joharry (2016) to explore how bundles function in Malaysian learners’ argumentative writing and that of Mohammad Arshi Saloot (2018) to evaluate the performance of the Malay language Tweet normalisation architecture. Thus, using a corpus-driven approach, language functions can be investigated in various circumstances. The following theme was 'language use,' which is concerned with analysing the application of language. Three researchers (Siti Aeisha Joharry, 2016; Chareonkul & Wijitsopon, 2020; Goyak et al., 2021) expressed such purposes by aiming, for instance, to check the use of linguistic items in two corpora, to investigate the use of present perfect in general corpora, and to analyse the use of mental verbs in the Diachronic corpus. This provided an additional function of the corpus-driven approach. Following that, the concept of 'language assistance' refers to the use of a corpus-driven approach as a sub-supporting instrument to accomplish a primary research objective that was not linguistic. This was utilised by Tsuchiya and Handford (2014), Norsimah Mat Awal et al. (2019), Nor Fariza Mohd Nor et al. (2021), and Muhammad Imran Shah and Singh (2021) for different research purposes, namely examining the effects of turn-taking in a multiparty professional ELF meeting, looking at issues concerning water in Malaysian Parliamentary debates, investigating the portrayal of mental health issues, and exploring the effects of data-driven language learning on Pakistani EFL undergraduates. Thus, rather than serving as the primary paradigm for language studies, a corpus-driven approach could be used as part of the methodology for various research purposes.

Finally, the themes of 'language relationship' (*f*=2), 'language content' (*f*=1), and 'language design' (*f*=1) were discovered to have been used with a relatively low frequency in the research purposes examined. In terms of ‘language relationship’, the purposes included showing how metaphors reflect the current political ideology, the socio-cultural milieu, the historical background, and even the physical and natural environment (Awab & Norazit, 2013); and investigating differences and similarities using multimodality theory in relation to products in Muslim fashion shops in Indonesia, Malaysia, the U.K., and the U.S. (Pristian, 2016). The relationship between a language and its many facets was researched and investigated to gain fresh insights, which would enrich the language with new information. In terms of ‘language content’, Huang (2015) explored the content of Malaysian weblogs written in English by both males and females. The findings were expected to reveal some of the most frequently discussed concerns of Malaysians. Meanwhile, the theme of ‘language design’ involved designing Malay language Tweet normalisation architecture to convert Malay language Tweet informal Malay into standard Malay (Mohammad Arshi Saloot, 2018). The discussion revealed that the corpus-driven approach could be applied to various applications other than the most well-known linguistic patterns. This expands the options and prospects open to corpus linguists and researchers in terms of their use of the corpus-driven approach in various future research purposes, both linguistic and non-linguistic in character.

4.4 Research Design

The fourth category of the taxonomy covers articles detailing the research design of each study (n= 21/21). This section reports in detail the various research designs used for corpus-driven studies.

Table 6. Research design

Author(s)/Year	Research Design
Awab & Norazit (2013); Cheng & Lam (2013); Tsuchiya & Handford (2014); Huang (2015); Leung (2016); Pristian (2016); Siti Aeisha Joharry (2016); Chan (2017); Ang & Tan (2018); Mohammad Arshi Saloot (2018); Nor Fariza Mohd Nor et al. (2019); Norsimah Mat Awal et al. (2019); Siti Aeisha Joharry & Hairani Wahab (2019); Nor Fariza Mohd Nor & Adlyn Syahirah Zulcafli (2020); Chareonkul & Wijitsopon (2020); Goyak et al. (2021); Nor Fariza Mohd Nor et al. (2021); Liu & Muhammad Afzaal (2021); Muhammad Imran Shah & Singh (2021)	Mixed-method (*f*=19)
Ang & Tan (2019); He et al. (2021)	Quantitative method (*f*=2)

As shown in Table 6, 19 researchers employed mixed-methods in their corpus-driven approach (Awab & Norazit, 2013; Cheng & Lam, 2013; Tsuchiya & Handford, 2014; Huang, 2015; Leung, 2016; Pristian, 2016; Siti Aeisha Joharry, 2016; Chan, 2017; Ang & Tan, 2018; Mohammad Arshi Saloot, 2018; Nor Fariza Mohd Nor et al., 2019; Norsimah Mat Awal et al, 2019; Siti Aeisha Joharry & Hairani Wahab, 2019; Nor Fariza Mohd Nora & Adlyn Syahirah Zulcafli, 2020; Chareonkul & Wijitsopon, 2020; Goyak et al., 2021; Nor Fariza Mohd Nor et al., 2021; Liu & Muhammad Afzaal, 2021; Muhammad Imran Shah & Singh, 2021). Instead of quantifying language using computational linguistic methods, they incorporated qualitative data into their conclusions to achieve more reliable, credible, and legitimate results via triangulation.

However, Ang and Tan (2019) and He et al. (2021) chose the quantitative method as the primary research design due to their distinct research objectives of, respectively, analysing the characteristics of a type of discontinuous phraseological sequence and compiling lists of academic phraseological expressions by deriving a pedagogically useful list of phrase frames for a specific discipline. These two research purposes could be accomplished entirely through quantitative methods. In summary, the review suggests that between 2013 and 2021, the mixed-methods approach was preferred by the majority of researchers undertaking corpus-driven investigations. Thus, corpus linguists and researchers are encouraged to use a mixed-methods approach as the primary research strategy in future corpus-driven investigations.

5. Conclusion

Since its inception, research on corpus-driven approaches has increased. However, the research conducted has several shortcomings that must be considered and addressed. Corpus-driven approaches, whether in educational settings or across disciplines, have gained increasing global attention and acceptability. This article contributes to the body of work on corpus-driven approaches that have been classified. Numerous intriguing categorical patterns appeared through the review of the corpus-driven research. The articles analysed were classified according to the corpus linguistic software used, the corpus used, the research goal, and the research design. The majority of difficulties cited are associated with the deployment of corpus-driven approaches for various research purposes, the corpus size, the use of corpus-driven approaches in a respective course of study, as well as other similar issues. This study is groundbreaking and will serve as a significant reference for future scholars.

6. Suggestions

The present review article has been conducted utilising a collection of articles spanning the years 2013 to 2021. It is strongly recommended that one engages in a comprehensive examination of corpus studies that have been published within the timeframe spanning from 2021 to 2023. Moreover, it is worth considering the possibility of relocating the setting from Asia to alternative geographical locations in order to gain a more comprehensive and enlightened perspective. The potential area of emphasis lies within the realm of corpus-based pedagogy.

7. Co-Author Contribution

The authors have explicitly stated that no conflict of interest exists within the context of this article. Chen and Yin in their scholarly capacity, undertook the task of crafting the introductory section and diligently curated a selection of articles to be subjected to critical examination and evaluation. The systematic literature review was conducted by Wong. Charanjit and Warid conducted a comprehensive analysis and subsequently presented their findings. Yoong, in her scholarly capacity, successfully concluded the study and proceeded to refine its findings.

8. Acknowledgements

This article is a preliminary result of the following research projects:
1) "Empirical Study on the Improvement Path of In-Service EFL Teachers' Content Knowledge to Instruct English Early Reading in Light of National Curriculum" [Grant Number: 2022SC006] from the Chinese Society for Tao Xingzhi Studies;
2) "Improving EFL students' Language Abilities Via Phonics Instruction" [Grant Number: QJZD20220628] from China Basic Foreign Language Education Research & Training Center;
3) "A Study on the Impacts of Cross-Regional Instructional researches upon Teachers' Professional Development in Nationality Areas" [Grant No. PDTR2017-1[from Sichuan Primary and Secondary School Teacher Development Research Center.

9. References

Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies*. Routledge.

Alsahlanee, A., & Jaganathan, P. (2022). Lexical complexity in the writings of Iraqi, English L2, and English L1 writers. *Asian Journal of University Education*, *18*(4), 1105-1118.

Ang, L. H. & Tan, K. H. (2018). Specificity in English for Academic Purposes (EAP): A Corpus analysis of lexical bundles in academic writing. *3L: The Southeast Asian Journal of English Language Studies, 24*(2), 82 – 94.

Ang, L. H. & Tan, K. H. (2019). From lexical bundles to lexical frames: Uncovering the extent of phraseological variation in academic writing. *3L: The Southeast Asian Journal of English Language Studies, 25*(2), 99 – 112.

Awab, S. & Norazit, L. (2013). 'Challenging' Times or 'Turbulent' Times: A study of the choice of metaphors used to refer to the 2008 economic crisis in Malaysia and Singapore. *Intercultural Pragmatics, 10*(2), 209-233.

Barlow, M. (2004). *Collocate 1.0 Software*.

Bibri, S. (2019). *Big data science and analytics for smart sustainable urbanism*. Springer.

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

Chan, T. (2017). The umbrella movement in the media: A Corpus-driven analysis of newspapers in Hong Kong and China. *Journalism and Discourse Studies*, 1-26.

Chareonkul, Ch. & Wijitsopon, R. (2020). The English present perfect in authentic use and textbooks: A corpus-driven study. *Journal of Language Teaching and Learning in Thailand, 60*, 275-308.

Cheng, W. & Lam, P. (2013). Western perceptions of Hong Kong ten years On: A corpus-driven critical discourse study. *Applied Linguistics, 34*, 173-190.

Chung, C. J., Lai, C. L., & Hwang, G. J. (2019). Roles and research trends of flipped classrooms in nursing education: A review of academic publications from 2010 to 2017. *Interactive Learning Environments*, 1–22.

Cohen, L. (2017). *Research methods in education*. Routledge.

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.

Dawson, C. (2019). A-*Z of digital research methods*. Routledge.

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

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2019). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning*, 1-24.

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,

Fassnacht, C. & Woods, D. (2002). *Transana, Version 2.12-Win*. University of Wisconsin-Medison.

Fetzer, D. (2018). *Development of a MATLAB/simulink framework for phasor-based power system simulation and component modeling based on state machines*. Kassel University Press GmbH.

Fletcher, W.H. (2002). *KfNgram software*. USNA.

Goyak, F., Mazura Mastura Muhammad, Farah Natchiar Mohd Khaja, Muhamad Fadzllah Zaini & Ghada Mohammad. (2021). Conversational mental verbs in English song lyrics. *Asian Journal of University Education, 17*(1), 222-239.

Friginal, E. (2018). *Corpus linguistics for English teachers*. Routledge.

Ge, Z.-G. (2018). The impact of a Forfeit-Or-Prize gamified teaching on e-learners' learning performance. *Computers & Education, 126*, 143-152.

Ghavifekr, S., & Wong, S. Y. (2022). Technology leadership in Malaysian schools: The Way forward to education 4.0 – ICT utilization and digital transformation. International *Journal of Asian Business and Information Management (IJABIM), 13*(2), 1-18.

Hannis, M., Mkpong-Ruffin, I. & Hamilton, D. (2022) Student educational learning experience through cooperative research. *Lecture Notes in Networks and Systems, 310,* 22-27.

Hatice, K., Nurseven, K., Gamze, G. Y. & Abdulkadir, G. (2018). Students' reflections on vocabulary learning through synchronous and asynchronous games and activities. *Turkish Online Journal of Distance Education, 19*(3), 53-70.

He, M. Y., 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

Hempel, S., Gore, K., & Belsher, B. (2019). Identifying research gaps and prioritizing psychological health evidence synthesis needs. *Medical care, 57 Suppl 10 Suppl 3* (10 Suppl 3), S259–S264.

Huang, W. J. (2015). *A corpus-driven analysis of male and female weblog users in an ESL context*. [Doctoral dissertation, University of Malaya, Malaysia]. University of Malaya Student Repository.

Hussein, K. (2017). *A corpus-driven approach to stylistic analysis of a lexical richness curve*. GRIN Verlag.

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.

Hilton, A. (2020). *Learning to research and researching to learn*. Cambridge University Press.

Hung, H. T. (2018). Gamifying the flipped classroom using game-based learning materials. *English Language Teaching Journal, 72*(3), 296-308.

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

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.

Lange, C. (2019). *Corpus linguistics for world Englishes*. Routledge.

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

Leung, R. (2016). Representation of gamblers in the Singaporean press since casino legalization: A corpus-driven critical analysis. *International Journal of Applied Linguistics and English Literature, 5*, 51-63.

Liu, K. & Muhammad Afzaal. (2021). Translator's style through lexical bundles: A corpus-driven analysis of two English translations of Hongloumeng. *Frontier of Psychology, 12*, 633422.

Lytras, M. (2020). *Big data research for social sciences and social impact*. MDPI.

Lui, M. & Baldwin, T. (2012). Langid.PY: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 25-30). Association for Computational Linguistics.

Meakins, F. (2018). *Understanding linguistic fieldwork*. Routledge.

Mohammad Arshi Saloot. (2018). *Corpus-driven Malay language tweet normalization*. [Doctoral Dissertation, University of Malaya, Malaysia]. University of Malaya Student Repository.

Muhammad Imran Shah & Singh. (2021). Effect of Data Driven Language Learning (DDLL) on EFL learners: A corpus-driven language learning approach. *Journal of Linguistics and English Language Teaching, 1*(1), 35-55.

Ngulube, P. (2019). *Handbook of research on connecting research methods for information science research*. IGI Global.

Nor Fariza Mohd Nor, Anis Anis Nadiah Che Abdul Rahman, Azhar Jaluddin, Imran Ho Abdullah & Tiun, S. (2019). A corpus driven analysis of representations around the word 'Ekonomi' in Malaysian Hansard Corpus. *GEMA Online® Journal of Language Studies, 19* (4), 66-95.

Nor Fariza Mohd Nor & Adlyn Syahirah Zulcafli. (2020). Corpus driven analysis of news reports about covid-19 in a Malaysian online newspaper. *GEMA Online® Journal of Language Studies, 20*(3), 199-220.

Nor Fariza Mohd Nor, Jeffree & Hilwa Abdullah@Mohd Nor. (2021). Health is wealth: A corpus-driven analysis of the portrayal of mental health in Malaysian English online newspapers. *GEMA Online® Journal of Language Studies, 21*(2), 46-71.

Norsimah Mat Awal, Azhar Jaludin, Anis Nadiah Che Abdul Rahman & Imran Ho Abdullah. (2019). "Is Selangor in deep water?": A corpus-driven account of air/water in the Malaysian Hansard Corpus (MHC). *GEMA Online® Journal of Language Studies, 19*(2), 99-120.

Oakes, M.P. (1998). *Statistics for corpus linguistics*. Edinburgh University Press

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher. D. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(n71). https://doi.org/10.1136/bmj.n71

Pristian, V. M. (2016). *The portrayal of online Muslim fashion shops in Indonesia, Malaysia, UK and USA: Corpus linguistics and multimodality approaches*. [Master Thesis, Universitas Airlangga, Indonesia]. Universitas Airlangga Student Repository.

Ramos, F. P., & Guzmán, D. (2021). Examining institutional translation through a legal lens: A comparative analysis of multilingual text production at international organizations. *Target, 33*(2), 254-281.

Rayson, P. (2001). *Wmatrix: A web-based corpus pro-cessing environment*. Computing Department, Lancaster University.

Romero-Hall, E. (2020). *Research methods in learning design and technology*. Routledge.

Roslim, N., Azizul, A. F., Nimehchisalem, V., & Tew Abdullah, M. H. (2021). Exploring movies for language teaching and learning at the tertiary level. *Asian Journal of University Education*, *17*(3), 271.

Satrovic, E. (2019). *Emerging trends in trade in Turkey*. GRIN Verlag.

Scott, M. (2012). *WordSmith Tools* (Version 6.0). Lexical Analysis Software.

Sengupta, S. (2022). Possibilities and challenges of online education in India during the Covid-19 pandemic. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 17*(4), 1-11.

Shneikat, B. (2021). *Global perspectives on recruiting international students*. Emerald Group Publishing.

Singh, C.K.S., Singh, T.S.M., Ja'afar, H., Tan, W. H., Subramaniam, G.J. & Shu, M.H.B.A. (2019). Developing a prototype speaking game for engineering students at polytechnic in Malaysia. *Journal of Engineering Science and Technology*, 9-17.

Siti Aeisha Joharry. (2016). *Malaysian learners' argumentative writing in English: A contrastive, corpus-driven study*. [Doctoral Dissertation, University of Sydney, Australia]. University of Sydney Student Repository.

Siti Aeisha Joharry & Hairani Wahab. (2019). Auditing awareness: A corpus-driven approach. *e-Academic Journal*, *8*(1), 1-16.

Soderland, Stephen, Lehnert & Wendy. (1994). Corpus-driven knowledge acquisition for discourse analysis. *Proceedings of the National Conference on Artificial Intelligence*, *1*, 827-832.

Tan, L. (2018). Meaningful gamification and students' motivation: A strategy for scaffolding reading material. *Online Learning*, 22.

Tichý, O. (2021). Corpus driven identification of lexical bundle obsolescence in Late Modern English. *Studies in Language Companion Series, 218*, 101-129.

Tsuchiya, K., & Handford, M. (2014). A corpus-driven analysis of repair in a professional ELF meeting: Not 'Letting It Pass'. *Journal of Pragmatics*, *64*, 117-131.

van der Meij, M. G., Broerse, J. E. W., & Kupper, F. (2017). Supporting citizens in reflection on synthetic biology by means of video-narratives. *Science Communication, 39*(6), 713–744.

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

Zainuddin, Z., Kai, S.W.C., Shujahata, M. & Perera, C.J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, *30*, 100326.

Zhang, X. & Yang, H. (2021). Gender voices in Chinese university students' English writing: A corpus study. *Linguistics and Education, 64*, 100935.

Zufferey, S. (2020). *Introduction to corpus linguistics*. John Wiley & Sons.