SPECIALISED LEARNER CORPUS RESEARCH: A REVIEW FOR FUTURE DIRECTIONS OF THE GLOBAL AND MALAYSIAN CONTEXTS

Radika Subramaniam[a], Sheena Kaur[b]
([a]radika5352@yahoo.co.uk; [b]sheena@um.edu.my)

Universiti Malaya, Malaysia
Jalan Universiti, 50603, Wilayah Persekutuan Kuala Lumpur, Malaysia

Abstract: Literature survey concerning the Learner Corpus Research (LCR) in the last 20 years has shown a paucity of studies involving specialised/discipline-specific text. Since the use of discipline-specific academic writing learner corpora is useful in determining the language pattern within the English for Specific Academic Purposes (ESAP) context, this paper presents a review of specialised LCR based on journal articles from the Web of Science database and the Google scholar, reference books and relevant websites to address the gap both globally and within the Malaysian context for future needs. The review suggests the need for more specialised/genre-specific or discipline-based learner corpus studies employing the non-native novice versus native novice and expert comparative method in both global and Malaysian contexts, with more studies executed in relation to specialised learner corpora in the latter. In relation to linguistic aspect, the understanding of lexical bundles' communicative function can be attempted through the analysis from the core word to the surrounding pattern or discourse. The use of learner corpus as a pedagogical instrument needs also to be well-accepted and executed to enhance the teaching and learning of academic writing in ESAP classrooms via Data-Driven Learning approach or corpus-related activities.

Keywords: learner corpus research, contrastive interlanguage analysis, native versus non-native, novice versus expert, ESAP

DOI: http://dx.doi.org/10.15639/teflinjournal.v34i1/176-193

The Learner Corpus Research (LCR) era began with the establishment of the first learner corpus in the 1980s- the International Corpus of Learner English (ICLE) (Flowerdew, 2015a; Gilquin, 2021; Granger et al., 2020). The ICLE was initiated as part of the International Corpus of English (ICE) project that contains corpus data (of both spoken and written) of a regional variety of English, that is the English language which serves as the "majority first language" or "official additional language" of a particular country (Greenbaum & Nelson, 1996, p. 3). The ICLE, which is coordinated by Sylviane Granger, based at the University of Louvain, is a joint effort from various universities around the world as resource centres for supplying international learner corpora data. The inclusion of learner writing has gradually developed in the ICLE. The essays (i.e., argumentative essays and literature examination papers) collected from upper intermediate and advanced-level learners from at least 25 different L1 backgrounds have contributed to this 5.5-million-word corpus (Granger et al., 2020).

In researching the native/non-native variety of essay-oriented genres, many studies can be identified utilising learner data from the International Corpus of Learner English (ICLE) (e.g., Gilquin & Granger, 2011; Parkinson, 2015). Other than that, the Louvain Corpus of Native English Essays (LOCNESS) is also among the prevalent ones in serving as the reference or comparative corpus to aid second language writing research (e.g., Nam & Park, 2020; Sung, 2020). It comprises 324,304 words which are compiled from native British and American students' argumentative essays from pre-university to university level (Granger, 1998). The LONGDALE, on the other hand, is a Longitudinal Database of Learner English which contains both spoken (i.e., informal interviews) and written (i.e., argumentative and narrative essays) data collected from learners of different native language backgrounds (Meunier, 2016). As for studies exploring the use of English as a lingua franca in Asian countries (e.g., Chen, 2019; Lin & Lin, 2019), ICNALE, the International Corpus Network of Asian Learners plays a major role being the data resource with more than 3.5 million words. It is formed by speech and essay collections written by college to graduate students from nine Asian countries namely China, Indonesia, Japan, Korea, Pakistan, Philippines, Singapore / Malaysia, Taiwan and Thailand, and a region (i.e., Hong Kong) (Ishikawa, 2013).

Despite the increase in the number of learner corpora in aiding the English for General Academic Purposes studies with learner text collections that are more of argumentative type (Flowerdew, 2015a), research in Language for Specific Purposes (LSP) area necessitates the development of more specific discipline-oriented learner corpora to aid the English for Specific Academic Purposes (ESAP) studies. Discipline-specific corpus which can also be addressed as specialised corpus serves as a purposive collection of texts based on a topic or subject area (Sinclair, 2005) or genre (e.g., fiction, conversation, academic prose) (Biber, 1988) to reveal specific findings in relation to the target subject or genre of the language under study. Unlike general corpora which are most often used as reference corpora to study the "lexico-grammatical or discursive features" of language, specialised corpora are designed to study the language use within "certain discourse community" (Chang, 2011, p. 11) to see the function the language serves within the linguistic domain. Creating specialised corpora for analysing language in a specific domain is more useful than larger general corpora which may not be able to address the specific language needs (Nesi, 2016). In such cases, specialised genre-based or disciplinary academic writing studies very frequently employ the learner corpus data from the British Academic Written English Corpus (BAWE) (e.g., Durrant, 2017; Nesi, 2021) and the Michigan Corpus of Upper-level Student Papers (MICUSP) (e.g., Lei & Yang, 2020; Wang & Zhang, 2021). These two learner corpora (i.e., BAWE and MICUSP) are compiled based on disciplinary groups (e.g., Linguistics, Mechanical Engineering, Computer Science, Cybernetics and Electronic Engineering) and genres (e.g., research report, proposal, argumentative essay and response paper) of proficient student papers from various mother tongue backgrounds.

Corresponding to the need of growing LSP research, the VESPA (Varieties of English for Specific Purposes dAtabase) is another initiative from the Centre for English Corpus Linguistics (CECL) from the UCLouvain which provides collections of L2 university undergraduate disciplinary writings from various L1 backgrounds. The corpus texts (i.e., multi-million words) gathered from several countries in Europe (e.g., Belgium, Sweden and Norway) include various disciplines (e.g., linguistics, business, engineering, experimental sciences and environmental

sciences), genres (e.g., research papers, project reports, MA dissertations) and writer expertise (i.e., from first-year students to doctoral students) (Granger & Paquot, 2013). The VESPA which was initiated in 2008 is still under its development; however, few studies have employed the sub-set of the corpus as part of the study data (e.g., Larsson, 2017, 2019). Similarly, the Corpus of Academic Learner English (CALE) (Callies & Zaytseva, 2013) is another large-scale learner corpus that contains writing from learners of various native language backgrounds based on different genres and disciplines. The CALE project which commenced in 2014 at the University of Bremen, Germany is an effort to build a specialised corpus to assist ESAP studies (Kizil, 2020).

The development and availability of various general and specific learner corpora for both commercial and non-commercial uses (depending on the method of access based on individual corpus) depict the growing interest in LCR studies. Literature survey concerning the LCR in the last 20 years has resulted in the identification of a review of similar nature, that has addressed learner corpus studies as a whole with no specific reference, particularly to discipline or genre-specific text (e.g., Yang, 2023). Yang (2023), in a review of research articles from renowned linguistics journals from 2007 to 2021, among others, has found that learner corpus studies (regardless of learner's text type/genre whether general or genre/discipline-specific writing) are more dominant in the UK and the US compared to any other countries (i.e., China, Japan, Belgium, Spain, German). As for the focus (i.e., content) of the learner corpus studies, Yang (2023) identifies that there has been less research on grammatical items and discourse units. Future studies are recommended to compare phraseological units of L1 and L2 varieties of English (Yang, 2023).

Since the use of specialised or discipline-specific academic writing learner corpora is useful in determining the language pattern within the ESAP context, the current paper presents a review of specialised learner corpus studies, especially in terms of the method employed, which was not addressed in Yang's (2023) review, and linguistic features investigated, in contributing to the field of ESAP. The focus of the review entails LCR in the global and the local contexts (this refers to studies conducted by Malaysian researchers predominantly, using Malaysian learner data). As Malaysia has not been recorded in the list of top 10 countries in publishing research on learner corpora (especially the English language) (see Yang, 2023), this paper intends to establish the gap which could be addressed to shape the future directions of specialised LCR, both in the global and Malaysian contexts.

The literature survey commenced with the search for relevant resources (i.e., journal articles, reference books, and corpora websites) in relation to Learner Corpus Research. The journal articles selection was made through the Web of Science database. For other reference materials, Google Scholar (e.g., Joharry & Rahim, 2014) was employed. There was no restriction set to the year of publication as the researcher aimed at exploring the diachronic development of the learner corpus research since its inception with the first learner corpus, ICLE. The search was executed by inserting the keyword via the Web of Science database search. The most important keyword which has assisted the search and helped the discovery of learner corpus studies was *learner corpus*. As per the initial search, there were 778 research articles discovered involving LCR; however, the list was then filtered to learner corpus studies in relation to academic writing which later resulted in 207 articles in the last 20 years (as recorded in the

database). Studies involving specialised learner corpus were then shortlisted. The specialised corpora include genre-based or discipline-related written texts.

The initial screening stage included scrutinising the method section which provided the details of the corpora used along with the research design. The data analysis section was also identified to check the analysis procedures implemented in order to identify the target linguistic feature(s) under investigation. Since the preliminary observation has shown that a vast number of studies (from the past to present) have employed the Contrastive Interlanguage Analysis (CIA) as the most relevant approach in designing learner corpus studies, this review delimits itself to discuss the notion of CIA and debatable issues pertaining to it, especially studies comparing the non-native and native learner and/or expert writing exploring linguistic features ranging from word to discourse level, and the emergence of learner corpus as pedagogic corpus. In addition, learner corpus studies in the Malaysian context were also identified via Google Scholar and the Malaysian Corpus Research Network (MCRN) website to fill in the gap not only from the global perspective but also from the local context.

This paper is organised into three main sections. The first section of the review discusses the principles/features of the CIA approach in learner corpus studies which include the issues revolving around corpus comparability and possible recommendations/solutions as presented in previous scholarly works. The following section reviews specialised learner corpus studies using the CIA approach from vocabulary to grammatical and discourse levels from the global context and the role of learner corpus in pedagogy. The last section concludes with a review of LCR in the Malaysian context and the gaps identified as a recommendation for future studies.

CONTRASTIVE INTERLANGUAGE ANALYSIS (CIA) AND THE ISSUE OF CORPUS COMPARABILITY

The key principle of learner corpus research is to study language patterns through variation. As opined by Johansson (1991, p. 306), "variety is probably the area where corpus workers can make the most significant contribution". The variation is explored between the native and non-native variety of the same language by adopting the Contrastive Interlanguage Analysis (CIA) approach which according to Gilquin (2001, p. 95), explicates the "nativeness and non-nativeness of learner language by comparing it with native language". Hence, the CIA is perceived as relevant in revealing the English language features produced by non-native speaker learners to see not only the errors but also its overused and underused patterns compared to another variety of the same language (Granger, 1996). Granger (1996, p. 44) brings to light two types of comparison involved in the CIA approach (i.e., native language versus interlanguage and interlanguage versus interlanguage).

Elaborating the two approaches more vividly, the first type is a comparative method which studies the variation between the native learner language (L1 English) and non-native learner language (L2 English) whereas the second type of comparison includes the variation between the interlanguages, for instance, L2 English learners of French and L2 English learners of German. The first type of CIA requires a comparative native learner corpus which serves as a control or reference corpus to reflect the "overrepresentation" or "underrepresentation" of a linguistic phenomenon in the target non-native variety (Granger, 1996, p. 45). The second type,

according to Granger (2009, p. 18), is relevant to "assess the degree of generalisability of interlanguage features across learner populations and language situations". The method of analysis of this type includes a comparison between sub-corpora which are compiled based on several variables, for instance, the learners' age, native language background, level of proficiency and task type, learning setting and medium (Granger, 1998, p. 13) to study the association of these variables with the interlanguage productions.

Although CIA becomes the most prominent comparative measure in researching the learner corpora (Crosthwaite et al., 2017), there is always a need to objectify the type of comparison and the reason for such comparison to be executed, that is, whether using the native corpus as a "tool" for analysing non-native language or as a "support" to model the native-like proficiency (Seidlhofer, 2001, p. 144). As raised by Cobb and Horst (2015), one of the main challenges in CIA is corpus comparability (i.e., to identify the most compatible comparative control or reference corpus) to the interlanguage variety. The global acceptance of English as a lingua franca; however, places its competent users' texts (i.e., highly proficient users who are not necessarily native speaker learners) as part of a control or reference corpora indicating the standard variety (Cobb & Horst, 2015).

The issue of comparability has always been the central argument for many scholars who have different views on the virtue of comparison. Leech (1998) sees the comparison between learner texts to that of the native speaker learner texts as unreliable as the learner corpus is less likely a good model for its recognition as a reference corpus for the other group of learners to emulate. Diverting the model of comparison from native learner corpus to professional or expert writing, however, it has been confronted that this could be completely unrealistic as the "knowledge" and "understandings of academic conventions" of the novice and expert groups may differ (Hyland, 2012, p. 139). Alternatively, the method of "parallel corpora" is perceived as helpful to reveal "what different groups of language users actually do," processing from their own mental schemata (Hyland, 2012, p. 139).

The nature of a comparative model (L2 novice versus L1 novice or expert) ultimately depends on the objective of the study or the research question(s) in which the study intends to answer. According to Adel (2006, p. 207-208), both norms of comparison (i.e., "peer status" of native novice and "professional status") with the non-native learner text would render "additional perspective on the similarities and differences and their varying conditions of use". This suggests that a parallel native learner corpus is used to study the idiosyncratic pattern of the non-native learners when the aim of the comparison is "evaluative" and a more reliable (i.e., professional or expert) text is employed to achieve a "pedagogical" target to enable the learners to model the learning process (Gilquin, 2021, p. 5).

To reiterate, both the native novice texts and professional or expert texts serve as useful reference corpora in learner corpus research using the CIA approach provided that any comparative corpus is able to display the feature of "expert performances" (Bazerman, 1994, p. 131). Moreover, these corpora, if similar in their genre or text type, are comparable to reveal useful findings as contended by Adel (2006, p. 207), the comparison of "both native speaker student-essays and professional texts would give a broader picture of what the status of the learner essays is in relation to native-speaker texts".

SPECIALISED LEARNER CORPUS STUDIES IN THE GLOBAL CONTEXT

Learner corpus studies can be found ranging from lexis to discourse level (Granger et al., 2015). LCR using specialised or discipline-specific corpora have not been reported widely, however, there are some relevant studies that mark the use of learner corpora to study lexis, phraseology and pragmatics-discourse of the non-native variety in comparison to the native and/or expert variety, predominantly via the CIA approach (e.g., Chen & Baker, 2010; Lee & Chen, 2009; Leedham & Fernandez-Parra, 2017; Lei & Yang, 2020; Liardet & Black, 2019; Subramaniam & Kaur, 2021; Wang & Zhang, 2021). Hence, in this review section, only six main articles which have employed the CIA method to analyse linguistic features involving lexical units/ function words, phraseological units (i.e., lexical bundles), and discourse-pragmatic level are discussed.

Non-Native versus Native Novice and/or Expert CIA Studies from Lexis to Discourse Level

In one of the earlier studies, Lee and Chen (2009) employ the CIA approach in the identification of the nature of Chinese undergraduate writings (i.e., dissertations compiled as CAWE, the Chinese Academic Written English Corpus) from the English language and applied linguistics category against two comparative native speaker corpora (i.e., novice and expert writing) from the same disciplinary group. The native speaker learner data are extracted from the existing BAWE corpus, whereas the expert corpus is developed based on research articles downloaded from high-ranking linguistics and applied linguistics journals. The authors of these articles are recognised as 'experts', hence the native speaker status is excluded from the selection criteria. Keywords analysis is performed to locate the overused and underused items from the CAWE corpus. The findings reveal several keywords and n-grams which have been overused in the Chinese learner corpus compared to the other two comparative corpora. These words are further analysed qualitatively to see how they appear (the lexico-grammatical features) in different groups of writing. All three corpora contain different numbers of words (i.e., CAWE - 407,960 words, BAWE -L – 177,153 words and Expert Journal Articles, EXJA – 388,490 words). As the BAWE-L consists of a far lower number of words, the EXJA serves as the main comparative corpus, however, the novice native sub-corpus (BAWE) is used as a more "realistic reference yardstick" between the non-native learner and expert texts (Lee & Chen, 2009, p. 284).

Adopting a similar CIA approach to Lee and Chen (2009), more recently, Lei and Yang (2020) use the corpus compiled from Chinese doctoral students' research manuscripts to compare with the native speakers' undergraduate and postgraduate level research papers extracted from the MICUSP and published research articles corpus written by English native speakers (selected based on the names and affiliations of the first authors), to study lexical richness. Both the Chinese doctoral students' corpus and MICSUP (henceforth, native beginners) include texts from several science and engineering disciplines (e.g., biology, civil and environment engineering, industrial and operations engineering, mechanical engineering, natural resource environment, and physics), whereas the published research articles (i.e., native expert corpus) contains journals recommended by the science and engineering discipline experts. The analyses which indicate the lexical diversity, sophistication and density of Chinese

learner writing compared to the other two groups show that expertise plays a greater role in influencing lexical richness than nativeness in research paper writing.

Apart from analysing the content words, the CIA approach can also be seen applied in researching the function words. In a study comparing two non-native varieties with the native learner variety, Leedham and Fernandez-Parra (2017) extract the engineering writings produced by Chinese, Greek and English undergraduate to master-level learners from the BAWE corpus, to investigate the use of first-person pronouns (i.e., *we* and *I*). The functional classifications (i.e., guide, opinion, reflector, recounter and representative) of the pronouns explicate the association between the learner's cultural background and identity formation. This study sheds light on the EAP teaching and learning practices especially for instructors and engineering learners not to neglect the knowledge of these pronouns in their academic writing.

Research on lexical bundles, the recurring sequence of words, has also been identified as another point of interest among scholars in the last 20 years. In researching this type of phraseological unit, Chen and Baker's (2010) study has been one of the most commonly cited. This study has compared the novice (native and non-native) and expert writing via three sub-corpora (i.e., BAWE-English, BAWE-Chinese and Freiburg-Lancaster-Oslo/Bergen sub-corpus of academic prose). The n-gram list is generated, and all the 4-word bundles are identified across the corpora by eliminating overlapping occurrences. The analyses show that the verb phrase (VP) bundles appear more frequently in novice writing in comparison to native experts' writing. The L2 learners (non-native novice writers) display substantial use of VP-based bundles, however with less "passive + PP-fragment" as compared to the native peer and native expert. This study employs the CIA approach by making a comparison between novice writings (i.e., L1 versus L2), as well as novice versus expert. Studies on lexical bundles, despite undertaking the investigation into the general or specialised/discipline-specific corpora, have largely focused on the analysis of structural and functional categories (e.g., Adel & Erman, 2012; Bychkovska & Lee, 2017; Hyland, 2008) with no studies illustrating in-depth qualitative investigation on how the functional categories of lexical bundles are realised or what (e.g., structural form, lexico-grammatical pattern, semantic choice) motivates these communicative functions.

In another study comparing native and non-native novice writing with that of native expert writing, Liardet and Black (2019) use a total of 190 assignments (from the Macquarie University Longitudinal Learner Corpus – MQLLC) and 100 research articles to study the use of reporting verbs (RVs). For the learner corpus, the assignments (i.e., business reports, integrative summaries and persuasive essays) are written by first-year undergraduate students (from various L1 backgrounds, including English native speakers, majoring in Business and Economics, and Arts and Humanities) for the Academic Communication division which offers academic literacy units to students across the university. There are 65 texts written by non-native speaker learners whereas the L1 English texts form the remaining 125 texts. For the expert corpus, the research articles are downloaded from the top 20 high-ranked journals (i.e., 5 articles are selected from each journal) identified in SciMago 2015. The findings reveal variations in the choice of RVs between the two groups of learners. Apart from the frequency-based analyses, the qualitative analysis reveals the development of intertextuality within the text through the choice of RVs used in the text to communicate stance and engagement. This study is one of the examples which

elucidates the association between word-level analysis to the discourse-pragmatic realisation of a text.

Apprehending the concept of variation and the benefits it could bring to the realm of pedagogy, Wang and Zhang (2021) in a recent study, have explored the effect of language background and expertise on textual priming and semantic association of the multiword unit *according to* across two novice (non-native and native), and two expert (native) corpora. The learner corpora involved in this study are English language research papers written by the L2 English Chinese doctoral students and L1 English learners (American writing of various disciplinary backgrounds, extracted from the MICUSP). The expert corpora include research articles written by native English experts (identified from the first authors' English names) and Hyland Corpus which has been claimed as L1 English expert writing despite the content covering research articles written by experts regardless of nativity, as it is believed that the authors had had "sufficient academic training to publish in peer-reviewed international journals" (Wang & Zhang, 2021, p.51). The findings illustrate that variation in disciplinary groups does affect textual priming, and more than the writing expertise, the language background imposes a greater effect on the use of the MWS.

The Need for Future Directions

Generally, learner corpus studies depicting genre/discipline-specific writing, have garnered less attention, particularly using the CIA approach which takes the double comparative corpora (i.e., comparing non-native learner writing with that of the native learner and expert writing, at the same time). One of the studies which has executed his type of comparative method is Chen and Baker (2010), however, there is no specific disciplinary group highlighted as the focus of analysis and the texts selected for comparison between the corpora do not show any genre-specific comparison (i.e., essays versus journals and book sections) which may have impacted the results. This calls for a more precise comparative method by taking into consideration the genre and disciplinary group of the comparative corpora to establish that the variation between the texts occurs due to different groups of writers without the influence of any other elements (i.e., different types of texts and disciplines). In terms of the linguistic features studied in learner corpora, although lexical bundles have garnered more attention in relation to their structural and functional categories variation between non-native novice and native novice or expert writers, studies which explicate how the functions of bundles are discovered/postulated (with regard to lexico-grammatical pattern and semantic choice/preference) have not been recorded thus far.

The most common principle of LCR is the use of learner corpus as a source of data to understand the learner language. However, extending this primary role as a data source, the learner corpus is also deployed as a pedagogic corpus in teaching and learning to enhance learners' use of language (Chambers, 2015). The benefit the learner-pedagogic corpus could proffer is greater than the criticism against it as a possible source of erroneous language production among learners (Granger, 2002). The identification of gaps in learners' own writing and building awareness on what to improve and how it can be improved (by comparing with the native speaker writing) is certainly a promising feature of the application of learner corpus as a pedagogic instrument (Chambers, 2015).

Studies on the use of learner corpus for teaching and learning purposes using the Data-Driven learning (DDL) approach are gradually rising, however more studies are needed within the ESAP field (e.g., Flowerdew, 2015b; Friginal, 2013). In Friginal's (2013) study, students' research paper drafts are brought in for one of the DDL lessons to identify the use of reporting verbs to assist the hands-on concordancing activity. Flowerdew (2015b, p. 61) in guiding the postgraduate science and engineering students to write the discussion section of a thesis highlights four elements in designing the corpus consultation learning (i.e., input, practice, consolidation and extension). While the input and practice stage involves retrieving the students' prior knowledge in the use of certain expressions, the consolidation and extension stage comprises the use of the student's own work to identify the errors and make corrections. The DDL approach, in other words, fosters self-discovery learning among students/learners.

LEARNER CORPUS AND STUDIES IN THE MALAYSIAN CONTEXT

The EMAS Corpus

Learner Corpus Research (in the English language) in the Malaysian context commenced in the year 2002 with the first learner corpus (i.e., The English of Malaysian School Students [EMAS] corpus) developed to study the English language use among Malaysian school students (Samad et al., 2002). Since its inception, the EMAS corpus has become the main source of data for Malaysian researchers cum linguists to investigate linguistic idiosyncrasies of Malaysian school students to further proffer pedagogical implications (e.g., Akbari, 2009; Ang et al., 2011; Zarifi & Mukundan, 2014). The EMAS corpus is a collective effort from Universiti Putra Malaysia researchers. The corpus which comprises 472,652 words includes both written (i.e., narrative, picture and school-based essays) and spoken data (i.e., interviews and verbal essays) extracted from Primary Five, Secondary One and Secondary Four students across Malaysia.

The MACLE

MACLE, the Malaysian Corpus of Learner English, developed by Universiti Malaya is also used in researching the language pattern of Malaysian learners (Knowles et al., 2006). The corpus comprises approximately 800,000 words of argumentative essays written by Universiti Malaya undergraduate students from various disciplines (second to the fourth year of study) between 2004 and 2005. However, a limited number of past studies concerning the corpus (e.g., Aziz, 2018; Don & Sriniwass, 2017) show that this corpus has not been much explored. As opined by Aziz (2018), the non-availability of the corpus for public use (available upon request of access) could have been one of the reasons. Another corpus which is similar to MACLE is the Malaysian English Corpus (MEC), a learner corpus that contains Universiti Malaya undergraduate students' written essays (Kaur & Shamsudin, 2010). However, there is not much information which could be collected regarding its development and use for research purposes.

The CALES

The Corpus of Archive of Learner English in Sabah and Sarawak (CALES) (Botley et al., 2005) is another corpus that needs to be mentioned as far as the learner corpus research in

Malaysia is concerned. The corpus project was initiated in the year 2003 with a collection of argumentative essays (approximately 400,000 words) from students at tertiary institutions in Sabah and Sarawak, West Malaysia. There are four institutions involved in supplying the needed data (i.e., Universiti Teknologi Mara Sarawak, Universiti Teknologi Mara Sabah, Universiti Malaysia Sabah and Universiti Malaysia Sarawak) (e.g., Botley & Dillah, 2007). Studies using the CALES involve spelling (e.g., Botley & Dillah, 2007), phraseological units between Malaysian, British and American students (e.g., Botley, 2010) and rhetorical features of arguments in written discourse (e.g., Botley, 2014).

The MCSAW

The Malaysian learner corpora can be seen largely as developed from argumentative essays. MCSAW, the Malaysian Corpus of Students' Argumentative Essays (Mukundan & Kalajahi, 2013) is a corpus of 565,500 words which is constructed based on argumentative essays written by secondary school (i.e., Secondary Four and Secondary Five) and first-year college students from four states in Malaysia (i.e., Selangor, Negeri Sembilan, Melaka and Kelantan) (Joharry, 2016). Researching the grammatical and phraseological patterns of the English language written by Malaysian secondary and college learners, a number of studies have been identified to date, using the MCSAW (e.g., Joharry, 2016; Kader et al., 2013; Manokaran et al., 2013). The MCSAW has become the main source of Malaysian written learner data in Joharry's (2016) study which investigates the individual keywords and key lexical bundles that appear in the L2 learner writing by comparing the features of *can* and *we*, as well as three to four-word lexical bundles together with their functional categories (i.e., referential, discourse organising and stance), with the native speaker learner writing via LOCNESS. This study is the only one (to date) which has investigated lexical bundles in Malaysian ESL learner writing apart from the current PhD study of the researcher who published a corpus-driven comparative study on the use of passive verb bundles between L1 and L2 English speakers in academic writing (see Subramaniam & Kaur, 2021).

The ICE-Malaysia

Literature has also shown the use of the International Corpus of English (Malaysia), the Malaysian English variety (Rahim & Awab, 2012, as cited in Rahim, 2014), a subset of a large internationally acknowledged English language corpus (ICE) which consists of spoken and written data in learner corpus study. In a recent study, Ong and Rahim (2021) deploy the non-professional writing sub-section of the ICE-Malaysia (which contains 15 essays of undergraduates from Malaysian universities that make up a total of 31,854 tokens) to analyse the light verb construction of Malaysian students while comparing it with the British students' writing via ICE-Britain.

Self-Compiled Learner Corpus for Research Purpose

The Written English Corpus of Malaysian English Learners (WECMEL), a 470,000-word self-complied learner corpus (Abdullah & Noor, 2013), serves as an argumentative-type essay corpus which was collected in 2013 to study the verb-noun collocational patterns among the

Malaysian pre-degree learners (a pre-entrance course to Law programme) at Universiti Teknologi Mara in comparison to LOCNESS. The use of this corpus has been mentioned in one of the recent studies researching on collocations (e.g., Abdullah et al., 2021). Studying the rhetorical moves of arguments in argumentative essays written by pre-university students, Kanestion and Kaur (2021) compile the Corpus of MUET Writing Argumentative Essays (COMWArE).

Similar to WECMEL and COMWArE, there are other self-developed learner corpora which have been constructed by Malaysian researchers for research purposes. This shows the enduring effort by Malaysian researchers to collect students' written work to study the learner language from the primary to tertiary level of education (e.g., Abdullah & Noor, 2013; Chau, 2015; Kaur, 2009). Kaur (2009) has compared the vocabulary choices between Malaysian children (boys versus girls) and British children (boys versus girls) by building her own corpora for both groups. The Malaysian children corpus, MCCW (the Malaysian Corpus of Children's Writing) contains free-writing entries (of children from 8 to 12 years) extracted from a local English newspaper and open-topic essays produced by students (within the same age range) from one of the urban schools in Malaysia. On the other hand, in attempting a second language acquisition study, Chau (2015) studies the L2 learner language development (in relation to function words and narrating structures) at four different points in time. The longitudinal corpus which is designated as the Longitudinal Corpus of Developing Language User Narrative Texts (LoCDeLUNT) comprises 496 texts (116,399 words) collected from 124 students studying at one of the secondary schools in Malaysia in May 2007, November 2007, November 2008 and June 2009 within a period of 24 months.

Apart from argumentative and narrative texts which are common in learner corpus research, Malaysian learner language pattern has also been studied using descriptive texts. Focusing on a rural secondary school in one of the northern parts of Malaysia, Ang et al. (2020) collect both descriptive and narrative texts (i.e., 128 essays altogether) written by Secondary Two students. The cleaned corpus which comprises 24,037 words, however, has not been designated with any specific name. The compiled corpus is used to identify the subject-verb agreement and copula *be* errors in the writing of lower secondary school students in Malaysia. Ang et al.'s (2020) study uses the EA (error analysis) approach to identify the misused items in students' writing, hence there is no comparison made to the native learner variety using the CIA approach.

LCR in the Malaysian context has been attempted not only from primary to undergraduate levels but also extended to postgraduate students to study the discourse-pragmatic features. More recently, Lo et al. (2021) study the use of boosters in the drafts of written research proposals. The 64,500-word corpus is compiled from eight first-year ESL doctoral candidates from four areas of study in education.

The Need for Specialised Learner Corpus Studies

By and large, the literature survey has shown that learner corpus studies in Malaysia witness continuous development. This means that numerous scholarly works have been executed exploring students' language use (from vocabulary, phraseology, and grammar to discourse level) ranging from primary to college level of education. Nevertheless, there are very few

studies which have reported the comparison between the native and non-native learner variety (e.g., Abdullah & Noor, 2013; Aziz, 2018; Botley, 2010; Joharry, 2016; Kaur, 2009) and as far as the researcher is concerned there is no comparison made between learner and expert language. Furthermore, despite the growth in the number of LCR, EAP or ESP studies using the specialised or discipline-specific learner corpora can be seen as an under-researched area in the Malaysian context. Yunus and Awab (2011) although utilise Malaysian undergraduate Law students' discipline-related text (i.e., legal contract genre) in analysing the collocational competence of prepositions, this is not classified as a corpus-related analysis as the texts (i.e., 40,600 words) are manually analysed without using any corpus tool.

Hitherto, there are two studies using the specific-text learner corpora that have been identified in Malaysian setting (e.g., Kaur & Shamsudin, 2011; Shauki & Kaur, 2022). Kaur & Shamsudin (2011) develop the Business and Management English Language Learner Corpus (BMELC), a written corpus which contains essay writing assignments, journal writing activities and media invitations produced by undergraduate students who enrolled in business and management courses at two universities (i.e., public and private) in Malaysia. The BMELC is a corpus compiled specifically to study the noun forms (i.e., neutral, singular, plural and proper nouns) and their distribution pattern in these students' writing. The need to develop a corpus as a pedagogical instrument of sales email writing in Business English classrooms, in one of the most recent studies, Shauki and Kaur (2022) elaborate on the development of Corpus of Entrepreneurship Emails (COREnE) from pedagogical sales email written by undergraduate students of the first and second year in Universiti Malaya Kelantan. Based on industry informants' and subject specialists' interviews, the moves in the corpus are tagged and coded using the Integrated Moves Approach (moves and part of speech tagging). The corpus analyses result in generating a framework of sales email that can be used to address pedagogical needs in an ESP-based classroom at the tertiary level in Malaysia.

This section concludes that despite a growing trend in learner corpus research in Malaysia, there are three aspects which are still far under-explored. The first is the corpus data itself, in which more specialised/discipline-specific text needs to be aimed at as the corpus under investigation to explore academic writing in ESAP. Next is the linguistic feature in focus, where the lexical bundles, a recurring sequence of words that "helps to shape meanings" and "contribute to our sense of coherence in a text" (Hyland, 2008, p. 4) have not been extensively studied (e.g., Joharry, 2016), and the other is the analysis of interlanguage features via comparison with native novice and expert writing.

CONCLUSIONS

This review highlights several aspects which can be taken into account to contribute to the existing body of knowledge involving the Learner Corpus Research. The issue of comparability can be resolved via a justified comparison between the non-native learner language and the native novice or expert language. To aim for pedagogical needs, a comparative or reference corpus which can be modelled as a reliable source regardless of nativity but expertise (i.e., non-native expert/ proficient users of English as a lingua franca) can be considered. Despite having a plethora of learner corpus studies exploring word or phraseological units towards the

paradigmatic understanding of the discourse, more genre-specific and discipline-based comparisons need to be attempted to contribute to the area of ESAP. For example, the Electrical and Electronic engineering learner research report is compared to research articles from the same sub-discipline of engineering as these two genres have similar text organisation.

In addition, treating the learner corpus as more than just a source of erroneous language is another important aspect to welcome the acceptance of learner corpus as a pedagogical instrument to assist the DDL process in improving target language production. In the Malaysian context, although there have been many learner corpus studies exploring language use among ESL students, studies at the undergraduate level are still limited especially using learner writing based on specific disciplinary categories. In terms of the linguistic aspect, the lexical bundles, recurring sequence of words, though have been profoundly studied in LCR in terms of structural and functional categories, the analysis of this phraseological unit can be initiated from the word level (i.e., the core lexical item) within the bundle and further moved to an extended set of patterns which includes its co-occurring syntagmatic combinations (i.e., lexicogrammar) that convey the discourse function of the unit (Sinclair, 1996). The semantic preference associated with the discourse function is interesting to be discovered by studying the paradigmatic semantic relation of the unit to further understand the context of use (pragmatics) (Cortes & Hardy, 2013). This type of study which illustrates a comprehensive qualitative analysis of a phraseological unit from word to discourse level may aid pedagogical activities using corpora in ESAP classrooms.

REFERENCES

Abdullah, S., Aziz, R. A., & Kamaruddin, R. (2021). Lexical verbs in verb-noun collocations: Empirical evidence from a Malay ESL learner corpus. *3L: The Southeast Asian Journal of English Language Studies*, *27*(4), 144-156. http://doi.org/10.17576/3L-2021-2704-11

Abdullah, S., & Noor, N. M. (2013). Contrastive analysis of the use of lexical verbs and verb-noun collocations in two learner corpora: WECMEL vs. LOCNESS. *Learner Corpus Studies in Asia and the World*, *1*, 139-160. https://doi.org/10.24546/81006680

Adel, A. (2006). *Metadiscourse in L1 and L2 English*. John Benjamins Publishing.

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81-92. https://doi.org/10.1016/j.esp.2011.08.004

Akbari, O. (2009). *A corpus-based study on Malaysian ESL learners' use of phrasal verbs in narrative compositions* [Doctoral dissertation, Universiti Putra Malaysia]. https://core.ac.uk/download/pdf/43000861.pdf

Ang, L. H., Rahim, H. A., Tan, K.H., & Salehuddin, K. (2011). Collocations in Malaysian English learners' writing: A corpus-based error analysis. *3L: The Southeast Asian Journal of English Language Studies*, *17*(Special issue), 31–44. http://www.ukm.my/ppbl/3L/3LHome.html

Ang, L. H., Tan, K. H., & Lye, G. Y. (2020). Error types in Malaysian lower secondary school student writing: A corpus-informed analysis of subject-verb agreement and copula be. *3L: The Southeast Asian Journal of English Language Studies*, *26*(4), 127-140. http://doi.org/10.17576/3L-2020-2604-10

Aziz, R. A. (2018). *A corpus-based study of the use of "BE" in Malay ESL learner essays* [Doctoral dissertation, Universiti Malaya]. http://studentsrepo.um.edu.my/id/eprint/8962

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In A. Freedman & P. Medway (Eds.), *Genre and the new rhetoric* (pp. 79–101). Taylor and Francis.

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Botley, S. (2010). A corpus-based comparison of idiom use by Malaysian, British and American students. *Proceeding of International Conference on Science and Social Research*, 139–144. Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/CSSR.2010.5773752

Botley, S. (2014). Argument structure in learner writing: A corpus-based analysis using argument mapping. *Kajian Malaysia*, *32*(Supp. 1), 45-77. http://web.usm.my/km/vol32_supp1_2014.html

Botley, S., De Alwis, C., Metom, L., & Izza, I. (2005). CALES: A corpus-based archive of learner English in Sarawak. *Final project report.* Unit for Research, Development and Commercialisation, Universiti Teknologi MARA.

Botley, S., & Dillah, D. (2007). Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research*, *3*(1), 74-93. https://melta.org.my/journals/MAJER/downloads/majer03_01_03.pdf

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52. https://doi.org/10.1016/j.jeap.2017.10.008

Callies, M., & Zaytseva, E. (2013). The Corpus of Academic Learner English (CALE): A new resource for the assessment of writing proficiency in the academic register. *Dutch Journal of Applied Linguistics*, *2*(1), 126-132. https://doi.org/10.1075/dujal.2.1.11cal

Chambers, A. (2015). The learner corpus as a pedagogic corpus. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 445-464). Cambridge University Press.

Chang, J. Y. (2011). *The use of general and specialized corpora as reference tools for academic and technical English writing: A case study of Korean graduate students of engineering* [Doctoral dissertation, Seoul National University].

Chau, M. H. (2015). *From language learners to dynamic meaning makers: A longitudinal investigation of Malaysian secondary school students' development of English from text and corpus perspectives* [Doctoral dissertation, University of Birmingham]. http://etheses.bham.ac.uk/id/eprint/6087

Chen, A. C. H. (2019). Assessing phraseological development in word sequences of variable lengths in second language texts using directional association measures. *Language Learning*, *69*(2), 440-477. https://doi.org/10.1111/lang.12340

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30-49. http://dx.doi.org/10125/44213

Cobb, T., & Horst, M. (2015). Learner corpora and lexis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 185-206). Cambridge University Press.

Cortes, V., & Hardy, J. A. (2013). Analysing the semantic prosody and semantic preference of lexical bundles. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 180-201). University of Michigan Press.

Crosthwaite, P., Cheung, L., & Jiang, F. K. (2017). Writing with attitude: Stance expression in learner and professional dentistry research reports. *English for Specific Purposes*, *46*, 107–123. https://doi.org/10.1016/j.esp.2017.02.001

Don, Z. M., & Sriniwass, S. (2017). Conjunctive adjuncts in Malaysian undergraduate ESL essays: Frequency and manner of use. *Moderna Spark*, *1*, 99-117. https://ojs.ub.gu.se/index.php/modernasprak/article/view/3611

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*(2), 165–193. https://doi.org/10.1093/applin/amv011

Flowerdew, L. (2015a). Learner corpora and language for academic and specific purposes. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 465-484). Cambridge University Press.

Flowerdew, L. (2015b). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, *20,* 58-68. https://doi.org/10.1016/j.jeap.2015.06.001

Friginal, E. (2013). Developing research report writing skills using corpora. *English for Specific Purposes*, *32*(4), 208–220. https://doi.org/10.1016/j.esp.2013.06.001

Gilquin, G. (2001). The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast*, *3*(1), 95–123. https://doi.org/10.1075/lic.3.1.05gil

Gilquin, G. (2021). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, 1–13. https://doi.org/10.1017/S0261444821000094

Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee, & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 55-78). John Benjamins.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund University Press.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). Longman.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). John Benjamins.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). John Benjamins.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English*. *Version 3*. Presses universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Granger, S., & Paquot, M. (2013). Language for Specific Purposes learner corpora. In C. A. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 3142–3146). Blackwell Publishing Ltd.

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. *World Englishes*, *15*(1), 3-15. https://doi.org/10.1111/j.1467-971X.1996.tb00088.x

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2012). Undergraduate Understandings: Stance and voice in final year reports. In K. Hyland & C. S. Guinda (Eds.), *Stance and voice in written academic genres* (pp. 134-150). Palgrave Macmillan.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118. http://www.lib.kobe-u.ac.jp/repository/81006678.pdf

Johansson, S. (1991). Times change, and so do corpora. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 305-314). Longman.

Joharry, S. A. (2016). *Malaysian learners' argumentative writing in English: A contrastive, corpus-driven study* [Doctoral dissertation, University of Sydney]. https://ses.library.usyd.edu.au/handle/2123/16770?show=full

Joharry, S. A., & Rahim, H, A. (2014). Corpus research in Malaysia: A bibliographic analysis. *Kajian Malaysia*, *32*(Supp. 1), 17-43. http://web.usm.my/km/vol32_supp1_2014.html

Kader, M. I. A., Begi, N., & Vaseghi, R. (2013). A corpus-based study of Malaysian ESL learners' use of modals in argumentative compositions. *English Language Teaching*, *6*(9), 146-157. https://eric.ed.gov/?id=EJ1077048

Kanestion, A., & Kaur, M. (2021). A corpus-based investigation of moves in argument stage of argumentative essays. *EPRA International Journal of Multidisciplinary Research*, *7*(9), 199-204. https://doi.org/10.36713/epra8475

Kaur, M., & Shamsudin, S. (2010). Corpus linguistics: Syntactical analysis of learner corpus anyone? *Proceedings of the 7th International Language for Specific Purposes Seminar on Globalisation of New Literacies*, 1-28. Universiti Teknologi Malaysia.

Kaur, M., & Shamsudin, S. (2011). Extracting noun forms: A lesson learnt. *International Journal of Language Studies*, *5*(4), 19-32. https://journaldatabase.info/articles/extracting_noun_forms_lesson_learnt.html

Kaur, S. (2009). *A corpus-driven contrastive study of girls" and boys" use of vocabulary in their writing in Malaysia and the United Kingdom* [Doctoral dissertation, Lancaster University].

Kizil, A. S. (2020). Corpus of Academic Learner English (CALE): A new corpus at the intersection of corpus linguistics and English for academic purposes. *The Literacy Trek* *6*(2), 41-54. https://doi.org/10.47216/literacytrek.791664

Knowles, G., M. Don, Z, M., Jariah, M. Jan, J. M., Sargunan, R., Yong, J., Sathiadevi, et al. (2006). The Malaysian Corpus of Learner English: A bridge from linguistics to ELT. In H.

Azirah & H. Norizah (Eds.), *Varieties of English in Southeast Asia and beyond* (pp. 257-267). University of Malaya Press.

Larsson, T. (2017). The importance of it is important that or importantly? The use of morphologically related stance markers in learner and expert writing. *International Journal of Corpus Linguistics*, *22*(1), 57–84. https://doi.org/10.1075/ijcl.22.1.03lar

Larsson, T. (2019). Grammatical stance marking across registers: Revisiting the formal-informal dichotomy. *Register Studies*, *1*(2), 243–268. https://doi.org/10.1075/rs.18009.lar

Lee, D. Y. W., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, *18*(4), 281–296. https://doi.org/10.1016/j.jslw.2009.07.003

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Longman.

Leedham, M., & Fernandez-Parra, M. (2017). Recounting and reflecting: The use of first person pronouns in Chinese, Greek and British students' assignments in engineering. *Journal of English for Academic Purposes*, *26*, 66-77. https://doi.org/10.1016/j.jeap.2017.02.001

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

Liardet, C. L., & Black, S. (2019). "So and so" says, states and argues: A corpus-assisted engagement analysis of reporting verbs. *Journal of Second Language Writing*, *44*, 37- 50. https://doi.org/10.1016/j.jslw.2019.02.001

Lin, C. H., & Lin, Y. L. (2019). Grammatical and lexical patterning of make in Asian learner writing: A corpus-based study of ICNALE. *3L: The Southeast Asian Journal of English Language Studies*, *25*(3), 1-15. http://doi.org/10.17576/3L-2019-2503-01

Lo, Y. Y., Othman, J., & Lim, J.W. (2021). Mapping the use of boosters in academic writing by Malaysian first-year doctoral students. *Pertanika Journal of Social Sciences & Humanities*, *29*(3), 1917-1937. DOI: https://doi.org/10.47836/pjssh.29.3.23

Manokaran, J., Ramalingam, C., & Adriana, K. (2013). A corpus-based study on the use of past tense auxiliary 'Be' in argumentative essays of Malaysian ESL learners. *English Language Teaching*, *6*(10), 111-119. https://eric.ed.gov/?id=EJ1077151

Meunier, F. (2016). Introduction to the LONGDALE Project. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment* (pp. 123–126). Peter Lang.

Mukundan, J., & Kalajahi, S. A. R. (2013). *Malaysian Corpus of student argumentative writing*. Australian International Academic Centre.

Nam D., & Park, K. (2020). I will write about: Investigating multiword expressions in prospective students' argumentative writing. *PLoS ONE*, *15*(12), 1-13. https://doi.org/10.1371/journal. pone.0242843

Nesi, H. (2016). Corpus studies in EAP. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 206-217). Routledge.

Nesi, H. (2021). Sources for courses: Metadiscourse and the role of citation in student writing. *Lingua*, *253*, 1-17. https://doi.org/10.1016/j.lingua.2021.103040

Ong, C. S. B., & Rahim, H. A. (2021). Analysis of light verb construction use in L1 and L2: Insights from British and Malaysian student writing. *TESL-EJ*, *25*(2), 1-18. http://www.tesl-ej.org/wordpress/issues/volume25/ej98/ej98a1/

Parkinson, J. (2015). Noun-noun collocations in learner writing. *Journal of English for Academic Purposes*, *20*, 103-113. https://doi.org/10.1016/j.jeap.2015.08.003

Rahim, H. A. (2014). Corpora in language research in Malaysia. *Kajian Malaysia*, *32*(Supp. 1), 1-16. http://web.usm.my/km/vol32_supp1_2014.html

Samad, A. A., Hassan, F., Mukundan, J,. Kamarudin, G., Rahman, S. Z. S. A., Rashid, J. M., & Vethamani, M. E. (2002). *The English of Malaysian school students (EMAS) corpus*. Universiti Putra Malaysia.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, *11*(2), 133–158. https://doi.org/10.1111/1473-4192.00011

Shauki, N. B. I., & Kaur, M. (2022). Developing a corpus of entrepreneurship emails (COREnE) for business courses in Malaysian university using integrated moves approach. *Sains Humanika*, *14*(1), 1-9. https://doi.org/10.11113/sh.v14n1.1885

Sinclair, J. (1996). The search for units of meaning. *Textus*, *9*(1), 75–106.

Sinclair, J. (2005). Corpus and Text – Basic Principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxbow Books.

Subramaniam, R., & Kaur, S. (2021). A corpus-driven study on the use of passive verb bundles in academic writing: A comparison between L1 and L2 English speakers. *GEMA Online Journal of Language Studies*, *21*(4), 64-87. https://ejournal.ukm.my/gema/issue/view/1440

Sung, M. C. (2020). Underuse of English verb–particle constructions in an L2 learner corpus: Focus on structural patterns and one-word preference. *Corpus Linguistics and Linguistic Theory*, *16*(1), 189-214. https://doi.org/10.1515/cllt-2017-0002

Wang, M., & Zhang, Y. (2021). 'According to…': The impact of language background and writing expertise on textual priming patterns of multi-word sequences in academic writing. *English for Specific Purposes*, *61*, 47–59. https://doi.org/10.1016/j.esp.2020.08.005

Yang, S. (2023). A review of research on learner corpora —Taking overseas core journals in Linguistics from 2007 to 2021 as an example. *Theory and Practice in Language Studies*, *13*(2), 417-423. https://doi.org/10.17507/tpls.1302.16

Yunus, K., & Awab, S. (2011). Collocational competence among Malaysian undergraduate Law students. *Malaysian Journal of ELT Research*, *7*(1), 151-202. https://myjurnal.mohe.gov.my/public/article-view.php?id=14634

Zarifi, A., & Mukundan, J. (2014). Creativity and unnaturalness in the use of phrasal verbs in ESL learner language. *3L: The Southeast Asian Journal of English Language Studies*, *20*(3), 51-62. http://ejournals.ukm.my/3l/index