Learner corpus research in Hong Kong: past, present and future

Kanglong Liu,[1] Joyce Oiwun Cheung[2] and Nan Zhao[3]

Abstract

As a field of research closely connected with second language acquisition, teaching and learning, learner corpus research (LCR) has garnered interest among language teachers and researchers in Hong Kong, where English is one of the two official languages (alongside Chinese) and also one of the chief mediums of instruction in education. In view of this unique situation, this paper provides a comprehensive overview of LCR within different teaching contexts in Hong Kong and identifies some major research trends and issues. Through this survey of the development of LCR in the region, we find that great advances have been made over the past three decades. Specifically, the object of analysis has shifted from cherry-picked, isolated textual features to operationalised parameters such as metadiscourse markers, lexical diversity, and syntactic complexity to study learners' language output. Despite the progress that has been achieved so far, there remain a number of important questions for LCR in the context of Hong Kong. In particular, some researchers tend to broadly apply the term 'learner corpus' even to the language output of expert-level L2 speakers. Yet, whether this group of speakers can be treated as L2 learners, and their language output as a learner corpus, remains contested. In addition, existing learner corpora are also limited in their scope by genre, with the majority being compiled from letters and essay writings. This paper concludes with suggestions on how these limitations can be addressed in future research.

Keywords: corpus linguistics, Hong Kong EFL, learner corpus research, second language acquisition

[1] Room AG518d, The Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong.
[2] AG518, The Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong.
[3] Department of Translation, Interpreting and Intercultural Studies, Hong Kong Baptist University, Kowloon Tsai, Hong Kong.
Correspondence to: Kanglong Liu, *e- mail*: kl.liu@polyu.edu.hk

Corpora 2022 Vol. 17 (SI): 79–97
DOI: 10.3366/cor.2022.0248

www.euppublishing.com/cor

1. Introduction

Being one of the two official languages (alongside Chinese) in Hong Kong, English is the main medium of communication in workplace and professional settings. Due to its colonial history, English remains the dominant language in the fields of education, law and business, and it is taught as a compulsory subject for all students beginning in primary school. However, whether English is treated as an ESL (English as a Second Language) or an EFL (English as a Foreign Language) is rather complicated and still much debated. Based on Kachru's (1985) three-circle model of English, some researchers (e.g., Bolton [2003]) have accorded Hong Kong English the status of an ESL, while others (e.g., O'Brien [2004]) have increasingly viewed English as a foreign variety in Hong Kong. This latter position has emerged because 'a mixed code of Cantonese and English has been the predominant style of presentation' (O'Brien, 2004: 1), instead of English alone (see Li [2017]). We believe that instead of categorising Hong Kong English as a clear-cut EFL/ESL dichotomy, it is more fruitful to view it as one situated on a continuum expressing features of both learner contexts, as proposed by Gilquin and Granger (2011: 76), and perhaps more towards the EFL pole of the cline, as is often assumed in much of the LCR in Hong Kong.

Motivated by a desire to improve English language pedagogy, linguists in Hong Kong have long used learner corpora to compare learners' language production with that of native English speakers. The first large-scale learner corpus research project in Hong Kong can be traced back to the development of The Hong Kong University of Science and Technology Learner Corpus (HKUST-LC), which consists of Chinese undergraduate students' academic essays (Milton and Tsang, 1993). Based on this corpus, a number of comparative studies were conducted to investigate the unique features of Hong Kong learners' English (Flowerdew, 1998; and Hyland and Milton, 1997). As corpus technology develops, more researchers have gradually chosen to compile their own corpora and adopted more sophisticated linguistic parameters to examine learner English. For almost three decades, this line of research has continued to attract the interest of various research groups working in the fields of Applied Linguistics, Corpus Linguistics and Second Language Acquisition. This paper aims to review the development of Hong Kong LCR to bring the picture of LCR into sharper relief.

Before defining what is absent or problematic in LCR, it will be useful to briefly overview what is present and positive about the learner corpus. Learner corpora, defined by Granger *et al*. (2015: 10), are 'electronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria'. In the opening issue of the International Journal of Learner Corpus Research, Callies and Paquot (2015: 1) described learner corpora as 'systematic collections of authentic, continuous and contextualised language use (spoken or written) by L2 learners stored in electronic format'

providing empirical data across a range of disciplines. Following the above definitions, Hong Kong LCR can be approached using various contextual factors including synchronicity/diachronicity, spoken/written tasks, various disciplines, education levels, L2 proficiency, first language and genres.

To review the major empirical LCR studies undertaken in Hong Kong, we searched the string 'learner corp*' in two academic databases, namely Scopus and Web of Science (WoS), which retrieved all articles containing either 'learner corpus' or 'learner corpora' in their titles or abstracts. After careful screening, we found a total of ninteen articles on Hong Kong LCR have been published from 2006 to 2021. We also surveyed a few earlier works on LCR that were not indexed by these two databases or failed retrieval because they did not use the particular terms, 'learner corpus/corpora'. Altogether, we identified twenty-six articles on Hong Kong LCR (see Appendices A and B).

2. Overview of learner corpus research in Hong Kong

Since the initiation of the HKUST-LC project (Milton and Tsang, 1993), learner corpus research in Hong Kong has undergone significant changes. With the caveat that natural variability renders categorisations arbitrary, we can tentatively divide these decades into three research phases: the infancy phase (early 1990s to 2000), the developmental phase (2001 to 2010) and the maturity phase (2011 to 2020). During the early 1990s to 2000, Hong Kong LCR tended to employ the HKUST-LC. Between 2000 and 2010, some researchers began to utilise self-compiled learner corpora to examine how learners differ from native speakers in their language use, both linguistically and stylistically. From 2011 to 2020, new perspectives emerged in Hong Kong LCR in which the focus shifted to learner accomplishments and explicating language learning due to cross-linguistic interference. Some comprehensive linguistic parameters were also being applied during this phase (lexical complexity and syntactic complexity).

Early LCR in Hong Kong was pioneered by Milton and his colleagues, using the HKUST-LC. For example, Milton and Tsang (1993) found that English writing by Hong Kong learners is characterised by an overuse of logical connectors in comparison to native English speaker use. Hyland and Milton (1997) found that the Hong Kong EL2 learners utilised syntactically simpler hedging and booster constructions than their EL1 counterparts by adopting a more limited range of expressions and showing greater problems in conveying a precise degree of certainty. Also using the same learner corpus, Flowerdew (1998) compared the underuse and overuse of cause/effect markers between Hong Kong learners and L1 expert users, finding that writing by the former was characterised by an underuse of prepositions to express causality. This shows that the emergence of Hong Kong LCR was more or less in line with global LCR, initiated by Granger's

(1996) International Corpus of Learner English (ICLE). During this phase, the goal had chiefly been focussed on examining how Hong Kong learners' use of English 'deviates' from that of native speakers.

Starting from the early 2000s, researchers began to use self-compiled corpora to conduct LCR in Hong Kong. The following are several studies illustrating the then trend of computer-aided error analysis based on self-compiled corpora. By comparing learner and professional corpora, Hyland (2002a) found that Hong Kong learners tended to underuse the authorial pronoun, which according to him was 'problematic'. Later, Flowerdew (2006) also compiled a learner corpus of argumentative essays and claimed that the most frequent category of errors made by L1 Cantonese students was colligation, where students often used incorrect prepositions after signalling nouns. Moreover, Fung and Carter (2007) constructed a learner corpus of spoken English, which enabled them to discover that Hong Kong learners demonstrated a limited use of discourse markers in comparison to native speakers. They argued that the limited use of discourse markers and the prevalence of specific markers in the learner corpus was connected to the unnatural linguistic input that ESL students are exposed to. Specifically, traditional grammar-orientated pedagogy prioritised propositional meanings rather than pragmatic usage in spoken language. Yeung's (2009) hypothesis that Hong Kong learners tended to overuse certain connectives was confirmed through her self-compiled corpus of argumentative essays, which revealed that Hong Kong students overused connectives like *besides*, *due to* and *moreover*, yet underused the causal connector *because*. Some corpus linguists observed that comparing learners' L2 performance to that of native speakers may not be as beneficial to pedagogy as had been assumed. When examining the Hong Kong Corpus of Spoken English (HKCSE, composed of EL1/EL2 dialogue exchanges in Hong Kong), Cheng and Warren (2007) found that even real-life expert-level English differed from the 'standard' English prescribed in local textbooks. Such incongruence is also supported by relevant corpus research. For instance, Seto (2009) found that there were significant differences between textbook English and genuine English used in expressing agreement. These studies increasingly call for the improvement of English textbooks by emphasising everyday language use.

Starting from the 2010s, researchers tended to investigate to what extent learners can actually advance in their second language acquisition. Studies in this line of enquiry have rectified the tendency of blaming the learners (e.g., learner language error) towards appreciating learners' improvement and predicting learners' progress, in a more holistic manner. This shift in perspectives and theorisations in LCR has resulted in a number of breakthroughs. Researchers began to adopt a more descriptive approach in characterising the language output of Hong Kong learners. This highlights the perception and reality of an English variety resulting from an interaction with the Cantonese L1, rather than a non-standard variety. Hong Kong LCR experienced this shift in theorisation mainly because international LCR scholars have increasingly reflected on the potential problem of LCR

describing interlanguage as an 'incomplete version' of native English (Granger, 2004). In this regard, Fan (2009) concluded from a learner corpus study that Hong Kong students' use of collocations, lexis and grammar were 'affected by their Cantonese L1'. Also, by comparing the Hong Kong and British components of the International Corpus of English (ICE-HK and ICE-GB), Yao (2016) found that *it*- and *wh*-clefts used by Hong Kong learners diverged from those by native speakers in ways that suggest the influence of contact, including distributional patterns, use of relativisers, and the syntactic function of the cleft element. She reasoned that the grammar of Hong Kong English is shaped by the transfer of gradient grammatical rules from the substrate language and it, thus, should be treated as a regional variety in its own right. Using learner corpus techniques to make sense of Hong Kong learners' language production, Yao has demonstrated that the deviations between learner English and native English should be explained with reference to cross-linguistic influence, rather than language misuse by learners without regard for situational relevance.

Another major breakthrough of LCR in Hong Kong has been spearheaded by Crosthwaite (2016) and Crosthwaite and Jiang (2017), who examined learners' improvement over the course of their learning. These two papers reported a longitudinal study of Hong Kong university students' English proficiency based on corpora of written assignments and examination answers. By comparing the assignment corpus with the examination corpus, Crosthwaite (2016) found several positive outcomes, including the use of fewer first-person pronouns and more nominalisations in students' academic essay writing. Crosthwaite and Jiang (2017) also reported substantial improvement in student writing at the discourse level – for example, using more hedge words and fewer boosters to avoid making unsubstantiated claims or sweeping statements. These two studies demonstrated how learner corpora can help to track students' progress in language learning. In applying learner corpus techniques to a discipline-specific context, Hafner and Wang (2018) built a learner corpus of legal academic writing, and Wong (2018) compiled a separate one of students' peer comments and self-reflections from an online news writing project. The former study revealed that senior-year law students used fewer boosters than their junior peers, showing a heightened awareness of conforming to disciplinary expectations. Likewise, the latter study found that students showed a clear understanding of news writing structure, layout and style, despite occasional grammatical mistakes.

Apart from these changes in research perspectives, Hong Kong LCR has also undergone transitions in the scope of research, moving away from individual linguistic markers to more comprehensive sets of linguistic parameters such as meta-discourse markers, lexical diversity and syntactic complexity. These holistic approaches offered greater insights into learner language output. For example, Yan (2019) cross-compared spoken and written native English with spoken and written Hong Kong English, revealing a clear distinction between spoken and written English by native speakers but a less discernible one by Hong Kong learners. The indicators used

in the study included a range of holistic measures such as vocabulary size, mean word length, lexical density and lexical coverage. Taking meta-discourse markers as indicators, El-Dakhs (2020) demonstrated that Hong Kong students overused hedges compared to native speakers, but underused attitude markers, self-mention markers and interactional markers. With the advancement of corpus software and tools, the use of such parameters has become more popular with LCR researchers. Lee *et al*. (2021) recently discovered that examination grades in writing by secondary-school students are more strongly correlated with lexical and syntactic complexity. This innovative study took LCR to the next level by exploring writing quality using linguistic complexity metrics.

3. Corpus design

Throughout the decades of LCR in Hong Kong, corpus size has varied to a large extent, ranging from merely a thousand words (e.g., Fan [2009]) to 14.7 million words (e.g., Li and MacGregor [2010]). Contrary to the general inclination, LCR in Hong Kong does not use only large-scale corpora. Studies making use of the Hong Kong Corpus of Spoken English (HKCSE) varied in corpus size from 88,077 words (Yeung, 2009) to 920,000 words (Cheng and Warren, 2007) to 2 million words (Seto, 2009), depending on the sub-sections of the corpus they used and any data they added to the existing corpus. Although corpus size can be as small as 1,327 words (Fan, 2009), the usual corpus size in Hong Kong LCR ranges between 10,000 and 100,000 words. Researchers have rarely constructed corpora of over one-million words, but tend to resort to readily available corpora such as the HKUST-LC, when larger language samples are required. Clearly, LCR in Hong Kong has attached more emphasis to representativeness of the content rather than corpus size, even though both are considered to be crucial criteria in corpus research.

Instead of corpus size, researchers in Hong Kong have tended to be concerned with corpus design and the construction of their own corpora. Out of the twenty-six LCR research papers we surveyed that explicitly investigated the language output of Hong Kong learners, nineteen used a self-compiled corpus, three added new data to an existing corpus, and four relied solely on readily available corpora. Depending on their study goals, researchers might create their own corpora or use existing ones. For instance, Hyland (2002a,b, 2004) tended to tailor-make corpora to study different cohorts of learner language (undergraduates and postgraduates) one at a time; alternatively, El-Dakhs (2020) simply used the Hong Kong ESL sub-corpus and the Japanese EFL sub-corpus from the larger learner corpus project, the International Corpus Network of Asian Learners of English (ICNALE), in order to compare non-native EL2 learner types. Using sub-corpora from the same parent corpus facilitated El-Dakhs' (2020) comparative research, since both sub-corpora were collected using the same sampling frame. LCR

in Hong Kong has used all forms of data collection, with the compilation of a new corpus being the most popular. This practice may be due to the high flexibility of self-compiled corpora, which can be shaped to answer specific research questions.

Of the twenty-two LCR papers which either studied written or spoken English (but not both), eighteen analysed written corpora, while only four analysed spoken, reflecting a preference in Hong Kong LCR for written data. One might speculate that this trend is due to the difficulties involved in the construction of spoken corpora. While written corpora are typically compiled from essays and letters (discussed further below), compiling a spoken corpus typically requires more time and manpower. For example, Fung and Carter's (2007) corpus of learner group discussions involved recording the sessions, transcribing the audio, and annotating the transcripts. This degree of effort and resources required in the compilation of spoken corpora made it unviable, particularly in the early stages of LCR in Hong Kong. Such difficulties are reflected in the under-availability of spoken corpora worldwide (Granger, 2004). It should be noted, though, that such difficulties should have been greatly reduced after two decades, particularly with the use of sophisticated speech recognition and annotation tools, which can be used to automatically transcribe recorded speech in seconds, and annotate the corpus with enhanced tag sets. Although quality control still necessitates some manual labour, various corpus tools have, in recent years, helped to reduce error rate to a tolerable range. Despite these additions to performance and also the virtue of spoken corpora in better reflecting learners' spontaneous and unrepaired language production (Myles, 2015), Hong Kong LCR has largely ignored the use of spoken corpora and is thus predominantly skewed towards written forms.[4]

As previously stated, written learner corpora in Hong Kong are typically made up of essays or assignments written by university students (Flowerdew, 1998, 2006; Hyland, 2002a,b; and Crosthwaite and Jiang, 2017) or secondary-school students (Hyland and Milton, 1997; Fan, 2009; and Lee *et al*., 2021). When we examined the essays and letters in these corpora more thoroughly, we noticed that a large proportion of the essays were argumentative/persuasive, while the majority of the letters concerned business and advice. Compared to the limited variety of genres addressed by Hong Kong LCR in written English, spoken corpus research appears to cover a broader range of genres. Fung and Carter's (2007) learner corpus, for example, documents role-modelling meetings in which students acted as employees of a toy firm. Ng's (2015) court interpretation corpus also stands out for its uniqueness. The result shows that LCR spoken corpora are relatively more diverse than written corpora in Hong Kong.

---

[4] The Hong Kong Corpus of Spoken English (HKCSE) compiled by Cheng *et al*. (2005) consists of naturally occurring speech produced by adults in professional instead of learner settings. Thus, most studies based on this corpus which do not investigate learner English are not covered in this review paper.

More importantly, the prevalence of essay and letter genres in the data also point to the type of language learner typical in Hong Kong LCR. Among the LCR studies we investigated in this review, fourteen used learner corpora comprising university students' L2 production, eight used corpora of business setting communications, while only four used corpora of senior secondary students' L2 production. Both university students and adult professionals have an advanced command of the English language. Since The Hong Kong Diploma of Secondary Education requires undergraduates to pass the English test, even the many LCR studies based on first year university students would present high levels of English language output, regardless of discipline (Bolton *et al*., 2002; Flowerdew, 2006; Crosthwaite, 2016; Ma and Wang, 2016; and Crosthwaite and Jiang, 2017). Given that this primary target group of tertiary students are already quite fluent in English, LCR in Hong Kong is skewed towards intermediate to advanced learners. For example, Crosthwaite (2016) and Crosthwaite and Jiang (2017) specifically pointed out that the students involved usually have a C1 level of proficiency in the Common European Framework of Reference for Languages (CEFR). The corpora by Qian and Pan (2019) and El-Dakhs (2020) were based on intermediate learners whose English level amounted to B1 or B2 in the CEFR. Yet this phenomenon in Hong Kong is aligned with the practice of LCR internationally in targeting intermediate to advanced level learners (Granger, 2004) while overlooking the beginner and pre-intermediate language learners.

Finally, the practice of using the HKCSE in LCR shifts the focus away from English learners. Cheng and Warren (2007), who created the HKCSE, explicitly warned that their corpus was not a learner corpus, but one comprised of competent speakers of English communication, having both completed higher education and gained recognition for their English proficiency in academic, business, conventional and public settings. Nevertheless, a number of studies that use the HKCSE appear to treat the language in the data as learner output, without acknowledging the uniqueness of the language users. While there is no simple solution, it is important for LCR in Hong Kong to spell out the background of 'learners' so that the findings can be contextualised and categorised appropriately. Appendix A summarises the list of studies in HK LCR.

4. Methodological issues

We have also witnessed some progress concerning LCR methodology in Hong Kong, notably in: (*1*) the increased use of automated software over manual analysis; (*2*) the development and application of comprehensive and sophisticated measurements of linguistic markers; (*3*) the use of quantitative statistical methods instead of frequency counting; and, (*4*) the selection of appropriate reference corpus in the research design.

During the first two decades of LCR in Hong Kong, manual annotation with a moderate amount of computer assistance using tools was common (Milton and Tsang, 1993; Flowerdew, 1998; Hyland, 2004; and Fan, 2009). Among the tools, Wordsmith has been quite popular among researchers for its user-friendly interface, and its compatibility with other software such as Wmatrix (Qian and Pan, 2019), which has online compatibility for keyword analysis (Wong, 2018), and UAM, which facilitates the use of user-specified tagsets (Crosthwaite and Jiang, 2017). The use of Computerized Language Analysis (CLAN) is also becoming popular because CLAN files are compatible with the Child Language Data Exchange System convention (see Lee *et al*. [2021]). The variety of corpus tools that have been employed demonstrates that LCR in Hong Kong is keeping up with international trends. The last decade demonstrates the use of more automated software such as multi-dimensional analysis tagger (e.g., Crosthwaite [2016]) and L2 syntactic complexity analyser (e.g., Lee *et al*. [2021]) to directly extract linguistic features from corpora. The application of more comprehensive and sophisticated indicators has seen a shift from focussing on isolated and cherry-picked features to more systematic analysis of linguistic trends in learners' L2 production. These indicators can be divided into three categories based on their linguistic levels: lexis, syntax and discourse.

For LCR studies taking lexis as an entry point, researchers often start with frequency counts. For example, Li and MacGregor (2010) assessed the validity of Vocabulary Levels Test (VLT) as a test of university students' vocabulary essential to their fields of study. The study found that VLT lexis frequencies did not match real-life language use. Wong (2018) also counted word frequencies used in commenting by university students. Frequency counting appears to be a convenient and intuitive way of studying lexis in LCR, yet recent LCR using lexical metrics has explored more fruitful research questions, including lexical complexity in different grades of essay compositions (Lee *et al*., 2021) and speech (Yan, 2019). Over time, the focus of lexis-type research has shifted from small units (e.g., collocations) to clauses and sequences, and finally to overall syntactic complexity in L2 learner output. Fan (2009) first studied collocations between words (limited to five) that were commonly used by learners; later, Yao (2016) focussed on clefts (i.e., the separation of clause elements for emphasis); while Qian and Pan (2019) explored modal sequences (i.e., expressing mood, possibility and obligation). In another recent study, Lee *et al*. (2021) used fourteen syntactic complexity indices from Lu (2010) in various grades of essay writing and found significant differences between high- and low-scored texts. These studies highlight that LCR in Hong Kong has started to use a diverse range of syntactic indicators to acquire better insights into learner language.

Discourse indicators have also been used in Hong Kong LCR to assess how learners' language output differs from that of native speakers, in terms of the organisation and structuring of textual content. Discourse indicators in LCR are significantly more diverse than lexical and syntactic

ones, reflecting a variety of research foci including discourse markers (Fung and Carter, 2007), such as connectors (Milton and Tsang, 1993; Bolton *et al*., 2002; Ma and Wang, 2016; and Yeung, 2009), and metadiscourse markers (El-Dakhs, 2020) such as self-mentioning (Hyland, 2002a; and Crosthwaite and Jiang, 2017) and self-repetitions (Fung, 2007). Some researchers have even looked at broad interactional and interpersonal strategies (Cheng and Warren [2007] and Seto [2009], respectively).

In Hong Kong, Hyland's (2005) metadiscourse framework has had a significant impact on LCR studies. Subsequent research has focussed on students' use of stance markers (Crosthwaite and Jiang, 2017) and boosting devices (Hafner and Wang, 2018), or even whole taxonomies (El-Dakhs, 2020). Other frameworks that have been used in LCR in Hong Kong include Norrick's (1987) taxonomy of Same Speaker Repetitions, and Milton and Tsang's (1993) twenty-five connectors used by Hong Kong students. More sophisticated linguistic metrics have also been used in Hong Kong LCR, including Biber's (1988, 1989) Multidimensional Analysis, McKee *et al*.'s (2000) lexical diversity 'VocD' measure, and Lu's (2010) L2 Syntactic Complexity indices. It should be noted that researchers have developed automated software which can extract the linguistic features required for analysis under these frameworks very quickly, for example, Crosthwaite (2016) directly used Nini's Multidimensional Analysis Tagger and Lee *et al*. (2021) used Lu's (2010) Syntactic Complexity Analyzer.

In terms of data analysis, Hong Kong LCR is no longer limited to the early obsession with frequency counts. More recent research has used statistical testing to determine whether the observed differences are statistically meaningful. This has improved the scientific rigor of LCR in Hong Kong. Nevertheless, on the one hand, Hong Kong LCR exhibits both descriptive and inferential statistics, such as testing significance with log-likelihood (e.g., Crosthwaite and Jiang [2017] and Qian and Pan [2019]), chi-square test (Yan, 2019), ANOVA tests (El-Dakhs, 2020); on the other hand, analytical techniques, such as multi-factorial modelling and multi-variate exploratory tools which have already gained currency in international LCR (Gries, 2015), were not found in the current review. Therefore, Hong Kong LCR is still lagging behind the development of international LCR in terms of data analysis.

The final methodological issue is the selection of a reference corpus. LCR in Hong Kong has tended to use existing large-scale native English corpora as a reference corpus since Milton and Tsang (1993) first compared local students' language using BROWN and LOB. Yet this has become more diversified since the turn of the millennium. Bolton *et al*. (2002) used ICE-GB as a reference corpus; Fung and Carter (2007) used parts of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) to achieve a similar word count to the sample corpus, Yeung (2009) adapted the Collins Birmingham University International Language Database (COBUILD), and Yan (2019) included the British National Spoken Corpus of English. Concerned more with comparability, Ma and Wang (2016) chose

the Louvain Corpus of English Essays, since it was comprised of writing genres similar to their own self-compiled corpus. Yao's (2016) use of the International Corpus of English (ICE) incorporates comparable attention to genre and speaker profile, and thus is arguably an advancement in LCR. One way of managing comparability has been the creation of Hong Kong reference corpora. Cheng and Warren (2007) collected English conversations between Hong Kong and native speakers of English, ensuring that both groups were conversing in the same context. Fan (2009) used sixty essays written by native speakers of the same educational level as her L2 learners as a reference corpus for her study on the writing styles of Hong Kong students. Generally speaking, Hong Kong LCR researchers show awareness of the importance of using appropriate reference corpora, with some creating their own to facilitate comparability (see Appendix B).

To summarise, Hong Kong's LCR has advanced significantly over the last two decades. This can be seen in the employment of increasingly sophisticated indicators and systematic linguistic frameworks, enhanced corpus tools and software, refined statistical methodologies, and suitable reference corpora to improve comparability.

5. Summary and future directions

The first critical issue concerning LCR in Hong Kong is corpus design, specifically the selection of the target learners. The early reliance on language samples in the HKCSE is misplaced, as highlighted by the corpus creators, Cheng and Warren (2007), since it was comprised of professional communications with advanced L2 fluency. As such, any findings based on the HKCSE may not be applicable to less advanced learners, such as secondary school students or primary school children. We believe that LCR researchers in Hong Kong should exercise caution when using the term 'learner corpus' in these situations: if researchers sample data from proficient speakers or writers, they should avoid using the term, or at least justify why the collected data merits that status.

Second, the majority of LCR studies in Hong Kong have focussed on university students, while few studies examined the language output by senior secondary-school pupils (an exception is Lee *et al*. [2021]) and none touched upon junior secondary-school or primary-school students. This leaves an important research gap in LCR in Hong Kong. The inclusion of language output by young learners can also help lay a foundation for longitudinal studies on learners' language progression in EFL/ESL settings. Another gap in LCR in Hong Kong is learner language production across disciplines. The majority of LCR studies have recruited university students based on their academic year, rather than on their field of specialisation. As students' language proficiency can vary considerably across disciplines or even courses, the generalisability of findings might be

limited. The preference for L2 production by university students has also resulted in another critical issue in corpus design: lack of genre diversity. Since university students are commonly required to write lengthy essays or term papers, many learner corpora in Hong Kong are composed of argumentative/persuasive and organisational correspondence. More creative writing genres such as fiction, short stories, poetry, plays, or even microblogs and vlogs, need to be included in the corpus designs in order to be truly representative of English learner use in Hong Kong.

Despite the limitations mentioned above, we can see that LCR in Hong Kong has made significant progress in uncovering the nature of learners' English language output in the Hong Kong context. Entering yet another decade brings with it some exciting initiatives that connect LCR with other disciplines such as translation and interpreting studies in Hong Kong (e.g., Liu [2015]). This indicates that LCR research, with its methodological and scientific robustness, can cross-fertilise and inform a wide range of disciplines and pedagogical research (Liu, 2020). Furthermore, Hong Kong has a significant number of non-Chinese ethnic minoritie, and future LCR can also look into the use of English as well as Chinese by this group of people rather than simply focussing on the English output by learners with L1 Chinese.

Acknowledgments

This research was funded by a General Research Fund (GRF) grant (Ref: 15605520) from the Research Grants Council of Hong Kong.

References

Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. 1989. 'A typology of English texts', Linguistics 27 (1), pp. 3–43.

Bolton, K. 2003. Chinese Englishes: A Sociolinguistic Survey. Cambridge: Cambridge University Press.

Bolton, K., G. Nelson and J. Hung. 2002. 'A corpus-based study of connectors in student writing: research from the International Corpus of English in Hong Kong (ICE-HK)', International Journal of Corpus Linguistics 7 (2), pp. 165–82.

Callies, M. and M. Paquot. 2015. 'Learner corpus research: an interdisciplinary field on the move', International Journal of Learner Corpus Research 1 (1), pp. 1–6.

Cheng, W. and M. Warren. 2007. 'Checking understandings: comparing textbooks and a corpus of spoken English in Hong Kong', Language Awareness 16 (3), pp. 190–207.

Cheng, W., C. Greaves and M. Warren. 2005. 'The creation of a prosodically transcribed intercultural corpus: the Hong Kong Corpus of Spoken English (prosodic)', ICAME Journal 29, pp. 47–68.

Crosthwaite, P. 2016. 'A longitudinal multidimensional analysis of EAP writing: determining EAP course effectiveness', Journal of English for Academic Purposes 22, pp. 166–78.

Crosthwaite, P. and K. Jiang. 2017. 'Does EAP affect written L2 academic stance? A longitudinal learner corpus study', System 69, pp. 92–107.

El-Dakhs, S.D.A. 2020. 'Variation of metadiscourse in L2 writing: focus on language proficiency and learning context', Ampersand 100069, pp. 1–8.

Fan, M. 2009. 'An exploratory study of collocational use by ESL students – a task based approach', System 37 (1), pp. 110–23.

Flowerdew, L. 1998. 'Integrating "expert" and "interlanguage" computer corpora findings on causality: discoveries for teachers and students', English for Specific Purposes 17 (4), pp. 329–45.

Flowerdew, J. 2006. 'Use of signalling nouns in a learner corpus', International Journal of Corpus Linguistics 11 (3), pp. 345–62.

Fung, L. 2007. 'The communicative role of self-repetition in a specialised corpus of business discourse', Language Awareness 16 (3), pp. 224–38.

Fung, L. and R. Carter. 2007. 'Discourse markers and spoken English: native and learner use in pedagogic settings', Applied Linguistics 28 (3), pp. 410–39.

Gilquin, G. and S. Granger. 2011. 'From EFL to ESL: evidence from the International Corpus of Learner English' in M. Hundt and D. Mukherjee (eds) Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap, pp. 57–80. Amsterdam: 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 and M. Johansson (eds) Languages in Contrast: Text-based Crosslinguistic Studies, pp. 37–51. Lund: Lund University Press.

Granger, S. 2004. 'Computer learner corpus research: current status and future prospects' in U. Connor and T.A. Upton (eds) Applied Corpus Linguistics: A Multidimensional Perspective, pp. 123–45. Leiden: Brill.

Granger, S., G. Gilquin and F. Meunier. 2015. 'Introduction: learner corpus research – past, present and future' in S. Granger, G. Gilquin and F. Meunier (eds) The Cambridge Handbook of Learner Corpus Research, pp. 1–5. Cambridge: Cambridge University Press.

Gries, St.Th. 2015. 'Statistics for learner corpus research' in S. Granger, G. Gilquin and F. Meunier (eds) The Cambridge Handbook of Learner Corpus Research, pp. 159–81. Cambridge: Cambridge University Press.

Hafner, C.A. and S.H. Wang. 2018. 'Hong Kong learner corpus of legal academic writing in English: a study of boosters as a marked language form in an English-medium instruction context', TESOL Quarterly 52 (3), pp. 680–91.

Hyland, K. 2002a. 'Authority and invisibility: authorial identity in academic writing', Journal of Pragmatics 34 (8), pp. 1091–112.

Hyland, K. 2002b. 'Directives: argument and engagement in academic writing', Applied Linguistics 23 (2), pp. 215–39.

Hyland, K. 2004. 'Graduates' gratitude: the generic structure of dissertation acknowledgements', English for Specific Purposes 23 (3), pp. 303–24.

Hyland, K. 2005. Metadiscourse: Exploring Interaction in Writing. New York, NY: Continuum.

Hyland, K. and J. Milton. 1997. 'Qualification and certainty in L1 and L2 students' writing', Journal of Second Language Writing 6 (2), pp. 183–205.

Hyland, K. and P. Tse. 2005. 'Hooking the reader: a corpus study of evaluative that in abstracts', English for Specific Purposes 24 (2), pp. 123–39.

Kachru, B. 1985. 'Standards, codification, and sociolinguistic realism: the English language in the Outer Circle' in R. Quirk and H. G. Widdowson (eds) English in the World: Teaching and Learning the Language and Literatures, pp. 11–30. Cambridge: Cambridge University Press.

Lee, C., H. Ge and E. Chung. 2021. 'What linguistic features distinguish and predict L2 writing quality? A study of examination scripts written by adolescent Chinese learners of English in Hong Kong', System 102461, pp. 1–19.

Li, D.C. 2017. Multilingual Hong Kong: Languages, Literacies and Identities. Gewerbestrasse: Springer.

Li, L. and L.J. MacGregor. 2010. 'Investigating the receptive vocabulary size of university-level Chinese learners of English: how suitable is the Vocabulary Levels Test?', Language and Education 24 (3), pp. 239–49.

Liu, K. 2015. 'Investigating corpus-assisted translation teaching: a pilot study' in P. Sánchez-Gijón, O. Torres-Hostench and B. Mesa-Lao (eds) Conducting Research in Translation Technologies, pp. 141–62. Bern: Peter Lang.

Liu, K. 2020. Corpus-assisted Translation Teaching: Issues and Challenges. Singapore: Springer.

Lu, X. 2010. 'Automatic analysis of syntactic complexity in second language writing', International Journal of Corpus Linguistics 15 (4), pp. 474–96.

Ma, Y. and B. Wang. 2016. 'A corpus-based study of connectors in student writing: a comparison between a native speaker (NS) corpus and a non-native speaker (NNS) learner corpus', International Journal of Applied Linguistics and English Literature 5 (1), pp. 113–18.

McKee, G., D. Malvern and B. Richards. 2000. 'Measuring vocabulary diversity using dedicated software', Literary and Linguistic Computing 15 (3), pp. 323–38.

Milton, J.C. and E.S.C. Tsang. 1993. 'A corpus-based study of logical connectors in EFL students' writing: directions for future research' in R. Pemberton and E.S.C. Tsang (eds) Studies in Lexis, pp. 215–46. Hong Kong: The Hong Kong University of Science and Technology Language Center.

Myles, F. 2015. 'Second language acquisition theory and learner corpus research' in S. Granger, G. Gilquin and F. Meunier (eds) The Cambridge Handbook of Learner Corpus Research, pp. 309–31. Cambridge: Cambridge University Press.

Ng, E. 2015. 'Teaching and research on legal interpreting: a Hong Kong perspective', Monografias de Traduccion e Interpretacion 7, pp. 243–70.

Norrick, N.R. 1987. 'Functions of repetition in conversation', Text 7 (3), pp. 245–64.

O'Brien, T. 2004. 'Writing in a foreign language: teaching and learning', Language Teaching 37 (1), pp. 1–28.

Qian, D.D. and M. Pan. 2019. 'Politeness in business communication: investigating English modal sequences in Chinese learners' letter writing', RELC Journal 50 (1), pp. 20–36.

Seto, A. 2009. '"I agree with you" – a corpus-based study of agreement', Language, Linguistics, Literature 15 (1), pp. 41–67.

Wong, D. 2018. 'A corpus-based study of peer comments and self-reflections: how did ESL learners use peer comments in an online newswriting project?', International Journal of Online Pedagogy and Course Design 8 (4), pp. 65–90.

Yan, Y. 2019. 'A corpus-based comparative study of Lexis in Hong Kong and native British spoken English', Glottometrics 47, pp. 66–82.

Yao, X. 2016. 'Cleft constructions in Hong Kong English', English World-Wide 37 (2), pp. 197–220.

Yeung, L. 2009. 'Use and misuse of "besides": a corpus study comparing native speakers' and learners' English', System 37 (2), pp. 330–42.

Appendix A (*continued on following page*): Corpus designs in Hong Kong LCR.

Study	Corpus	Tokens	Education level	Language proficiency	Genre
Milton and Tsang (1993)	Self-compiled (later HKUST-LC)	4,084,000	University year 1 and senior secondary	Not mentioned	Written assignments Examination scripts
Hyland and Milton (1997)	Self-compiled	500,000	Senior secondary	TOEFL 450–600	Examination scripts
Flowerdew (1998)	HKUST-LC	40,000	University year 1	Not mentioned	Student assignments on environmental pollution
Bolton *et al*. (2002)	Self-compiled (later ICEHK)	46,460	University undergraduates (across disciplines)	Not mentioned	Examination scripts
Hyland (2002a; 2002b)	Self-compiled	630,000	University final year (across disciplines)	Not mentioned	Project reports
Hyland (2004)	Self-compiled	35,000	University postgraduates (across disciplines)	Not mentioned	Dissertation acknowledgement
Hyland and Tse (2005)	Self-compiled	105,000	University postgraduates (across disciplines)	Not mentioned	Dissertation abstracts
Flowerdew (2006)	Self-compiled	110,000	University year 1 (across disciplines)	Not mentioned	Argumentative essays
Fung and Carter (2007)	Self-compiled	14,157	Senior secondary (aged 17–19)	Not mentioned	Group discussions on toy manufacturing proposal
Fung (2007)	Addition to existing corpus	52,200	Expert users	Not mentioned	Business meetings
Cheng and Warren (2007)	Self-compiled	920,000	Expert users	Not mentioned	Mixed spoken
Seto (2009)	Addition to existing corpus	2,000,000	Expert users	Not mentioned	Mixed spoken
Yeung (2009)	Self-compiled	88,077 (Hong Kong learners)	University (across years)	Not mentioned	Argumentative and expository essays
Fan (2009)	Self-compiled	1,327 (Hong Kong learners) 2,782 (British speakers)	Senior secondary (year 10)	Not mentioned	Narrative essays

Appendix A (*continued from previous page*): Corpus designs in Hong Kong LCR.

Study	Corpus	Tokens	Education level	Language proficiency	Genre
Li and MacGregor (2010)	Addition to existing corpus	14,700,000	Expert users	Not mentioned	Mixed
Ng (2015)	Self-compiled	Not mentioned (>100 hours recording)	Expert interpreters	Not mentioned	Court interpretation
Crosthwaite (2016)	Self-compiled	213,408	University year 1 (across disciplines)	CEFR C1 / IELTS 6.5–8	Essays and reports
Yao (2016)	ICEHK	Not mentioned	Adults (aged above 18 and completed secondary school)	Not mentioned	Mixed
Ma and Wang (2016)	Self-compiled	48,721	University year 1 (across disciplines)	Not mentioned	Argumentative essays
Crosthwaite and Jiang (2017)	Self-compiled	213,408	University year 1 (across disciplines)	CEFR C1 / IELTS 6.5–8	Essays and reports
Wong (2018)	Self-compiled	15,741 (comments) 16,095 (self-reflections)	University (major/minor in English Studies)	Not mentioned	Peer comments self-reflections
Hafner and Wang (2018)	Self-compiled	1,037,387	University (major in law)	Not mentioned	Legal academic writing
Qian and Pan (2019)	Self-compiled	81,637 (Hong Kong learners) 72,827 (Shanghai learners)	University year 3 (across disciplines)	CEFR B2	Business letters
Yan (2019)	ICEHK	495,019 (spoken corpus) 440,332 (written corpus)	Adults (aged above 18 and completed secondary school)	Not mentioned	Mixed
El-Dakhs (2020)	ICNALE	14,158 (Hong Kong learners) 13,408 (Japan learners)	Not mentioned	CEFR B1, B2	Argumentative essay
Lee *et al*. (2021)	Self-compiled	64,060	Senior secondary (aged 17–18)	IELTS 4.79–7.77	Letters and reports for advice on social issues

Appendix B (*continued on following page*): Methodological details of Hong Kong LCR.

Articles	Reference corpus	Linguistic features	Methods	Statistical testing
Milton and Tsang (1993)	American BROWN and LOB, science textbooks	Logical connectors	Concordancing	Frequency counting
Hyland and Milton (1997)	GCE A level General Studies scripts	Modal expressions	Hand-coding	Frequency counting
Flowerdew (1998)	MCC	Cause/effect markers	Concordancing	Frequency counting
Bolton *et al*. (2002)	ICEGB	Connectors	Not mentioned	Frequency counting
Hyland (2002a; 2002b)	Professional research articles, textbooks	(a) author pronouns (b) directives	Wordpilot 2000 Concordancing	Frequency counting
Hyland (2004)	N/A	Generic structures	Winmax pro Hand-coding	Frequency counting
Hyland and Tse (2005)	N/A	Evaluative that	Winmax pro Concordancing	Frequency counting
Flowerdew (2006)	N/A	Signalling noun errors	Hand-coding	N/A
Fung and Carter (2007)	CANCODE	Functional paradigm of discourse markers (Interpersonal Referential Structural Cognitive)	Wordsmith (most frequent wordlist) Hand-coding	Contrastive frequency analysis
Fung (2007)	Business meetings of HK-based airline native speakers	Norrick's (1987) taxonomy of same speaker repetitions	Not mentioned	Frequency counting
Cheng and Warren (2007)	native speakers in HKCSE	Speaker-initiated forms Hearer-initiated forms	Not mentioned	Frequency counting
Seto (2009)	N/A	Expressions of agreement	Concapp Hand-coding	Frequency counting
Yeung (2009)	Cobuild Database	Use of *Besides*	Not mentioned	Frequency counting
Fan (2009)	Narrative essays by native speakers in Northern England	Collocations of noun, verbs, adjectives and adverbs	Concapp Concordancing	Frequency counting

Appendix B (*continued from previous page*): Methodological details of Hong Kong LCR.

Articles	Reference corpus	Linguistic features	Methods	Statistical testing
Crosthwaite (2016)	N/A	Biber's multidimensions	Nini's Multidimensional Analysis Tagger	Mann-Whitney U test
Yao (2016)	ICEGB	It-clefts, wh-clefts	Concordancing	Chi-square test
Ma and Wang (2016)	LOCNESS	Milton and Tsang's (1993) 25 connectors used by Hong Kong students	Concordancing	Frequency counting
Crosthwaite and Jiang (2017)	N/A	Hyland's metadiscourse stance markers	Uamcorpustool	Mixed-effects linear regression
Wong (2018)	N/A	Five categories: contents, organisation, grammar, style and tone, layout and display	Wmatrix	Built-in keyword analysis
Hafner and Wang (2018)	N/A	Hyland's metadiscourse boosting devices	Concordancing	Multiple regression
Qian and Pan (2019)	N/A	Modal sequences	Wmatrix Powergrep Wordsmith SPSS	Log-likelihood
Yan (2019)	BNC Spoken and Written	Lexis (vocabulary size, mean word length, lexical density, and lexical coverage)	Microsoft Visual Foxpro SPSS	Chi-square test One-way ANOVA
El-Dakhs (2020)	Essays by native speakers	Hyland's taxonomy of metadiscourse	Hand-coding	One-way ANOVA two-way ANOVA t-test
Lee *et al*. (2021)	N/A	Qin and Uccelli's (2016) lexical diversity, frequency of academic vocabulary, frequency of polysyllabic words, rating and frequency of abstract words Lu (2010) 14 syntactic complexity indices	Computerized Language Analysis (CLAN) L2 syntactic complexity analyser (L2SCA) Hmisc and MASS package in R	One-way ANOVA Bonferroni test (post-hoc analyses)