ReCALL (2019), 31: 3, 255–275
doi:10.1017/S0958344019000077

ARTICLE

Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning

Peter Crosthwaite
The University of Queensland, Australia (p.cros@uq.edu.au)

Lillian L.C. Wong
The University of Hong Kong, Hong Kong (lillianwong@hku.hk)

Joyce Cheung
The Hong Kong Polytechnic University, Hong Kong (cheungoiwun@gmail.com)

Abstract

Data-driven learning (DDL; Johns, 1991), involving students' hands-on use of corpora for self-guided language learning, is a methodology now increasingly used in many tertiary contexts to enhance the teaching of disciplinary postgraduate thesis writing. However, there are still few studies tracking students' actual engagement with corpora for DDL. This mixed-methods study reports on the tracking of students' corpus use via a purpose-built corpus query and data visualisation platform integrated into a large postgraduate disciplinary thesis writing program at a university in Hong Kong. Data on corpus usage history (e.g. times of access, duration of use), query syntax (e.g. query lexis/phraseology and use of wildcards and part-of-speech tags), query function (e.g. frequency lists/distribution, concordance sorting and collocation) and query filters (e.g. searches by faculty, discipline, or thesis section) were collected from 327 students spanning over 11,000 individual corpus queries. The results show significant interdisciplinary and inter-/intra-user trends and variation in the use of particular corpus functions and query syntax adopted by corpus users. Students varied in the type of knowledge (e.g. domain-specific, language-specific) they were accessing, and frequently went beyond the exemplars of the DDL course materials to generate unique queries under their own initiative. Qualitative case study data from three corpus users' activity logs also show distinctive individual corpus engagement by query frequency and function. These data provide a clearer insight into what students actually *do* during DDL and the different directions and trajectories that individual users take as a result of DDL. All accompanying DDL tasks are also included as supplementary materials.

Keywords: corpora; data-driven learning; L2 writing; disciplinary writing; English for academic purposes

1. Introduction and aims of the study

Data-driven learning (DDL; Johns, 1991) is a methodology now increasingly used to enhance the teaching of English for academic purposes (Chen & Flowerdew, 2018; Cotos, 2014; Crosthwaite, 2017; Lee & Swales, 2006; Yoon & Hirvela, 2004). DDL involves the investigation of language corpora through printed concordance materials or students' direct, hands-on use of corpus query tools, and has been used for a range of purposes including language acquisition, genre awareness, and understanding discipline specificity. DDL is typically facilitated through structured tasks that

Cite this article: Crosthwaite, P., Wong, L.L.C. & Cheung, J. (2019). Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning. *ReCALL* 31(3): 255–275. https://doi.org/10.1017/S0958344019000077

© European Association for Computer Assisted Language Learning 2019.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

require students to consult corpora and interpret the output data with a view to "noticing" (Schmidt, 1990) certain statistical patterns of contextualised language in use, including frequency, collocation, and keyness. As a pedagogical approach, DDL creates plentiful opportunities for focus on form (Long, 1991) where tasks draw learners' attention to the language features present within authentic corpus data. The statistical (and increasingly visual) nature of corpus output also facilitates constructivist/connectionist approaches to language learning such as "chunking" (Millar, 2011), where data on collocates and multi-word units are clearly presented in lists or charts to facilitate learning. The self-guided nature of students' corpus engagement for DDL is claimed to result in improved learner autonomy for resolving language-related problems (Leńko-Szymańska & Boulton, 2015). DDL is also claimed to be an increasingly relevant pedagogy for modern digitally oriented learners (Boulton, 2015; Kilgarriff & Grefenstette, 2003) looking for alternatives to dictionaries or translation websites. Medium to large effect sizes for the effectiveness of DDL for language learning have been found across a wealth of DDL studies and thousands of research participants in recent meta-analyses (e.g. Boulton & Cobb, 2017; Cobb & Boulton, 2015; Lee, Warschauer & Lee, 2018).

To date, there has been relatively little research focusing on DDL with postgraduate research students (Chen & Flowerdew, 2018; Flowerdew, 2016) and fewer still investigating students' uptake of DDL for disciplinary thesis writing, with most studies focusing on undergraduate writers within single disciplines only. Studies on disciplinary academic writing have found substantial cross-disciplinary variation in the language features employed across the hard sciences, social sciences, and arts and humanities (e.g. Hyland, 2000), as well as the kind of language reference resources employed by these different groups (e.g. Steel, 2012). These differences may potentially cause significant variation in the uptake and usage of corpora for DDL. Research involving disciplinary writing and DDL include Lee and Swales (2006), where students consulted the British National Corpus and a corpus of research articles before generating their own discipline-specific corpora, with students reporting more engagement with the self-built disciplinary corpora. Charles (2007, 2014) implemented self-built disciplinary corpora with her research students, noting improved longitudinal uptake in the disciplinary corpora over general corpora. Within the same context as the present study, Crosthwaite (2017) conducted postgraduate DDL training courses focusing on error correction, finding excellent qualitative perceptions of DDL and significant positive effects for the correction of lexical and collocation errors.

However, because DDL is still largely treated as an extracurricular activity conducted out of regular class hours or during vacation non-credit courses, most DDL studies are relatively small in scope. The average number of participants in DDL studies ranges between 20 and 49, with "large" studies involving little more than 50 (Boulton & Cobb, 2017). One of the largest DDL studies (Chen & Flowerdew, 2018) focusing on Hong Kong postgraduates boasts 547 participants, although the study involved only short (3½ hour) DDL workshops and the data included only students' post-training perceptions of DDL. Large-scale DDL studies focusing on corpus use across multiple disciplines are still relatively rare, and very little is currently known about postgraduate students' disciplinary corpus use or query habits (Hafner & Candlin, 2007). These data are vital in understanding the affordances of DDL for postgraduate disciplinary writing courses in that "what students report to be doing or what we assume they are doing when we observe them might be quite distant from what they are actually doing" (Pérez-Paredes, Sánchez-Tornel, Alcaraz Calero & Jiménez, 2011: 235).

Documenting students' corpus queries is typically done manually (e.g. Chambers & O'Sullivan, 2004; Frankenberg-Garcia, 2005) or through the collection of computer logs (e.g. Gaskell & Cobb, 2004; Hafner & Candlin, 2007; Pérez-Paredes *et al.*, 2011; Yoon, 2008). For manual data collection, Chambers and O'Sullivan (2004) tracked corpus use and revisions made to the academic writing of eight advanced learners of French, asking students to enter the corpus queries they made and the derived results into a separate worksheet. Frankenberg-Garcia (2005) asked Portuguese-English translation students to manually log the resources (including a selection of online

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

concordancers) used to resolve mistranslations. For automated data collection, Gaskell and Cobb (2004) used a purpose-built platform to track students' error analyses for second language (L2) writing, logging the number of corpus queries and IP addresses of users. Although useful, data regarding the syntax of the queries made or individual corpus usage habits were not available. Studies that did collect query syntax and individual corpus usage data include Hafner and Candlin's (2007) study involving students within the law discipline, which used a purpose-built platform to track user IDs, date and time of access, search queries, and the corpora queried. The most encompassing study on corpus usage for DDL to date, Pérez-Paredes *et al.* (2011), collected data on the number of actions performed by individuals, corpus activities completed, corpus queries, and query syntax. Despite providing a valuable window into actual corpus use by DDL students, the number of users in both studies was small ($n = 37$), and the wide range of corpora used in the latter study made it difficult to track usage across all available platforms.

Tracking issues often feature in DDL studies as many involve established and popular online or offline corpus query interfaces such as AntConc (Anthony, 2014), Sketch Engine (Kilgarriff, Rychly, Smrz & Tugwell, 2004), or Sketch Engine for Language Learning (Baisa & Suchomel, 2014). Other studies (e.g. Hafner & Candlin, 2007) use in-house online concordancers with limited functionality. These may be excellent teaching resources, but they do not easily facilitate the collection of data on corpus usage, nor – more importantly – data on the actual corpus queries made or who made them. Without these data, there is a large gap in what we currently know about what postgraduate students actually *do* when consulting corpora for DDL on thesis writing courses and a lack of information on disciplinary corpus use within multidisciplinary student cohorts. The present study therefore addresses the following research questions through the tracking and characterisation of students' corpus query usage for postgraduate disciplinary thesis writing at the cohort, discipline, and individual levels:

RQ1. How do postgraduate students engage with corpora in terms of their actual corpus usage, query function preferences, and query syntax for DDL?
RQ2. What is the extent of disciplinary variation in the usage of and engagement with corpora for postgraduate DDL involving a multidisciplinary corpus platform?

2. Method

2.1 Research context and duration

The data were collected from students attending disciplinary thesis writing courses for both humanities- and science-related disciplines at a leading university in Hong Kong where the second author was the coordinator of the graduate school English program. Courses run for 24 hours across eight sessions, and aim "to enhance students' awareness of the language features and skills [. . .] so as to approach writing more systematically and with greater confidence" (Centre for Applied English Studies, 2017: 14). Courses ran between the 1st September 2017 and 7th December 2017, with the 7th being the date of the final taught class of the final course group. We continued to collect data until the date immediately prior to the next semester's cohort (21st January 2018) to track continued corpus usage outside of the mandated in-class period.

2.2 Participants

A total of 327 postgraduate PhD and MPhil students were enrolled on the thesis writing courses during the data collection period, with 89 in the humanities and related disciplines classes and 238 in the sciences and related disciplines classes. Individual class groups comprised a maximum of 30 students taught by one teacher, with multiple class groups running throughout the semester at different times. The vast majority (> 90%) of the students are monolingual Mandarin or

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 1. Word counts per faculty in the HKGC

Faculty	Number of words	Proportion (%)
Arts	2,260,154	20.8
Education	2,108,264	19.4
Medicine	1,877,060	17.3
Social sciences	1,217,307	11.2
Law	912,170	8.4
Engineering	811,025	7.5
Architecture	788,474	7.2
Science	438,082	4
Business and economics	401,057	3.7
Dentistry	55,793	0.5

Cantonese speakers with an International English Language Testing System (IELTS) band score of at least 6.5 required for enrolment for a PhD. Within the sciences, medicine and engineering students constitute the two largest cohorts, and students from education are the largest group within the humanities. All students were invited to complete a post-course questionnaire designed by the second author, which was administrated online (see supplementary materials). The first section focused on how students currently searched for information about language, with the second section on the use of the corpus platform. Ninety-three students responded, stating that when writing their thesis drafts, they "often" used dedicated language learning apps on smartphones (53%), social media (41.7%), dictionaries (47.1%), and spellcheckers or grammar checkers (33%). When asked if they used corpora prior to our DDL training, only 3.1% reported they "often" used them, whereas 57.4% said they had never heard of a corpus or had never used one for writing. These data are in line with that found in the same research context in Crosthwaite (2017) and with that found for L2 learners in other context in Steel and Levy (2013).

2.3 Corpus construction

To meet students' diverse discipline-specific writing needs, the authors were awarded a grant to produce a multidisciplinary thesis writing resource to accompany the writing courses comprising a searchable corpus of high-level theses with accompanying DDL tasks that would raise students' awareness of the key linguistic features of disciplinary theses, and aid students in applying this knowledge in their own writing. All 10 faculties of the university were approached to recommend 10 "excellent" completed PhD theses within the last 10 years from each respective school/department, with "excellent" defined as the grade assigned to the thesis following examination. PDF versions of the recommended theses were downloaded from the university library and converted to plain text using AntFileConverter (Anthony, 2017) before being manually stripped of the cover page, acknowledgements, tables, graphs, charts, mathematical equations, diagrams, appendices, and references, leaving only running text and a metadata header containing information on department/faculty to allow filtering of corpus queries. Following this process, the complete corpus (the *Hong Kong Graduate Corpus*, HKGC) comprises 10,869,386 words from theses spanning 52 departments/schools (see Table 1). Certain faculties (e.g. science, business and economics, dentistry) have smaller student cohorts for research degrees, accounting for their smaller presence in the corpus.

Each file was also manually annotated to label sections of the thesis. Due to great diversity in thesis structure across disciplines, we decided on a final annotation structure of ABSTRACT,

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 2. Word counts per section

Section	Number of words	Proportion (%)
Abstract	203,727	1.9
Introduction	1,096,634	10.1
Literature review	2,832,241	26.1
Methodology	1,561,161	14.4
Results and findings	2,182,279	20.1
Discussion	2,299,395	21.2
Conclusion	693,949	6.4

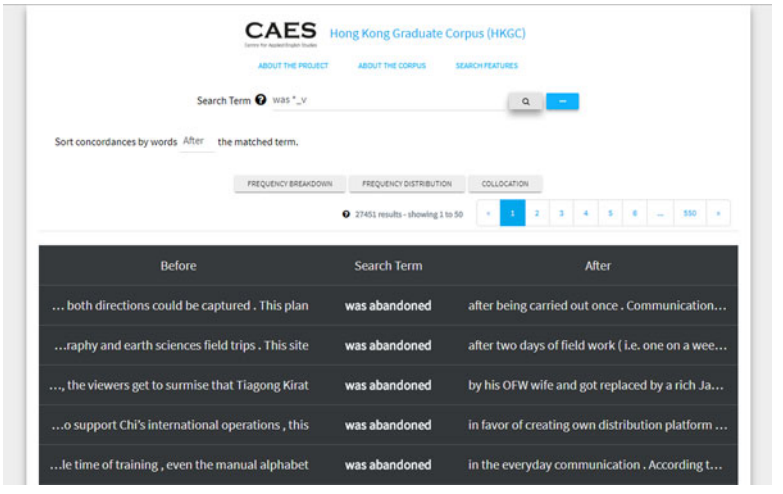


Figure 1. Concordances of "was *_v"

INTRODUCTION, LITERATURE REVIEW, METHODOLOGY, RESULTS AND FINDINGS, DISCUSSION, and CONCLUSION sections (see Table 2). The authors met periodically to agree on the labelling of problematic cases (primarily in medicine and maths), and thesis sections were allowed multiple annotation labels in case of disagreement (e.g. "results & discussion" sections that could be either RESULTS AND FINDINGS or DISCUSSION).

2.4 Corpus platform functionality

In collaboration with computer scientists working in the university's Technology-Enhanced Learning Initiative, we developed an online corpus query platform for the HKGC, with access for staff and students through the Moodle Learning Management System (see supplementary materials).

Via the platform, users can perform simple lexical or phrasal searches as well as searches including wildcards (*) and/or part-of-speech (POS) tags (e.g. "was *_v" would bring up concordances for passive constructions; see Figure 1) and can filter any returned hits by section of the thesis, by faculty, or by specific discipline.

An additional function is frequency breakdown, which provides a bar chart with the most frequent hits for certain queries. (Refer to supplementary materials for the results of the wildcard/POS query "was *_v".)

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press



Figure 2. Right/left collocates of "was"

The frequency distribution function presents a visual display of the distribution of corpus results by section, faculty, and discipline (see supplementary materials).

Finally, the collocation function presents a visual list of collocates, with the search query located in the centre, and collocations one word to the left/right of the query term shown on the left-hand/right-hand side of the centre circle respectively (see Figure 2). Positioning to the left/right is determined by corpus frequency, with frequent hits shown closer to the centre in orange, with less frequent hits found on the periphery and in green. Presenting students with multiple sources of information is claimed to aid learning under a constructivist approach (Flowerdew, 2015) in that "the more possible starting points a corpus offers for exploitation, the more likely it is there exists an appropriate starting point for a specific learner" (Widmann, Kohn & Ziai, 2011: 168). We also emulated Charles (2015), who incorporated concordances together with additional tools to derive word lists, collocates, and concordance plots (i.e. images showing the location of query hits within the whole text or corpus). The highly visual nature of the data also ensures those with more visual learning styles can engage with DDL (Flowerdew, 2015).

2.5 Learner behaviour tracking parameters and analysis

The corpus platform tracked learners' corpus use according to the following parameters: (1) user ID; (2) time, date, and duration of user login to the platform; (3) individual corpus query syntax; (4) any filters applied to corpus query results (i.e. searches by thesis section/faculty and subdiscipline); and (5) corpus function used (i.e. concordance, frequency breakdown, etc.). The platform produced some of this data in a visual format (e.g. number of users, queries, query terms by frequency), while all five tracking parameters were output as. csv files, which were analysed and summarised, as seen in the Results section, by the third author, a bilingual speaker of English and Cantonese working as a research assistant on the project. As with Yoon (2008) and Pérez-Paredes *et al.* (2011), these parameters are considered "precise" for corpus consultation as they allow for individual use to be tracked and not just groups. These data were intended to help the research team better understand which information learners are focusing on when querying the corpus, learn how they navigate the corpus to solve their language problems, and suggest improvements to course content, the interface, and its functionality in future iterations of the course. The third author also ensured that the teachers' usage histories and corpus queries were not included in the final results.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Step 3: Point out the importance of knowing more about that topic AND/OR a gap in current knowledge

- The importance of one's research can be indicated by a reference to a particular problem or the limitation(s) of existing research.
- To do so, gap or problem statements, which are very common in RA/theses across topics and fields, are often used. They often include negative words and expressions such as *little research, few studies, no work*, or words beginning with in- (e.g., *incomplete, insufficient*).
- A second important feature that frequently occurs in gap statements is the contrastive signal word - e.g., *However, While, Although*. These words introduce contrasts or problems in relation to the part of the thesis chapter which has preceded them.
- A third common feature of gap statements is that they often occur at the beginnings of new paragraphs

Figure 3. Example source content on gap statements

2.6 DDL materials

The thesis writing course curriculum covers variation in thesis structure; language and discourse features used in reviewing research literature, identifying the research gap, explaining methodology, reporting, and discussing results and findings; writing of abstracts, introductions, conclusions, and thesis titles; and use of signposts and verb tenses across the thesis. Following the creation of the corpus platform, the next step was to determine which activities from the previous course materials could be augmented or replaced with DDL activities. Following Hafner and Candlin (2007), DDL activities were placed after the introduction of relevant content sections, lists of language features of general and disciplinary thesis writing, or suggested process writing/drafting practices in the materials, so as to make the presentation of these forms more interactive and to promote both "top-down" and "bottom-up" learning where students combine analyses of longer sequences of texts with corpus-based investigations of grammar and lexis (Charles, 2014). This approach is shown in the following example focusing on presenting the "gap" in current research for the literature review. Here, a list of guiding bullet points regarding gap statements in prose (see Figure 3) are then accompanied by a corpus task that asks students to explore the expression "little research" using the concordance, frequency breakdown, and frequency distribution functions (see Figure 4). This activity also serves to train students in the use of these functions in preparation for later tasks.

DDL materials were also used to replace traditional gap-fill activities, specifically activities where course teachers had reported that students previously did not engage with the activity or struggled to complete it. Figure 5 presents a gap-fill activity from the previous year's course materials, and Figure 6 describes the corpus task that replaced it (answers to the task are provided in red font next to the target query term).

In total, two to three individual corpus tasks requiring multiple queries and analysis for completion were built into each of the eight required course units, spanning 22 full tasks in total. The materials were identical for the science-focused and humanities-focused versions of the writing course. Certain activities were intended to be conducted in class, although students were generally invited to complete longer tasks out of class due to time constraints. Activities ranged from awareness-raising tasks (e.g. In what discipline is "my" more frequently used?), sentence completion via copying concordance lines (e.g. This research [has six major purposes]), sentence completion via wildcards (e.g. List five verbs that appear using the search "research has * that"),

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Enter the term 'little research', while checking 'introduction' in the search field.

1) Write 5 sentences containing the words 'little research' in the middle
   a) [In Hong Kong]_____ little research _____[has been conducted in relation to gay culture.]
   b) [there is]_____ little research _____[addressing this issue.]
   c) [there has been]_____little research _____[related to forgiveness or reconciliation.]
   d)[there has been]____ little research _____[on the pooling effect.]
   e) [very]_____ little research _____[has examined these variables.]

2) Enter the term '* little research' then click 'frequency breakdown'. What *collocates* of 'little research' are common? 'Very'.

3) Enter 'little research' and click 'frequency distribution'. Is the phrase 'little research' commonly used in your discipline? Note: it mostly occurs in the Division of Learning, Development and Diversity and the Division of Policy, Administration and Social Sciences Education.

Figure 4. Corpus task (potential answers to the left/right of "little research")

The following methodology section is abridged from the one in a paper entitled "Relationship between impulsive sensation seeking traits, smoking, alcohol and caffeine intake, and Parkinson's diseases". Complete the passage with the appropriate verb form.

METHODS

Patients

Consecutive outpatients of Caucasian descent fulfilling Queen Square Brain Bank criteria for PD[18] _____ **(undergo)** a Mini-Mental State Examination[19] (MMSE) administered by the examining physician and _____ **(invite)** to participate if the MMSE score was >26. We _____ **(exclude)** patients with significant cognitive decline because of the requirement to complete the behavioural and depression rating scales. The Unified Parkinson's Disease Rating Scale (UPDRS)[20] part II _____ **(rate)** for the "on" state and patients provided a list of all current medications and their dosages. Demographic data including age, sex, and age at onset of symptoms of PD _____ **(also collect)**.

Figure 5. Previous gap-fill activity

understanding frequency counts and distributions of specific search terms (e.g. Compare the frequency of "will" and "may" in these sections – what do you notice about how "will" and "may" are used?), and using the frequency breakdown/distribution and collocation functions. The complete set of corpus tasks students were asked to complete have been added as supplementary materials for the reader.

Sequencing of the tasks and detailed rubrics provided for each task also served as an indirect method of training students in the use of the corpus platform and its associated functions, rather than devoting extensive class time to direct training in corpus queries. Screenshots showing

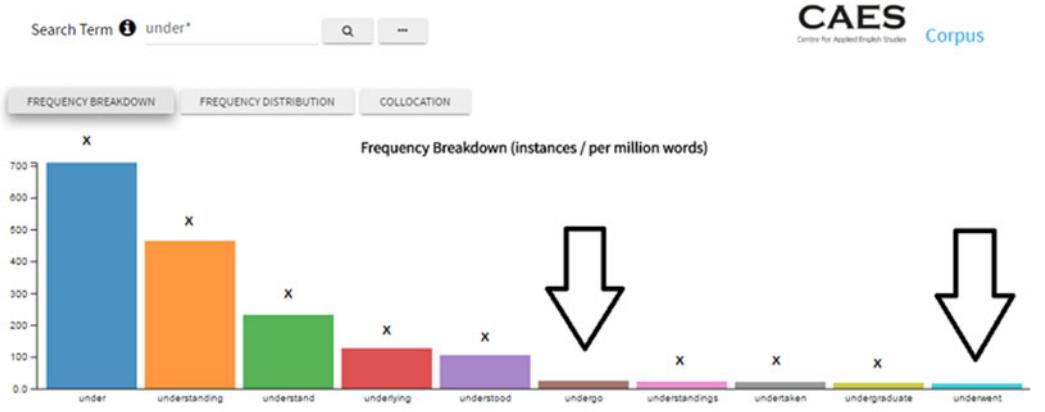
https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

p.74 - Tense and voice in the methodology chapter

Using corpora to resolve morphosyntax

Use the wildcard * after a word's base form as well as the frequency breakdown and collocation functions to get an idea about which tense to use for each verb in the extract. Note that some of these verbs will be in the passive voice, requiring the addition of the correct 'be' verb. Also use the wildcard * before the search item to represent the verb-to-be if necessary, and check the 'methodology' section. The first has been done for you:

1) Under*, click 'frequency breakdown'. Two phrases are possible from the available options. Click on either, and use the concordance lines to make a decision.



2) [were] Invite [invited]
3) Exclude [excluded]
4) Also collect [Look for words before 'Also' as well as the right form of 'collect'] [also collected]
5) Provide [provided]
6) [Were] give [given]
7) Include [included]
8) Collect [collected]
9) Assess [assessed]
10) Range [ranged]
11) [Was] estimate [was estimated]
12) [Was] convert [was converted]

Figure 6. Replacement corpus task

students where to find corpus functions or showing the answer for worked examples were also added to the materials for the students' benefit. Our approach to corpus training is characterised in Pérez-Paredes *et al.* (2011) as a series of *unguided* tasks with training delivered solely through the materials. This is in contrast to *guided* corpus consultation tasks involving explicit teacher-led instruction on the functions of the corpus platform, with the guided condition in their study suggested to be preferable to the unguided condition.

2.7 Teacher preparation

Twelve teachers from both the humanities- and science-focused courses were invited to a three-hour pre-course workshop to trial the updated activities. All teachers had previously taught these courses. Although some had previous experience using corpora, none had specifically used DDL materials in class. The workshop progressed until each felt comfortable using the platform to complete the activities, and each was invited to contact the research team if they encountered

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 3. Corpus queries by function

Function	Queries	Proportion
Basic search/concordance	6,250	51.2%
Frequency breakdown	2,005	16.6%
Frequency distribution	1,715	14.2%
Collocation	1,466	12.1%
Sort concordances by left context[1]	625	5.9%

[1]As the "default" setting is to have concordances sorted to the right, we are unable to separate these results from when students *specifically* sorted the concordances to the right.

difficulties. A few teachers stated during the end-of-course meeting that they often ran out of time to do the activities in class, or that they had used the HKGC themselves only rarely. They did comment, however, that the design and sequencing of the activities was sufficient for most students to be comfortable completing the activities with little guidance from the teacher.

3. Results

This section addresses students' engagement with corpora in terms of their actual corpus usage at the whole cohort and individual levels before determining the extent of disciplinary variation in corpus use in the following section.

3.1 Overall platform use by cohort

The usage statistics of the whole data collection period (1st September to 21st January) were output into an Excel file by the analytics resource platform built into the software. This included 258 unique users, 11,436 accumulated searches, 449 accumulated site visits, and 2,498 searched (unique) queries. Of the 327 students enrolled, 69 (21%) had not attempted to query the corpus. As the corpus tasks did not form part of the course assessment, it is possible these participants failed to find the corpus activities appealing, although it is equally likely these students did not complete any non-assessed in-class activities. It is also possible that some teachers did not have time to complete corpus activities in class, and students rejected completing activities as homework. The data suggest that the frequency of unique queries is far higher than those featuring as exemplars in the course materials and is indicative of substantial variation and innovation among users in the queries made.

Additionally, there are indications of continued corpus use beyond the final taught class and the data collection cut-off date. During this period, the corpus was visited 23 times by 14 student users conducting 197 unique queries, thus accumulating 365 searches in total. Although only representing a small fraction of the users/queries during the taught period, the finding that some users later returned to the platform under their own initiative is encouraging.

Regarding corpus queries by function across the cohort, the HKGC platform requires a basic search to be conducted prior to the other four functions, which accounts for its high proportional use (see Table 3).

Comparing the other four main query functions (excluding basic search/concordance), frequency breakdown accounts for 34% of queries, with frequency distribution at over 29%, collocation at 25%, and sort concordance by left context at 10%. In the materials, students are specifically prompted to use the frequency breakdown function on six occasions, frequency distribution on three occasions, and the collocation function twice, which partly explains the overall proportional use of these functions. However, given that the number of searches made involving each of

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

these functions far exceeds those mandated in the study materials, the data are therefore indicative of students' preferences for the specific output displayed by each function, as well as their intentions regarding the kind of information that they are specifically querying this corpus for. Primarily, it appears that students most frequently like to query the corpus to derive suggested words or phrases from the data, and feel comfortable including wildcards and POS tags in their query syntax to do so. Students seek to determine the distribution of query terms by section or faculty in the second instance, and choose to seek collocates of query terms in the third instance. However, we are not able to determine from the way the data are structured whether these differences are statistically significant. Unlike the other functions, students were not explicitly guided to "sort" concordances in the DDL materials, although students could easily notice the option to do so on the corpus front end. This finding is potentially suggestive that sorting concordances to the left/right of the query term provides a very specific kind of output that students either do not understand how to make use of, or that does not provide the kind of information that they are typically using the corpus to search for.

Table 4 describes the most frequent corpus query terms used by all users during the data collection period. Query terms explicitly featured in the DDL materials are distinguished in Table 4 from unique terms generated by the users outside of the tasks, with the latter shown in bold.

Although the most frequent query terms were obviously exemplars from the course materials, many query terms were of the students' own making. For example, a number of terms related to quantity and quality (e.g. *few research*, *limited studies*, *no studies*, *no research*) appear in the search history. Some of these terms are listed in the course materials when exemplifying how to create a "research gap" as part of writing the literature review (i.e. pointing out the gap in others' contributions) and as part of writing the limitations section of the reports (i.e. pointing out the gap left by the author's own contributions), but there are no specific DDL activities requesting the students to use these terms. This strongly indicates that students are choosing to query the corpus when encountering these terms as they complete the course's non-corpus-based "top-down" activities (requiring extensive reading and involving Swalesian move structure-like analysis) so as to gain more information about their function, usage, and distribution. They are doing this without explicit prompting in an autonomous fashion – a key tenet of the affordances of DDL for language learning.

We also see flexibility in the use of wildcard and POS queries for unique queries outside those in the DDL materials, with wildcards following words so as to determine morphological variants (e.g. "deriv*"), in place of words (e.g. "as * as"), and combined with POS tags (e.g. "*_adj risk"). We also see indications of some erroneous queries as well, such as "studies have*that" where the spacing for the wildcard is incorrect, and "find_v" or "found_v", which can only be verbs. That these erroneous, unique queries are found frequently in the data may indicate either mistakes in explicit guided training (Pérez-Paredes *et al.*, 2011) or that students often experiment with a variety of similar corpus query terms before coming to the correct syntax. Their high frequency is also potentially indicative of students sharing these erroneous query terms with each other, although we are unable to provide further proof.

Table 5 describes the individual corpus query frequency, filters, function, and syntax used by the top 10 most frequent users of the corpus, ranked in order of the total number of corpus queries. At an individual level of detail, even among these top 10 users, there is significant variation in their corpus usage, including variation in the frequency and range of different corpus functions employed, with certain users (e.g. Ranks 2, 4, 6) mainly using the collocation function alongside basic searches, others (e.g. Ranks 1, 3, 10) preferring to use the frequency breakdown function, and variations in whether users sorted concordance results.

This variation may suggest that certain users were looking for different information from the corpus, or that they better understood the format of the output they were receiving from the given function, or it may indicate certain users' willingness to experiment with the full range of corpus

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 4. Most frequent corpus query terms

Query syntax	No. of queries	Query syntax	No. of queries
this research	316	indicate	30
Studies have * that	231	will	29
my	221	**It is hoped that**	28
little research	175	null hypothesis	28
This chapter	172	show	26
suggest	106	may	24
describe	99	***_adj risk**	24
our	84	**found_v**	23
research shows	81	**Show***	22
argue	77	*variable	22
few research	60	strongly	22
find_v	58	**substantially**	22
Possibly	55	**show_v**	22
no studies	55	**studies have*that**	21
research question	49	**no research**	21
research	47	**hopefully**	19
as * as	44	studies have*	19
*little research	43	**research has***	18
studies have shown that	42	**claim**	18
research has*	39	studies have	18
we	37	**deriv***	17
*have shown that	35	**the questionnaire *_v**	17
limited studies	33	***_adj studies**	16
***_adv understood**	32	few studies	16

functions available to them. There is also apparent variation in the typical query syntax employed among users, with some searching without POS tags or wildcards (Rank 7), others employing wildcards at the end of words to look up information on the lemma/morphology (e.g. Ranks 2, 6, 8), others interested in particular POSs (e.g. adjectives, Rank 9), and others using both wildcards and a range of POS tags in their corpus queries (Ranks 3, 10). Despite all students receiving the same input from the course materials, the top 10 most frequent users had learned from the exemplars and were able to move beyond these to explore the HKGC for their own needs. However, they were also generally searching for different information and querying the corpus for the type of information they were explicitly interested in or were comfortable with using.

Tables 6 to 8 present samples of the corpus query usage patterns and habits of three individual users from the list in Table 5 (Ranks 1, 4, and 7). These users were selected by the third author after checking the top 10 users' query and usage histories for their idiosyncratic approaches to individual corpus use. The selected samples represent the duration of corpus use and a qualitative summary of actual corpus queries based on an Excel output file containing each individual query.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 5. Ten most frequent corpus users' usage history

Rank by query frequency	Query frequency	Faculty filter?	Concordance	Collocation	Frequency breakdown	Frequency distribution	Sort left?	Remarks
1	296	Business, Medicine, Education	212	12	52	21	0	Used wildcards and POS _v. Queries include *margin*, *correlation*, *negate*, *theory*, *accumulation*.
2	239	Science, Medicine	170	48	14	7	18	Only used wildcards to derive suffixes. Queries include *large amount*, *residue*, *sentence*, *quotation*, *quantities*.
3	198	Medicine	104	37	47	10	0	Combined use of wildcards and a range of POS tags. Queries include *questionnaire*, *consumption*, *proportion*.
4	197	Education, Social sciences	130	39	14	14	44	Used wildcards and POS _v _n. Queries include *integration*, *creation*, *learn*, *entangled*, *provoke*.
5	181	All	122	21	25	13	0	Used wildcards and POS _v _n. Queries include *questionnaire*, *survey*, *derive*, *future work*, *knowledge gap*.
6	168	Default	108	55	3	2	0	Only used wildcards to derive suffixes. Queries include *judicial*, *classical literature*, *monograph*, *weakness*.
7	152	Education	121	15	9	7	25	No use of wildcards/POS tags. Queries include *little*, *thorough*, *word choice*.
8	151	All	146	1	2	2	1	Only used wildcards to derive suffixes. Queries include *STEM*, *self-directed*, *schools*.
9	150	Science, Medicine	115	8	18	9	0	Used wildcards and POS _adj _v. Queries include *majority*, *consumption*, *risk*.
10	143	Education, Engineering	99	6	31	7	0	Combined use of wildcards and a range of POS tags. Queries include *questionnaire*, *motivation*, *excretion*.

Table 6. The persistent user (Rank 1)

Date	Activity
17/09/17	Spent 30 minutes querying exemplars including "studies have * that" and "little research" in the introduction section. Typically made a query before using the frequency breakdown function to generate the next query in an ongoing loop.
18/09/17	Spent over an hour using wildcards and frequency breakdown on terms including "*correlation" in the literature review, filtering results by Medicine.
19/09/17	Spent 10 minutes using frequency breakdown for "margin" in Business and Economics. Also explored "few studies suggest" in the introduction.
20/09/17	Spent one hour querying "this research *" in Education and "this chapter" in the introduction and literature review. Other queries included "studies have shown that", "little research", etc.
22/09/17	Spent 17 minutes querying "research has shown" in the literature review. Attempted to use POS tags such as "find *_v" as well as the frequency distribution function for the first time.
23/09/17	Spent an hour checking contrastive phrases, such as "negate", "deny", "reject", "challenge the view", "question the view", etc., all within the literature review.
26/09/17	Spent 20 minutes querying "variables" and "null hypothesis" in the methodology.
28/09/17	Spent 45 minutes querying the results section using wildcards and using terms including "the table present" and "significant difference".
19/10/17	Returned to the corpus using the frequency distribution and frequency breakdown functions for "theory *".
21/10/17	Used the corpus for three hours for terms including "slave trade" and "thirteen colonies", presumably relevant to other assignments.
28/12/17	Returned after two months to check the difference between "table *", "graph *", and "chart *".
29/12/17	Spent 12 minutes querying "accumulation", "adherence to *", and "alignment with", this time using the collocation function.
10/01/18	Returned after two weeks to check collocations and frequency breakdown results for "subject".

The Persistent User (Table 6) followed the provisions of each DDL task, as evidenced by the order in which the different query terms and sections of the theses were queried. They also took the initiative to conduct their own queries beyond those included in the materials. Though the user specifically filtered queries by "Education" and "Business" early on, they were particularly interested in querying "Medicine". They typically spent between a couple of minutes to an hour using the corpus, occasionally idling the corpus for up to three hours. The logs also show longitudinal development in the corpus functions used, with their earlier queries typically involving the *frequency breakdown* function to derive new vocabulary (and recycling these data for new queries), before primarily using the *collocation* function in later queries. They were also one of the returnees to the corpus beyond the taught component of the course, continuing to use the corpus to resolve language issues for their assignments for other courses.

The Search Guru (see Table 7) typically spent less than 40 minutes using the corpus, yet quickly become competent in using all possible corpus functions and all available filtering options. The user found all three functions useful in their own right, in contrast with the Persistent User who used a comparatively limited range of query forms and functions. The range of DDL tasks appears sufficient for some users to fully explore each of the corpus platform's functions, and the length of training required appears relatively short for users of this type.

The Quitter (see Table 8) rejected using the corpus after their first real session with it, perhaps turned off by the in-class activities or the corpus platform. However, based on the time and duration of their later visits, we assume that the user may have had other assignments due for

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 7. The search guru (Rank 4)

Date	Activity
28/09/17	Spent 12 minutes exploring "this research" in the introduction and "this study" in the literature review within Education. All three functions as well as wildcards were used.
29/09/17	Checked "there" and "help" in the morning, returning to check collocations of "integration" and "semantic" in the literature review within Education.
30/09/17	Spent 40 minutes using a combination of wildcards and POS tags; e.g. "learn *_n" in the methodology. Frequently filtered results via Information and Technology Studies.
01/10/17	Briefly queried the terms "action" and "actionable" within Education.
10/10/17	Quickly checked examples of "learning" and "support" in the abstract section.
18/10/17	Quickly queried "entangled" in the Social Sciences. First attempt to sort concordances by left context.
31/10/17	Returned after a lengthy break to find collocations of "provoke" in Education and the Social Sciences.
07/11/17	Spent 26 minutes querying cohesive devices "therefore", "as a result", and "as a consequence", frequently sorting the resulting concordances by left context.
15/11/17	Briefly queried words related to "pedagogy".
16/11/17	Spent 30 minutes querying "due to" and "in which" using the collocation, frequency breakdown, and frequency distribution functions.
16/12/17	Returned a month later querying "principal research question".
18/12/17	Spent 20 minutes querying "on the one hand" and "on the other hand" in the literature review, and used collocation, frequency breakdown, and frequency distribution for "as a result".

Table 8. The quitter (Rank 7)

Date	Activity
13/10/17	Explored the three major functions (i.e. collocates, frequency breakdown, and frequency distribution) for the term "my". Sorted concordances by left and right context.
09/12/17	Returned after two months to spend 26 minutes extensively querying "little" and "fill" within Education, using collocation, frequency breakdown, and frequency distribution functions
12/12/17	Spent three hours querying items including "thorough", "introduction", "review", "e.g.", and "research question", but without filtering queries by section or faculty.
13/12/17	Checked random queries including "implication", "contribution", and "session" within the literature review, reading only concordances.

mid-December. At this point, when facing a language problem that they were unable to resolve by other means, they returned to use the corpus extensively to resolve their issue.

3.2 Discipline-specific corpus use

Table 9 outlines the frequency of corpus queries filtered by specific faculty. Filtered queries by specific faculty/faculties (including "all" faculties) made up roughly 39% of all queries on the HKGC platform.

Overall, there is greater use of the corpus in the sciences as compared with the arts and humanities/social sciences, with medicine and engineering alone accounting for almost 50% of the total

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 9. Queries filtered by faculty[1]

Faculty/Faculties	Queries	Proportion by faculty (in this list only)	Proportion of all HKGC queries
Medicine	1,098	24.6%	9.6%
Engineering	1,032	23.1%	9.0%
Education	819	18.4%	7.2%
All*	517	11.6%	4.5%
Arts	208	4.7%	1.8%
Dentistry	173	3.9%	1.5%
Architecture	144	3.2%	1.3%
Law	139	3.1%	1.2%
Business and economics	125	2.8%	1.1%
Education + Social science	111	2.5%	1.0%
Medicine + Science	92	2.1%	0.8%

[1]Not all selected options are shown.

queries, although this is in line with student enrolment ratios. However, the proportion of queries for the arts and humanities/social sciences is higher than their 27% enrolment ratio, accounting for around 34% of queries in all. Regarding queries filtered by specific subdisciplinary groups, queries involving the science subdisciplines (computer science, mechanical engineering, public health, physics, etc.) are much more frequent than those involving arts and humanities and social sciences, accounting for each of the top 10 queried subdisciplines. The next disciplines outside the sciences are a broad range of subdepartments within the Faculty of Education, including the Division of Chinese Language and Literature, the Division of English Language Education, and the Division of Information and Technology Studies. These account for 178 filtered searches, or just 1.60% of all HKGC queries. This does not reflect the enrolment distribution for arts and humanities/social sciences majors of 27%, leaving queries by subdiscipline heavily skewed towards the sciences.

There is also a degree of variation across disciplines in terms of the kind of information users wish to receive from the platform (see Figure 7). Although the frequency of basic searches is the highest across all disciplines[1], those searching within the arts or education subcorpora tend to conduct simple searches that provide only concordance output more often than those from other disciplines, and do not make frequent use of the other corpus functions involving frequency or collocation. Users querying the architecture or engineering subcorpora frequently use the frequency breakdown function to select the correct word or phrase from the list of options this function provides. Users querying the law subcorpus more frequently search for collocates of query terms than those in arts or education, while users querying the education and social sciences subcorpora frequently sorted the output concordance context to the left of the query term than users querying other subcorpora. The best way to summarise this variation is that those in the physical/life sciences prefer to query the corpus for statistical information, whereas those in arts and humanities and social sciences disciplines prefer to query the corpus for textual information, although we cannot provide inferential statistics for this claim due to the way the data are structured.

[1]This is because a basic search must be carried out first before another function can be employed.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

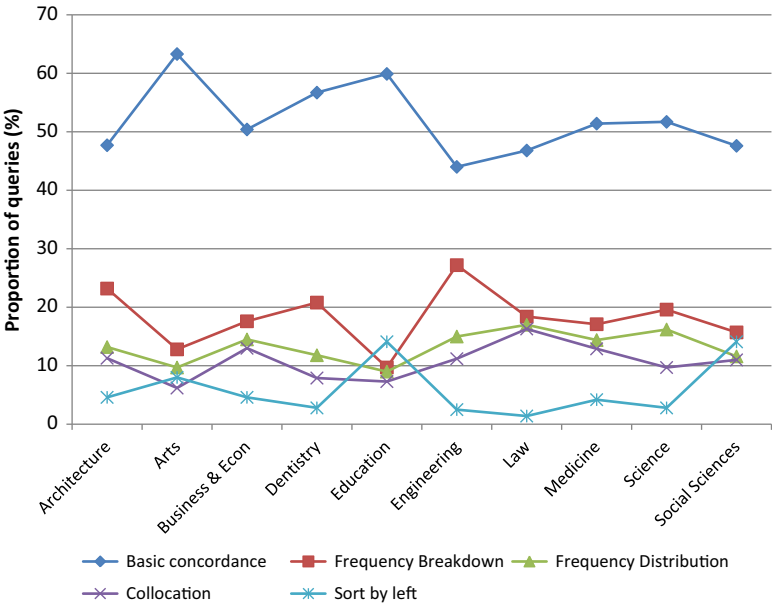


Figure 7. Proportional use of query functions employed for queries filtered by faculty

The data also suggest interdisciplinary variation in the query syntax and lexis used by those filtering their searches by faculty (see Table 10). None of the items in the table were present in the DDL activities.

Discipline-specific lexis is present in queries involving the faculties of architecture, business, education, engineering, and social sciences, as well as wildcards and POS tags within disciplinary searches, as students presumably used the corpus to aid them with other ongoing assignments in their disciplinary courses. For example, those searching within medicine and science were interested in phrases involving *questionnaires*, likely because they were soon to conduct an assessment requiring one. However, it is possible that some of the tasks requiring disciplinary searches may have influenced the search behaviour shown in Table 10.

4. Discussion and conclusion

The present study is a comprehensive analysis of the corpus usage characteristics of students engaged in DDL for disciplinary thesis writing. Regarding RQ1, a range of analytics have been presented outlining how postgraduate students engage with corpora in terms of their actual corpus usage, query function preferences, and query syntax for DDL. The detailed cohort and individual activity logs allowed us to determine users' attempts at using the corpus, as well as the modifications they made to their corpus query habits and usage over time. We are encouraged by the short length of time required for corpus uptake via our unguided approach to the DDL materials, as previous DDL research in the same context (Crosthwaite, 2017) and Pérez-Paredes *et al.*'s (2011) study relied on explicit teacher-led guidance on corpus consultation. The DDL students also often went beyond the provisions of their assigned tasks to freely experiment under their own autonomy, following Hafner and Candlin (2007), where corpus consultation became an "integrated and self-directed part of the students' ... writing process, unconstrained by the imposition of specific data-driven learning tasks set by the teacher" (p. 306). The user-friendly and visual nature of the HKGC platform's user interface and the structured, inductive focus-on-form approach taken to the corpus materials design appears to have facilitated sustained

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Table 10. Top five unique queries by faculty (cut-off frequency = 5)

Faculty	Query syntax	Freq.	Faculty	Query syntax	Freq.
Architecture	cultural landscape	16	Engineering	describe	47
	paradigm	12		possibility	28
	landscape	7		machine learning	28
	show_v	7		novel	18
	"garden City"	6		simulations	7
Arts	show*	12	Law	show *_v	13
	clean	6		include*	9
				data have shown	7
				thesis	6
Business	marketing	16	Medicine	no studies	24
	*margin	8		it is hoped that	23
	blockchain	8		the questionnaire *_v	22
				consum*	18
				limited studies	16
Dentistry	describe	20	Science	describe	34
	hypothesis	6		possibly due	25
				no studies	14
				hopefully	10
				The *of the questionnaire	7
Education	understanding	24	Social	tackle	13
	learn *_n	16	sciences	gender	9
	emotion	11		tourism	9
	show*	10		entangled	6
	evaluation_n	9		gap	6

and autonomous corpus use both during and after the writing course for a large number of users, although not everyone was involved.

However, there is still a need for studies investigating the longitudinal effects of corpora and DDL on writing or language development in the medium to long term (Luo, 2016), with most DDL studies (including ours) lasting for a semester at most. It would also be helpful to extend the study to explore the use of the corpus during actual drafting and revision stages rather than focusing on completion of the DDL activities within the coursebook. Claims regarding DDL's impact on writing development (e.g. accuracy, complexity, fluency), while valid within the context of individual studies, do need to be supported with further evidence from longer-term studies. The analytics from the corpus platform are still being collected as the thesis writing courses continue with new cohorts; at the time of writing, the current number of unique users since the platform was launched now stands at 512, with 1,882 accumulated platform visits, 4,402 unique queries, and 21,479 accumulated searches across two semesters of instruction. We intend to revisit these usage statistics after three full years of implementation so as to determine with greater validity and power how corpus users are engaging with our platform and how this may be affecting their writing.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Regarding RQ2 on the extent of disciplinary variation in the uptake and usage of corpora for postgraduate DDL using our multidisciplinary corpus platform, we have determined interdisciplinary variation in the usage of particular corpus functions and query syntax, such as the disproportional frequency of queries filtered within the arts and humanities/social sciences disciplines and those filtered within the physical/life science subdisciplines, as well as whether students accessed domain- or language-specific knowledge outside of the course materials. Our materials took a one-size-fits-all approach, but future iterations of the corpus materials need to ensure that the corpus platform and activities are more flexible in meeting the needs of students across different disciplines in terms of the specific issues (disciplinary domain, or language/composition) they may be having with their writing. It may also be necessary in future to provide more activities that make use of particular corpus platform functions in line with some of the disciplinary preferences we have outlined in this research. We also see the need to move beyond corpus activities that focus solely on "low-level" phenomena involving grammar and lexis, with corpora that facilitate analysis of "higher-level" phenomena at the discourse or genre level (Boulton, Carter-Thomas & Rowley-Jolivet, 2012: 3). While our course materials feature both "low" and "high" level analysis by combining corpus tasks with non-corpus-based activities on larger sections of text, the corpus tasks and platform only currently facilitate queries for local grammar and disciplinary lexis. As we are now beginning to see the use of genre-annotated corpora for DDL (e.g. Cotos, Link & Huffman, 2017), a logical extension would be to create discipline-annotated corpora and employ these for DDL.

In terms of areas for future research, there is a need to triangulate our findings with questionnaire and interview data regarding student and teacher perceptions of the DDL materials, corpus platform, and the effectiveness of DDL for disciplinary writing. Such data is important in understanding why 20% of enrolled students never actually used the platform at all, and why a number of students refrained from using the corpus after a short period. We did in fact collect extensive accompanying survey and interview data from participants during the study period, but these data will necessarily be the subject of a forthcoming article. A potential limitation of this study lies in our interpretation of corpus query logs and usage histories as indicative of students' actual intentions and processes behind their corpus use, rather than any online procedure. Future studies may combine our tracking parameters with automatic screen recording or EEG devices to triangulate students' actions and cognitive processes during DDL. Another limitation is that our platform's data structure requires modification to allow for inferential statistics regarding cross-discipline analyses. As one anonymous reviewer of the paper also suggested, it might also be useful to look at the breakdown of corpus queries and usage per activity type so as to see expected search behaviour versus triggered search behaviour. Running multiple course groups at any one time during the semester meant that many students were conducting different corpus activities in real time. As the platform only records queries, ID, and filter analytics, matching queries to corpus activities would have to be done manually, which is impossible given the data structure of user analytics, the multiple class groups, and the high number of corpus searches. Further studies investigating expected search behaviour versus triggered search behaviour for individual tasks are therefore required.

Nonetheless, we are still confident the findings of this study represent the largest and most detailed insight into students' disciplinary corpus use for DDL while highlighting the affordances of such use for those looking to implement DDL into their own practice. It is not just a case of "users vary" – by understanding individual and disciplinary variation in corpus usage, we can do better in designing corpus tools and DDL materials that work for everyone.

Supplementary materials. To view supplementary material for this article, please visit https://doi.org/10.1017/S0958344019000077

Ethical statement. All aspects of this study were approved by the Human Ethics Research Committee of the University of Hong Kong, with all participants informed that their data would be used for research purposes.

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Funding statement. The project was supported by a Teaching Development Grant awarded by the University of Hong Kong for the period of 1st September 2016 to 31st December 2017 (Project number: 101000623). Lillian Wong was the Principal Investigator, Peter Crosthwaite was one of the two co-investigators and Joyce Cheung was the Research Assistant.

References

Anthony, L. (2014) *AntConc (Version 3.4.4)*. Tokyo: Waseda University. http://www.laurenceanthony.net/software

Anthony, L. (2017) *AntFileConverter (Version 1.2.1)*. Tokyo: Waseda University. http://www.laurenceanthony.net/software

Baisa, V., and Suchomel, V. (2014) SkELL: Web interface for English language learning. In Horák, A. & Rychlý, P. (eds.), *RASLAN 2014: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 63–70). Brno: NLP Consulting.

Boulton, A. (2015) Applying data-driven learning to the web. In Leńko-Szymańska, A. & Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 267–295. https://doi.org/10.1075/scl.69.13bou

Boulton, A., Carter-Thomas, S., and Rowley-Jolivet, E. (eds.) (2012) *Corpus-informed research and learning in ESP: Issues and applications*. Amsterdam: John Benjamins.

Boulton, A., and Cobb, T. (2017) Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2): 348–393. https://doi.org/10.1111/lang.12224

Centre for Applied English Studies (2017) Introduction to Thesis Writing. Hong Kong: The University of Hong Kong.

Chambers, A. & O'Sullivan, Í. (2004) Corpus consultation and advanced learners' writing skills in French. *ReCALL*, 16(1): 158–172. https://doi.org/10.1017/S0958344004001211

Charles, M. (2007) Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4): 289–302. https://doi.org/10.1016/j.jeap.2007.09.009

Charles, M. (2014) Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, 35: 30–40. https://doi.org/10.1016/j.esp.2013.11.004

Charles, M. (2015) Same task, different corpus: The role of personal corpora in EAP classes. In Leńko-Szymańska, A. & Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 131–153. https://doi.org/10.1075/scl.69.07cha

Chen, M., and Flowerdew, J. (2018) Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes*, 50: 97–112. https://doi.org/10.1016/j.esp.2017.11.004

Cobb, T., and Boulton, A. (2015) Classroom applications of corpus analysis. In Biber, D. & Reppen, R. (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 478–497. https://doi.org/10.1017/CBO9781139764377.027

Cotos, E. (2014) Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2): 202–224. https://doi.org/10.1017/S0958344014000019

Cotos, E., Link, S., and Huffman, S. (2017) Effects of DDL technology on genre learning. *Language Learning & Technology*, 21(3): 104–130.

Crosthwaite, P. (2017) Retesting the limits of data-driven learning: Feedback and error correction. *Computer Assisted Language Learning*, 30(6): 447–473. https://doi.org/10.1080/09588221.2017.1312462

Flowerdew, L. (2015) Data-driven learning and language learning theories: Whither the twain shall meet. In Leńko-Szymańska, A. & Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 15–36. https://doi.org/10.1075/scl.69.02flo

Flowerdew, J. (2016) English for specific academic purposes (ESAP): Making the case. *Writing & Pedagogy*, 8(1): 5–32. https://doi.org/10.1558/wap.v8i1.30051

Frankenberg-Garcia, A. (2005) A peek into what today's language learners as researchers actually do. *International Journal of Lexicography*, 18(3): 335–355. https://doi.org/10.1093/ijl/eci015

Gaskell, D., and Cobb, T. (2004) *Can learners use concordance feedback for writing errors? System*, 32: 301–319. https://doi.org/10.1016/j.system.2004.04.001

Hafner, C. A., and Candlin, C. N. (2007) Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, 6: 303–318. https://doi.org/10.1016/j.jeap.2007.09.005

Hyland, K. (2000) Disciplinary discourses: Social interactions in academic writing. Harlow: Longman.

Johns, T. (1991) Should you be persuaded: Two examples of data-driven learning materials. In Johns, T. & King, P. (eds.), *Classroom concordancing: English Language Research Journal 4*. Birmingham: Centre for English Language Studies, University of Birmingham, 1–16.

Kilgarriff, A., and Grefenstette, G. (2003) Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3): 333–347. https://doi.org/10.1162/089120103322711569

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004) The Sketch Engine. *Information Technology*, 105: 116–127.

Lee, D., and Swales, J. (2006) A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1): 56–75. https://doi.org/10.1016/j.esp.2005.02.010

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press

Lee, H., Warschauer, M., and Lee, J. H. (2018) The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amy012

Leńko-Szymańska, A., and Boulton, A. (eds.) (2015) *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.69

Long, M. H. (1991) Focus on form: A design feature in language teaching methodology. In de Bot, K., Ginsberg, R. B. & Kramsch, C. (eds.), *Foreign language research in cross-cultural perspective*. Amsterdam: John Benjamins, 39–52. https://doi.org/10.1075/sibil.2.07lon

Luo, Q. (2016) The effects of data-driven learning activities on EFL learners' writing development. *SpringerPlus*, 5(1): 1255. https://doi.org/10.1186/s40064-016-2935-5

Millar, N. (2011) The processing of malformed formulaic language. *Applied Linguistics*, 32(2): 129–148. https://doi.org/10.1093/applin/amq035

Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J. M., and Jiménez, P. A. (2011) Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, 24(3): 233–253. https://doi.org/10.1080/09588221.2010.539978

Schmidt, R. W. (1990) The role of consciousness in second language learning. *Applied Linguistics*, 11(2): 129–158. https://doi.org/10.1093/applin/11.2.129

Steel, C. (2012) Fitting learning into life: Language students' perspectives on benefits of using mobile apps. In Brown, M., Hartnett, M. & Stewart, T. (eds.), *Future challenges, sustainable futures. Proceedings ASCILITE*. Wellington: Massey University, 875–880.

Steel, C. H., and Levy, M. (2013) Language students and their technologies: Charting the evolution 2006–2011. *ReCALL*, 25(3): 306–320. https://doi.org/10.1017/S0958344013000128

Widmann, J., Koh, K., and Ziai, R.(2011) The SACODEYL search tool: Exploiting corpora for language learning purposes. In Frankenberg-Garcia, A., Flowerdew, L. & Aston, G. (eds.), *New trends in corpora and language learning*. London: Continuum, 167–178.

Yoon, H. (2008) More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology*, 12(2): 31–48.

Yoon, H., and Hirvela, A. (2004) ESL student attitude toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4): 257–283. https://doi.org/10.1016/j.jslw.2004.06.002

About the authors

Peter Crosthwaite is a lecturer in applied linguistics at the School of Languages and Cultures, The University of Queensland, Australia. His research interests focus on English for general and specific academic purposes, second language writing, corpus linguistics, and the direct use of corpora for education purposes. He has published widely on these issues in a number of leading journals.

Lillian L. C. Wong is a senior lecturer in the Centre for Applied English Studies at The University of Hong Kong. She researches innovation and change in English language education, Information and Communication Technologies, teacher professional development and English for Academic and Specific Purposes. She is Chair of the Research Professional Council (2019-2020), served on the Board of Directors (2012-15) and was the Professional Development Committee Chair (2005-06) of TESOL International Association.

Joyce Cheung is a research associate at The Hong Kong Polytechnic University, working on an engineering-specific, corpus-driven academic English enhancement program. She holds a master's degree in English language studies and a bachelor's degree in communication. Her interest is to apply a corpus-linguistic approach to discourse analysis to study culture.

Author ORCiD. Peter Crosthwaite, https://orcid.org/0000-0002-1482-8381
Author ORCiD. Lillian Wong, https://orcid.org/0000-0002-0141-5908
Author ORCiD. Joyce Cheung, https://orcid.org/0000-0002-2110-5720

https://doi.org/10.1017/S0958344019000077 Published online by Cambridge University Press