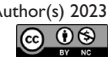
Article

LANGUAGE
TEACHING
RESEARCH

Enhancing student writing with do-it-yourself corpora on a PhD pre-sessional programme

Language Teaching Research
1–18



DOI: 10.1177/13621688231221350
journals.sagepub.com/home/ltr


Dana Therova
Durham University, UK

Andrew McKay
Durham University, UK

Abstract

As an effective writing course should focus on disciplines and their unique characteristics, practitioners of English for academic purposes (EAP) are often faced with the challenge of addressing the different needs of learners from various fields of study. This article reports on how a data-driven learning (DDL) approach can be applied to enhance student written production in a multidisciplinary classroom in a 10-week PhD pre-sessional programme at a British University. The participants were six international students who used a do-it-yourself (DIY) corpus in weekly DDL sessions to familiarize themselves with discipline-specific academic writing conventions and applying them in their writing. The effectiveness of this approach was investigated through a 'talk around texts' technique employed in semi-structured interviews with individual students and their supervisors on programme completion. The findings show that a DDL approach utilizing a DIY corpus has the potential of enhancing PhD student writing in a multidisciplinary classroom on a pre-sessional programme. This article suggests that DDL could be successfully implemented not only in PhD pre-sessional programmes, but also in wider EAP contexts.

Keywords

academic writing, data-driven learning, do-it-yourself corpus, pre-sessional programme, technical vocabulary

Corresponding author:
Dana Therova, Durham Centre for Academic Development (DCAD), Durham University, Teaching and Learning Centre, Durham, DH1 3LU, UK.
Email: dana.therova@durham.ac.uk

I Introduction

Language corpora (i.e. large collections of electronic texts) have played a significant role in language learning and teaching for decades. One application of corpora was given prominence by Johns (1990), who coined the term 'data-driven learning' (DDL) to describe an approach to language learning in which learners act as researchers exploring samples of language obtained from a corpus. This approach is underpinned by the assumption that effective language learning occurs through discovery whereby corpus data offer a unique resource for self-directed explorations of authentic language. This makes DDL a learner-centred approach utilizing inductive learning strategies (Johns, 2012). These characteristics of DDL make this a valuable approach to teaching writing in contexts of English for academic purposes (EAP), where practitioners must often navigate between a wide range of genres and writing conventions representing various disciplines as learners wish to acquire the specialized terminology of their field of study.

However, it may be challenging for EAP practitioners, who are unlikely to be specialists in the target field, to address the learners' needs, particularly in contexts where a writing course is taken by a heterogeneous group of learners from diverse disciplines (Anthony, 2016). The value of a DDL approach in EAP settings thus lies primarily in the learners' ability to conduct autonomous language searches drawing on a corpus relevant to their learning contexts. This approach is particularly valuable in discipline-specific writing courses as it caters for students from various academic domains and their specific learning needs without the need for EAP practitioners being discipline specialists (Anthony, 2019).

II Literature review

There is a substantial body of research into DDL encompassing course design, use, and effectiveness. This can usefully be divided into two areas. First, qualitative studies focusing on student perceptions of the efficacy of DDL (e.g. Chambers, 2005; Chang, 2014; Charles, 2012, 2014, 2022; Charles & Hadley, 2022; Geluso & Yamaguchi, 2014; Mizumoto et al., 2016; Sun, 2007; Therova & McKay, 2022; Yoon & Hirvela, 2004). Second, quantitative studies investigating student production of linguistic features associated with academic writing (e.g. Ackerley, 2017; Boulton, 2009; Bridle, 2019; Cotos, 2014; Daskalovska, 2015; Friginal, 2013; Geluso & Yamaguchi, 2014; Lay, 2020; Li, 2017; Mizumoto & Chujo, 2015, 2016; Shin et al., 2018; Smith, 2020; Wu, 2021). As a quantitative analysis of student written production is not part of this study, this review will focus on qualitative research on the effectiveness of DDL in EAP instruction in a university context.

Yoon and Hirvela (2004), using a survey and follow-up interviews, investigated 23 second language (L2) students' perceptions of an academic writing course that incorporated the use of the Collins COBUILD corpus. The participants were primarily from China and Korea and studying technical disciplines at a major US university. Students were very positive about the effectiveness of corpus use in developing their academic writing skills and increasing their confidence. Student perception of the effectiveness of a DDL approach was also part of Chambers' (2005) study into the use of corpora by

14 undergraduate and postgraduate students on language courses in English, French, German, Irish, and Spanish. Students received training in consulting small corpora compiled by the course designers and were encouraged to use them to analyse their own writing. Whilst the participants were more familiar with a traditional deductive approach using dictionaries and grammar books for reference purposes, they were generally positive about the value of corpora. Specifically, they commented on the authenticity of the language and the motivational aspects of an inductive approach. They did, however, also note the amount of training necessary for effective use of corpora. Sun (2007), using questionnaires and interviews, investigated the reactions of 20 doctoral students in technical disciplines at a research university in Taiwan to an academic writing template incorporating a discipline-specific concordance tool to support research article writing. Student response to the tool was highly positive. The main reported benefits were noted in areas such as sentence structure, organization, signposting, and word choice. Positive student evaluations of corpus use were also reported in Chang's (2014) case study of corpus use by five master's and five doctoral engineering students at a Korean university, and in Flowerdew's (2015) study on the effectiveness of a workshop course on corpora use for postgraduate engineering and science students in Hong Kong.

It should be noted that the above studies into student perceptions of the value of a DDL approach used either generally available reference or designer/instructor compiled corpora. These studies thus neglect the value of self-compiled do-it-yourself (DIY) corpora, which lies primarily in their direct relevance to the learners' needs. The value of self-compiled corpora has been the subject of several studies. Lee and Swales (2006) conducted what they termed an 'experiment' in 'technology enhanced rhetorical consciousness raising' by training four doctoral students in the use of concordance tools. The participants were introduced to existing specialized corpora, then compiled two more corpora: one of their own writing and the other of 'expert' writing in their own field of study. During the 13-week course students investigated linguistic features such as language patterns, reporting verb use, and active and passive voice use. In interview, students reported that they felt using corpora helped them in recognizing disciplinary variations in language use, built their confidence, and that they enjoyed the greater autonomy it afforded them.

Following Lee and Swales' (2006) rationale, Charles (2012) conducted a larger study of 50 graduate students into the value of do-it-yourself discipline-specific corpora. Using a questionnaire, participants were asked about their use of the corpus, whether they would use it in the future, and its advantages and disadvantages. The vast majority (90%) of students reported that they found it easy to build their own corpus, felt it helped their writing, and that they would would consult it in the future. A year later, a follow-up study on long-term use found that 70% of respondents had used their corpus. Out of these participants, 38% used it once a week or more, and 33% once a month or less (Charles, 2014). Encouraging though these figures appear, a larger study of 182 graduate students spanning 9 years found that although 63% of students used their corpus regularly after completing a six-week course, this had decreased to 36% after one year (Charles, 2022).

Self-compiled corpora were also part of a large-scale project with 473 postgraduate research students in six Hong Kong universities where students were introduced to corpus use in 20 3.5-hour workshop sessions (Chen & Flowerdew, 2018). Student

responses to an evaluation questionnaire were highly positive with 95.56% of respondents stating they would recommend the workshop to others. Our earlier, albeit small scale report on the first iteration of our course also found students were very positive about the benefits of their self-compiled corpus, particularly for acquiring discipline-specific vocabulary, usage, and error correction. They also commented on the ease of finding relevant results from a self-compiled corpus as opposed to a larger general corpus (Therova & McKay, 2022).

In summary, the above reviewed studies of the effectiveness of DDL show that a DDL approach results in positive student outcomes. However, most studies into student perceptions of the value of a DDL approach used either generally available reference or designer/instructor compiled corpora and students were already embedded within their departments. Further, whilst the student responses in previous studies are encouraging for proponents of DDL use, it might be argued that previous studies on the effectiveness of DDL approaches have not considered the views of supervisors. Given their role in assessing student work we believe their opinions are valuable. This is particularly important in the context of PhD pre-sessional courses given that the purpose of these courses is to adequately prepare students who have not reached a standard to allow them to directly enrol for PhD study. Despite the diversity of previous studies into the implementation of DDL in various contexts, there is a lack of studies investigating DDL in the context of PhD pre-sessional programmes. We thus believe that further research in this area is justified, particularly with a cohort of doctoral students whose entry to the institution is conditional on successful completion of the programme. Therefore, this study seeks to address the following research questions:

- Research question 1: What are the students' perceptions of a DDL approach using self-compiled DIY corpora for the investigation of disciplinary academic writing conventions on a PhD pre-session programme?
- Research question 2: To what extent do international PhD students find a DDL approach effective in informing their written production?
- Research question 3: What is the PhD supervisors' evaluation of their students' writing produced on completion of a DDL-assisted pre-sessional programme?

III Methodology

1 Context

The present study is set in the context of a pre-sessional EAP programme for PhD students at a British university located in the North-East of England. Typically, the University offers pre-sessional courses for international undergraduate and postgraduate students from various disciplines who do not meet the requirements for direct entry to the University's degree programmes. These pre-sessional programmes run over a period of 6, 10 and 20 weeks. Following a successful introduction of a pre-sessional programme for students aspiring to pursue doctoral study at the University in the summer of 2021, the PhD pre-sessional programme was further developed and repeated in the summer of 2022. The programme was developed by the authors and delivered by the

first author. Following the Covid-19 pandemic, the programme took the mode of distance learning and had the following aims:

- to improve students' level of English proficiency;
- to further develop their transferrable academic skills;
- to raise their awareness of British academic conventions and academic culture;
- to develop their subject-specific and topic-specific knowledge relating to their PhD research area; and
- to prepare them for other aspects of a PhD life (e.g. their role as student researchers, the required level of independence, managing the PhD process) in order to prepare them for the demands of doctoral study.

To reflect the specific disciplinary needs of the students, the programme was designed in consultation with the students' PhD supervisors, who supplied discipline-specific sources in the form of reading texts and pre-recorded lectures. These sources were utilized by the students throughout the programme for acquisition of discipline-specific vocabulary, disciplinary academic writing conventions as well as knowledge and concepts relevant to the students' area of research.

2 Participants

Prior to approaching potential participants, ethics approval was obtained from the University's Human Research Ethics Committee, followed by seeking informed consent from all participants in this study. In the summer of 2022, six international students attended the 10-week online PhD pre-sessional programme and agreed to participate in the current study. The students (5 female and 1 male) were aged 25–40 years ($M = 35$, $SD = 5.3$) from three nationalities: Saudi Arabia ($n = 4$), China ($n = 1$) and Turkey ($n = 1$); and two disciplines: Computer Science ($n = 3$; Students A, B, C) and Mathematical Science ($n = 3$; Students D, E, F). Despite only two disciplines divided equally between the participants, one of the challenges of the programme was to cater for these two disciplinary backgrounds. This is because academic writing practices are not universal across disciplines and academic literacy needs of students thus vary reflecting the differences in disciplinary writing conventions. A further challenge was to enable the students to expand their topic-specific knowledge relating to their varied and very individual research areas. This can be particularly challenging for EAP practitioners who are often not specialists in the students' target fields and may thus feel ill-equipped to cope with the specialized terminology of the students' subject domain. Given the very specific nature of the students' research area and that submitting a written thesis is a prerequisite of a PhD award, meeting individual students' academic literacy needs was therefore of vital importance in the context of the PhD pre-sessional programme. This was intended to be achieved by enabling the students to work with topic-specific content through a data-driven approach to learning (DDL) utilizing the students' self-compiled do-it-yourself (DIY) corpora.

In addition to student participants, the students' PhD supervisors also agreed to take part in this study. In total, there were five supervisors among whom the supervision of the six student participants was shared (with three students being supervised jointly by

two supervisors). Supervisor A is a Full Professor of Computer Science, has been supervising for 6 years, and currently has 3 supervisees. Supervisor B is a Full Professor of Computer Science, has been supervising for 17 years, and currently has 12 supervisees. Supervisor C is a Full Professor of Mathematical Sciences, has been supervising for 30 years, and currently has 21 supervisees. Supervisor D is an Associate Professor, has been supervising for 6 years, and currently has 18 supervisees. Supervisor E is an Assistant Professor, has been supervising for 4 years, and currently has 15 supervisees.

3 Data-driven learning

a Do-it-yourself (DIY) corpus. In the first DDL session in week 1 of the programme, the students were introduced to the concept of 'corpus' (i.e. a collection of texts in an electronic format) and built their own discipline-specific corpus containing reading sources relating to their research topic. This meant that the input materials were different due to the varied topics of the proposed theses, and this was reflected in the composition of the students' individual corpora. The recommended types of sources to include in their DIY corpus included journal articles, e-books, and PhD e-theses from their discipline. The aim of this was to enable each student to work with their own corpus of authentic texts related to their research topic which could be interrogated for the academic writing conventions in their disciplines. This introductory corpus-building session resulted in each student having their own personal discipline-specific corpus ranging between 1 and 4 million words in size ready for subsequent DDL sessions focusing on various aspects of academic writing in their fields of study, including discipline-specific terminology.

b #LancsBox. The corpus-building was followed by an introduction to the corpus software. For the purpose of the DDL sessions, #LancsBox (Brezina et al., 2020) was utilized on the programme. This corpus tool was selected for its desktop-based user-friendly interface, a range of functionalities providing useful insights into texts, and its ability to work with self-compiled corpora containing files in different file formats. The introduction to #LancsBox comprised a tutor demonstration of the tool and its functionalities relevant to the purposes of the programme, followed by series of online tutorials made available by the #LancsBox developers (Brezina et al., 2020), which the students watched as part of their independent study time outside of class time.

The main #LancsBox features used on the pre-sessional programme included the Key Word In Context (KWIC) function generating a list of all instances of a search term in a corpus in the form of a concordance, which can subsequently be sorted or filtered to obtain the desired output. This feature can be used to search for individual words or phrases as well as grammatical categories such as nouns, verbs, or adverbs in the form of 'smart searches,' which can also include searches for complex linguistic structures such as passives or noun phrases. In addition to KWIC, the GraphColl feature was used, which generates collocates (i.e. words which systematically co-occur) of the search term, identifies colligation (i.e. co-occurrence of grammatical categories), visualizes collocations and colligations and identifies shared collocates of a word or phrase. Following an introduction to #LancsBox and a demonstration of its selected features, this tool was utilized in weekly DDL workshops throughout the programme.

c Data-driven learning workshops. The DDL approach was applied by drawing on the students' self-compiled DIY corpora in recurring weekly sessions over eight weeks running from week 2 to week 9 of the course (excluding week 1 which was an introductory week, and week 10 which was an assessment week). These weekly sessions were 90 minutes long and included a series of practical workshops comprising Listening Workshops, Reading Workshops and Vocabulary Workshops. The aim of these workshops was to develop the students' topic-specific knowledge, to expose them to writing conventions in their disciplines, to expand their repertoire of technical (i.e. discipline- and topic-specific) vocabulary, and to become familiar with the usage of these vocabulary items in their specific disciplinary contexts. Due to the distance learning nature of the programme, these workshops took the form of flipped learning whereby the students prepared for these workshops in advance of the sessions by completing several tasks, as follows:

Ahead of the Listening Workshops the students watched a pre-recorded lecture supplied by their PhD supervisor, and they then prepared to share the following with the class:

- a written summary of the content of the lecture (of no more than 200 words);
- what they had learned from the lecture in terms of discipline-specific and topic-specific concepts;
- how this can inform their own research;
- what they found particularly interesting/difficult/challenging (e.g. in terms of the content/subject knowledge, language use, delivery);
- a list of 3–5 new words which they learned from the lecture.

Similarly, prior to the Reading Workshops the students had read one of the sources supplied by their supervisors or another source from their discipline-specific corpus and had prepared to share the following with the class:

- a written summary of the content of the reading text (of no more than 200 words);
- what they had learned from the source text in terms of discipline-specific and topic-specific concepts;
- how this can inform their own research;
- what they found particularly interesting/difficult/challenging (e.g. in terms of the content/subject knowledge, language use, delivery);
- a list of 3–5 new words which they learned from the lecture.

In addition to this, the Reading Workshops were also used for exploration of various features of academic writing covered each week using #LancsBox. The linguistic features explored during these sessions included noun phrases, academic tone, hedging and boosting, passive voice, or tense, for instance.

The Listening and Reading Workshops served as a basis for the subsequent Vocabulary Workshops, during which the students explored the vocabulary items acquired from the lectures and reading texts. In preparation for the Vocabulary Workshops, the students considered various aspects of the newly acquired vocabulary items. These included the meaning of the words, which they would look up using an online dictionary such as

Longman dictionary of contemporary English online (Pearson, 1996–2022), *Oxford learner's dictionaries* (Oxford University Press, 2022) or *Cambridge dictionary* (Cambridge University Press, 2022) for general academic words complemented by discipline-specific online dictionaries such as *A dictionary of computer science* (Oxford University Press, 2016), *A dictionary of statistics* (Oxford University Press, 2014) or *The concise Oxford dictionary of mathematics* (Oxford University Press, 2021) for discipline-specific terms. However, the primary focus of the Vocabulary Workshops was on the investigation of the unfamiliar words through concordance lines using the KWIC function and their collocations using the GraphColl function in #LancsBox. Following this discovery learning, the students used the newly acquired vocabulary items in sentences to practise productive usage of these words.

It is noteworthy that while the Listening and Reading Workshops focused primarily on content and topic knowledge relating to the students' research areas, with the Reading Workshops additionally exploring various aspects of disciplinary academic writing, the Vocabulary Workshops were aimed at expanding the students' receptive and productive vocabulary knowledge. These sessions were hence in line with the main characteristic of the DDL approach to learning where learners act as researchers drawing on language data supplied by a corpus (Johns, 1986, 1991, 2012). Further, the Listening, Reading and Vocabulary Workshops intended to promote facilitative learning whereby the students provided the lesson content, and the tutor's primary role was in giving feedback on their learning and guidance on future development and application of new knowledge.

4 Data collection

To address the research questions investigating a DDL approach using students' personal DIY corpora for the investigation of disciplinary academic writing and its effectiveness in informing their written production, the present study was motivated by the Academic Literacies model (Lea & Street, 1998) which places emphasis on exploring student writing beyond their texts by focusing on the nature of academic writing practices in various disciplinary contexts (Lea & Street, 1998, 2006). Academic Literacies thus aims to understand student writing by taking into consideration the complex nature of writing practices at universities with broader institutional discourses. Accordingly, insights into the nature of academic writing are often gained by exploring the understanding that both academic staff and students have regarding their own literacy practices. This is enabled by drawing on ethnographically oriented data as the primary empirical methodology to inform research utilizing a wide array of data including textual data, interviews, discussions and observations of the practices involved in the production of texts and participants' perspectives on texts and practices.

In line with the Academic Literacies approach, two types of data were collected for the purpose of this study: textual data comprising the student written production in the form of their final written assignment (Section III.4.a), and interview data collected from students and their supervisors (Section III.4.b).

a Textual data. The textual data comprised a 2,000-word (±10%) Critical Literature Review completed by individual students and submitted electronically to the University for assessment purposes in the last week of the programme (i.e. week 10). The purpose of this assignment was to allow the students to develop, practise and demonstrate their

ability to read, understand and process complex ideas and select from these to provide a critical, academic response relating to their research topic. In this assignment, the students had to demonstrate several skills, mirroring the processes they are likely to experience in their department on completion of the pre-sessional programme, including:

- locating relevant sources and assessing their suitability and relevance to their topic;
- reading the selected identified sources and understanding the main ideas, arguments, evidence and supporting information;
- critically evaluating the information presented in the selected sources;
- planning and writing a clear response relating to their research area;
- explaining and referencing arguments and supporting information from sources in their own words using academically appropriate and acceptably accurate language; and
- observing appropriate disciplinary academic conventions.

Since all participants were going to research a different topic on their PhD on successful completion of the pre-sessional programme, the topic for this assignment was determined by individual students. These texts were drawn on during the interview data collection (Section III.4.b) to facilitate 'talk around texts' (Lillis, 2001) commonly adopted in both the Academic Literacies tradition (Lea & Street, 1998, 2006) as well as in English for academic purposes (EAP) research contexts (Lillis, 2008) to gain insights into literacy practices in specific disciplinary contexts.

b Interview data. The collected textual data in the form of individual Critical Literature Reviews were complemented by interview data obtained from the students and their supervisors. For the purpose of this study, the semi-structured format of interviews was used for both sets of interviews as its guidelines allows flexibility to enable extensive follow-up of the participants' responses (Hyland, 2016). The interviews also utilized the 'talk around texts' technique (Lillis, 2001). The aim of utilizing this technique in the present study was to gain insights into the writing processes employed by the students during the process of composing their texts with the assistance of #LancsBox (Section III.3.b). This was achieved by interviewing the student participants with reference to their texts submitted to the University at the end of the pre-sessional programme (Section III.4.a). In addition, the 'talk around texts' was intended to generate insights into the supervisors' perceptions of the students' writing to establish whether the objectives of the pre-sessional programme had been reached not only from the point of view of the pre-sessional tutor, but also from the perspective of the academic departments to which the students will progress. This was achieved by interviewing the supervisors with reference to their prospective students' texts. We believe this is valuable as it will allow us to both establish the effectiveness of our approach and to identify areas where we can develop our course to better match supervisor expectations.

c Student interviews. The student interviews were conducted with individual participants online via Microsoft Teams on completion of the programme after submission and marking of the Critical Literature Review assignment. The timing of the interviews was

intended to enable the student participants to reflect on their learning experience, and to minimize the issue of reactivity referring to the effects of the researcher on the nature of the collected data (Hammersley & Atkinson, 2007). To further reduce the potential problem of reactivity, the interviews were carried out by a tutor who was not involved in the delivery of the programme or the marking of the assignments. It is believed that this increased the objectivity of the participant responses.

The student interviews were approximately 30 minutes long and covered a wide range of topics relating to the students' reflections on the pre-sessional programme. However, since the implementation of DIY corpora had not previously been explored on the University's pre-sessional programme in relation to the students' written production, this article reports on the students' perspectives on the application of their self-compiled corpora during the process of composing their writing.

This was explored through the following questions focusing on the DDL sessions and the students' DIY corpus, followed by 'talk around texts': Can you tell me about the Reading/Listening/Vocabulary Workshops:

- Did you find them useful/beneficial? In what way?
- What did you learn from these sessions?
- What was the main benefit of these sessions for you?
- How would you reflect on your use of your discipline-specific corpus?
- Did you find it useful/beneficial? How / why / in what way?
- What did you mostly use your discipline-specific corpus for?
- What was the size of your corpus? / What did it contain?
- Which aspects of your Critical Literature Review did your corpus help you with? / Which aspects of your Critical Literature Review did you consult your corpus for?

d Supervisor interviews. The interviews with the students' supervisors were also carried out online via Microsoft Teams at the end of the pre-sessional programme after submission and marking of the Critical Literature Review assignment. Five supervisors were interviewed with two supervisors jointly supervising three of the six students on the programme. These two supervisors were interviewed together. The interviews were approximately 20 minutes long and focused on the supervisors' perceptions of their students' written production. This was explored through the following 'talk around texts' questions focusing primarily on various linguistic aspects of the students' writing:

- Do you find the language use suitable? Why / why not?
- To what extent does the language use correspond to the conventions in your discipline?
- To what extent do the students include appropriate discipline-specific terminology?
- Is the language use what you would like to see your students use in their written production?
- How does the students' writing compare with work from international students who did not complete the Pathway to PhD pre-sessional programme?
- Are there any other aspects of the students' writing that you would like to comment on?

5 Data analysis

The textual data served the sole purpose of a prompt during the interviews with the student participants and their supervisors. The collected interview data were analysed drawing on thematic analysis as it offers a theoretically flexible approach to qualitative data analysis (Braun & Clarke, 2006). Following this approach, the interview data were explored using a deductive approach to identify various themes relating to the phenomena under investigation. Several analytical steps were necessary to interrogate the interview data, including the production of initial themes reflecting the various aspects of the data that were of relevance. This was followed by several phases of further defining and refining of the themes leading to a final set of themes relating to the phenomenon under study (Braun & Clarke, 2006).

The interview data and identified themes were reviewed by both authors to reduce any potential bias. In addition, the issue of reactivity (Hammersley & Atkinson, 2007) was also considered during the thematic analysis, referring to the effects of the researcher on the participants, potentially resulting in the participants telling the researcher what they think they want to hear, for example (Zahle, 2023). As far as the student interviews are concerned, the issue of reactivity is likely to have been reduced by the fact that the interviews took place on completion of the programme and the students were interviewed by the second author who was not their tutor. Moreover, eliminating reactivity is not always a prime consideration provided that the researcher is aware of how their presence may have shaped the interview data, which ought to be interpreted accordingly (Hammersley & Atkinson, 2007). As for supervisor interviews conducted by the first author, reactivity was not considered to be an issue as the supervisors were not directly involved in the programme.

IV Findings and discussion

The thematic analysis of the collected interview data generated several themes reflecting the students' perceptions of the DDL approach on the Pathway to PhD pre-sessional programme (Section IV.1). The interview data also offered valuable insights into the supervisors' perspectives of their prospective students' disciplinary writing (Section IV.2).

1 Students' perspectives

Overall, the students' reflections on the DDL sessions were found to be very positive with two overarching themes resulting from the analysis of the interview data. These include the perceived benefits of the DDL sessions utilizing specialized do-it-yourself corpora for the development of the students' receptive and productive vocabulary knowledge (Section IV.1.a) as well as other features of disciplinary academic writing (Section IV.1.b).

a Value of DDL sessions for vocabulary knowledge. The benefits of the DDL sessions in the form of Vocabulary Workshops were noted by all six participants, particularly in relation to their acquisition and development of receptive and productive vocabulary:

> These workshops gave me a chance to know many many new words using my corpus . . . to know the collocations, the phrases, nouns that's all related to the discipline. (Student A)

> Vocabulary workshops were beneficial. I learned words I didn't know from my field and when we came to the last week, I started to have difficulty in finding keywords that I didn't know in the articles, because I learned most of the words. Because of that the vocabulary workshops were useful . . . I found #LancsBox very useful. I didn't learn about this programme before this course and learning the noun phrases from my discipline was the most beneficial part for me. (Student B)

> I think vocabulary workshops were useful. I think it can help me understand noun phrases and how to use it and I think I will be using #LancsBox to find the usage of the corresponding vocabulary in academic articles. I think I learned a lot of words in this course and also how to use this #LancsBox tool. I think it can help me deeply understand my professional terms. (Student C)

> I can say this is one of the most beneficial [sessions] because I get some more vocabulary. Before, I didn't look for the family of this vocabulary and how to use it in different areas and how I can find it in the paper in which they use it. (Student D)

> The vocabulary workshop was beneficial for finding meaning, and for some new vocabulary for me . . . I tried to go to #LancsBox to find the meaning for a word by trying to understand the sentence, the whole sentence and to find the collocations for the word, like what preposition comes before or after this word, which is very important to me. (Student E)

> The programme [#LancsBox] helped me how learn vocabulary in my discipline and how to make a connection between my work and the profession . . . I find this programme [#LancsBox] useful for the words in my discipline with collocations. (Student F)

These quotes highlight two important areas concerning language acquisition. First, all six students noted that the use of their corpus enabled them to acquire new vocabulary related to their work. This is closely related to the notion of 'noticing', which is an important first step in the process of vocabulary acquisition that occurs when learners give attention to a vocabulary item as they become aware of its usefulness (Nation, 2001). This may be affected by several factors such as the salience of the item in a textual input, as is often the case with corpus searches. The students' reflections thus indicate that the DDL workshops led to the noticing of new vocabulary. Moreover, four students (Students A, B, C, F) referred specifically to the relevance of their corpus results to their discipline. These students' reflections thus suggest that the corpus searchers of their DIY corpus led to acquisition of 'technical' vocabulary, defined as words relating to a specific topic or subject area (Nation, 2001). This is an important reflection as it suggests that the DDL sessions achieved one of the aims of the programme, whereby the students were introduced to some of the specific disciplinary needs of their future field of study in the form of frequently used specialized vocabulary items. This link between their corpus searches and their own discipline can, therefore, be regarded as a key step in the development of their linguistic repertoire.

Second, the students' reflections underline the importance of considering the various features of word use, which are important aspects of receptive knowledge of a word. These are:

- grammatical function relating to the patterns in which the word typically occurs;
- collocation referring to words which typically co-occur with the word; and
- constraints on use concerting where, when, and how often one would expect to meet this word (Nation, 2001).

This becomes clear from Students A, E and F, who specifically noted collocations and words which typically occur with the newly acquired vocabulary items, and Students B and C who noted the acquisition of noun phrases characteristic of academic writing (Biber & Gray, 2016). Hence, the students' developing knowledge of the newly acquired vocabulary items in their contextual environments can be seen as another major step in the development of their vocabulary knowledge.

In sum, the students' reflections on the interrogation of their self-compiled specialized corpus discussed above highlight two major benefits of utilizing a DDL approach: acquisition of technical vocabulary and the importance of seeing newly acquired vocabulary items in their contextual environments underlying the vital role of phraseologies and collocations, which are an important aspect of knowing a word.

b Value of DDL sessions for written production. A further benefit of the DDL sessions was noted in relation to the students' own writing whereby the corpus searchers of the students' self-complied discipline-specific corpus enabled them to notice not only how particular linguistic features are used in expert writing representing their discipline, but also how the students themselves can draw on these corpus findings during their own written production:

> . . . how to use the collocations when it's used usually for these words especially. (Student A)

> I didn't directly open and use #LancsBox while writing my literature review, but I had already learned the terms related to my field while preparing presentations in the previous weeks in the vocabulary workshop. I used that in my literature review . . . I mostly used this for noun phrases and academic tone. (Student B)

> The most useful, I think, if I worry about writing if I am not sure how to use something I will use #LancsBox to search it and see how other people use it before I use this in writing . . . I think this tool is very useful for me for example I used it to write this long and difficult sentence which made it difficult for readers to understand my article, to understand my sentence. So, when I wrote the literature review, I used a lot of noun phrases to replace the long difficult sentences to make my sentences clear. Maybe I used it for reporting words and others I have forgotten but I used it a lot. (Student C)

> It's very useful. In terms of vocabulary workshops, we are looking for vocabulary in my corpus. When I am looking, I find there is a lot of vocabulary which is used in the papers which I can use in my dissertation. (Student D)

> I used [my corpus] for a lot of things: To find how to use passive in my discipline, the way they use it in the introduction section, in the discussion section or in the conclusion section; I tried to find a lot of things like academic phrases and to use it in my critical literature review. Some other grammar structures and grammar tense which is used a lot in my discipline and in which section to use it. Because of this, it is very beneficial. (Student E)

Student F noted that they used #LancsBox for their literature review assignment but 'didn't know how to explain that'.

These quotes illustrate the perceived value of consulting a DIY corpus for productive purposes during the students' writing. This is a further important step in the process of vocabulary acquisition (following the initial noticing), referred to as 'generative use' (Nation, 2001). During this stage of the vocabulary learning process, previously met lexical items are encountered or used in ways which are different from the previous meeting with the item. Specifically, the generative use involves the production of the vocabulary in new ways and contexts; that is, in the students' Critical Literature Review assignments in the context of this study. This generative use of new vocabulary is also closely linked to several aspects of word use relating to productive knowledge of a word, similar to the aspects of receptive knowledge of a word discussed above (i.e. grammatical function relating to the patterns in which the word typically occurs; collocation referring to word which typically co-occur with the word; and constraints on use concerting where, when and how often one would expect to meet this word), as can be seen from Students A, B, C and E who note the usefulness of their corpus for phraseologies including collocations, academic phrases and noun phrases.

In addition to productive vocabulary knowledge, the students referred to other aspects of their writing for which they consulted their corpus. These include other key features of academic writing which novice writers need to familiarize themselves with, such as the passive voice and tense (Student E), academic tone (Student B), or reporting words (Student C). This reported focus on these aspects of academic writing by the participants highlights the wider benefits of a DDL approach utilizing a specialized DIY corpus, which lies in the potential of informing various aspects of disciplinary academic writing, other than the lexical features characteristic of a particular discipline.

In brief, from the students' reflections on the value of their specialized DIY corpus discussed above, two main benefits become apparent:

- the ability to draw on a corpus during written production with regard to the deployment of vocabulary and related phraseologies including collocations and noun phrases; and
- the ability to consult a corpus for other characteristics of disciplinary writing in order to inform the students' writing.

2 Supervisors' perspectives

While the students' reflections on their linguistic development during the 10-week programme were largely positive, the response from the supervisors were somewhat mixed.

> I think this is still quite a general level in terms of the technical terms . . . I would say more technical so-called jargon can go in the Literature Review because it's supposed to be academic work. It seems [the student] looked at more so-called general still quite laymen's terms but I think this is fine, it's something [the student] is going to learn. (Supervisor A)

> I liked the style; I thought it was very appropriate. (Supervisor B)

> The main issue we found, which is different from other subjects, is that students try to find alternative terms to statistical terminology . . . trying to come up with different words to describe statistical terminology, and you can't just change statistical concepts which are well-known; you have to use the same terminology. (Supervisors C and D)

> I find [his/her] writing satisfactory, or excellent; [s/he] included lots of jargon – very professional. (Supervisor E)

The supervisors' responses highlight the gap between the supervisors' expectations of student researchers at doctoral level of study and what is achievable on a 10-week pre-sessional programme in terms of improving students' productive knowledge of disciplinary language. It is also interesting to note the gap between the students' perceptions of their linguistic abilities and the supervisors' assessment of their language use. That is, the students were overall positive about their improvement of the technical vocabulary of their discipline, whereas some supervisors found this insufficient. This result may indicate the students' lack of awareness of the expectations of doctoral students in terms of the level of English proficiency. Although the supervisor interviews generated inconsistent results regarding the evaluation of the students' disciplinary written production, the interview data suggest that the 10-week pre-sessional programme introduced students to disciplinary language and enabled them to use it productively in their writing. However, the result also shows that this level of exposure and practice is likely to be insufficient to equip the students with the lexical repertoire of their discipline. This emphasizes the importance of a continuous development beyond the summer pre-sessional programme. Overall, the supervisor responses generated mixed results, which may indicate a varied level of student preparedness as well as different individual supervisor's expectations of novice student researchers.

V Conclusions

The aim of the present study was to examine the effectiveness of a DDL approach utilizing self-compiled discipline-specific corpora for the investigation of disciplinary academic writing on a PhD pre-sessional programme in the context of a British University and to evaluate the extent to which this approach can enhance the students' disciplinary writing. One of the significant findings to emerge from this study is the participants' perceived value of this approach for the development of their vocabulary knowledge together with building awareness of various features of disciplinary writing. The second major finding was that the students found the DDL approach beneficial in informing their own written production, particularly in relation to the usage of technical vocabulary and related phraseologies as well as other characteristics of writing in the disciplines.

Despite these benefits of DDL reported by the students, this study has also found that on completion of the pre-sessional programme the supervisors' expectations of their students' language abilities were not always met, whereby not all supervisors found the students' use of technical vocabulary sufficient.

Overall, nonetheless, our findings highlight the benefits of a DDL approach utilizing self-compiled corpora on a short pre-sessional programme, which suggests that DDL can be successfully implemented on a pre-sessional programme preparing students from various disciplines for doctoral study. We believe that the insights gained from this study can be transferrable to other EAP contexts and may hence be of assistance to not only EAP practitioners delivering pre-sessional courses to students from a range of disciplinary backgrounds, but also more broadly to EAP provision. We, therefore, argue that DDL can be usefully implemented in wider EAP settings where it could become integral part of EAP provision catering for students from a range of disciplinary backgrounds.

Several limitations of this study need to be acknowledged, however. First, this study is limited to a distance learning context. It is, therefore, not possible to determine whether the same findings generalize to face-to-face learning contexts. Second, the findings are based on a relatively small sample size. Next, the interviews with the student participants were conducted on completion of the pre-sessional programme. Hence, the students' reflections may not be an accurate account of the different uses of their DIY corpus during the programme due to the time that had elapsed. Further, the participants were exploring the writing conventions of their specific field of study, potentially neglecting awareness building of disciplinary differences. Further research could, therefore, usefully investigate this approach on a greater number of pre-sessional programmes in both distance learning as well as face-to-face contexts and with a greater number of participants. Valuable insights could also be obtained from interviewing students on a regular basis throughout the programme to explore the specific ways in which they utilize their DIY corpora during the writing process. Future studies could also focus on determining the usefulness of this approach in raising students' awareness of disciplinary differences in academic writing conventions.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Dana Therova https://orcid.org/0000-0001-8079-5728

References

Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning and Technology*, *21*, 195–216. Available at: http://llt.msu.edu/issues/october2017/ackerley.pdf (accessed December 2023).

Anthony, L. (2016). Introducing corpora and corpus tools into the technical writing classroom through data-driven learning (DDL). In Flowerdew, J., & T. Costley (Eds.), *Discipline-specific writing: Theory into practice* (pp. 162–180). Routledge.

Anthony, L. (2019). Tools and strategies for data-driven learning (DDL) in the EAP writing classroom. In Hyland, K., & L.C.W. Lillian (Eds.), *Specialised English: New directions in ESP and EAP research and practice* (pp. 179–194). Routledge.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, *21*, 37–54.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). *#LancsBox v. 5.x: Lancaster University corpus toobox*. [software]. Lancaster University. Available at: http://corpora.lancs.ac.uk/lancsbox (accessed December 2023).

Bridle, M. (2019). Learner use of a corpus as a reference tool in error correction: Factors influencing consultation and success. *Journal of English for Academic Purposes*, *37*, 52–69.

Cambridge University Press. (2022). *Cambridge dictionary*. Cambridge University Press. Available at: https://dictionary.cambridge.org (accessed December 2023).

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning and Technology*, *9*, 111–125.

Chang, J.Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, *26*, 243–259.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*, 93–102.

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, *35*, 30–40.

Charles, M. (2022). The gap between intentions and reality: Reasons for EAP writers' non-use of corpora. *Applied Corpus Linguistics*, *2*, 1–9.

Charles, M., & Hadley, G. (2022). Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, *56*, 1–12.

Chen, M., & 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.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*, 202–224.

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning (CALL)*, *28*, 130–144.

Flowerdew, L. (2015). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, *20*, 58–68.

Friginal, E. (2013). Developing research report writing skills using corpora. *English for Specific Purposes*, *32*, 208–220.

Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, *26*, 225–242.

Hammersley, M., & Atkinson, P. (2007). *Ethnography: Principles in practice*. 3rd edition. Routledge.

Hyland, K. (2016). Methods and methodologies in second language writing research. *System*, *59*, 116–125.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System*, *14*, 151–162.

Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Australia*, *10*, 14–34.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *ELR Journal*, *4*, 1–16.

Johns, T. (2012). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Odlin, T. (Ed.), *Perspectives on pedagogical grammar* (pp. 293–313). Cambridge University Press.

Lay, K.J. (2020). Data-driven learning of academic lexical bundles below the C1 level. *Language Learning and Technology*, *24*, 176–193. Available at: http://hdl.handle.net/10125/44741 (accessed December 2023).

Lea, M.R., & Street, B.V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*, 157–172.

Lea, M.R., & Street, B.V. (2006). The 'academic literacies' model: Theory and applications. *Theory into Practice*, *45*, 368–377.

Lee, D., & 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*, 56–75.

Li, S. (2017). Using corpora to develop learners' collocational competence. *Language Learning and Technology*, *21*, 153–171. Available at: http://llt.msu.edu/issues/october2017/li.pdf (accessed December 2023).

Lillis, T. (2008). Ethnography as method, methodology, and 'deep theorizing': Closing the gap between text and context in academic writing research. *Written Communication*, *25*, 353–388.

Lillis, T.M. (2001). *Student writing: Access, regulation, desire*. Routledge.

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, *22*, 1–18.

Mizumoto, A., & Chujo, K. (2016). Who is data-driven learning for? Challenging the monolithic view of its relationship with learning styles. *System*, *61*, 55–64.

Mizumoto, A., Chujo, K., & Yokota, K. (2016). Development of a scale to measure learners' perceived preferences and benefits of data-driven learning. *ReCALL*, *28*, 227–246.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Oxford University Press. (2014). *A dictionary of statistics*. Oxford University Press. Available at: https://www.oxfordreference.com/view/10.1093/acref/9780199679188.001.0001/acref-9780199679188 (accessed December 2023).

Oxford Univeristy Press. (2016). *A dictionary of computer science*. Oxford University Press. Available at: https://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975-e-963?rskey=VE6PT0&result=1 (accessed December 2023).

Oxford University Press. (2021). *The concise Oxford dictionary of mathematics*. Oxford University Press. Available at: https://www.oxfordreference.com/view/10.1093/acref/9780198845355.001.0001/acref-9780198845355-e-1776?rskey=O50y2X&result=1 (accessed December 2023).

Oxford University Press. (2022). *Oxford learner's dictionaries*. Oxford University Press. Available at: https://www.oxfordlearnersdictionaries.com (accessed December 2023).

Pearson. (1996–2022). *Longman dictionary of contemporary English online*. Pearson. Available at: https://www.ldoceonline.com (accessed December 2023).

Shin, J., Velázquez, A.J., Swatek, A., Staples, S., & Partridge, R.S. (2018). Examining the effectiveness of corpus-informed instruction of reporting verbs in L2 first-year college writing. *L2 Journal*, *10*, 31–46.

Smith, S. (2020). DIY corpora for accounting & finance vocabulary learning. *English for Specific Purposes*, *57*, 1–12.

Sun, Y-C. (2007). Learner perceptions of a concordancing tool for academic writing. *Computer Assisted Language Learning*, *20*, 323–343.

Therova, D., & McKay, A. (2022). Introducing data-driven learning on a PhD pre-sessional programme. *Journal of Academic Language and Learning*, *16*, 91–104.

Wu, Y.-j. A. (2021). Discovering collocations via data-driven learning in L2 writing. *Language Learning and Technology*, *25*, 192–214. Available at: http://hdl.handle.net/10125/73440 (accessed December 2023).

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, *13*, 257–283.

Zahle, J. (2023). Reactivity and good data in qualitative data collection. *European Journal for Philosophy of Science*, *13*, 1–18.