BIROn - Birkbeck Institutional Research Online

A critical review of corpus-based pedagogic perspectives on thesis writing: Specificity revisited

Abstract

Thesis writing (used here as an umbrella term to cover both master's and doctoral postgraduate-level writing) is a high-stakes genre for postgraduate students. This important student genre has been well-researched from a corpus-based perspective. Corpora of theses and also research articles have been used for data-driven learning (DDL) of this key genre. The purpose of this article is to critically examine key DDL initiatives, some of which take a 'research into practice' orientation. Importantly, the discussion is framed around the notion of 'specificity' in the context of needs analysis, and whether the initiatives take a wide-angle, narrow-angle or move from a wide-angle to a narrow angle approach. Accounts which focus on DIY (do-it-yourself) mini-corpus compilation and use by students are also reviewed. The final section of the article presents a critique of current pedagogic applications, taking a closer look at the issue of 'specificity' within the wider context of needs analysis and mapping out areas for future consideration. It is suggested that an ethnographic perspective may be particularly useful for conceptualising specificity relating to students' present situation needs. The article also considers the impact of AI/ChatGPT on future corpus-based pedagogy of thesis writing.

Keywords: thesis and dissertation writing, needs analysis, specificity, data-driven learning, ethnographic approach, AI/ChatGPT

1. Introduction

Thesis writing, used here as an umbrella term to cover both master's and doctoral postgraduate-level writing, is an extended piece of writing completed at the end of a master's or doctoral studies programme. It is a high-stakes genre for postgraduate students, as witnessed by the number of publications devoted to this area. These include textbooks for student thesis writers (e.g., Bitchener, 2010; Swales & Feak, 2000) and resources and guides for thesis supervisors and their students (e.g., Carter et al., 2020; Paltridge & Starfield, 2019). Thesis writing has also been well-researched from a genre- and corpus-based perspective (see Thompson, 2016, 2024 for overviews).

A few early corpus-based studies examined the generic structure of sub-sections of the thesis, which paved the way for future corpus work on specific sub-sections. For example, Kwan (2006) investigated the schematic structure of literature reviews in Applied Linguistics and Bunton (2002, 2005) the generic structure of PhD introduction and conclusion chapters across a range of disciplines. Other corpus research studies are in a somewhat different vein. Most of these are of a contrastive nature, which can be broadly classified according to two separate dimensions: (1) the grounds on which the comparison has been done (i.e., student writing vs. expert writing, different disciplines, different levels of study, and L1 vs. L2 writing), and (2) the linguistic foci of the comparison and contrast and whether the starting point for the analysis is lexico-grammatical patterns or rhetorical strategies. The majority of studies have compared postgraduate writing with expert writing of research articles, such as Hyland's (2004, 2008) research on the use of metadiscoursal devices and Koutsantoni's (2006) investigation of strategies from the perspective of power relations in engineering. Other studies comparing postgraduate writing with research articles, all in the field of Applied Linguistics, include Su Zhang and Chau (2022) on exemplification, Zhang and Su (2023) on the (mis)use of definitions, Zhou and Jiang (2023) on the presentation of limitations in conclusions sections and Zhou and Liu (2021) on the use of *If*-constructions. Disciplinary differences are another of the main foci of Hyland's (2004, 2008) research, as is the study by Sun and Crosthwaite (2022), who analysed *negation* across various hard and soft disciplines in the limitations section of PhD theses. A few studies have examined L1 vs. L2 writing (see Li, Jiang & Ma, 2023) and various cross-linguistic/cross-cultural features (see Lee & Casal, 2014; Can & Cangir, 2019; Dontcheva-Navratilova, 2023). The findings from these research studies show variation in the thesis genre across linguistic and disciplinary dimensions, making thesis writing highly complex and demanding for students to master.

In contrast to this large and diverse body of corpus-based research on thesis writing, the use of corpora of theses and research articles for pedagogic applications, i.e., to prepare master's and doctoral students for thesis writing, has received far less attention. Such work is invaluable for sharing good practice in course development in ESP contexts. Using corpora for data-driven learning (DDL) has several advantages. One of the affordances is its uniqueness of showing what can loosely be termed phraseological patterning involving lexico-grammatical phrases, as such features tend not to be easily accessible in either dictionaries or grammars. Another affordance of corpus-based pedagogy is that the 'hands-on' activities promote a discovery-based inductive approach to learning.

Corpus-based pedagogic initiatives on this key genre, like corpus-based research, first surfaced in the ESP (English for Specific Purposes) literature in the early 2000s. However, the various reports on DDL for thesis writing are scattered throughout the literature, and, to date, little attempt has been made to discuss these in-depth. There is thus a need to present a coherent, systematic overview of these various initiatives to provide a useful pointer to researchers and practitioners, alike. This is the major aim of our study in addition to providing a critical review of these endeavours to advance this important area of DDL.

Our review will be framed within the concept of 'specificity', one of the pillars of ESP, in the context of needs analysis (see Brown, 2016; Flowerdew, 2024 for an overview of needs analyses in ESP). In our review, strategies that practitioners adopt for exploiting corpora to address specific needs of students even in mixed-discipline classes are highlighted. A review of key initiatives is followed by a critique of these pedagogic applications, in which we take a closer look at the issue of 'specificity' within the wider context of needs analysis and map out areas for future consideration. Of significance is the impact of accelerating technological developments of AI/ChatGPT, which has profound implications for corpus-based pedagogy for thesis writing.

The following two sections provide an overview of the concept of specificity in the context of corpus-based writing and a background section on our review.

2. Specificity revisited in the context of needs analysis

2.1 Specifying needs for corpus-based writing pedagogy

Needs analyses are the mainstay of second language writing programmes, and here, corpus construction and research have an important role to play (see Flowerdew, 2021). Nesi (2015) has identified the need to compile corpora which focus on the types of texts learners will engage in, underscoring the value of specificity in corpus design: 'Useful corpus categories for teaching and learning are those which shed light on communicative function and context rather than simply indicating a broad domain such as "Academic" … or "Business" …' (p. 11). Basturkmen (2010) presents a helpful model for classifying ESP courses along a continuum of low to high specificity (see Section 3.2). Hyland (2018) affirms the value of corpus research for pedagogy: 'Armed with evidence of language use from corpora, we can make principled

decisions about which features of target genres are essential to disciplinary communication, and which are less so' (p. 396).

However, corpora of expert research genres such as journal articles are not necessarily the most useful resource for pedagogic applications in regard to student writing. Corpus analyses of successful student writing, as indicated by high marks awarded to master's theses or doctoral theses deemed of high quality by subject specialists, can provide insights into features of writing valued by disciplinary lecturers (see Petrić, 2007). The closer the corpora in terms of the genre students are engaged in writing, the more pedagogically useful. Such analyses of successful student writing can be complemented by corpus analyses of learner language providing insights into problematic features of student writing to address, thereby fostering a learner-centred targeted approach (see Gilquin, 2023; Gilquin, Granger & Paquot, 2007). However, needs analysis need not be solely the domain of the ESP tutor. As Jeaco (2020) notes, learners can engage in needs analyses of their own by building corpora of their own writing and comparing vocabulary and phrases of interest to them with those found in specialised corpora of the target genre (see Section 4.4 on student do-it-yourself (DIY) corpus compilation and use).

2.2. Specifying needs for corpus-based thesis writing

To what extent a corpus-based thesis writing course can be considered 'specific' largely depends on whether the corpus used for instruction is discipline-specific, mirroring the writing from the student's own discipline. As Hyland (2002) states: 'Disciplines have different views of knowledge, different research practices, and different ways of seeing the world, and, as a result, investigating the practices of those disciplines will inevitably take us into a greater degree of specificity' (p. 389). In addition to specificity of discipline, there is also the question of specificity of genre (see Hyland, 2016). In this respect, Tribble (2011) refers to instructional corpora as either 'exemplar' or 'analogue'. An exemplar corpus would exactly parallel the pedagogic corpus in terms of genre and discipline, i.e., a corpus of successfully defended theses from the student's own discipline. An 'analogue' corpus, on the other hand, is as close as possible in terms of genre, discipline etc. to the student writing.

A greater degree of specificity can be achieved if the ESP tutor can carry out their own research beforehand by comparing a corpus of their own students' writing with an expert corpus. This 'research into practice' approach can determine which specific items to focus on in the instructional materials and the tutor-compiled corpus can also be used for hands-on DDL activities (see Eriksson, 2012). While the concept of needs analysis is inherently present in such corpus initiatives, it is not explicitly referred to, unlike in other reports in the literature on thesis writing (see Ch. 8 in Basturkmen, 2010; Link, 2018). Alternatively, training students to compile their own personal mini-corpora of writing from their specific disciplines (see Aston, 2002), and a mini-corpus of their own writing, i.e., DIY corpora, is an ideal means of incorporating discipline-specific concerns within a more general thesis writing course (see Section 4.4). This aligns with de Chazal's (2012) thinking, whose view is that it is ultimately the student's responsibility to acquire familiarity with a specific discipline: 'This stance makes excellent practical sense. In practice, there are so many disciplines that it is absolutely unrealistic for EAP teachers to gain a working knowledge of how they all operate academically and culturally' (p. 141). Also, providing students with tools-of-the-trade, so to speak, and involving them in their own corpus building and analyses can help to promote student autonomy and learning outside the classroom (see Charles, 2014; Charles & Hadley, 2022).

3. Background of the review

3.1 *Literature search and rationale*

As mentioned previously, thesis writing is a high-stakes genre for both master's and doctoral students. It is therefore surprising that compared with the considerable body of corpus-based research on this key genre, so few pedagogic applications have been implemented. Flowerdew (2015) remarked on the lack of corpus-pedagogic applications for PhD students and Chen and Flowerdew (2018a) have also made a similar comment on the limited literature in this area. To investigate this situation in a more in-depth and systematic way, we conducted a literature search from the year 2000 in four databases – ERIC (Education Resources Information Centre), Scopus (Elsevier's abstract and citation database), LLBA (Linguistics and Language Behavior Abstracts) and Google Scholar, using the following keywords in various combinations: *thesis writing*, *dissertation writing*, *doctoral*, *master's*, *data-driven learning*, *DDL*, *corpora* and *corpus-based*. We also trawled the contents of several key journals in which articles on thesis

writing and DDL have been published, with *English for Specific Purposes* and the *Journal of English for Academic Purposes* the most prominent for such work. Other relevant journals were searched – two journals dedicated to academic writing (*Journal of Second Language Writing* and *Writing & Pedagogy*) and three journals specialising in the use of technology for language learning (*Language Learning & Techology*, *System* and *ReCALL)*. We also consulted Chen and Flowerdew's (2018b) critical review of research and practice in DDL in the academic writing classroom and edited collections on corpora for language teaching and specialised genres.

Fifteen key studies were found to be pedagogically-oriented (rather than research-based or reporting an experimental DDL study) and on the topic of thesis writing for postgraduate students. One influential author prominent in the literature search was Maggie Charles with eight references. Except for one initiative (Charles, 2007), all her other accounts include students' use and evaluation of DIY corpora and have been accorded a separate sub-section for discussion. The references section of these fifteen studies were perused and, as one might expect, cross-references to previous studies were made in more recent accounts, with the exception of Nordrum and Eriksson (2015). Although this initiative relates to students' self-assessment, it is included in our review as it is pedagogically-motivated and references three of the other key studies reviewed.

Our searches have thus revealed that the lack of pedagogic initiatives in this area still persists today. Moreover, these few accounts are scattered across the literature and, to date, there has been no systematic review of these. It is therefore opportune to take stock of the limited initiatives with a view to identifying areas for future exploration in order to advance this important pedagogic field.

3.2 Scope and classificatory framework

Both master's and doctoral level theses are covered in the studies for review, which, in fact, are two related but distinct genres differing in scope, length and complexity. While both are written for assessment purposes, they differ in terms of readership and purpose. Master's theses are, in the main, written to demonstrate knowledge gained during a course, while doctoral theses are expected to display more novelty and originality and may be intended for a wider audience beyond their supervisor and examiners (see Paltridge & Starfield, 2020, pp. 70-72). The accounts cover both master's and PhD level students with a few initiatives targeting both

cohorts within the same class. As there are so few reports in the literature with some addressing a mixed cohort, for the sake of expediency, we review these pedagogic applications together but clarify the intended level.

There are also references in the literature on using corpora, comprising discipline-specific research articles, with PhD students for developing research writing (see Chen & Flowerdew, 2018a; Cortes, 2011; Cotos et al., 2016; Dong & Lu, 2020). The main objective of these initiatives, also limited in number, is to help novice scholars write quality research articles for publication, although Chen and Flowerdew's (2018a) project also had a wider brief in that it aimed at disseminating a DDL learning approach to teaching research writing in general among PhD students. These key studies can undoubtedly be considered of relevance to thesis writing. However, we confine our review only to those initiatives which are targeted at master's or doctoral level students mainly for help with thesis writing and which also refer to the compilation of theses for instruction. It could well be the case, though, that students are also working on writing research articles and may be using corpora for help with both genres.[1]

The vast majority of literature reviews present information in a chronological order. While our review has a chronological dimension as it discusses key studies over a 20-year period, it is thematic in nature, based on Basturkmen's (2010, p.55) continuum of specificity. Basturkmen conceives specificity as ranging along a continuum from 'wide-angle' to 'narrow-angle' approaches, as shown in Figure 1.

Figure 1

Figure 1: the wide- and narrow-angled continuum (Basturkmen, 2010: 55).

We have plotted the reports in the literature along this continuum, taking into account that the issue of 'specificity' can be related to both discipline and genre, as stated in Section 2. Figure 2 below provides an overview of the accounts discussed in the following section.

Figure 2

---

[1] We thank one of the reviewers for raising this point.

Figure 2: Current studies plotted along Basturkmen's (2010) continuum.

4. Corus-based pedagogy for thesis writing: Looking back

For ease of discussion, the studies below are seen as primarily (1) wide-angle (2) moving from a low to a more high level of specificity during progression of the course (3) narrow-angle, and (4) DIY highly specific. Other aspects are considered such as to what extent 'hands-on' concordancing tasks are used and how these are integrated with other activities, and whether the studies incorporate student and teacher evaluations of the corpus consultation, a key component of ESP course design.

4.1 Taking a wide-angle approach

Two accounts of a wide-angle approach to corpus-inspired thesis writing in the literature involving mixed-discipline classes are reported in Charles (2007) and Flowerdew (2015). Both combine bottom-up lexico-grammatical searches with a more Swalesian (2004) top-down genre perspective. The instructional cycle reported in Charles (2007) for the function of *Defending your Research against Criticism* is part of an extensive programme on using corpora for thesis writing, involving 40 international graduate students, the majority of whom were taking a doctoral degree, working in 27 different research fields. Following a 'function-first' approach, as advocated by Durrant and Matthews-Aydinli (2011), Charles introduces her students to the two-part structure (*Anticipated criticisms* and *writer's defence*) of this key rhetorical function through initial pen-and-paper tasks using text extracts from successful theses. This paves the way for follow-up 'hands-on' bottom-up corpus consultation using WordSmith Tools Version 4 (Scott 2004) and two corpora of successfully-defended theses written by L1 speakers of English: MPhil theses (about 190,000 words) in politics/international relations and eight doctoral theses (approximately 300,000 words) in materials science. With guidance from the tutor, students were instructed to concordance on salient items, e.g., *while,* and notice from the expanded concordance lines how this is used for constructing a concession, followed by writer's defence, e.g., ***While I acknowledge*** *that in some cases the distinction between institutions and groups may seem rather arbitrary, such political actor subjects are not the focus of interest in this thesis* (p. 294). Charles's module is specific in that corpora of theses were used but these were from only two disciplines and not necessarily related to the students' field of study (but see Charles's work in Section 4.4, which presents a more narrow-angle approach in other modules of her course).

The account reported in Flowedew (2015) describes a two-part workshop spanning 4.5 hours in total to aid science and engineering students, mostly doctoral students but also a few M.Phil students, in writing the Discussion section of their thesis. Similar to the design reported in Charles (2007), students first analysed printed extracts of Discussion sections of theses to identify prototypical move structure patterning. These pen-and-paper tasks were followed by hands-on activities using the freely available 5.6-million-word Hong Kong PolyU Corpus of Research Articles (CRA), comprising 780 research articles spanning 39 different disciplines (Lin & Evans, 2012) and the Michigan Corpus of Upper-level Student Papers (MICUSP), in other words, 'analogue' corpora (Tribble, 2011). The sub-corpus of discussion sections, comprising 2.3 million words, was used for tasks designed to familiarise students with search strategies for identifying useful lexico-grammatical patterns for various specific rhetorical functions. The functions and lexico-grammar selected for corpus enquiries were based on previous corpus-based studies in the area, such as the research conducted by Hyland on thesis writing (2004, 2008). In this case, consulting previous research conducted in the field, as advocated by Woodrow (2018, p. 25), ensures some degree of 'specificity' as regards what is to be taught. Most of the tasks were of a hypothesis-testing nature, rated as the most useful type by Eriksson's (2012) students. For example, for the function of 'Showing comparison with previous work', students were first asked to write their own phrases beginning with *This finding…* and compare their suggestions with those in the corpus, which provided a range of useful phrases (e.g., *This finding is consistent with…; This finding is in agreement with…; This finding is similar to…; This finding is supported by…*).

This workshop could thus be viewed as occupying the wide-angle end of the ESP continuum as the focus was on non-topic specific, i.e., 'common core', lexico-grammatical phrases. The common core vs. specific debate has incited much discussion in the literature in early work in ESP. Even though one might agree, in principle, with Hyland's (2002) viewpoint on specificity, it is not always feasible to implement a highly specialised course in contexts involving mixed-discipline classes where lack of resources and time constraints impinge on course preparation and design. In fact, Hyland (2016) has now reframed this debate, viewing it as a 'dilemma rather than a conflict' with the 'increased understanding of both the complexities of instructional contexts and the characteristics of academic language' (p. 17).

4.2 Moving from a wide-angle to a narrow-angle approach

Three early pedagogic applications, targeting either students from broad disciplines or a mixed discipline group, commence from a wide-angle approach moving to a more narrow-angle, learner-centred perspective. Two of these focus on citation practices. Thompson and Tribble (2001) report a four-stage procedure to help postgraduate humanities students, both doctoral and master's students, familiarise themselves with this important aspect. Stage 1 introduces learners to a general range of citation forms and functions, based on those identified in a previous corpus research study of theses from the fields of agricultural botany and agricultural economics. For Stage 2, it is suggested that an analogue micro-corpus of 22 extracts from one academic journal (*Language and Literature*) be extracted from the British National Corpus (BNC) as this would be appropriate for Humanities students' field of study. In Stages 3 and 4, students review their own citation practices in self-compiled corpora of their own work and in their peers' corpora of assignments set by professors from their own department, thereby addressing students' own specific writing needs.

Another case study on citation use, specifically how to report the research of others in 'Writing a Literature Review', can be found in Starfield (2004), who takes a critical pedagogies approach. Starfield first introduces her arts and social sciences PhD students to concordances of phrases for referring to the work of others (e.g., *according to…*), using extracts from Thurstun and Candlin's (1998) textbook *Exploring Academic English: A workbook for student essay writing,* which draws on the one-million-word *Microconcord Corpus of Academic Texts* for its concordance examples. Using these corpus examples as a springboard into more disciplinary investigations, she then used a whole 'authoritative' *TESOL Quarterly* article for worksheet exercises designed to move students 'beyond the concordance line' to examine how an author was not just reporting the work of others but setting up a niche in which to insert their own work. Instead of having students examine this rhetorical function in discipline-specific corpora, students were asked to bring to class an authoritative text from their own discipline to deconstruct how writers positioned themselves with respect to the work of others. Follow-up interviews with students revealed that they were self-motivated to 'take up concordancing' outside the classroom, which, they said, created a sense of empowerment, Starfield concluded by saying that this technique provided 'a way into powerful, potentially exclusionary, discourses' (p. 154).

A landmark corpus consultation course for doctoral students is reported in Lee and Swales (2006). While only a limited number of six students from different disciplines attended this 13-week optional course, a close look at the three components of the course illustrates how the

concept of 'specificity' is operationalised across various components. For the first 10 weeks students attended weekly 2-hour lab sessions in which they were inducted into the 'corpus approach' and introduced to the concordancing program WordSmith Tools (Scott, 1996). Three main corpora were used for this general introduction, including *Hyland's Research Article Corpus* (1.3 million words taken from published journal articles from eight disciplinary areas), and academic sub-components from the *British National Corpus* (BNC), including textbooks, doctoral dissertations and journal articles totalling about 15 million words. The hands-on concordancing tasks generally followed a bottom-up, lexico-grammatical approach and were of an inductive nature, for which corpus searches are ideally suited. For example, to reinforce the useful pattern *ADJ/N/V/ for V-ing* structure and its associated semantic prosody (e.g., *appropriate for modelling…; devices for storing…; suited for studying…*) students were given a worksheet, on which one of the tasks asked them the following: (1) guess and rank the frequencies, guess some exemplar word forms, and then check what the corpus says (p. 63). This component of the course could generally be regarded as wide-angle. However, it also allowed for specific queries raised by students for follow-up corpus consultation, such as the difference between *totally* and *in total,* and included tasks focusing on disciplinary differences.

A more narrow-angle approach addressing specificity was incorporated into the other two components of the course by having students build two additional corpora: one of their own writing (dissertation drafts, term papers etc.) and one of 'expert' writing of published papers in their own field. Lee and Swales stressed the motivational value to the participants of having '*their own* precisely targeted databases' (p. 61), an aspect especially important for the participants whose disciplines (pharmacology, biomedical statistics, educational technology) were not represented in the instructional corpora (see Section 4.4 for other accounts of student self-compiled corpora). The second component of the course included optional individual consultations in which assistance was given with corpus compilation and feedback on written work which 'proved useful by giving us clues about specific areas of weakness that the use of corpora could help remedy' (p. 58). In the third component of the course students gave oral presentations on their chosen linguistic devices for investigation based on the findings and analyses from their self-compiled corpora. In the end-of-course feedback participants reported that they found the use of corpora confidence-building and empowering, as was the case with Starfield's students. Importantly, students commented that corpus searches allowed them to see disciplinary differences, which, as Lee and Swales point out, is important for some linguistic

structures: 'Overall, they [the participants] felt that the exemplification was often much closer to their contextual and textual circumstances' (p. 71).

4.3 Taking a narrow-angle approach

It goes without saying that a more narrow-angle approach can be taken when students are from the same discipline. One of the first discipline-specific pedagogic initiatives is reported in Hewings and Hewings (2002). Similar to the pedagogic framework presented in Thompson and Tribble (2001), this study is underpinned by a 'research into practice' approach and was one of the first to use a learner corpus for research and materials preparation. The study explores one grammatical feature of metadiscourse, interpersonal clauses with an anticipatory 'it' and extraposed subject (e.g., 'It is interesting to note that no solution is offered'). The learner corpus used in the study was made up of 15 dissertations written by L2 speakers of English as a final component of their MBA programme. The 'expert' corpus for comparison consisted of 28 journal articles from three business journals. The researchers are mindful that the journal corpus is an 'analogue' one: 'the two genres are not identical with respect to content, organisation, and readership, and this needs to be borne in mind in making comparisons' (p.371), an observation which seems of particular importance when investigating interpersonal features as is the case here. However, it could be argued that the issue of 'specificity' has been considered in that the articles were selected from three specific journals whose articles occurred frequently on the reading lists for the MBA students. Based on their manual and concordance analysis of *it*-clauses in both corpora, Hewings and Hewings drew up a classificatory framework consisting of four categories (hedges, attitude markers, emphatics, attribution). The findings were then used to prepare pen-and-paper concordance tasks for class discussion, asking students to compare, for example, instances of *it seems* in the two sets of corpora (student dissertation: *It seems that different studies have shown different results* compared with published articles: *It seems likely that the eighties and nineties will be known as decades of large-scale disaggregation*). Thus, the value of incorporating learner corpus findings in materials serves to pinpoint specific problematic areas, which postgraduate students might not appreciate if they were just exposed to working with expert corpora alone (see Cotos, 2014 for an experimental study using learner corpora, which was found to have a positive impact on DDL-based instruction).

The case studies reported in Eriksson (2012) and Nordrum and Eriksson (2015) reside very much at the narrow-angle end of the continuum. Eriksson's (2012) workshop for doctoral

students of biochemistry also bridges the research/practice divide, with the instructional cycle moving through all stages of ESP course design from initial needs analysis through to evaluation. In contrast to the top-down move structure patterning combined with bottom-up lexico-grammatical approaches reported in Charles (2007) and Flowerdew (2015), Eriksson takes a bottom-up approach starting with a specific type of phraseology, i.e., lexical bundles (see Cortes, 2023). Lexical bundles, extracted computationally from a corpus, are '…sequences of words that commonly go together in natural discourse' and have been described as 'the building blocks of academic writing' (Biber, Johansson, Leech, Conrad, & Finegan, 1999, p. 990). Materials were based on Eriksson's findings from his comparison of three-word lexical bundles (e.g., *the absence of*) from his two small, specialised self-compiled corpora (around 100,000 words each), an expert corpus of articles from biochemistry and a learner corpus of his student writing from the same discipline. These two lists were then compared with Hyland's (2008) list of academic bundles. Eriksson's pre-instruction lexical bundle analysis thus acts as a needs analysis to determine which bundles to teach and also provides corpus evidence for disciplinary specificity and by extension the value of using specialised corpora as more than half of the top fifty bundles in his two self-compiled biochemistry corpora did not appear on Hyland's academic list. A set of activities focused on hands-on corpus searches, followed by tasks in which students were asked to identify the functions of specific bundles (e.g., comparison and contrast) and wrote example sentences showing how they would use bundles in the context of their own thesis writing.

The study reported in Nordrum and Eriksson (2015) differs from others in several respects. Whereas the studies discussed previously are concerned with using corpora for teaching thesis writing, the focus of this proposed initiative is on the use of small, specialised corpora for formative self-assessment practices on one specific rhetorical function – data commentary in science writing from the field of applied chemistry. Data commentaries were extracted from a learner corpus of master's theses and a corpus of published research papers in the field of applied chemistry and tagged manually for move structures (background information; presentation of visual; comment on result) and sub-moves. Of note is that the learner corpus included theses representing different levels of quality to highlight problem areas, which is possible here as master's theses are awarded marks/grades. The proposed assessment activities asked students to evaluate the effectiveness of data commentaries in the corpus of master's theses, as compared against those in the expert corpus. Nordrum and Eriksson state that 'The corpus-informed activities we have in mind can be integrated in ESP courses or workshops at

master's and PhD levels at technical universities in Sweden and elsewhere' (p. 71). They also view their initiative as a blueprint for investigating data commentaries in other fields and looking at disciplinary differences. Swales and Feak (2023) explore the task evolution of data commentaries in published EAP writing materials, from Information Transfer to Data Commentary, with a trend towards Critical Commentary. It would therefore seem that the corpus-inspired tasks proposed by Nordrum and Eriksson designed to foster critical awareness align with the movement towards critical thinking, 'one in which graduate students/apprentice researchers increasingly position themselves as perspicacious and perceptive analysts of the presented data' (Swales & Feak, 2023, p. 8).

In contrast to the above small-scale initiatives, Wong (2019) reports the implementation of a large-scale data-driven learning project for postgraduate thesis writing to cater to the needs of both PhD and master's students. This project, compared with Flowerdew (2015) workshop, follows a narrow-angle approach, and is highly specific in terms of both genre and discipline. The instructional corpus, namely, the Hong Kong Graduate Corpus (HKGC), was constructed using high standard theses of various disciplines and sub-disciplines recommended by ten faculties at a university in Hong Kong. The HKGC comprises 431 postgraduate theses (345 Ph.D. and 86 M.Phil. theses) from 52 departments, consisting of about 11 million words in total. The implementation of this project involved collaboration among EAP practitioners, corpus/DDL specialists and software developers 'who worked together to create a purpose-built, in-house tool which would be easy to use, provide a large corpus, rival established corpus query interfaces and have the added ability of tracking learner usage habits across a long period of instruction' (Wong, 2019, pp. 196-197; see also Crosthwaite, Wong & Cheung, 2019). The interface thus allowed students in heterogeneous classes to perform targeted searches for lexis and phraseology not only in their own disciplines but also in various sections of the thesis.

4.4 Implementing a student DIY approach

As illustrated by the above studies, 'specificity' in terms of both genre and discipline in a corpus-based thesis writing course or module is achievable if the class tutor builds an exemplar expert corpus of successfully defended theses (somewhat time-consuming and difficult on account of sourcing theses). An alternative more practical solution is to compile an analogue corpus of research articles from the students' discipline as in the account reported in Hewings and Hewings (2002). However, this works best if the students are a fairly homogeneous group

from the same discipline. In the case of mixed discipline classes, this problem can be resolved by putting the onus on students for compiling their own DIY personal corpora of expert disciplinary writing of research articles for comparison with a small self-compiled corpus of their own writing, as in the study by Lee and Swales (2006). The advantages of having students build their own corpora are that it not only meets students' disciplinary needs but can also address the question of *interdisciplinarity*. Charles (2012) notes that out of her 158 students enrolled for her thesis writing course, 43% regarded their study as inter-disciplinary. Another consideration is that having students compile a corpus of their own writing for comparison with an expert corpus can address their own specific issues and points of interest to them.

The focus of this section is to provide an overview of this approach, which has mainly been developed by Charles (2012, 2014, 2017, 2018). Charles (2018) provides some guidelines for corpus building by students for which they had used research articles and used the AntConc software (see Anthony, 2019) for queries. Based on data available for 90 students, Charles found that the purpose of most students' searches was to check lexico-grammar and extend their phraseological knowledge. While Charles acknowledges that the distinction between 'discipline-specific' and 'general academic' is not always easy to determine, she notes that a third of the student queries were of a discipline-specific nature. By way of example, she cites the query of an Italian student of chemistry, who wanted to find a synonym for the phrase *under harsh reaction conditions*. In her DIY corpus of 200,000 words consisting of 51 research articles, the student found 8 instances of the phrase *forcing conditions*, leading her to conclude that this phrase was indeed used in her own field. Charles also makes the important point that unless an EAP teacher had specialist training in chemistry, they would not be able to suggest the adjective *forcing* as an appropriate collocate for *conditions*; nor would they be able to verify its appropriate usage in this specialised context, thus concurring with de Chazal (2012). While the studies reported in the previous sections are generally concerned with using corpora at the initial stages of thesis writing, Charles (2018) describes a module in which students used their own individual specialised corpora (i.e., a corpus of research articles of expert writing in their own field and a corpus of their own writing of thesis chapters) for editing their theses at the final writing stage. Charles illustrates how students can use other tools besides concordancers, such as collocates and concordance plots, not only to address their own lexico-grammatical queries but also to check content and organisational issues. For example, one of her students used the concordance plot, which shows the distribution of a search term throughout a file and a corpus, to see the frequency and location of important content words in each chapter and

across the thesis, thus indicating how a specific aspect of the content has been developed throughout the thesis.

Five studies focus specifically on students' evaluation of using personal DIY corpora (Charles 2011, 2012, 2014; Zhang, Zheng, & Li, 2017; Charles & Hadley, 2022). Charles's (2012) evaluative study reports that students generally perceived the relevance of the self-compiled expert corpus for their own work, commenting that it gave them examples of the use of subject-specific terms difficult to find in more general corpora (see also Charles, 2017). Charles and Hadley's (2022) study on students' long-term use revealed that the DIY corpora compiled by students were specialised not just by discipline or sub-discipline but even more narrowly by the student's research topic (see Section 5 for other considerations regarding specificity). For example, one doctoral student working in the discipline of geography/environment on the topic of the role of transport and travel behaviour compiled a highly specialised corpus in this domain, which he found useful for key lexico-grammar. Disadvantages reported by students in Charles (2012) included the small size of the self-compiled specialised corpora such that the findings may not be representative of the discipline or not return enough or no hits for low frequency items. One student working in the interdisciplinary area of biophysics reported that she had difficulty deciding which articles to include in her corpus as she considered herself to be working in two quite different fields. A few students questioned the reliability of the English in certain articles, commenting that it was not up to standard even in good articles.

A similar evaluative study to that of Charles (2012) on DIY corpora is reported in Zhang et al. (2017). Of interest is that while the 25 first-year postgraduate students surveyed were medical students, their studies spanned 13 different disciplines, e.g., anatomy, histology and toxicology, thus lending weight to the value of having students compile their own specialised corpora. In this study questionnaire data were collected immediately after completion of the course and also six months after the course. Similar to Charles's findings, students were generally positive and used the corpora for checking lexical and grammatical issues, but also raised issues relating to size, reliability and non-standard English (see Wu, Mauranen, & Lei, 2020 for a study on syntactic complexity in English as a lingua franca in research articles). However, it was found that the use of personal corpora dropped dramatically in the intervening six-month period, mainly due to disciplinary demands on their time. In contrast, Charles's (2014) survey of 40 postgraduate students' use of personal corpora one year after her course revealed students did continue to use their personal corpora and incorporated its use into their writing practices (see also Charles & Hadley, 2022).

In the following section, with reference to the above thematic review, we present a critique of current initiatives. This involves taking a closer look at the issue of 'specificity' within the wider context of needs analysis and mapping out areas for future consideration, which are underexplored in current studies.

5. Corpus-based pedagogy for thesis writing: Moving ahead

The key pedagogic initiatives reported above range along a continuum from a wide-angle to a narrow angle approach in terms of discipline, sub-discipline and even topic, although the pedagogic corpus may not exactly mirror the instructional genre of thesis writing. Moreover, while the approach taken in all the studies allows for a high degree of specificity in terms of the target lexico-grammatical and discursive features, it does not necessarily ensure specificity in terms of the level of study or of the specific institutional thesis requirements. Those initiatives at the wide-angle end of Basturmen's continuum include both master's and doctoral level students while those at the narrow-angle end are predominantly aimed at either doctoral or master's level of student. Also, institutions may have different thesis requirements in terms of thesis length, structure and assessment-related expectations (e.g. level of criticality and originality). Such aspects are more context-related, residing at the very narrow-angle end of Basturkmen's continuum (see Figure 1) and best gleaned through ethnographic means. None of the reports above, however, make reference to the thesis requirements of their particular institution. Pedagogic initiatives may thus vary along these four dimensions of specificity (discipline, genre, study level, institution) in that an initiative may have a high level of specificity on one dimension but a low level on another. The study level and thesis requirements could therefore usefully be considered in future work. These case studies also prompt a deeper consideration of the level of postgraduate students targeted, the phraseologies investigated by students and the corpora used for instruction.

Most of the pedagogic accounts reviewed in Section 4 involve doctoral students only (e.g., Eriksson, 2012; Lee & Swales, 2006; Starfield, 2004) or mostly Ph.D students with a few M.Phil students (Charles, 2007; Flowerdew, 2015). It is only the initiative by Hewings and Hewings (2002) which specifically targets master's students, with Wong's (2019) large-scale project targeting both levels. Future pedagogic initiatives could therefore direct more attention to master's level students, especially as successful completion of a master's degree is a gateway

for acceptance onto a doctoral programme (see Maher and Milligan (2019) who note that the master's thesis is an understudied field).

The studies discussed in Section 4 address phraseological aspects of language used to realise various rhetorical devices (e.g., metadiscourse, citation practices) or genre moves (e.g., comparison with previous research in the discussion section). On the futures of EAP genre studies, Swales (2019) suggests that: '…more attention could usefully be given to a) syntactic and phraseological patterns and uses, and b) to local cohesive elements that will increase the 'flow' of student texts' (p. 81). Arguably, the corpus-based initiatives reviewed in this article achieve the former, but a question mark remains to what extent they sensitise students to the 'flow' of text as the phraseological realisations tend to be examined within individual genre moves or rhetorical features. One initiative reported in Charles (2021) seeks to compensate for this shortcoming. Students first analysed the moves within self-compiled corpora of literature reviews from their own fields, based on Kwan's (2006) model. They subsequently looked at wider stretches of text to examine the co-text of key lexis such as *studies* and *research*, in order to '… help them gain a sense of the logical flow of this subgenre' (p. 249). In an early paper, Swales (2002) has commented on the lack of attention paid to the macro-level in corpus studies. This important aspect therefore merits further enquiry, as exemplified in the recent corpus study by Golparvar et al. (2023). They examined how cohesion, a neglected area in corpus research, is realised in the different rhetorical sections of research articles. Also of relevance for future pedagogic applications of phraseological items is Wu et al.'s (2024) corpus study of such items in commentaries in visuals in research articles. They investigated these from a discourse-based functional perspective, noting that such research '…opens up a multimodal orientation to corpus-driven genre-analysis studies' (p. 35). Having students investigate phraseological items at a more macro-level would enhance further the affordances of corpus-based pedagogy.

Most of the implementations make use of an 'analogue' corpus of research articles for pedagogy on account of the challenges in compiling corpora of theses. Nevertheless, a certain degree of specificity is achieved, either by discipline or rhetorical section - Hewings and Hewings (2002) used research articles from journals which occurred frequently on the reading lists of their MBA students while Charles's students compiled small corpora of research articles related to their own field of study. Flowerdew (2015), meanwhile, used a corpus of research article discussions for the workshop on this section of writing a thesis. Such constraints and the compromises that have to be made are a recognised limitation on implementing ESP needs

analysis findings (see Brown, 2016, who classifies constraints into three broad categories, situational, stakeholder and theoretical). Projects such as the large-scale initiative reported in Wong (2019) require funding, collaboration among ESP practitioners, corpus/DDL specialists and software developers, and liaison with disciplinary faculty, a quite different situation from the one in which most individual ESP tutors find themselves.

Specificity in course design should also be seen in relation to different types of needs and whether these involve the present situation or target situation needs (Dudley-Evans & St. John, 1998; Hutchinson & Waters, 1987). Ethnographic research may also be particularly useful for conceptualising specificity related to present situation needs. By way of example, in the case of thesis writing Basturkmen (2010, pp. 124-125) notes the complex picture that emerged when investigating students' needs for 'open access' elective workshops on thesis writing. Consultation with students revealed that they were at different stages in their research and reported a variety of difficulties at a linguistic level, so their immediate needs differed. Another complicating factor was that, as reported by supervisors, students varied greatly in their awareness of the form and content of the various sections of a thesis, which can vary not only in terms of discipline but also the type of study (quantitative, qualitative, descriptive, or experimental). The DIY corpus-based pedagogic initiatives reported in Section 4.4 could be seen as a means of accommodating these differing needs to some extent. Moreover, needs analyses are not set in stone but ongoing, dynamic processes, morphing into programme evaluation (see Brown, 2016, p. 205). Feak (2016) notes that as a result of ongoing surveys and interviews with her post-graduate students in a thesis writing course some needs were found to remain constant while new challenges emerged, thus stating that it is incumbent on the class tutor to go 'beyond translating research into useful materials' … 'in order to remain responsive, current and flexible' (p. 494).

Another feature of the above pedagogic endeavours is that the corpora used contain only one version of a student text. Neither do they include detailed information on proficiency level of the learner texts, although Nordrum and Eriksson (2015) included writing of different levels of quality in their learner corpus of master's theses. In this regard, the MUCH (Mälmo-University-Chalmers Corpus) project aims at looking at 'writing as process' in the spirit of Curry and Hewings's (2002) and Matzler's (2021) framework of writing as a networked activity, thus according greater attention to writing context and the ethnographic dimension (Lillis, 2008). To this end the longitudinal monitor corpus, in the sense that the corpus maps student writing

over time and will be added to periodically, will include several drafts of a paper, and instructor and peer feedback (see Wärnsby et al., 2016).

The findings from analyses of the target and/or student corpus are just one means of determining students' specific needs and can usefully be supplemented by more contextual information, as outlined above. Hyland's (2004, 2008) corpus research is the exception for the inclusion of students' views and supervisors' comments on the use of metadiscoursal devices in L2 thesis writing. Future studies, as well as consulting corpus-based research on thesis writing, could also consult ethnographically-oriented research to gain a more holistic picture. For example, the fine-grained needs analyses of L1 and L2 doctoral students' experience in writing thesis discussion sections reported in Shen et al. (2019) and the longitudinal study by Harwood & Petrić (2018) on students' perceptions of master's supervision provide insightful findings. Also of relevance is the special issue 'Thesis and dissertation writing in a second language: Context, identity and genre' in the *Journal of Second Language Writing* (Paltridge & Starfield (Eds.) 2019). An overview of research focusing on insiders' views, including their roles, responsibilities and expectations in thesis and dissertation writing, is provided in Paltridge, Starfield and Tardy (2016). Further insights into students' experiences (e.g., emotions, perceived challenges, coping strategies) of thesis writing at universities across Europe are provided in Petrić and Castelló (in press). This ethnographic dimension is thus another important consideration for defining the conceptualisation of 'specificity' in ESP course design as revealed by students' views and supervisors' comments.

All the accounts in Section 4 are concerned with L2 writing (but see Petrić & Castelló, in press, for studies concerning L1 thesis writing in various languages as well as writing in English as L2). It cannot be assumed that postgraduate students with English as their L1 have mastered this high-stakes genre; witness Friginal's (2013) report of a corpus-based course for developing research report writing skills with EL1 students. Another feature of the above pedagogic activities is that they all involve monolingual corpora. In their methodology paper, Moreno and Swales (2018) have underscored the value of multilingual corpora for pedagogic purposes, which are now coming on stream for thesis writing, as described below.

While a few corpus studies looking at cross-linguistic/cross-cultural issues in thesis writing have been undertaken, these are mainly research-oriented, offering insightful implications for pedagogy. It is only very recently that the creation of a multilingual corpus of theses, namely the MoReThesisCorpus, set up for both research into language from a cross-cultural, cross-

disciplinary discourse perspective and for pedagogic purposes, has been undertaken. Bondi and Di Cristofaro (2023) describe the ongoing compilation of this corpus of approximately 86-million words consisting of 1,063 theses in English, 2,772 theses in Italian with a few in French (7), German (8) and Spanish (16) across a wide range of disciplines submitted to the University of Modena and Reggio Emilia. The sophisticated interface will allow students at the university to carry out detailed searches according to a specific discipline, sub-discipline and thesis sub-section. Its added value, however, is that it would also allow students, and not just researchers, to examine lexico-grammar from a cross-linguistic perspective. A key technological innovation for corpus compilation is that this project is aimed at producing a corpus-linguistics-ready version of theses submitted each year to the university, thereby representing 'an initial laboratory for the definition of replicable procedures meant to be automatically implemented and run every year' (p. 24).

One key topical issue which has profound implications for writing instruction in general and corpus-based writing pedagogy in particular is the increasingly pervasive influence of AI/ChatGPT, briefly addressed below. ChatGPT is a deep learning model that is trained on a vast corpus of text and generates responses through prediction. A useful overview of the pedagogic applications of ChatGPT is provided by Dr Emily Nordmann at the University of Glasgow (see https://sway.office.com/p3A96vxXue9qqeFF?ref=Link0), who recommends its applications to help students in editing their work and to aid teachers in giving feedback on drafts based on marking criteria provided.

A pressing question to ask is how much of a threat AI/ChatGPT poses to corpus-based DDL, which mainly relies on generation and analysis of concordance data. In this respect, Lin (2023) discusses the impact of ChatGPT on the field of corpus linguistics, in particular its potential as a concordance. Lin (2023) tested some of her own research queries previously answered by concordance data from a corpus using ChatGPT instead. Based on her testing, Lin, like Nordmann, acknowledges its value in generating output when suitable prompts are given. However, she also notes disadvantages in that ChatGPT failed to perform a collocation analysis, a basic function of standard concordancing programmes. Crosthwaite and Baisa (2023) also explore the potential advantages and disadvantages of GenAI (generative artificial intelligence) vis-à-vis the use of corpora. They note that 'One of the main advantages of corpora is that we know exactly the domain of texts from which the corpus data is derived, something that we cannot track from current large language models underlying applications like ChatGPT' (p. 1), a significant drawback from the perspective of ESP corpus pedagogy. In

common with Lin (2023), they note that collocations can be difficult for GenAI to handle and question the authenticity of the language generated, which may not be contextually or register appropriate. There are thus strong arguments for the continued use of corpus-based pedagogy for thesis writing, especially those of a DIY nature, to meet students' specific needs in terms of appropriate lexico-grammar for the discipline, thesis sub-section and topic. At the same time, though, Crosthwaite and Baisa make a case for bringing GenAI into DDL, including the reduction in levels of technical knowledge required on the part of the student and teachers and ease in preparing teaching materials.

Specific needs are usually discussed in relation to the learner and their immediate environment, but they can also be impacted by government and university policies at the national level (see Brown, 2016). With respect to GenAI, the Russell Group universities in the UK have a mandate to support students and staff to become AI-literate 'The transformative opportunity provided by AI is huge and our universities are determined to grasp it' (https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/).This accelerating technological innovation is thus opening up a new frontier for corpus-based research and pedagogy, which cannot be ignored.

  To conclude, this article has first presented a thematic review of corpus-based thesis writing pedagogy from its inception in the early 2000s through the lens of the notion of 'specificity' in the context of needs analysis. The second part provides a critique of these studies, mapping out areas for future consideration, which are underexplored in present initiatives. Moreover, new thesis formats will present challenges for corpus compilation and the analysis of multimodal data. These include practice-based theses where students complete a non-verbal piece of work (e.g. film, music, software) accompanied by a piece of writing or portfolio-based theses (see Paltridge et al. 2012a, 2012b for a study on doctoral writing in the visual and performing arts). Paltridge and Starfield (2020) have also remarked on the growing trend towards 'article compilation' theses for doctoral students, posing the question '…what will the 'dissertation of the future' look like, in what ways will it be different from what students are writing today?' (p. 14). As well as addressing the changing nature of theses, another pertinent question for corpus-based pedagogy for thesis writing is: What effect will AI/Gen have on current concordancing software and will it turn out to be a friend or a foe?

References

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

Aston, G. (2002). The learner as corpus designer. In B. Kettemann, & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 9-25). Rodopi.

Basturkmen, H. (2010). *Developing courses in English for Specific Purposes.* Palgrave Macmillan.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson.

Bitchener, J. (2010). *Writing an applied linguistics thesis or dissertation. A guide to presenting empirical research.* Palgrave Macmillan.

Bondi, M., & Di Cristofaro, M. (2023). MoReThesisCorpus. Documenting academic language as used in the theses submitted to the university of Modena and Reggio Emilia. In C. Argondizzo, & F. Poppi (Eds.), Special Issue: ICT Affordances and Higher Education. *Iperstoria*, *21*, 9-30.

Brown, J.D. (2016). *Introducing needs analysis and English for Specific Purposes.* Routledge.

Bunton, D. (2002). Generic moves in Ph.D thesis introductions. In J. Flowerdew (Ed.), *Academic discourse* (pp. 57-75). Pearson Education.

Bunton, D. (2005). The structure of PhD conclusion chapters. *Journal of English for Academic Purposes*, *4*(3), 207-224.

Can, T. & Cangir, H. (2019). A corpus-assisted comparative analysis of self-mention markers in doctoral dissertations of literary studies written in Turkey and the UK. *Journal of English for Academic Purposes*, *42*, 1-14.

Carter, S., Guerin, C., & Aitchison, C. (2020). *Doctoral writing: Practices, processes and pleasures*. Springer.

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.

Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing classes. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 26-43). Bloomsbury.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus building. *English for Specific Purposes*, *3*,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. (2017). Do-It-Yourself corpora in the classroom: Views of students and teachers. In L. Wong, & K. Hyland (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 107-123). Routledge.

Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes*, *36,* 15-25.

Charles, M. (2021). Combining genre analysis and corpus consultation in class: Using do-it-yourself corpora to explore the literature review. In K. Lin, I. Mwinlaaru, & D. Tay (Eds.), *Approaches to specialized genres* (pp. 243-258). Routledge.

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*, 101095.

Chen, M., & Flowerdew, J. (2018a). 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.

Chen, M., & Flowerdew, J. (2018b). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, *23*(3), 335-369.

Cortes, V. (2011). Genre analysis in the academic writing class: With or without corpora? *Quaderns de Filologia Estudis lingüístics, 16*, 65-80.

Cortes, V. (2023). Lexical bundles in English for academic purposes. In R. Jablonkai, & E. Csomay (Eds.), *The Routledge handbook of corpora in English language teaching and learning* (pp. 222-233) Routledge.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL, 26*(2), 202-224.

Cotos, E., Link, S., & Huffman, S. (2016). Studying disciplinary corpora to teach the craft of Discussion. *Writing & Pedagogy, 8*(1), 33-66.

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3), 100066.

Crosthwaite, P., Wong, L., & Cheung, J. (2019). Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning. *ReCALL*, *31*(3), 255-275.

Curry, M.J., & Hewings, A. (2002). Approaches to teaching writing. In C. Coffin, M.J. Curry, S. Goodman, A. Hewings, T. Lillis, & J. Swann (Eds.), *Teaching academic writing: A toolkit for higher education* (pp. 19-44). Routledge.

De Chazal, E. (2012). The General – Specific debate in EAP: Which case is the most convincing for most contexts? *Journal of Second Language Teaching and Research, 2*(1), 135-148.

Dong, J., & Lu, X. (2020). Promoting discipline-specific genre competence with corpus-based genre analysis activities. *English for Specific Purposes, 58*, 138-154.

Dontcheva-Navratilova, O. (2023). Self-mention in L2 (Czech) learner academic discourse: realisations, functions and distribution across masters theses. *Journal of English for Academic Purposes, 6*, 101272.

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for Specific Purposes.* Cambridge University Press.

Durrant, P., & Matthews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58-72.

Eriksson, A. (2012). Pedagogical perspectives on bundles: teaching bundles to doctoral students of biochemistry. In J. Thomas, & A. Boulton (Eds.), *Input, process and products. Developments in teaching and language corpora* (pp. 195–211). Masaryk University Press.

Feak, C. (2016). EAP support for post-graduate students. In K. Hyland, & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes* (pp. 489-501). Routledge.

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

Flowerdew, L. (2021). Needs analysis for the second language writing classroom. In: The encyclopedia for second language teaching (Ed. J. Liontas). https://doi.org/10.1002/9871118784235.eelt0523.pub2

Flowerdew, L. (2024). Needs analysis and curriculum development in ESP. In: S. Starfield, C. Hafner (Eds.), The Handbook of English for Specific Purposes (2nd ed.), Wiley-Blackwell (in press).

Friginal, E. (2013). Developing research report writing using corpora. *English for Specific Purposes, 32*(4), 208-220.

Gilquin, G. (2023). Written learner corpora to inform teaching. In R. Jablonkai, & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching* (pp. 281-295). Routledge.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*(4), 319-335.

Golparvar, S., Crosthwaite, P., & Ziaeian, E. (2024). Mapping cohesion in research articles of applied linguistics: A close look at rhetorical sections. *Journal of English for Academic Purposes, 67*, 101316.

Harwood, N., & Petrić. B. (2018). *Experiencing master's supervision: Perspectives of international students and their supervisors.* Routledge.

Hewings, M., & Hewings, A. (2002). 'It is interesting to note that…': A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, *21*(4), 367-383.

Hutchison, T., & Waters, A. (1987). *English for Specific Purposes: A learning-centred approach.* Cambridge University Press.

Hyland, K. (2002). Specificity revisited: How far should we go now? *English for Specific Purposes*, *21*, 385-395.

Hyland, K. (2004). Disciplinary interactions. Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing*, *13*, 133–151.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*, 41–62.

Hyland, K. (2016). General and specific EAP. In K. Hyland, & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes* (pp.17-29). Routledge.

Hyland, K. (2018). Sympathy for the devil? A defence of EAP. *Language Teaching, 51*(3), 383-399.

Jeaco, S. (2020). DIY needs analysis and specific text types: Using the Prime Machine to explore vocabulary in readymade and homemade English corpora. In M. Dodigovic, & M. Agustin-Llach (Eds.), *Vocabulary in curriculum planning: Needs, strategies and tools* (pp. 199-223). Palgrave Macmillan.

Koutsantoni, D. (2006). Rhetorical strategies in engineering research articles and research theses. Advanced academic literacy and relations of power. *Journal of English for Academic Purposes*, *5*, 19-36.

Kwan, B. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes*, *25*(1), 30-55.

Lee, D., & Swales, J.M. (2006). A corpus-based ESP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, *25*(1), 56–75.

Lee, J.J., & Casal, J. (2014). Metadiscourse in results and discussion chapters: A cross-linguistic analysis of English and Spanish thesis writers in engineering. *System*, *46*, 39–54.

Li, X., Jiang, K., & Ma. J. (2023). A cross-sectional analysis of negation used in thesis writing by L1 and L2 PhD students. *Journal of English for Academic Purposes, 64*, 101264.

Lillis, T. (2008). Ethnography as method, methodology and 'deep theorising': Closing the gap between text and context in academic writing research. *Written Communication*, *25*(3), 353-388.

Lin, P. (2023). Chat GPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, *3*(3), 100065.

Lin, L., & Evans, S. (2012). Structural patterns in empirical research articles: a cross-disciplinary study. *English for Specific Purposes, 31*(3), 150-160.

Link, S. (2018). Scaling up graduate writing workshops: From needs assessment to teaching practices. *Journal of Writing Research*, *10*(2), 357-399.

Maher, P., & Milligan, S. (2019). Teaching master thesis writing to engineers: Insights from corpus and genre analysis of introductions. *English for Specific Purposes*, *55*, 40-55.

Matzler, P. (2021). *Mentoring and co-writing for research publication purposes: Interaction and text development in doctoral supervision*. Routledge.

Michigan Corpus of Upper-level Student Papers (2009). The Regents of the University of Michigan. Available at: http://micusp.elicorp.info/

Moreno, A., & Swales, J.M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, *50*, 40–63.

Nesi, H. (2015). ESP corpus construction: a plea for a needs-driven approach. *Asp La Revue du Geras, 68*, 7-24.

Nordmann, E. (2023). Introduction to ChatGPT. Available at: https:// https://sway.office.com/p3A96vxXue9qqeFF?ref=Link0).

Nordrum, L., & Eriksson, A. (2015). Data commentary in science writing: Using a small, specialized corpus for formative self-assessment practices. In M. Callies, & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 59–83). John Benjamins.

Paltridge, B., Starfield, S., Ravelli, L., & Nicholson, S. (2012a). Doctoral writing in the visual and performing arts: Two ends of a continuum. *Studies in Higher Education*, *37*(8), 989-1003.

Paltridge, B., Starfield, S., Ravelli, L., & Tuckwell, K. (2012b). Change and stability: Examining the macrostructures of doctoral theses in the visual and performing arts. *Journal of English for Academic Purposes*, *11*(4), 332-344.

Paltridge, B., & Starfield, S. (Eds.) (2019). Special issue: 'Thesis and dissertation writing in a second language: Context, identity, genre'. *Journal of Second Language Writing*, *43*, 1-102.

Paltridge, B., & Starfield, S. (2019). *Thesis and dissertation writing in a second language: A handbook for students and their supervisors* (2nd ed.). Routledge.

Paltridge, B., & Starfield, S. (2020). Change and continuity in thesis and dissertation writing: The evolution of an academic genre. *Journal of English for Academic Purposes*, *48*, 100910.

Paltridge, B., Starfield, S., & Tardy, C. (2016). *Ethnographic perspectives on academic writing*. Oxford University Press.

Petrić, B. (2007). Rhetorical functions of citations in high- and low-rated master's theses. *Journal of English for Academic Purposes, 6(3)*, 238-253.

Petrić, B., & Castelló, M. (Eds.) in press. *The thesis writing journeys of Bachelor's and Master's students: A transnational European perspective.* Multilingual Matters.

Russell Group Universities (2023). New principles on use of AI in education. Available at: https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Shen, L., Carter, S., & Zhang, L. (2019). EL1 and EL2 doctoral students' experience in writing the discussion section: A needs analysis. *Journal of English for Academic Purposes*, *40*, 74-86.

Scott, M. (1996). *WordSmith tools*, version 1. Oxford University Press.

Scott, M. (2004). *WordSmith tools*, version 4. Oxford University Press.

Starfield, S. (2004). Why does this feel empowering? Thesis writing, concordancing and the corporatizing university. In B. Norton, & K. Toohey (Eds.), *Critical pedagogies and language learning* (pp. 138–157). Cambridge University Press.

Su, H., Zhang, Y., & Chau, M. H. (2022). Exemplification in Chinese English-major MA students and expert writers' academic writing: A local grammar-based investigation. *Journal of English for Academic Purposes*, *58*, 101120.

Sun, S. A., & Crosthwaite, P. (2022). "The findings might not be generalizable": Investigating *negation* in the limitations section of PhD theses across disciplines. *Journal of English for Academic Purposes*, *59*, 101155.

Swales, J.M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (pp. 150-164). Pearson Education.

Swales, J.M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Swales, J.M. (2019). The futures of EAP genre studies: A personal viewpoint. *Journal of English for Academic Purposes, 38*, 75-82.

Swales, J.M., & Feak, C. (2000). *English in today's research world*. University of Michigan Press.

Swales, J.M., & Feak, C. (2023). Task evolution in English for Academic Purposes writing materials: The case of "Information Transfer" to "Critical Commentary". *Journal of Second Language Writing*, *61*, 101017.

Thompson, P. (2013). Thesis and dissertation writing. In B. Paltridge, & S. Starfield (Eds.), *The handbook of English for Specific Purposes* (pp. 283-299). Wiley-Blackwell.

Thompson, P. (2016). Genre approaches to theses and dissertations. In K. Hyland, & P. Shaw (Eds.), *The Routledge handbook of English for Academic Purposes* (pp. 379-391). Routledge.

Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning & Technology, 5*(3), 91-105.

Thurstun, J., & Candlin, C. (1998). *Exploring Academic English: A workbook for student essay writing*. NCELTR.

Tribble, C. (2011). Revisiting apprentice texts: Using lexical bundles to investigate expert and apprentice performance in academic writing. In F. Meunier, S. de Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora* (pp. 85-108). John Benjamins.

Wärnsby, A., Kauppinen, A., Eriksson, A., Wiktorsson, M., Bick, E., & Olsson, L-J (2016). Building interdisciplinary bridges. MUCH: The Mälmo-University-Chalmers Corpus of Academic Writing as Process. In O. Timofeeva, Gardner, A.C., Honkapohja, A., & Chevalier, S.(Eds.), *New approaches to English linguistics* (pp. 197-211). John Benjamins.Wong, L. (2019). Implementing disciplinary data-driven learning for postgraduate thesis writing. In K. Hyland, & L. Wong (Eds.), *Specialised English. New directions in ESP and EAP research and practice* (pp. 195-213). Routledge.

Woodrow, L. (2018). *Introducing course design in English for Specific Purposes.* Routledge.

Wu, J., Zhao, C., Lu, X., & Jin, T. (2024). A rhetorical function and phraseological analysis of commentaries on visuals. *English for Specific Purposes, 73*, 33-45.

Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. *Journal of English for Academic Purposes*, *43*, 1-13.

Zhang, F., Zheng, Y., & Li, L. (2017). Using medical academic English corpus for postgraduate students' academic writing training. *Theory and practice in language studies*, *7*(10), 868-873.

Zhang, Y., & Su, H. (2023). (Mis)use of definition in Chinese EFL postgraduate students' academic writing: A local grammar-based investigation. *Journal of English for Academic Purposes*, *65*, 101283.

Zhou, H., & Jiang, K. (2023). 'The study has clear limitations': Presentation of limitations in conclusion sections of PhD dissertations and research articles in applied linguistics. *English for Specific Purposes, 71*, 34-47.

Zhou, H., & Liu, Y. (2021). If it had been conducted with a larger database…: A comparison of *If*-constructions in Chinese L2 learners' theses and published research articles. *English for Specific Purposes, 64*, 37-48.