Research Methods in Applied Linguistics 3 (2024) 100140

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Building a corpus of student academic writing in EMI contexts: Challenges in corpus design and data collection across international higher education settings

Dana Gablasova [a,*], Luke Harding [a], Raffaella Bottini [a], Vaclav Brezina [a], Haoshan (Sally) Ren [a], Giovanni Iamartino [b], Yingyu Li [c], Tanjun Liu [d], Laura Poggesi [e], Kristof Savski [f], Anuchit Toomaneejinda [g], Angela Zottola [e]

[a] Department of Linguistics and English Language, Lancaster University, County South, Lancaster LA1 4YL, United Kingdom
[b] University of Milan, Italy
[c] Xi'an Jiaotong University, China
[d] Xi'an Jiaotong-Liverpool University, China
[e] University of Turin, Italy
[f] Prince of Songkla University, Thailand
[g] Thammasat University, Thailand

ARTICLE INFO

Keywords:
Corpus construction
Written academic English
Corpus design
English as a Medium of Instruction
EMI
Corpus data collection

ABSTRACT

The article discusses methodological procedures and challenges in a project requiring multi-site, transnational data collection for the construction of a corpus of academic writing in EMI higher education contexts. Drawing on our decision-making experiences as a research team, together with empirical data generated through data collection logs recorded by a network of researchers involved in the project, we reflect on key issues in conducting the project and the solutions we found to address specific challenges. After describing the background to the project and the current status of the corpus, we focus on four broad challenges: (1) selecting partners and managing a multi-site project; (2) defining a working construct of academic writing; (3) categorising data according to disciplinary areas; and (4) managing data collection "on the ground". Throughout, we provide descriptions of our solutions to the challenges identified, and we conclude with a call for further publication of *corpus construction records* to provide greater transparency and detail around decisions and judgements made at all stages of a corpus construction project.

Introduction

In this article, we discuss a set of challenges encountered in the construction of a large-scale transnational corpus of student academic writing produced in English as a Medium of Instruction (EMI) higher education contexts (henceforth, the EMI Corpus). Corpora, curated datasets of language samples, form the foundation of corpus-based research (McEnery & Brezina, 2022). From start to finish, the process of corpus development involves a number of choices and decisions, which in turn determine the usability of the

* Corresponding author.
E-mail address: d.gablasova@lancaster.ac.uk (D. Gablasova).

https://doi.org/10.1016/j.rmal.2024.100140
Received 29 January 2024; Received in revised form 18 July 2024; Accepted 18 July 2024

corpus. While previous studies have reported on the composition of corpora in terms of data, metadata and corpus structure, less information is generally available about the process of corpus construction. As a result, there have been calls for greater transparency about methodological decisions in corpus construction, particularly as such decisions have implications for corpus representativeness and generalisability (Brezina, 2018; Egbert et al., 2022). To achieve a greater degree of transparency, it is crucial that "the design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken" (Adolphs & Knight, 2010, p. 40). This article therefore contributes to such practice by discussing the interplay of a set of particularly challenging theoretical, methodological and practical considerations involved in the design and data collection of the EMI Corpus.

EMI has become a major pedagogical trend in higher education world-wide, reflecting and shaping the status of English as a global language. In university-level EMI, the ability to use academic English is closely related to students' learning outcomes. However, there is surprisingly little empirical evidence about the actual use of English in these educational contexts (Jablonkai, 2021; Molino et al., 2022). The corpus discussed in this article—the EMI Corpus—was designed to address this gap and provide evidence about the use of English in students' academic writing across different EMI settings and disciplinary areas, with data collected from seven universities in China (2), Italy (2), Thailand (2) and the UK (1)[1]. The EMI Corpus seeks to complement existing corpora of student academic writing (e.g., the British Academic Written English corpus), contributing to a better understanding of the linguistic demands faced by students in EMI educational settings (an overview of the EMI Corpus is provided in the section below).

There are various inherent challenges involved in building a corpus of student academic writing (e.g., Alsop and Nesi 2009, Krishnamurthy and Kosem 2007, Römer and O'Donnell 2011, Stevens et al. 2020). These challenges are related both to (i) theoretical aspects (e.g., defining and operationalising the construct of student academic writing and classifying texts), and (ii) practical challenges involved in collecting the samples of writing. Both types of challenge are amplified in a large-scale, multi-site international project that involves data collection across different educational contexts (institutions and countries). While multi-site research allows for the creation of rich datasets capable of representing complex reality (Moranski and Ziegler, 2021), it also places greater demands on researchers in terms of planning, standardising procedures, negotiating cross-cultural differences, and developing theoretical frameworks that can be applied in different contexts.

To explore these challenges, this article reports on the first stages of constructing the EMI Corpus: corpus design and data collection. We first provide an overview of the corpus, including details about the motivation and aims of the project. Then, we focus on four key issues encountered in the construction of the EMI Corpus: (i) selecting partners and managing a multi-site project, (ii) defining the construct of academic writing, (iii) developing a framework for classification of texts, and (iv) dealing with practical challenges in data collection. Throughout, we discuss how a principled approach to multi-site research helped us address these challenges that emerged in the process of corpus construction.

Background: the EMI Corpus

Rationale for the corpus: understanding writing in EMI contexts

EMI refers to the educational practice in which academic subjects are taught through the medium of English. While some researchers stress that EMI refers to education in countries where English is not the primary language of communication (e.g., Macaro et al. 2018, Rose et al. 2021), others adopt a broader definition which also includes educational contexts in English-speaking countries (Fenton-Smith et al., 2017; Pecorari & Malmström, 2018). In this project we adopted the broader definition of EMI as it allowed us scope to explore, empirically, the extent of potential similarities/differences across a wide range of different educational settings. The adoption of EMI has been especially prominent in higher education (Sahan et al., 2021), with further growth predicted given direct governmental support for EMI in countries such as China and Japan (Galloway & Ruegg, 2020). Recent studies exploring the experiences of university students in EMI programmes in various countries (e.g., Korea, Turkey, China and Japan) reported that students find it difficult to engage with academic subjects due to the English language demands of their courses, with writing highlighted as particularly challenging (e.g., Kamaşak et al. 2021, Molino et al. 2022, Zhou et al. 2021).

Despite a considerable amount of EMI research, there is currently a need for more evidence about the nature of students' disciplinary writing in English at university level in non-English speaking countries. Most information about students' language experiences in EMI settings is based on self-reported data obtained through surveys and interviews (Jablonkai, 2021; Macaro et al., 2018). By contrast, a smaller number of studies have collected and analysed actual EMI language use. Among these, there has been a stronger emphasis on spoken discourse either in corpus-based studies (e.g., Dimova et al. 2024, Mauranen et al. 2010, Seidlhofer et al. 2011) or classroom-based research (e.g., Iliovits et al. 2022). A more limited number of corpus projects have included a focus on EMI writing (e.g., Paquot et al. 2022, Stevens et al. 2020), making it difficult to compare and synthesize findings across institutions and the wide range of countries where EMI is prevalent. Empirical evidence about written language use in EMI university-level study and how it varies across different disciplines and educational settings would be of vital importance for developing (i) a broad theoretical understanding of global academic written English, and (ii) policies and pedagogical approaches for EMI and targeted English for Academic Purposes (EAP) provision.

[1] See below for the rationale for including data from the UK in the EMI Corpus.

EMI Corpus overview

The EMI Corpus project was funded by the British Council Future of English research programme (www.britishcouncil.org/future-of-english). The project began in 2022, with data collection running since January 2023. Data collection has been managed from Lancaster University (UK), in collaboration with project partners in six participating universities: Thammasat University (TU) and Prince of Songkla University (PSU) (both in Thailand), Xi'an Jiaotong University (XJTU) and Xi'an Jiaotong-Liverpool University (XJLTU) (both in China), and Università Degli Studi di Milano (UniMi) and Università di Torino (UniTo) (both in Italy); further details about each institution and the rationale for their selection are provided below. Ethics approval was granted by the Faculty of Arts and Social Sciences Research Ethics Committee at Lancaster University in September 2022. Participating institutions often had additional requirements for ethics clearance and permissions that we met collaboratively during the early stages of the project before data collection could begin at each research site (see *Gathering written samples: The reality on the ground* for examples of challenges in navigating different institutional procedures).

Currently, the corpus contains over 3.1 million words from more than 1200 texts. When complete, the corpus is expected to reach approximately 3.5 million words from over 1500 texts. The corpus contains assessed writing from three disciplinary areas with considerable EMI provision world-wide (Sahan et al., 2021): (i) Social sciences and Humanities, (ii) Science and Technology and (iii) Business and Management. Each disciplinary area has been further subdivided to ensure the inclusion of key subject areas (see *Categorising data according to disciplines*). The corpus primarily focuses on the postgraduate (Master's) level of study; however, a decision has been taken to collect both undergraduate (Bachelor's) and postgraduate level writing for one disciplinary area – Business and Management – to allow for a focused cross-sectional comparison of student writing at these two levels. The undergraduate data were collected from final-year students. Table 1 provides an overview of the planned corpus structure, illustrating the balance across disciplinary fields, and including examples of core subjects/subject areas in each field.

In addition to samples of student writing, the following types of metadata were collected through a questionnaire: (i) student demographic data (e.g., gender, age, L1, English proficiency); (ii) students' academic habits and experience (e.g., how much and what type of academic writing in English they regularly do), and (iii) information about submitted texts (e.g., the mark received, the writing task instructions). Further metadata (e.g. the genre of the written texts) will be added following post-processing and further analysis of the texts. These metadata — collected from the participants or added as part of post-processing - can be used to further contextualise and interpret the findings from the corpus and the corpus will be searchable according to the relevant variables. Thus, for example, it will be possible to distinguish the disciplinary area, students' L1 and the specific EMI context (e.g. whether the data was collected in a predominately English-dominant or a predominately non-English dominant country, or the type of UG or PG programme it was collected from). Comprehensive information about the composition of the corpus, the metadata, and corpus access will be made available in corpus documentation upon the completion of the corpus.

Contextualising the corpus among other corpora of student academic writing

Corpus-based research on academic writing in English has a long tradition, with different types of corpora used to explore novice and expert texts. Two types of corpora are especially relevant to research on student writing in academic settings (Krishnamurthy & Kosem, 2007; Nesi, 2016). The first type represents texts produced by L1 and/or L2 writers on general academic topics, collected as part of regular EAP classes that focus on developing academic English writing skills or for specific research projects. For example, these corpora include the Hong Kong University of Science and Technology Learner corpus (Milton & Tsang, 1993), the International Corpus of Learner English (ICLE; Granger, 2003), the Corpus of Academic Learner English (CALE; Callies & Zaytseva, 2013) and the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023). The majority of these corpora contain writing on academic topics following typical academic genres (e.g., essays, argumentative writing). While the construction of these corpora often involved multi-site international data collection, the potential diversity in academic writing practices in these contexts was controlled, with topics and genres kept constant. The more controlled approach allowed researchers to study variation in writing due to variables such as the L1 background or L2 proficiency of the writers, with the data produced primarily for this purpose, rather than to capture a broader variety of disciplinary writing.

The second group of corpora represent disciplinary academic writing produced by students studying for an academic degree. Two major corpora of such writing have been collected at academic institutions in English-speaking countries. The 6.5-million-word British Academic Written English (BAWE; Alsop & Nesi, 2009) corpus contains over 6000 pieces of assessed, high-quality student writing from three UK universities. In the US context, the 2.6-million-word Michigan Corpus of Upper-level Student Papers (MICUSP; Römer &

Table 1
Corpus structure overview.

Academic level	UG	PG	PG	PG
Disciplinary areas	Business & Management	Business & Management	Humanities & Soc. Sciences	Science & Technology
Core subject areas	Business studies, Finance, Management science, Accounting, Administration		History, Literature, Sociology, Education, Linguistics	Life sciences (Chemistry, Biology), Physical sciences, Engineering, Computer Science
Proportion (%)	20	20	30	30

Note. UG = undergraduate, PG = postgraduate.

Swales, 2010) provides over 800 A-graded papers written by upper-level students (i.e., postgraduates and final-year undergraduates) at the University of Michigan. Both corpora contain writing by L1 and L2 users of English. In non-English-speaking contexts, the Varieties of English for Specific Purposes dAtabase (VESPA) corpus (Paquot et al., 2022) represents student disciplinary writing in ELF and EMI contexts from five European countries. It contains two million words and over 900 texts from L2 writers (undergraduate and postgraduate students) from a variety of L1 backgrounds. Currently, most of the texts in the corpus represent writing in linguistic courses (nearly eighty per cent), with further texts representing two other subject areas – literature and business communication. The Corpus of Chinese Academic Written and Spoken English (CAWSE; Stevens et al., 2020) is another corpus representing English use in education in a non-English speaking country. CAWSE is a multimodal corpus which includes a 1.5-million-word collection of English language writing by L1 Chinese students (exam scripts and coursework) at the University of Nottingham Ningbo China.

The EMI Corpus seeks to contribute to the second group of corpora, representing disciplinary writing in English collected in naturalistic settings. However, the corpus will be innovative in two ways. First, while the BAWE and MICUSP corpora contain writing from both L1 and L2 users, their academic context represents that of an English-speaking country. By contrast, the EMI Corpus will consist of data drawn mostly from academic institutions conducting English medium instruction in non-English dominant contexts. Second, corpora representing English writing in non-English dominant contexts have so far focused on a single institution (e.g., CAWSE) or currently contain a relatively limited number of disciplines (e.g., VESPA). The EMI Corpus will provide additional insights into the nature of English used by students in their academic studies by collecting data from a wide range of disciplines and from multiple national/institutional contexts representing distinct approaches to EMI higher education.

Methodological considerations and challenges in corpus construction

Having provided background information about the planned EMI Corpus and its rationale, in this section we discuss issues encountered and addressed during the design and data collection stages. We first consider general principles and challenges involved in conducting a multi-site research project involving diverse contexts and offer a rationale for the inclusion of specific research sites. Then, we provide examples of three further specific challenges that the multi-site research design posed for the theoretical and practical aspects of the project and the strategies we employed to address them.

A multi-site approach: principles, benefits and challenges

A multi-site, transnational approach to data collection was adopted in line with the project goal: to collect student academic writing in EMI contexts across different institutions and countries. Multi-site research is relatively common in social sciences such as psychology, sociology, and educational research that seek to capture the complexity of social reality and practices across different settings (e.g., Marcus 2009, LAARC et al., 2016). As discussed above, multi-site research has also been undertaken in several corpus projects to date.

A multi-site approach to research offers advantages at different stages of a project, from data collection to data analysis, informing discussion of findings and implications, due to the combination of cross- and within-site perspectives. First, the major advantage of this research design is that data collected at multiple sites has the potential to capture a more comprehensive picture of the observed phenomenon. In our project, this helped to enhance the representativeness of the sample by capturing the diversity and complexity of academic writing practices in different educational contexts. The diversity of such data also increases the external and ecological validity of the findings compared to single-site research, which is usually characterised by more homogeneous demographics (e.g., cultural and linguistic background; Moranski & Ziegler, 2021). The multi-site approach and the ability to triangulate findings across different sites of academic practice increases generalisability of the findings and the overall representativeness of the dataset. Since the target EMI domain is vast and multi-faceted, the boundaries of the operational domain for this corpus were defined as academic texts in English produced and used at the selected universities in the selected disciplinary areas. This helped us identify relevant categories of texts for the sampling process (cf. Egbert et al., 2022, p. 58-60). Findings from such varied datasets can inform (pedagogical) practice across a wider variety of contexts. Further, large samples of language use can be gathered more efficiently from multiple sites, contributing to the generalizability of findings and statistical power at the analysis stage (Moranski & Ziegler, 2021).

The second key benefit of multi-site projects is related to the collaborative nature of such research and the ability to draw on the collective expertise of team members and their insights into local research sites (Kwon et al., 2018). Knowledge sharing is crucial at different stages of the project: (i) at the conceptualisation stage, when planning the practical aspects of the data collection and agreeing on the theoretical frameworks that are applicable to and inclusive of practices at different research sites; (ii) during data collection, enabling collaborators to share experiences when issues arise, and to identify strategies more quickly, and (iii) at the data analysis and interpretation stage, when the combined experience and expertise of the team members can lead to "a more holistic understanding of findings" (Moranski & Ziegler, 2021, p. 223). For example, considerable knowledge of local contexts was required during the conceptualisation of the project when developing a construct of 'student academic writing' that was theoretically sound but also inclusive of the varied writing practices across participating universities/departments (see *Corpus design: Defining and operationalising 'academic writing'*). In another example, a solution developed to deal with a particular issue with data collection at one research site was shared with other researchers who were able to adapt the strategy in their contexts (see *Gathering written samples: The reality on the ground*).

Due to its complexity, large-scale multi-site projects also present considerable challenges related to the logistics of carrying out the research while ensuring the integrity and rigour of procedures. Three principles have been identified to play a key role in conducting effective multi-site projects (e.g., LAARC et al., 2016, Moranski and Ziegler 2021). First, it is crucial for data collection to be managed centrally and to follow standardised procedures to ensure methods are applied consistently at each site, enhancing the validity and

comparability of the data. Second, previous studies have also stressed the need for flexibility and adaptability in methodological procedures, allowing for localised and context-appropriate (e.g., culturally sensitive) solutions (LAARC et al., 2016). To reconcile these seemingly contradictory principles, it is necessary to establish which procedures/instruments need to be strictly adhered to and which methodological aspects can be localised so they will not compromise the reliability, validity and comparability of the data. In our project, one example of a centralised strategy which allowed for localised approaches was the questionnaire used to collect information about students' backgrounds and academic writing habits. While the same set of core questionnaire items was used across participating universities, the most suitable format of the questionnaire for each site (electronic, paper-based, or a combination of both) was determined following piloting and discussions with researchers representing each university. Finally, the third principle concerns communication among researchers at different sites. Effective communication is necessary for an in-depth understanding of each site (e.g., the setting, values and principles common in the environment) in order to guide the conceptual framing of the project as well as practical steps in carrying it out.

Following recommendations for conducting multi-site projects (e.g., LAARC et al., 2016, Moranski and Ziegler 2021), the selection of the research sites for the current project was guided by the theoretical and practical goals of our project, and the suitability of and access to the data. To reflect a broad range of EMI practice, we selected four institutions located in ODA[2] countries (China [2] and Thailand [2]) and three in Europe (Italy [2], UK [1]), representing university settings with different cultural and historical traditions, and different types of HE provision (e.g., public, private, transnational). More specifically, participating institutions in each country were selected based on criteria including: (i) the size and range of their EMI provision, to ensure data availability and to maximise both representativeness and comparability, and (ii) their tradition and experience with EMI. Two institutions for each country were included, to enable the evaluation of cross-country and cross-institutional differences in EMI writing practices. (Only one institution in the UK was included, as existing datasets such as the BAWE corpus can be used as an additional reference point for those who wish to perform comparative analyses).

The following universities were ultimately selected as the data collection sites. *Thammasat University (TU)* and *Prince of Songkla University (PSU)* are large universities in Thailand. TU is situated in the capital city Bangkok, while PSU is the largest academic institution in the southern region of Thailand. They are both leading teaching and research institutions in Thailand offering EMI academic programmes. TU is a public research university and the second oldest university in Thailand, attended by over 33,000 students, while PSU, also a public university, is attended by 35,000 students across five campuses. *Xi'an Jiaotong University (XJTU)* and *Xi'an Jiaotong-Liverpool University (XJLTU)* are two large universities located in the north-western and eastern parts of China respectively. XJTU, attended by 44,300 students, is a public university in Xi'an, Shaanxi. It is a research-oriented university and a member of the C9 League, an alliance of nine prestigious universities in the country. Its international degree programmes started in the 1950s, making it one of the first Chinese universities to accept international students. XJTLU, with 25,000 students, is a private, transnational university located in Suzhou. It was founded in 2006 as a partnership between the University of Liverpool (UK) and Xi'an Jiaotong University; today it is an independent university. Due to the partnership with Liverpool University, XJTLU has a very strong tradition of large-scale EMI provision in a wide range of subjects. Both *Università Degli Studi di Milano* (UniMi) and *Università di Torino* (UniTo) are leading, large public universities in the north of Italy, located, respectively, in the second and fourth largest cities. Both offer large EMI provisions. UniMi, founded in 1923, is attended by 62,000 students, while UniTo, founded in 1404, is attended by 81,700 students. *Lancaster University* (LU), situated in the north of England, is attended by 12,000 students. It is a research-oriented institution with a high internationally recognised standard of teaching and research.

Corpus design: defining and operationalising 'academic writing'

The decision about what language samples to include in a corpus constitutes a key component of corpus design, with implications for corpus representativeness and the type of research questions that can be pursued with the corpus (Egbert et al., 2022). The aim of the current project is to compile a corpus of student writing from different universities and countries; thus, it is crucial to establish a construct of academic writing that can be meaningfully applied across different higher education institutions. This section discusses the theoretical, methodological and practical reasons behind decisions about the construct of student academic writing adopted in the project.

Academic writing is a complex phenomenon, encompassing a varied set of writing practices that reflect the enormous diversity of academic actors, communicative aims, values, motivations and intellectual practices involved in academic study and research (Hyland, 2006). Writing practices can further vary according to academic values and norms in different disciplines, institutions and academic cultures typical of different countries (Paltridge, 2004). Student academic writing can be defined and operationalised with emphasis on different aspects of this practice, resulting in collections of different texts. For example, academic writing can be defined broadly as all writing completed by students in an academic setting, which would include assessed assignments alongside informal emails exchanged during a group project and notes taken during group discussions. A narrower definition may focus only on writing tasks set and assessed by instructors, thus excluding many of the previous examples.

A specific operationalisation of the construct of student academic writing—and its impact on the inclusion of texts—can be seen on the example of the BAWE corpus (Alsop & Nesi, 2009). The corpus includes (i) assessed student assignments (formative and summative) which (ii) were submitted electronically, (iii) met a certain academic standard (i.e., merit or distinction) and (iv) met a certain

[2] ODA refers to Official Development Assistance, a form of government aid. Countries receiving ODA are low and middle-income countries (based on their gross national income) who are recipients of governmental aid to promote their economic and social wellbeing.

standard of English proficiency. These criteria directly influence some of the characteristics of academic writing represented in the BAWE corpus. For example, electronically submitted assignments are likely to consist of texts for which the writers had the ability to plan and edit, and to check and incorporate information from references. These features of the writing process may in turn influence different aspects of academic writing such as the complexity and accuracy of language or referencing and in-text citation patterns. Further, the two criteria stipulating the quality threshold in terms of the mark and proficiency level are likely to exclude a considerable proportion of writing produced by students at the participating universities, and thus narrow down the construct of student academic writing represented in the BAWE corpus to what Ädel and Römer (2012, p. 5) referred to as "target writing" in EAP contexts.

In designing the EMI Corpus, we drew on operationalisations of academic writing used in existing corpora (such as the BAWE) and, following consultations with researchers at the participating universities, we extended them further to capture the variety of writing practices at these institutions. This allowed us to retain comparability with other corpora of student academic writing, while being inclusive of the different educational settings in the current project. The academic writing in our project was defined as disciplinary writing in the form of texts (e.g., assignments, exam papers) submitted for summative assessment as part of a degree programme, thus capturing the more formal and publicly oriented type of student writing. Both electronically submitted and hand-written pieces were included, seeking to capture writing that involved planning time as well as that produced under timed (exam) conditions. Hand-written texts still represent a major source of writing in some university contexts (e.g., in our data collection, the University of Turin and the University of Milan). Being inclusive of such text types can broaden the scope of understanding of the linguistic demands in EMI settings and help identify typical features of timed, hand-written texts.

Second, any assignment that received at least a pass mark was included. The rationale for extending the scope from "target writing" was two-fold: (i) Participating universities follow different marking criteria and we wanted to avoid excluding work of a similar standard that may have been marked differently at these institutions. The mark received for each assignment and the marking system that was used were documented so the marks can be transferred to a unified scale during the data processing stage once the data collection has been completed. (ii) While corpora representing "target writing", such as MICUSP and BAWE, are valuable for both research and pedagogical purposes, Krishnamurthy and Kosem (2007) stress that "without lower-grade student texts, there is no opportunity for monitoring progression, or for making comparisons with the higher-grade student writing" (p. 366). Therefore, both more and less successful pieces have been included in the EMI Corpus to gain insight into the wider range of writing produced in EMI environments, which can be, among others, used to inform EAP practice about systematic variation in writing of different quality.

Third, a relatively low threshold for the minimal length of texts was adopted, with the requirement of at least a hundred words per piece. This inclusive approach allowed for different writing practices across institutions and disciplines to be included in the corpus. This had a particular impact on writing from students in the Science and Technology domain whose assignments often consist of (a series of) relatively short written tasks (e.g., 200–400 words) and/or contain a large proportion of figures, code, equations or different types of visualisations. Excluding these texts would reduce the range of writing received from disciplines whose writing practices typically contain such features.

Finally, texts which resulted from different types of group-work were included alongside individually written production (with the information recorded in the metadata). While individually-written texts may be easier to classify and analyse in corpus studies (i.e., as they are associated with only one set of writer data, such as their L1), pair-work and group-work represent an important component of learning and training in many disciplines and occupational areas, with increasing emphasis being placed on the development of communicative skills (e.g., interpersonal and intercultural competence).

These decisions and criteria prioritised – as much as possible – an inclusive approach to the operationalisation of student academic writing designed to maximise the opportunities offered by access to multiple educational sites. As demonstrated in our project, broadening the inclusion criteria for corpus data collection has implications for theoretical and practical aspects of corpus design and construction. From the theoretical perspective, broader criteria allow researchers to systematically capture the complexity of EMI writing and provide empirical evidence for investigating characteristics of and variation in EMI production. However, on the methodological level, the greater variety or "messiness" of the data poses greater challenges for processing the data (e.g., dealing with equations, digitising hand-written texts). On the practical level, the collection, classification and processing of more varied text types have implications for the feasibility of the project in terms of the timeframe and funding, increasing the demands on time and resources. This demonstrates how theoretical considerations are often mitigated by practical concerns and resources available; the practical challenges are further described below (see *Gathering written samples: The reality on the ground*).

Categorising data according to disciplines

Another key decision to be made in planning the construction of the EMI Corpus is related to the classification of student writing according to its disciplinary affiliation. Such classification is closely linked to the questions of corpus composition and balance, comparability of a corpus with other relevant datasets, and the ability to offer insights into disciplinary variation in current EMI writing practices. Given the diverse and evolving nature of academic writing practices, finding meaningful and valid categories is not straightforward. Furthermore, the multi-institutional approach to data collection creates additional challenges due to the diverse traditions, practices and approaches to disciplinary and interdisciplinary classification within different universities and national educational contexts.

Disciplinary classification has traditionally been an important variable in corpus-based research on academic writing (e.g., Durrant 2017, Gardner et al. 2019, Hardy and Römer 2013), reflecting evidence that "communication practices are not uniform across academic disciplines but reflect different ways of constructing knowledge and engaging in teaching and learning" (Hyland, 2006, p. 8). Systematic discipline-related variation has been found with regard to a broad range of linguistic features and discourse practices which

have been associated with typical forms of communication needs in different academic fields and the occupational domains related to them (e.g., Gardner et al. 2019, Hyland 2006). Typically, disciplinary affiliation has been used to guide data collection and analysis in academic corpora. For example, when creating the MICUSP corpus, Ädel and Römer (2012, p. 6) stated that "it was an important goal in the compilation process to achieve a relatively balanced distribution with respect to discipline". However, despite its significance, often only limited information is made available about how disciplinary affiliation has been determined in corpora of academic writing. As Durrant (2017) argues, "the fact that these researchers provide little indication of how these categories were determined is striking because it is a decision which has important consequences for the analyses which follow" (p. 166). Ultimately, the disciplinary framework used to guide the construction of a corpus of student academic writing plays a central role in determining the nature of disciplinary-specific variation that can be observed, with major implications for EAP theory and practice in EMI contexts.

Corpora representing student writing have generally employed two disciplinary categories in their design: i) broad disciplinary areas, and ii) academic disciplines. For example, the MICUSP corpus lists student texts according to 16 disciplines and four "academic divisions" (Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences). The BAWE corpus used four major "disciplinary groups" (Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences), with 28 individual disciplines listed. These categorizations were guided by different principles, such as reflecting the local university organization (the MICUSP) or comparability with other corpora (the BAWE). Regarding the classification into individual disciplines, existing corpora mostly used the names of the departments in which the texts were produced as the primary disciplinary identifiers, although this practice sometimes proved problematic when establishing meaningful boundaries between disciplines. As Nesi (2015) observed, "boundaries between related fields of study are permeable, and within discipline-specific programmes there are often outlying modules, for example on the history of mathematics in a mathematics programme, or on business law for a degree in business" (p. 8). Further, using departments for the purpose of disciplinary classification in academic corpora has resulted in related academic subjects being subdivided or grouped into different disciplines. For example, if drawing on the existing departmental structure, 'Engineering' would represent a single discipline at Lancaster University, while in the MICUSP corpus it was subdivided into the following disciplines: 'Civil and Environmental Engineering', 'Industrial and Operations Engineering' and 'Mechanical Engineering'.

As demonstrated by these examples, defining the boundaries of academic disciplines and classifying student writing accordingly is not straightforward. Moreover, while disciplines have been traditionally defined in terms of their knowledge areas, objects of study and related procedures, methods and aims — all of which determine the shape and scope of teaching and research (Krishnan, 2009) - the nature of disciplinary identity has been undergoing considerable changes in the past two decades. While there are typically still recognisable core characteristics of individual disciplines, disciplines can also be internally relatively heterogeneous, with new practices, discourses and sub-disciplines emerging (Barrow et al., 2020; Trowler, 2012). Further, greater emphasis in research and teaching is increasingly being placed on addressing real-life problems, which has resulted in the emergence of more interdisciplinary academic units in higher education (Krishnan, 2009; Trowler et al., 2012). Such practices can be illustrated by the MA programme in "Environmental Change and Global Sustainability" offered by the Department of Environmental Science and Policy at the University of Milan[3]. The programme focuses on environmental sustainability as a central real-world challenge, and the description on its website states that "addressing this challenge requires a multidisciplinary approach that overcomes the usual boundaries of scientific disciplines". The programme therefore seeks to equip the students with "expertise in the hard- and life-science components of environmental studies as well as in their economic- and social-science components." As illustrated in Table 2, the courses in the programme draw on the theory and methodology of Environmental science, Economics, Management, Chemistry and Law. It is especially noteworthy that many courses appear to cross the boundaries of 'soft' and 'hard' sciences, which have been consistently found to result in different discursive practices (e.g., Durrant 2017).

Arriving at a meaningful disciplinary categorization becomes even more difficult when working with data from higher education institutions across different countries, with different traditions of disciplinary categorization. For example, at Lancaster University, an MA in Digital Humanities is offered by the Department of History, while at the University of Milan, a programme with the same title is taught in the Department of Computer Science. In both cases, the programme is delivered in collaboration with other departments.

Thus, due to the increasing interdisciplinarity and idiosyncrasies in how higher education institutes organize their departments, programmes, and courses, disciplines cannot be determined using departmental affiliation only. In the EMI Corpus project, we have therefore approached disciplinary categorization as a multi-step process, addressed at two stages of the corpus compilation: the data collection stage and the data processing stage.

First, following other large-scale corpora of student academic writing, broad disciplinary areas were employed as the guiding principle in determining the data collection strategy, with three areas used for this purpose: (i) Humanities and Social Science, (ii) Science and Technology, and (iii) Business and Management. Rather than drawing on the academic structure of a particular university – which would be problematic given the multi-institutional research design – these disciplinary areas were based on the categories identified in a major disciplinary framework, the Higher Education Classification of Subjects (HECoS) developed by the Higher Education Statistics Agency (HESA), to ensure that the classification would be applicable across different universities. Thus, we diverge from the BAWE and MICUSP corpora in including 'Business and Management' as a distinct disciplinary area, given (i) its current position as a major independent disciplinary field with a number of subdisciplines (as identified by HECoS), and (ii) the large EMI provision focused on this subject area (Sahan et al., 2021). Further, for each of these disciplinary areas, 'core' or 'prototypical' subjects were identified using the HECoS; for example, 'Science and Technology' was further sub-divided into 'Life sciences', 'Engineering' and

[3] https://www.unimi.it/en/education/master-programme/environmental-change-and-global-sustainability-ecgs

Table 2
Selected courses in MA in environmental change and global sustainability, University of Milan.

Agricultural and Natural Resource Economics and Policy
Applied Environmental and Resource Economics
Bioresource and Pollution Control Technology
Chemistry of Natural Processes and Technologies for the Environment
Environment Change and Public Health
Environmental Law
Environmental Geochemistry
Economic Botany and Zoology
Methods in Ecotoxicology
Sustainability Accounting and Management

'Computer Science.' Using this set of disciplinary categories, the coordinating team at Lancaster University and researchers at each partner institution discussed how this framework could be applied in their local contexts (e.g., how it can be 'translated' into the local academic units) in order to achieve the desired range and balance of data. Such discussion was particularly important in the case of departments/programmes pursuing interdisciplinary or multi-disciplinary research and teaching.

Second, in addition to using broader disciplinary categories to guide the data collection, we have also collected further information about each text to facilitate a systematic disciplinary classification of texts at the data processing stage. Specifically, as recommended by Krishnamurthy and Kosem (2007), and following the practice of other academic corpora (e.g. the BAWE), we have recorded three levels of information: the course (module), the degree programme, and the department that the data came from. This information will be used in the data processing stage to classify each text according to the HECoS disciplinary categorisation; the information will also allow for a direct comparison with other existing corpora of student academic writing, which may have used different approaches to disciplinary classification. In addition, to evaluate the validity and internal consistency of the disciplinary categories in the EMI Corpus, further bottom-up analyses will be performed (e.g., Durrant 2017, Gardner et al. 2019). Collecting and recording the information that would allow for a systematic disciplinary classification of texts was crucial; however, for the information to be meaningful, a considerable amount of further processing was required in order to clean and complement the data provided by students. For example, some students provided the names of the relevant modules/programmes in their local language (e.g., Chinese) or they just used a course code (e.g., LING428) which may not be meaningful to researchers outside of these universities. Thus, the process of text classification required an in-depth understanding of each institution and further emphasized that the local knowledge and close communication among researchers at different sites are crucial for successful multi-site research projects.

Gathering written samples: the reality on the ground

In this section, we discuss practical challenges (and solutions) encountered in our project "on the ground" when collecting written samples across the different educational settings described above. As each EMI site was unique – with internal institutional structures and hierarchies set within different cultural and linguistic environments – each presented new challenges which required locally-developed solutions when implementing the central data collection strategy. The challenges encountered in data collection have direct implications for shaping the content and structure of the corpus, and for issues of representativeness and balance. To elaborate

Problem (Aim)
Please use this section to describe and contextualize your issues/aims: 1. What was the aim you were trying to achieve? 2. What were some key/different aspects of this aim? 3. What were the challenges encountered?
Solutions
Please use this section to record the strategies you used, and why they worked or did not work. You may address questions such as: 1. What strategy/strategies have you used? 2. How did the strategy/strategies work for you – why did it work or didn't work? 3. What were the difficult aspects of solving the issue? 4. What helped you with dealing with this challenge?

Fig. 1. Template for data collection log entries.

on these issues, we draw on data collection logs completed at seven research sites as well as research notes and communication between researchers. The data collection logs were designed using a problem-solution framework to elicit reflections on challenges encountered in collecting written samples, and the strategies used to solve these problems locally. Fig. 1 shows the template that was used.

Data collection logs were analysed by members of the Lancaster team using a thematic analysis approach (Braun & Clarke, 2021) to identify common themes. The themes were then further complemented by information from notes and personal communication between researchers. In the section below, we describe three themes (and associated sub-themes) that emerged from this analysis.

Theme 1: Gaining access: navigating institutional structures

Gaining access to research sites and participants has long been recognised as a major challenge in conducting research with human participants. This challenge typically involves both a formal and a personal dimension: (i) *navigating institutional administrative requirements*, and (ii) *communicating effectively with institutional gatekeepers* (Feldman et al., 2003). Managing both dimensions proved crucial in gaining access to participants in all research sites in the EMI Corpus project.

On the first dimension, the initial stage of data collection typically required permission at the level of different academic units within an institution (e.g., faculty, department). Similar to issues recorded during the construction of other academic corpora (Alsop & Nesi, 2009; Stevens et al., 2020), institutional structures – internal systems, management hierarchies, and communication systems – posed challenges across all sites in our project. These differed in scope and type across the seven institutions involved, and were in some cases hard to anticipate (and therefore to plan for). Given that our data collection plans required access to different organisational units within each university (faculties, departments and classes), in some cases multiple permissions had to be gained within the same institution. For example, at one research site, to collect data in a particular faculty, the researchers needed to prepare a detailed document explaining the project for the faculty research unit and then obtain further approval from various units within the faculty before seeking final approval from the faculty dean. Unexpectedly, the same process then needed to be repeated for access to each faculty in the university where data collection was due to take place although permission at the university level had already been granted at the beginning of the project. In another example of the administrative differences across the institutions, while some institutions accepted the ethical approval granted by Lancaster University (where the project was initiated), two universities required the researchers to go through a local ethics procedure as well. While the researchers at one of the institutions were able to use some of the materials already prepared for the Lancaster University ethical approval, the other university required a new set and format of documentation. As a result, steering our way through these processes often took longer than anticipated as they involved unfamiliar procedures.

On the second dimension, gaining access required not only satisfying the administrative processes described above, but also obtaining permission from gatekeepers, such as deans, heads of department, or teachers (who would grant access to classes for recruitment messaging, or to exam sessions for data collection). At this personal level, permission could be delayed or denied not only due to administrative procedures, but also because of issues of trust or lack of clarity about the aims of the research. For example, at one site, the researcher had to attend a meeting in which they were "questioned extensively on the merits of the research" leading to a sense that "there seemed to be little trust in myself or the project team to handle the research." Issues of personal trust, familiarity with linguistic research, and perceived risks had to be carefully addressed by researchers at multiple sites.

The following strategies proved useful in dealing with both types of access-related challenge:

(i). Being prepared to communicate the goals of the project clearly to different audiences: Unfamiliarity with language-related research and its aims sometimes resulted in issues of trust and delays in granting access. This was especially the case with disciplinary areas that typically do not conduct research that requires human participants; when the gatekeepers had some experience or familiarity with language-related research, this often led to greater trust and cooperation. To address this issue, researchers need to allow time to repeatedly explain the goals of a project, as well as linguistic research more generally. In our project this was done, for example, through written FAQ documents and – similar to Kwon et al. (2018) – presentations to and discussions with the staff in different academic units (e.g., departmental sections). Showing examples of findings based on previous corpus projects conducted by the researchers also helped to establish the academic credentials of the project team and to address the perceived risk associated with giving permission for data collection.

(ii). Drawing on existing personal relationships: Personal relationships proved crucial for gaining access, especially in situations (as described above) where trust needed to be established. For example, one researcher's former Master's student was employed in a faculty where access was proving difficult. The researcher asked their former student to join the project, and this new recruit was able to anonymise participant data before rather than after the data was uploaded – something that the faculty had been concerned about and which affected the planned timeframe for the data collection. This provided a sufficient level of control over anonymity to satisfy the faculty administrators. In addition, a personal contact could also act to 'vouchsafe' for the researchers/the project when establishing new contacts and an introduction via a shared contact proved crucial in establishing trust and cooperation.

(iii). Prioritising personal, face-to-face communication: While it initially seemed efficient to establish contact and discuss the research via email channels, this often led to delays in gaining access (possibly due to the issues with establishing trust and communicating about perceived risks discussed above). A deliberate strategy of turning to personal, face-to-face meetings and cultivating trusted contacts often helped to resolve these issues and to speed up the process of gaining required permissions/access. For example, at one research site, scheduling a face-to-face meeting with the researcher's own faculty dean led to direct progress after two months of trying to make headway through email correspondence. Setting up personal meetings was more difficult in unfamiliar faculties; nevertheless, the researchers asked their small collection of known contacts for help, leading to a wider

network of connections, telephone numbers, and other leads. This personal approach was seen as fundamental for gaining access to classrooms (where recruitment often occurred). As one researcher in the study explained, a "deep personal connection" was important to gain access to a classroom, which is considered a "personal space".

Theme 2: Recruiting a sufficient number of students

The second broad theme relates to the challenges involved in recruiting a sufficient number of students to meet the aims of the study. Even when access and permissions were granted, gathering data at the scale required proved difficult, though to varying degrees, across the institutions involved in the study. The challenges involved two dimensions: (i) *establishing and maintaining contact* and (ii) *addressing students' concerns*.

On the first dimension, two main strategies were employed to establish and maintain contact with students following previous corpus construction projects (e.g., Alsop and Nesi 2009, Kwon et al. 2018, Römer and O'Donnell 2011, Stevens et al. 2020): (a) contacting students via their departments or teaching staff, and (b) reaching students individually. The first strategy required researchers to engage with programme directors, academic colleagues, and programme development advisors. The major benefits of this approach included more control over the balance of the data (e.g., the subject areas reached) and minimising issues of trust, as students perceived information about research projects sent by their department as more trustworthy. However, this approach was highly susceptible to the problems of delays in email communication described above. The second approach – contacting students directly offered greater efficiency and flexibility (while remaining fully compliant with ethical permissions) but presented more challenges with achieving balance in the data. Further, this method was time-intensive as researchers had to explain the project individually to participants with different levels of background knowledge and academic experience. In most cases, a combination of the two approaches was used. For example, one researcher described a situation in which the faculty dean at their institution had provided approval, but where a lack of response from a particular department within that faculty was causing a general delay. In this case, the researcher approached the university's international office at their institution, which could communicate directly with international students, and used their mailing list to send information about the project to a large number of participants across different subject areas without increasing workloads in individual departments.

Within both approaches, researchers tried a number of different strategies – some resembling those reported in previous corpus building projects (e.g. Alsop and Nesi 2009, Römer and O'Donnell 2011) – to attract students' attention. These required a high degree of flexibility and creativity at each site, in combination with a good understanding of local culture and values. First, using the financial incentives made available through project funding, researchers at different sites offered Amazon vouchers, honoraria, book tokens, coupons for coffee, McDonalds, KFC breakfasts, and movies. Second, posters with QR codes were put up around campuses. Third, researchers recruited student research assistants who distributed information about the project to students on university campuses . While some of these methods appeared to be successful in engaging students initially, many students often failed to follow through after first contact. In response, to maintain contact, researchers set up social media groups to which they invited students, shared links to the questionnaires, and encouraged participants to invite their friends. This all led to a gradual increase in participant numbers, but required considerable time and effort on the part of the research teams. In contrast to Römer and O'Donnell (2011), who noted that for the MICUSP corpus sending mass emails to students via mailing lists was the most efficient recruitment strategy, in the EMI Corpus project the researchers reported particular success from collaboration with student representatives across different majors. Student representatives were able to disseminate information about the project efficiently within their social groups. Strategies that proved effective at one site were shared between project research teams, who could then adapt these in their context.

Addressing students' concerns was the second dimension of the recruitment challenge. While students were often interested in the project and willing to contribute, they expressed reservations about what would happen with their data. Specific concerns appeared to relate to feeling embarrassed (if they submitted texts which had received a low mark), or concerns about intellectual property rights (if their work was very good and potentially suitable for publication). This latter issue was particularly common among postgraduate students; in at least one of the contexts, it was a degree requirement for postgraduate students to publish their work in a local or international journal. Students worried that if they contributed to the corpus, their work could appear somewhere in a repository prior to publication from where it might be plagiarised. Teachers held a similar concern at some research sites regarding, particularly, the inclusion of exam materials in the corpus.

Solutions to these problems were two-fold. In some cases, we avoided collecting pieces of writing that students viewed as sensitive (for the reasons discussed above). Although ideally it would have been useful to include these pieces of writing in the corpus, a compromise was required to ensure that trust between the researcher and the participants was maintained. In most other situations, students and teachers were reassured by further clarification about how data would be anonymised, who would have access, and how it would be made available in the final corpus (these subjects were addressed in the participant information sheet, which explained, for example, that other researchers would be able to access the data through corpus interfaces such as Sketch Engine (Kilgarriff et al, 2014)). Being prepared to explain and discuss our data management procedures with multiple audiences was therefore vital, to expand on points that had already been made in the participant information sheet, and to provide more detail about any specific cases. In this sense, a level of "corpus-literacy" was required both for researchers and for the participants to be able to negotiate trust and gain fully informed consent for project participation.

Theme 3. Achieving balance

The final theme related to the challenges involved in collecting data that reflected the planned structure of the corpus described in *EMI Corpus overview* above. Researchers at each site aimed to collect specific quotas of student writing at undergraduate and postgraduate level, and across broad disciplinary areas. While the planning stage involved an initial survey of EMI provision at each

participating university to establish the availability of the data from the target disciplinary areas, it was not always possible to collect the data due to (i) uneveness of EMI provision across institutions, and (ii) changing enrolment patterns/institutional policies. Furthermore, it took over a year from when the initial survey at each site was conducted for data collection to begin due to the time involved in submitting the research proposal and obtaining project funding. Consequently, one challenge in carrying out the planned data collection strategy related to changing enrolment patterns and changes to institutional policies over that period. For example, at one site, the researcher aimed to recruit 40 postgraduate students from Social Sciences, but found that a high number of EMI programmes within those discipline areas had not recruited any students that academic year. In a different example – highlighting the specific challenges of the EMI environment – the researcher found that a course on English and American Literature had been modified to focus its curriculum on Comparative Literature. As part of this change, the language of instruction for that course switched from English to the official national language. These issues were addressed by collecting student writing in a number of related programmes of study and finding suitable alternative disciplines in line with the central aims of data collection.

Conclusion

The aim of this article was to illustrate and reflect on different types of challenges that emerged during the construction of the EMI Corpus, with particular emphasis on the benefits and unique difficulties related to a multi-site research design. We reflected on these challenges with two goals in mind. First, we see value in documenting the challenges and the solutions to those problems to develop insights about the corpus construction process that will be of use to other corpus researchers involved in multi-site, international research, carried out in diverse (e.g. multicultural, multilingual) settings. Second, we see the acknowledgement and discussion of these issues as a vital step in illustrating the fine balance between pragmatism and idealism in developing a corpus that maintains ecological validity (capturing naturalistic language use in real-world written samples drawn from very different contexts). The process of documenting key challenges can help researchers to follow the principles of the structured approach proposed by Egbert et al. (2022), which considers the relationship between the target domain, operational domain and the sample itself. In particular, we hoped to demonstrate the close relationship between the theoretical and practical aspects of corpus construction and to highlight the fact that the quality of the final corpus (and the findings based on it) to a large extent depend on the quality of decision-making during the design and data collection stages, although this may often be less 'visible' in the information about the corpus that is provided. Thus, we call for more published reflections and explanations of decision-making processes in corpus construction, what we term *corpus construction records,* which can ultimately feed into a more comprehensive set of documentation for published corpora. This remains important as "in order to move forward, a discipline needs to continually reflect on its practices and draw on this reflection when developing new resources, methods and theories" (Gablasova et al., 2019, p. 126).

This article also highlights how the process of corpus construction can remain sensitive to a growing awareness of diversity around key phenomena in applied linguistics. There is an increasing acceptance that applied linguistics requires more context-sensitive, pluralistic approaches to capture complexity in language and to take account of the perspectives of language users in the periphery (see e.g. De Fina et al. 2023). We believe that, despite its inherent challenges, the multi-site, pluralistic approach to collecting corpus data highlighted in this article is particularly appropriate for understanding complex phenomena like EMI language use. We hope that our experience will encourage more researchers to embark on the construction of corpora that investigate similarly complex global phenomena, contributing to the continued diversification of applied linguistics knowledge.

CRediT authorship contribution statement

Dana Gablasova: Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization, Writing – review & editing. **Luke Harding:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Raffaella Bottini:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis. **Vaclav Brezina:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Haoshan (Sally) Ren:** Writing – review & editing, Writing – original draft, Data curation. **Giovanni Iamartino:** Data curation, Writing – review & editing. **Yingyu Li:** Writing – review & editing, Data curation. **Tanjun Liu:** Writing – review & editing, Data curation. **Laura Poggesi:** Writing – review & editing, Data curation. **Kristof Savski:** Writing – review & editing, Writing – original draft, Data curation. **Anuchit Toomaneejinda:** Writing – review & editing, Data curation. **Angela Zottola:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The study was supported by the British Council Future of English grants scheme which funded the creation of the EMI Corpus project and by the ESRC Grant ES/R008906/1.

References

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics, 17*(1), 3–34.

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics?. A. O'Keeffe & M. McCarthy. *The Routledge handbook of corpus linguistics* (pp. 38–52). Routledge.

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora, 4*(1), 71–83.

Barrow, M., Grant, B., & Xu, L. (2020). Academic identities research: Mapping the field's theoretical frameworks. *Higher Education Research & Development, 41*(2), 240–253.

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Callies, M., & Zaytseva, E. (2013). The Corpus of Academic Learner English (CALE): A new resource for the assessment of writing proficiency in the academic register. *Dutch Journal of Applied Linguistics, 2*(1), 126–132.

De Fina, A., Oostendorp, M., & Ortega, L. (2023). Sketches toward a decolonial applied linguistics. *Applied Linguistics, 44*(5), 819–832.

Dimova, S., Kling, J., & Margić, B. D. (2024). *EMI classroom communication: A corpus-based approach*. Taylor & Francis.

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

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge University Press.

Feldman, M., Bell, J., & Berger, M. (2003). *Gaining access: A practical and theoretical guide for qualitative researchers*. AltaMira Press.

Fenton-Smith, B., Humphreys, P., & Walkinshaw, I. (2017). *English medium instruction in higher education in Asia-Pacific*. Springer.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research, 5*(2), 126–158.

Galloway, N., & Ruegg, R. (2020). The provision of student support on English Medium Instruction programmes in Japan and China. *Journal of English for Academic Purposes, 45*, Article 100846. Article.

Gardner, S., Nesi, H., & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics, 40*(4), 646–674.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546.

Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 8*(2), 183–207.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Iliovits, M., Harding, L., & Pill, J. (2022). Language use in an English-medium instruction university in Lebanon: Implications for the validity of international and local English tests for admissions. *Journal of English-Medium Instruction, 1*(2), 153–179.

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Jablonkai, R. (2021). Corpus linguistic methods in EMI research: A missed opportunity?. J. Pun & S. Curle. *Research methods in English medium instruction* (pp. 92–106). Routledge.

Kamaşak, R., Sahan, K., & Rose, H. (2021). Academic language-related challenges at an English-medium university. *Journal of English for Academic Purposes, 49*, Article 100945. Article.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36.

Krishnamurthy, R., & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes, 6*(4), 356–373.

Krishnan, A. (2009). What are academic disciplines? Some observations on the disciplinary versus interdisciplinarity debate. NCRM Working Paper Series: ESRC National Centre for Research Methods. http://eprints.ncrm.ac.uk/783/.

Kwon, M. H., Partridge, R. S., & Staples, S. (2018). Building a local learner corpus: Construction of a first-year ESL writing corpus for research, teaching, mentoring, and collaboration. *International Journal of Learner Corpus Research, 4*(1), 112–127.

Language Reading Research Consortium (LAARC), Farquharson, K, & Murphy, KA (2016). Ten Steps to Conducting a Large, Multi-Site, Longitudinal Investigation of Language and Reading in Young Children. *Frontiers in Psychology*.

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching, 51*(1), 36–76.

Milton, J. C., & Tsang, E (1993). A corpus-based study of logical connectors in EFL students' writing: Directions for future research. R. Pemberton & E.S.C. Tsang. *Studies in Lexis: Working papers from a seminar* (pp. 215–246). Hong Kong University of Science and Technology Language Center.

Molino, A., Dimova, S., Kling, J., & Larsen, S. (2022). *The evolution of EMI research in European higher education*. Routledge.

Marcus, G. (2009). Multi-sited ethnography: Notes and queries. M. Falzon. *Multi-sited ethnography: Theory, praxis and locality in contemporary research* (pp. 181–196). Routledge.

Mauranen, A., Hynninen, N., & Ranta, E. (2010). English as an academic lingua franca: The ELFA project. *English for Specific Purposes, 29*(3), 183–190.

McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge University Press.

Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning, 71*(1), 204–242.

Nesi, H. (2015). ESP corpus construction: A plea for a needs-driven approach. *ASp La revue du GERAS, 68*, 7–24.

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

Paltridge, B. (2004). Academic writing. *Language Teaching, 37*(2), 87–105.

Paquot, M., Larsson, T., Hasselgård, H., Ebeling, S. O., De Meyere, D., Valentin, L., Laso, N. J., Verdaguer, I., & van Vuuren, S. (2022). The varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. *Research in Corpus Linguistics, 10*(2), 1–15.

Pecorari, D., & Malmström, H. (2018). At the crossroads of TESOL and English medium instruction. *TESOL Quarterly, 52*(3), 497–515.

Römer, U., & O'Donnell, M. B (2011). From student hard drive to web corpus: The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 6*(2), 159–177.

Römer, U., & Swales, J. M. (2010). The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes, 9*(3), 249.

Rose, H., Macaro, E., Sahan, K., Aizawa, I., Zhou, S., & Wei, M. (2021). Defining English Medium Instruction: Striving for comparative equivalence. *Language Teaching, 56*(4), 539–550.

Sahan, K., Mikolajewska, A., Rose, H., Macaro, E., Searle, M., Aizawa, I., & Zhou, S. (2021). *Global mapping of English-as-a-Medium-of-Instruction in higher education: 2020 and beyond*. British Council.

Seidlhofer, B., Breiteneder, A., Klimpfinger, T., Majewski, S., Osimk-Teasdale, R., Pitzl, M. L., & Radeka, M. (2011). *VOICE: Vienna-Oxford International Corpus of English, Literary and Linguistic Data Service*. http://hdl.handle.net/20.500.14106/2542.

Stevens, M. P., Chen, Y. H., & Harrison, S. (2020). The EMI campus as site and source for a multimodal corpus: Issues and challenges of corpus construction at a Sino-British university. A. Čermáková & M. Malá. *Variation in time and Space: Observing the world through corpora* (pp. 377–402). De Gruyter Mouton.

Trowler, P. (2012). Disciplines and interdisciplinarity: Conceptual groundwork. P. Trowler, M. Saunders, & V. Bamber. *Tribes and territories in the 21st century: Rethinking the significance of disciplines in higher education* (pp. 5–29). Routledge.

Trowler, P., Saunders, M., & Bamber, V. (2012). *Tribes and territories in the 21st century: Rethinking the significance of disciplines in higher education*. Routledge.

Zhou, S., McKinley, J., Rose, H., & Xu, X. (2021). English medium higher education in China: Challenges and ELT support. *ELT Journal, 76*(2), 261–271.