Developing a Multimodal Corpus of L2 Academic English from an English Medium Instruction University in China

Chen, Y-H., Harrison, S., Stevens, M. P. & Zhou, Q.

Author post-print (accepted) deposited by Coventry University's Repository

Original citation & hyperlink:
Chen, Y-H, Harrison, S, Stevens, MP & Zhou, Q 2024, 'Developing a Multimodal Corpus of L2 Academic English from an English Medium Instruction University in China', Corpora, vol. 19, no. 1, pp. 1-15. https://doi.org/10.3366/cor.2024.0295

DOI 10.3366/cor.2024.0295
ISSN 1749-5032
ESSN 1755-1676

Publisher: Edinburgh University Press

Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

This document is the author's post-print version, incorporating any revisions agreed during the peer-review process. Some differences between the published version and this version may remain and you are advised to consult the published version if you wish to cite from it.

Developing a Multimodal Corpus of L2 Academic English from an English Medium Instruction University in China

Yu-Hua Chen[1], Simon Harrison[2], Michael Paul Stevens[3], Qianqian Zhou[3]

Abstract

This paper describes the rationale for and design of a new multimodal corpus of L2 academic English from a Sino-British university in China, the Corpus of Chinese Academic Written and Spoken English (CAWSE). The unique context for this corpus provides language samples from Chinese students who use English as a second language (L2) in a preliminary-year programme, which prepares students for academic studies at university level, at a campus where English is used as the Medium of Instruction (EMI). Data were collected from a variety of settings including written (i.e. exam scripts and essays) and spoken assessments (i.e. interviews and presentations), covering the full range of grades awarded to those language samples, as well as from student group interactions during teaching and learning activities. The multimodal nature of the corpus is realised through the availability of selected audio/video recordings accompanied by the orthographically transcribed text. This open-access corpus is designed to help shed light on Chinese students' academic L2 English language use in a variety of written, spoken and multimodal discourses.

Keywords: corpus design; corpus construction; English Medium Instruction (EMI); English for Academic Purposes (EAP)

1 Introduction

English as the medium of instruction (EMI) in education continues to gain popularity worldwide and has been widely studied (e.g. Pun and Curle, 2022), but very few studies focus on corpus-based research of such EMI contexts, and an open-access EMI corpus

involving a specific second language population is also rare. The EMI title sometimes can be problematic because 'bilingual education' is often used as a catch-all term for any education system that offers at least some of its courses in English (Zhao and Dixon, 2017). However, the university in which our data was collected is a 'full' EMI institution with English specified in its publicly available policy as the only official language for all teaching and learning activities. We further argue that there is a need to research Chinese students' L2 English use because of the significant increase in Chinese students studying through the medium of English around the world. Take the numbers of Chinese students in L1 English-speaking countries in 2018 for example: 120,000 in the U.K., 370,000 in the U.S., 140,000 in Canada, 160,000 in Australia, and 111,000 in New Zealand (Hamp-Lyon & Jin, 2022). According to Universities UK (https://www.universitiesuk.ac.uk/), Chinese students have also become the largest group of international students in the U.K. after Brexit.

Drawing on the above contexts, the Corpus of Chinese Academic Written and Spoken English (CAWSE) offers a database of Chinese students' English language samples from the preliminary-year programme (a.k.a. the foundation year) at a Sino-British EMI university in China. The preliminary-year programme prepares students for academic studies at university level before they continue to more discipline-specific degree programmes at this EMI institution jointly established by a British university and a Chinese university. From 2016 to 2019, the project team collected students' L2 English samples in this context from a range of assessment tasks (written and spoken) and made video recordings of student group interactions in authentic classroom settings.

Given the availability of similar corpora of Academic English, e.g. MICUSP (Römer and O'Donnell, 2011), BAWE (Alsop and Nesi, 2009), ELF (English as a Lingua Franca) corpora, e.g. VOICE (Seidlhofer et al., 2013), CASE (Brunner et al., 2017), and learner corpora, e.g. ICLE (Granger et al., 2020), LINDSEI (Gilquin et al., 2010), we wish to

highlight several aspects in which the CAWSE corpus is different: 1) the EMI context, 2) the full range of grades covered, and 3) the multimodality of the corpus. Firstly, as mentioned earlier, the corpus data was collected from an EMI campus in China. As part of the entry requirements, domestic undergraduate students need to reach a minimum 'first level' in the subject of English from the National Matriculation Exam in China (GaoKao), which means no less than 115 out of 150. In the preliminary year, most students' L2 proficiency generally falls at the intermediate or upper-intermediate levels (approximately equivalent to the Common European Framework of Reference CEFR B1; see Council of Europe, 2001), as based on our experience as researchers and instructors and evidence from teaching and marking. Different from similar existing corpora, with the aim of facilitating the research on second language development as well as teaching and learning, the collection of our assessment data covers the full range of grades, from the bottom grade 'fail' (below 40) to the highest possible grade (70 and above). Another distinguishing feature of CAWSE is the inclusion of multimodal data in the form of audio-visual recordings. The available multimodal data, focusing on the manual gesture units (Kendon, 2004) in students' group discussion, are being made available together with multimodal transcripts prepared in ELAN annotation files.

An online version of the written and spoken subcorpus (in plain text files) is available for download via the project website (https://cawse.transcribear.com/index.html), and samples of multimodal data are available upon request through the channels indicated on our project website. So far at least a dozen Masters dissertations and PhD theses have been completed using part of the CAWSE corpus, and we have received requests from different areas and countries including Belgium, mainland China, Hong Kong, Singapore, the U.K., and the U.S. This indicates a demand for a dedicated L2 corpus, particularly for postgraduate students or junior researchers who often do not have the luxury of building their own corpus.

In recent years, an increasing number of publications have addressed the issues arising from the practicalities of building a corpus. For example, the challenges of data collection was described for BAWE (Alsop and Nesi, 2009), and the issue of defining ELF in corpus terms was discussed for VOICE (Breiteneder et al., 2006). Other similar methodological and practical accounts of corpus building can also be found for the Trinity Lancaster Corpus (Gablasova et al., 2019) and the Spoken BNC2014 (Love et al., 2017). Considering there are few open-access L2 or learner corpora currently available, this paper aspires to inform researchers who wish to embark on a similar project.

The primary objective for the CAWSE project is to provide open-access resources for teaching, learning and research purposes, and the corpus provides valuable resources for us to investigate distinctive linguistic characteristics in Chinese students' L2 academic English samples across different grades, assessment tasks, genres and many other contexts. The goals of the current paper are therefore to introduce the design of the corpus and explain our rationale for the empirical material and data collection. We will first introduce the overall contexts for our data collection and move on to the construction of the written and spoken subcorpus before continuing to the multimodal subcorpus. Some issues and challenges arising during the process of constructing this corpus in terms of ethics and open-access design will be described, and how we approach those issues and challenges will also be discussed.

2 Corpus development

2.1 Contexts: An EAP programme on an EMI campus in China

Sinclair (2005, p. 23) defines a corpus as 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research'. By 'external criteria', Sinclair

refers to the criteria used to select samples by examining the communicative function as opposed to internal criteria that are inherent in the language.

For the 'external criteria' mentioned by Sinclair, they are reflected in the data collection and selection of CAWSE, i.e. the academic language, regardless of our modes of recording (written, spoken-audio or audio-visual), required for students at the preliminary-year programme to communicate at an EMI university in China. This preliminary year is designed as an EAP and degree content introductory year, and the curriculum covers both English for general Academic Purposes (EAP) and English for Specific Purposes (ESP) although the ESP content still falls under the broad scope of EAP in this case. For example, two EAP modules offered are 'Reading and Writing in Academic Contexts' and 'Listening and Speaking in Academic Context's while one ESP module 'English in Specific Academic Contexts' in the second semester is divided into A: Arts and Social Sciences and B: Science and Engineering. A typical degree course in this EMI institution therefore starts with this preliminary year (where academic language, i.e. EAP, is the focus of teaching/learning), followed by three years of subject specific content (where the subject area, e.g. Economics, is the focus of teaching/learning). This is in line with a similar distinction between EAP and EMI defined by Kırkgöz and Dikilitaş (2018).

The written and spoken subcorpus composed of students' responses were all collected from the above EAP and ESP assessments (including the written samples from both the aforementioned ESP subgroups A and B), and the currently available multimodal data is based on the same preliminary-year programme. All the released corpus data, including participants' names or any other personally identifiable information, has been anonymised, and the development process will be explained in more detail in the next two sections.

2.2 The written and spoken subcorpus

Inclusion of a wide range of proficiency grade levels

For the written and spoken subcorpus, where the data came from the assessment of the preliminary year, we also consider the linguistic evaluation from a perspective of second language development. In second language research, learners' proficiency level is often determined by external criteria such as institution status or length of learning, and yet those external criteria are often not a reliable indicator of students' proficiency level (Callies, 2005; Thomas, 1994). An increasing number of studies, therefore, have turned to standardised language tests to determine L2 proficiency (e.g. Kennedy and Thorp, 2007; Staples et al., 2013). In a more recent study (Chen and Baker, 2016), the level was determined by a set of rigorous rating practices in a similar manner as high-stakes language exams.

Drawing on the above literature, we argue for the adoption of 'grade' as one key criterion in our data collection, which serves a crucial role for research into second language development using corpus approaches (Durrant, 2022). The parameter 'grade', however, is rarely used outside of exam-based learner corpora (e.g. the Trinity Lancaster Corpus mentioned earlier). As mentioned previously, both BAWE and MICUSP were created to represent British or American university students' writing and therefore only include proficient assessed student writing rather than the full range of grades, which serves a purpose different from second language research. For example, the BAWE corpus was designed to investigate genres in assessed writing in British Higher Education (Gardner and Nesi, 2013). For the CAWSE project, 'grade' is used as the key variable to ensure 'sampling and representativeness', as mentioned by McEnery and Wilson (1996), instead of typical demographic information such as age or gender as in some other corpora. Just as other British Higher Education (HE) institutions, the grades of our assessment data were allocated to students by professional teaching staff who followed a set of university-wide marking criteria

(in addition to moderation, second-marking, and occasional academic misconduct cases), which strengthens reliability. The samples were then collected from each of the score bands from 'fail' (below 40), 40-49, 50-59, 60-69, to Class I (70 and above, the highest band score in the British HE system, approximately equivalent to A or above in the North American system). Another reason why demographic information was not included is because of how assessment data was collected for the CAWSE, which will be explained in the next section.

Data collection

Instead of approaching individual students, the preliminary-year programme already has an existing mechanism of seeking students' consent at the beginning of each academic year regarding whether students would be willing to contribute their assessment samples for teaching, learning and research purposes. The collection of assessment data was therefore achieved indirectly through the preliminary-year programme, and the appointed members of staff from the programme served as the gatekeeper. All the assessment data, therefore, have undergone robust checks of whether consent from individual students was obtained. For the teacher-student interviews, separate consent was obtained from the tutors because they served as the interlocutors in the assessment of interviews.

As the data collection was achieved with the assistance of the staff from the preliminary-year programme, we therefore did not have the access to more detailed background information of the students. More detail regarding data in this regard will be discussed in Section 3. Even without detailed demographic information recorded for individual students, it has to be noted that the students enrolled in the preliminary-year programme are highly homogenous. Almost all of the data (over 97%) come from L1 Chinese students from Mainland China. Their age typically falls into the range of 18-19 and,

as required by the Ministry of Education of China, received English instruction throughout their education, with some local variations (e.g. Wei and Su, 2012).

For the spoken data, the aim was to achieve thirty samples for each of the score bands, considering the amount of transcription required for L2 speech. For the written data, we aimed for 100 scripts for each score band. As can be expected, the data collection easily reaches the thresholds from the score bands in the middle range such as 50-59 and 60-69, and yet much less is achieved for the bottom (below 40) and the top groups (70 and above). The threshold of thirty (speech) to one hundred (writing) was deemed manageable, and it is also comparable with similar corpora such as the BAWE corpus, where 32 samples per level of study were collected from most of the specified disciplines with exceptions of 16 or 64 samples in certain disciplines (Nesi, 2011).

An overview of the written and spoken subcorpus

For the written and spoken subcorpus, the sampling frame covers both the grades and the registers (written and spoken) as implemented in the assessments of the preliminary-year programme. As can be seen in Figure 1, the metadata available including the grades and assessment task types are reflected in the structure of the data available in individual folders that can be downloaded online. Here L refers to Listening, R for Reading, W for Writing while A is Arts and Social Sciences and B is Science and Engineering (cf. Section 2.1).

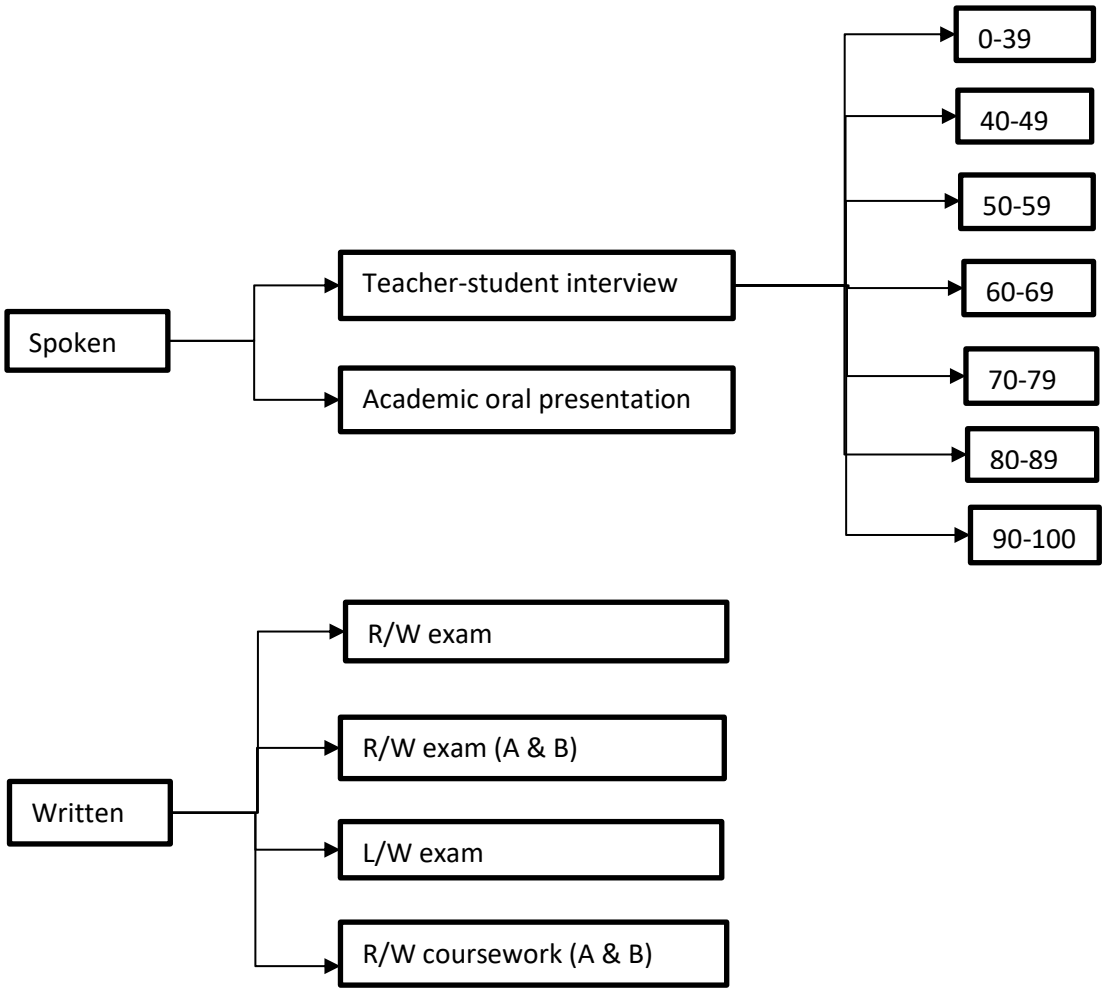

Figure 1 Structure of the written and spoken subcorpus with an example of the interview data in score bands

For the written component, the assessment tasks including hand-written exam scripts and electronic coursework (essays) are included. For the spoken component, the teacher-student interviews and student presentations are included. The corpus size altogether amounts to slightly over two million tokens for the written and spoken subcorpus (see Table 1). More details about this subcorpus, including a short description for each of the tasks and sample scripts, can be found on the project website.

Table 1 Composition of the CAWSE written and spoken subcorpus

Mode	**Written**		**No. of scripts / tokens**	**Spoken**	**No. of sessions / tokens/ duration**
Individual	Exam	Reading and summary writing (avg. 488 tokens)	365 (178,278 tokens)	10-minute presentation (avg. 1146 tokens)	184 (210,846 tokens/ approx. 30 hrs 40 mins)
		Reading and writing (A & B) (avg. 615 tokens)	612 (376,326 tokens)		
		Listening and writing (avg. 251 tokens)	515 (129,119 tokens)		
	Coursework	Writing assignment (A & B) (avg. 1529 tokens)	657 (1,004, 523 tokens)		
Teacher-student				10-minute interview (avg. 1160 words)	122 (141,487 tokens / approx. 20 hrs 20 mins)
Subtotal			**2,149 (1,688,246 tokens)**		**306 (352,333 tokens/ approx. 51 hrs)**

2.3 The multimodal subcorpus

Students' group discussion as the focus

Paired and group tasks became the focus of the multimodal subcorpus because the learning objectives of the preliminary year included preparing students for academic seminar discussions (see Shi, Irwin and Du, 2022; Rybarczyk and Stevens, 2022). Many activities were designed to help students develop 'interactional competence', features of which include turn management, topic management, non-verbal behavior, breakdown repair, and interactive listening (Galaczi and Taylor, 2018, p. 227). The students therefore practiced (and sometimes being reminded to use) a range of interactional language functions, such as agreeing, disagreeing, and requesting clarification.

Data collection

For the multimodal subcorpus, a total of thirty-five hours in raw data were collected from students' discussions in a number of classrooms and 'Chat-up' sessions. Designed to improve interactional competence, the extra-curricular Chat-ups were peer-led by students from higher years (year-2 to year-4).

Different from the written and spoken subcorpus, a separate written consent form was designed for the audiovisual recordings and clearly communicated with the participants prior to the start of any recording. After the consent was obtained, we joined several weekly sessions for a period of time agreed on with the staff and students, sometimes recording only one group of students during their discussions while other times recording multiple groups in the classroom at the same time. Some of such activities were typical classroom tasks such as 'discuss X with your partner'; others were often related to the teaching materials 'compare your answers to Exercise Y in small groups' etc.

Preparing the raw data

The raw data for the recordings was inevitably 'messy' and required careful attention to how the data would be organised. For instance, the start and end points of the recordings do not always correspond with the start and end points of the discussions. Students often refer to information that discloses their anonymity, or their name cards placed on the table are visible in the recording. The project team has therefore been working on identifying smaller portions of data that can be processed (screened for any ethical issues and segmented into manageable files) to be usable for specific research projects, which so far include a single-case analysis (Harrison and Chen, 2021; Chen, Harrison and Weekly, 2019), a PhD dissertation (Stevens. 2021) and several Masters dissertations (e.g. Stutzman, 2017). This is somewhat similar to Egbert et al. (2021)'s approach, where functional units of conversations were identified and

segmented for the BNC Spoken 2014. Another reason of this decision is because the transcription of multimodal data, particularly when it involves spontaneous speech and multiple speakers, requires tremendous amounts of resources.

An overview of available multimodal data

Drawing on the rationale discussed above, a smaller subcorpus LRE-MuCAWSE has just been completed on the basis of the well-established notion of language-related episodes (e.g. García Mayo & Zeitler, 2017; Basterrechea & Gallardo-del-Puerto, 2020). Language-related episodes or LREs were originally defined by Swain and Lapkin (1998) as 'any part of a dialogue where the students talk about the language they are producing, question their language use, or correct themselves or others' (p. 326). Such episodes are known to occur, for example, when students either begin struggling or making mistakes (Ohta, 2001), which makes a start point for the episode relatively easy to identify. As the episode begins, students subsequently collaborate by discussing and negotiating their understandings of the aspect of language in question (e.g. pronunciation, lexis, grammar), and then eventually, the students tend to resolve the issue or abandon it and move on (Leeser, 2004). Below is an example of a lexically-focused LRE presented with orthographic transcription.

S1 I think it have low ability to collect er open source oh to collect yeah yeah yeah
S2 collect open source? can you clarify what is collect open source?
S1 er a just 就收集的 {just collect}

The potential of CAWSE multimodal data for contributing to the study of language-related episodes is what guided the design and annotation of our sub-corpus LRE-MuCAWSE. This subcorpus currently consists of forty-eight language-related episodes (with an average duration of 57 seconds) that have been transcribed and annotated for aspects of speech and gesture (Table 2), and another similar subcorpus is currently being developed

based on portions of the raw data. More research projects or subcorpora with a different focus may be created in the future on the basis of the larger multimodal subcorpus.

Table 2 Composition of the CAWSE multimodal subcorpus

Context	**No. of sessions (raw data)**	**No. of LREs in LRE-MuCAWSE (transcribed and annotated)**
Chat-up	7	-
Classroom	87	48
Total	94 (35 hours)	48 (45 mins 28 secs)

3 Issues and Challenges

Several challenges were met in the creation of CAWSE, primarily involving stakeholders both human (students and staff) and institutional (university integrity). As the main research participants for CAWSE were students enrolled into the preliminary-year programme, the ethics application for this project was therefore discussed in close collaboration with senior members of the centre where the programme was delivered, including the Ethics Officer, Director of Research and Head before it was approved by the university ethics committee. Accordingly, it was agreed that the EAP support centre would be recognised in the application as the gatekeeper for the data. For the assessment samples described above, this allowed us to benefit from the centre-wide request for consent distributed to students during the start of session, when 42-44% of consent was obtained across different modules.

Having the support centre involved in the project yielded various degrees of access to the classroom and to assessment data. For instance, we had access to a semester's worth of classroom observation, as well as the assessment data. Although the assessment design in the preliminary-year programme also included a speaking task of group discussion, where students would discuss a given topic in a small group, we were unfortunately unable to use this data. As the gatekeeper pointed out, topics used in group discussion would be reused, and

it was therefore not possible to include them in an open-access corpus in case students might come across those confidential materials on the internet. We therefore made an effort to collect video recordings of rehearsals for these group discussions, to be used for multimodal data (not advertised as assessment data).

Similarly, certain information such as the exam prompts or marking criteria, despite the wealth of information they may provide for the corpus, are considered the intellectual property of the university where the data came from. After several rounds of internal discussion, it was decided not to include those as the metadata. Because the project lasted several years undergoing the turnover of some staff members, it was evident that effective and timely communication was the key in working with stakeholders such as the gatekeeping institution.

As the corpus was designed to be open access, we also faced the challenge of where to archive the corpus after two of the key team members left the university towards the end of the project. A commercial service such as Sketch Engine (https://www.sketchengine.eu/) was once considered; however, this option was not sustainable because of our limited funding. There were also free services available such as the Oxford Text Archive (https://ota.bodleian.ox.ac.uk/repository/xmlui/), but as we prefer to be able to monitor the use of the corpus, it was then decided to work with a computer scientist and set up a website with minimum maintenance required, where users can register and download the data. This proves to be the best option at the time, because the project team is able to determine which data can be released while more data are being processed and awaiting more funding.

4 Conclusion

This paper discusses the design of a multimodal corpus of L2 academic English in the context of an EMI university in China and describes the theoretical and methodological

considerations that our corpus design took into account as well as some issues and challenges that we were confronted with.

Even after decades of rapid development in Corpus Linguistics and large-scale L1 corpora such as COCA or BNC 2014, still very few learner or L2 corpora are freely accessible to the research community. As pointed out by Love et al. (2017, p. 320), 'a well-known problem afflicting corpus linguistics as a field is its tendency to prioritise written forms of language over spoken forms, in consequence of the much greater difficulty, higher cost and slower speed of collecting transcribed speech'. The CAWSE corpus takes steps towards redressing this imbalance, at least in the area of learner or L2 corpora, and further innovates by including the written, spoken and multimodal subcorpora.

It is hoped that the current paper has provided a state-of-the-art overview of building an L2 corpus in condensed format. The CAWSE corpus has the potential to become a resource for researchers and practitioners working in relevant areas, including various subfields of Applied Linguistics related to Second Language Acquisition (in the contexts of EAP, ESP, EMI, and ELF).

Acknowledgement

This project was supported by the Ningbo 3315 Overseas Individual Talent Award to Dr Yu-Hua Chen and the University Matched Funding. The work described in this paper was also partially supported by City University of Hong Kong and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11609221). We are grateful for the staff and students at the research site who agreed to be included in the corpus, and we particularly appreciate the invaluable input and support from the original project team. We would also like to thank the anonymous reviewers for their insightful comments on this manuscript.

Notes

1 School of Humanities, Coventry University, Coventry CV1 5FB, United Kingdom.

Correspondence to: Yu-Hua Chen, e-mail: yu-hua.chen@coventry.ac.uk

2 Department of English, City University of Hong Kong, Hong Kong, China.

3 School of English, University of Nottingham Ningbo China, Ningbo, China.

References

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

Breiteneder, A., Pitzl, M.-L., Majewski, S., Theresa, and Klimpfinger. (2006). VOICE Recording - Methodological Challenges in the Compilation of a Corpus of Spoken ELF. *Nordic Journal of English Studies*, 5(2), 166-188.

Brunner, M. L., Diemer, S., & Schmidt, S. (2017) ... okay so good luck with that ((laughing))? – Managing rich data in a corpus of Skype conversations. *Studies in Variation, Contacts and Change in English*, 19.

Callies, M. (2015). Using Learner Corpora in Language Testing and Assessment: Current Practice and Future Challenges. In Castello, E., Ackerley , K. and Coccetta, F. (Eds.), *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*. Peter Lang, pp. 21-35.

Chen, Y. H., Harrison, S. and Weekly, R. (2019). "*I don't have communicate ability*": Deviations in an L2 Multimodal Corpus of Academic English from an EMI University in China – Errors or ELF? In H. Parviainen, M. Kaunisto and P. Pahta (Eds), *Corpus Approaches into World Englishes and Language Contrasts* (eVarieng). https://varieng.helsinki.fi/series/volumes/20/chen_harrison_weekly/

Chen, Y. H. and Baker, P. (2016). Discourse Features across Second Language Writing Development: Lexical Bundles in Rated Learner Essays CEFR B1, B2, and C1. *Applied Linguistics*, 37(6), 849-880. https://doi.org/10.1093/applin/amu065

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Language Policy Division.

Durrant, P. (2022). *Corpus Linguistics for Writing Development: A Guide for Research* (1st ed.). Routledge. https://doi.org/10.4324/9781003152682

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41(5-6), 715-737.

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

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219-236.

Gardner, S. and Nesi, H. (2013). A Classification of Genre Families in University Student Writing. *Applied Linguistics*, 34(1), 25-52.

Gilquin, G. De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken English Interlanguage* (CD-ROM + handbook). Presses universitaires de Louvain, Louvain-la-Neuve.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020). *The International Corpus of Learner English. Version 3*. Louvainla-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Hamp-Lyons, L. & Jin, Y. (Eds). (2022). Assessing the English Language Writing of Chinese Learners of English. London: Springer Nature.

Harrison, S. and Chen, Y. H. (2021). From Dancing K-Pop with Chinese and "English in class please": English language policy negotiations as relational-languaging episodes during classroom interaction. *RELC Journal*, 52(2), 270-286. https://doi.org/10.1177/00336882211022123

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kennedy, C., and Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor and P. Falvey (Eds.), *IELTS collected paper: research in speaking and writing assessment*, pp. 316-378. Cambridge University Press.

Kırkgöz, Y. & Dikilitaş, K. (2018). Recent developments in ESP/EAP/EMI contexts. *Key Issues in English for Specific Purposes in Higher Education*, 1-10.

Love, R., Dembry, C., Hardie, A., Brezina, V., and McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, 22(3), 319-344. doi: 10.1075/ijcl.22.3.02lov

Nesi, H. (2011). BAWE: An introduction to a new resource. In Frankenberg-Garcia, A., Flowerdew, L., & Aston, G. (Eds.). *New Trends in Corpora and Language Learning*. A&C Black. pp. 213-228.

McEnery, T., and Wilson, A. (1996). *Corpus linguistics* (2nd ed). Edinburgh University Press.

Ohta, A. S. (2001). *Second Language Acquisition Processes in the Classroom: Learning Japanese*. NJ: Lawrence Erlbaum.

Pun, J. K. H. and Curle, S. M. (ed.) (2022). *Research Methods for English Medium Instruction*. Routledge.

Römer, U. and O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2), 159-177. doi: 10.3366/cor.2011.0011

Rybarczyk, M. and Stevens Pérez, M.P. (2022). Meaning and Intersubjectivity. In Li, T. (ed.) *The Handbook of Cognitive Semantics*. E.J. Brill.

Sinclair, J. (2005). Corpus and Text - Basic Principles. In M. Wynne (ed.), *Developing Language Corpora: A Guide to Good Practice*, pp. 5-24. Oxbow Books.

Shi, D., Irwin, D. and Du, P. (2022). Languaging dynamics in interactive lecturing: exploring an embodied approach to deep learning in L2 higher education contexts. *Classroom Discourse*, DOI: 10.1080/19463014.2021.1971543

Staples, S., Egbert, J., Biber, D., and McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, 12(3), 214-225.

Stevens, M. P., Chen, Y. H., and Harrison, S. (2020). The EMI campus as site and source for a multimodal corpus. In Čermáková, A. and Malá, M. (Eds) *Variation in Time and Space: Observing the World through Corpora*. De Gruyter, 377-401. https://doi.org/10.1515/9783110604719

Stevens, M. P. (2021). *The interactive ecology of construal in gesture: A microethnographic analysis of peer learning at an EMI university in China*. Doctoral thesis, University of Nottingham Ningbo China.

Stutzman, L. D. 2017. *Multimodal Corrective Feedback and Interactional Moves within Language-Related Episodes and Inscription-Related Episodes: An Analysis*. MA dissertation, University of Nottingham Ningbo China.

Swain, M., and Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, 82(3), 320-337.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, 44(2), p.p. 307–36.

VOICE. (2013). *The Vienna-Oxford International Corpus of English* (version 2.0 XML) (2013-01-22 ed.). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka.

Wei, R. and Su , J. (2012). The statistics of English in China. *English Today*, 28, pp. 1014.