Corpora, Locally Sourced: An Approach to Addressing the Specific Needs of ESL Writing Programs

JOSEPH J. LEE
Department of English, Dalarna University
Falun, Sweden

Abstract

This paper discusses the development and implementation of a locally-sourced corpus to address the specific needs of an ESL writing program. The paper begins with a description of the motivation and development of the Corpus of Ohio Learner and Teacher English (COLTE), a large in-house corpus consisting of assessed ESL student writing and teacher feedback from first-year writing courses. The paper illustrates research conducted within the local context using the COLTE and discusses how the program’s corpus-based approach influenced curricular revisions and instructional practice. It addresses the evaluation of these efforts, challenges encountered, and considerations for building and using corpora at the program level. The paper concludes by arguing that stakeholders within programs can work collaboratively to build and use locally-sourced corpora grounded in their specific setting to explore local practices and address the English language learning and teaching needs of their own particular context.

doi: 10.1002/tesq.3277

In recent decades, considerable research has focused on corpus-based approaches to English language teaching. However, most studies have been limited to individual classroom practitioners compiling corpora and assessing the effectiveness of such approaches (e.g.,
While these efforts have shed light on the value of corpus-based approaches, the integration of corpus research and classroom teaching at the program level for ESL writing pedagogy has been essentially absent in the literature.

Adopting the stance of a “learning organization” (Middlewood & Abbott, 2015), this paper discusses the development and implementation of a locally-sourced corpus by the Classroom Research Unit (CRU) of the Academic & Global Communication Program (AGC) at Ohio University (OU) to address the needs of its specific writing program. The paper briefly describes the AGC and CRU, and then focuses on the motivation and development of the Corpus of Ohio Learner and Teacher English (COLTE). This large in-house corpus consists of assessed ESL student writing and teacher feedback from first-year writing (FYW) courses designed for international and multilingual undergraduate students at OU.

The paper illustrates research conducted within the local context using the COLTE and discusses how the program’s corpus-based approach influenced curricular revisions and instructional practice. It addresses the evaluation of these efforts, challenges encountered, and considerations for building and using corpora at the program level. The paper concludes by arguing that stakeholders within programs can work collaboratively to build and use locally-sourced corpora grounded in their specific setting to explore local practices and address the English language learning and teaching needs of their own particular context.

THE CLASSROOM RESEARCH UNIT (CRU)

Before delving into the motivation behind establishing the CRU, it may be beneficial to provide an overview of OU’s AGC program. The AGC’s primary mission is to address the academic literacies needs of diverse matriculated international and domestic graduate and undergraduate students. It offers comprehensive instruction in advanced academic writing, oral communication, and critical reading. For graduate students, the program offers courses tailored to specific disciplines and genres, as well as general academic writing, reading, and oral presentation courses. While its main focus is on graduate-level instruction, the program oversees three FYW courses specifically designed to meet the writing demands of international and multilingual undergraduate writers. Additionally, the AGC provides tutoring services through its online and in-person writing, reading, and presentation labs to support students’ academic communication skills. By offering this comprehensive array of courses and support services, the AGC aims to facilitate the academic success of its diverse student population.
Motivation for the CRU

ESL research has offered valuable insights and recommendations for teachers and teacher education. However, studies have shown that teachers’ engagement with published research is often limited due to factors including perceptions of research, its relevance to classroom practice, and lack of institutional incentives (Sippel & Sato, 2022). Furthermore, teacher cognition research suggests that teacher education may have varying effects on practice (Borg, 2006), with evidence indicating that teachers may develop insufficient competencies during coursework (Leńko-Szymańska, 2017). Notably, contextual elements, including institutional ideology, culture, curriculum, and student needs, often have a more profound influence on teachers’ practice, even after teacher training programs (Borg, 2006).

Thus, to transform practice, Middlewood and Abbott (2015) argue that change needs to be initiated at the program level. They propose adopting the stance of a “learning organization,” which emphasizes reflexive inquiries to seek out candid answers to uncomfortable questions specific to the local context. As they contend, engaging in internal inquiries of local practice is crucial for programs to actualize purposeful transformations within their educational setting (e.g., Matsuda, Saenkhum, & Accardi, 2013).

Aligned with this stance, the CRU was established to drive purposeful changes within the AGC. Comprising program administrators, instructional faculty, and linguistics master’s students, the CRU’s mission is to conduct research within the local educational context to inform the AGC curriculum and teacher development, as well as to share the research findings with the broader TESOL community. The CRU focuses on addressing issues and questions most relevant for AGC students and teachers. Over the years, its various projects have resulted in curriculum and classroom practice changes. Examples include investigating the effects of collaborative student writing (Bikowski & Vithanage, 2016), exploring student engagement with interactive digital textbooks (Bikowski & Casal, 2018), and analyzing discourse practices in online writing tutorials (Casal & Lee, 2018). Notably, the development of the COLTE was a significant accomplishment within our context.
THE COLTE

Before starting the COLTE project, the CRU searched for existing corpora that could address the needs of ESL undergraduate students in FYW courses. However, available corpora at the time, like the MICUSP, BAWE, and ICLE, did not align with the specific writing produced by FYW students. The MICUSP and BAWE focus on disciplinary writing, mostly by L1 students, while the ICLE contains short EFL writing samples that do not represent the types of writing expected of undergraduate students at the university. Also, most corpus-based studies at the time centered on timed essays, EFL student writing, TOEFL written responses, or disciplinary writing. Recognizing these limitations, the decision was made in 2013 to develop a program-specific corpus, leading to the creation of the COLTE. This in-house corpus allows for a comprehensive understanding of the writing practices and specific needs of ESL FYW students within the specific context of the AGC.

Between 2013 and 2018, the CRU collected thousands of samples of assessed ESL student writing and teacher feedback. The COLTE comprises writing samples from three FYW courses: ENGD160, ENG1610, and ENG1610-Business. ENGD160 is designed for international students who meet certain criteria related to language proficiency (e.g., TOEFL iBT writing section score below 24). The standardized curriculum focuses on developing academic writing skills, including organization, coherence, idea development, summarizing, paraphrasing, grammar, vocabulary, and source use. It also serves as preparation for the subsequent FYW course (ENG1610 or ENG1610-Business).

ESL students who meet certain language requirements (e.g., TOEFL iBT writing section score of 24 or higher) qualify for ENG1610. It is worth noting that most students take ENGD160 before ENG1610. The standardized curriculum emphasizes higher-level academic writing abilities essential for success in disciplinary courses. This includes composing

1 The Michigan Corpus of Upper-level Student Papers (MICUSP) comprises 830 A-graded student-produced papers from a range of disciplines written by final-year undergraduate to third year graduate students at the University of Michigan: https://micusp.elicorpora.info/main.
2 The British Academic Written English (BAWE) corpus includes 2761 proficient-student papers from diverse disciplines written by first-year undergraduate to first-year graduate students at UK universities: https://warwick.ac.uk/fac/soc/ai/research/collections/bawe/.
4 Note that the Corpus & Repository of Writing (CROW) consists of L2 student writing in FYW, but it did not exist at the time the COLTE project was initiated in 2013. The CROW was launched in 2013: https://crow.corporaproject.org/page/researchers#citation.
effective papers for different audiences and purposes, analyzing audience and purpose in relation to various academic genres, conducting secondary research, integrating sources using paraphrasing, summarizing, and quoting with adherence to APA style, adopting appropriate academic style, and self-editing for grammatical accuracy. Alternatively, students can opt for ENG1610-Business, which follows a project-based approach and collaborates with a local business. This specialized section concentrates on business-related genres such as memos, SWOT (Strengths, Weaknesses, Opportunities, Threats) analyses, and recommendation reports. Completing either ENG1610 or ENG1610-Business fulfills institutional FYW requirements for graduation.

The COLTE contains a diverse range of text types written by 594 ESL FYW students with varying language backgrounds. It includes summaries, summary-responses, argumentative essays, problem-solution essays, research essays, observation reports, memos, SWOT analyses, recommendation report, and timed tests. The in-house corpus includes both assessed “one-shot” and multidraft texts, covering grades from A to F. In total, the COLTE consists of 4,707 texts, approximately 2.5 million words.5

LOCAL CORPUS INFORMING LOCAL CURRICULUM AND INSTRUCTION

Using the COLTE, nine corpus-based projects have been conducted thus far, focusing on diverse issues to gain a deeper understanding of ESL student writing practices. Due to space limitations, only three studies are highlighted to demonstrate how the findings have resulted in curricular revisions and teaching practices.

Stance and Engagement

In our initial investigation using the COLTE, we examined stance and engagement features in A- and B-graded ESL argumentative essays and how they compared to A-graded L1 essays in the MICUSP (see Lee & Deakin, 2016). Stance (hedges, boosters, self-mentions, attitude markers) and engagement devices (e.g., reader pronouns, directives) are crucial components of effective argumentation (Hyland, 2005). The objective was to determine the degree to which the two ESL

5 As noted, some of assignments are multidrafts such as the argumentative essay, recommendation report, and research essay. Excluding second/third drafts, the COLTE contains 3,326 texts, approximately 1.6 million words.
groups differed in their utilization of stance and engagement features. By comparing them with MICUSP writers, we aimed to identify the areas where ESL students required additional assistance. The findings showed that A-graded ESL writers’ use of stance and engagement devices closely matched that of the L1 students. However, the B-graded ESL essays exhibited significantly fewer instances of hedges, in particular, compared to both the A-graded ESL and L1 students.

These findings led to workshops for AGC faculty and teaching assistants (TAs), focusing on the importance of stance and engagement. We recognized the need to enhance students’ awareness of the communicative impact conveyed through the strategic use of devices like hedges and boosters in constructing successful arguments. As a result, the FYW curriculum was revised to place greater emphasis on these interpersonal features, incorporating direct instruction through mini-lectures and activities in all FYW courses. For example, activities include analyzing hedging and boosting devices in argumentative essays using the MICUSP platform, as well as explaining writers’ intentions and identifying other places in the texts these devices could be used. Another activity involves comparing and evaluating these interpersonal resources in sample texts (e.g., high- and low-rated ESL essays) in the COLTE. Further, in-class peer- and self-review activities include students providing feedback on these interpersonal features, among other important areas. Finally, stance and engagement features have been included as a criterion in the grading rubrics for writing assignments. These measures have raised students’ awareness, improved their understanding, and led to strategic integration of these features in their academic writing.

Lexical Bundles

In another study using the COLTE, we examined the (mis)use of lexical bundles (LBs) in argumentative essays by A-graded ESL students and compared them with L1 student writers in the MICUSP (see Bychkovska & Lee, 2017). LBs are the most frequently recurring multi-word units of three or more words in a specific register or genre (e.g., in the context of, at the end of the). The aim was to explore the differences in the structural and functional patterns of LBs employed by high-rated ESL and L1 students, while also identifying commonly misused bundles in ESL student writing. The findings revealed that ESL students used a broader range of LBs and at a significantly higher frequency compared to L1 students. Despite showing signs of adapting to academic writing conventions, high-rated ESL students employed more bundles characteristics of oral language. They used more verb-phrase bundles (e.g.,
there are three main) and stance bundles (e.g., is a good choice). ESL students also tended to misuse bundles particularly involving articles (e.g., on * other hand) and prepositions (e.g., *in the same time).

In response, we developed activities to raise students’ awareness of common LBs used in academic writing. In addition to mini-lectures on typical academic LB structures and functions, one activity involves searching for academic LBs on the MICUSP platform and identifying their functions. Additional activities were also designed to focus on highly productive frames such as in the + NOUN of and the NOUN + of the/a. Worksheets containing ESL student texts that featured LBs with missing articles or prepositions were also developed to practice constructing grammatically accurate bundles. Admittedly, while some improvements have been observed, many students still seem to face challenges in using appropriate academic LBs. Thus, this is an area that would require additional and consistent direct instruction to foster further progress.

Citation Practices

In the final sample study, we investigated citation behaviors in A-grade ESL student research papers, focusing on forms, functions, and stance, since citations are essential in academic writing as they contribute to effective persuasion (see Lee, Hitchcock, & Casal, 2018). Our analysis categorized citation forms (e.g., direct quotation, paraphrase), types (integral, non-integral), and reporting structures (e.g., verb + that clause, preposition phrase), while also analyzing citation functions (e.g., attribution, evaluation) and writer stance (e.g., distancing, contesting). The findings revealed that students heavily relied on paraphrasing, signaling the influence of instruction, but rarely synthesized multiple sources or used non-integral (i.e., parenthetical) citations. Their use of the verb + that clause was limited, with a restricted range of reporting verbs, including informal ones (e.g., say, talk). Regarding function and stance, they predominantly used the attribution function, indicating a tendency to retell existing knowledge rather than transform it. Their stance toward the cited material was mostly non-committal, showing deference to received knowledge. These findings underscored the need for additional support in synthesizing sources, using a diverse range of reporting structures and adopting a more critical orientation to cited material.
To address these issues, we developed activities to improve source integration, along with mini-lectures covering forms, functions, and stances. To raise awareness of reporting verbs, an activity utilizing the COCA\(^6\) platform was designed. Using the Chart function, the first part of the task involves distinguishing academic and non-academic reporting verbs. Another activity uses the KWIC (keyword in context) function to analyze the surrounding words for structural patterns of reporting verbs and establish general functional “rules” for these structures. For further citation development, excerpts from A-graded ESL and MICUSP texts are used in an activity in which students analyze integral/non-integral citations, author stance, and tense/aspect. Also, a synthesis activity was designed that involves grouping quotations by category and composing a paragraph synthesizing the information to practice paraphrasing, synthesizing, diverse reporting structures, and citation types. Peer- and self-review activities also focus on citation use to enhance students’ awareness of available citation choices. While students still appear to experience challenges with synthesizing multiple sources, these activities appear to have led to important improvements in students’ use of a broader range of formal, functional, and stance options.

**CHALLENGES AND CONSIDERATIONS IN BUILDING A PROGRAM-LEVEL CORPUS**

This section highlights three particular challenges we encountered in building and using a program-level corpus and how we addressed them. Initially, there was resistance from some faculty/TAs who were unfamiliar with the concept of a corpus or did not see the need for an in-house corpus since there were existing corpora available (e.g., COCA, MICUSP). Addressing these two points were rather straightforward. However, importantly, concerns were raised that the project might be used to evaluate their teaching performance, which was a significant point of resistance, especially since none of our faculty was tenured and TA positions were dependent on their performance.

To address this concern, we focused on three key strategies. First, we emphasized the importance of faculty/TAs’ participation and collaboration to create the in-house corpus. Second, we provided clarity on the purpose of the project—to gain insights into the writing practices of ESL students and the feedback provided by teachers in the local context. This information would enable us to develop the curriculum, design

---

more effective classroom activities, and improve teaching practices to better serve our students’ needs based on empirical evidence. The most crucial step was addressing the fear of evaluation. We assured teachers/TAs that the project was not intended for evaluation but rather to enhance our understanding of areas where students needed more support in academic writing. We stressed our commitment as a learning organization to collaboratively explore issues relevant to both teachers and students. To ease teachers/TAs’ concerns about privacy, we explained that all papers would be anonymized, providing illustrations to demonstrate how anonymized texts would appear. Since participation was voluntary, a couple of teachers initially opted not to contribute. However, after observing the project’s non-evaluative nature and potential benefits, these teachers eventually chose to participate. By emphasizing collaboration, clarifying the project’s purpose, and reassuring privacy and nonevaluative intentions, we overcame the challenges.

The second challenge revolved around logistics. The project spanned 5 years, requiring significant commitment from all stakeholders, particularly FYW teachers/TAs and CRU members. While obtaining Institutional Review Board approval was relatively straightforward, the real difficulty lay in collecting, organizing, and compiling a coherent and valuable in-house corpus. At the project’s initiation, AGC offered multiple sections of ENGD160 and ENG1610, each with 15–20 students per semester, including summer courses. Ensuring that we received and compiled papers was a formidable task. To address this, a strategic and coordinated effort was essential. At the start of each semester, we informed FYW students about the project and obtained informed consent forms. Since teachers/TAs collected and submitted papers electronically, participation from students was relatively high. Despite their busy schedules, teachers/TAs diligently submitted student papers, and we occasionally sent friendly reminders throughout the semester. Once papers were collected, identifying features were removed, and each text file was consistently named using codes for author, text type, assignment type, grade, etc. A spreadsheet with codes and contextual information (e.g., language background, and years of English study) was created. The files were systematically organized into folders and subfolders based on courses, text types, and grades. Admittedly, the first year was quite challenging, but we developed processes for collecting, compiling, and organizing the texts. This strategic coordination and effort allowed us to successfully build a streamlined process for building the COLTE over the 5-year period.

The final challenge was determining the most relevant issues to focus on in our context. While we were aware of broader field interests at the start of the project, adopting a top-down research program aligned with those interests would have contradicted our principle as a
learning organization and the goal of creating an in-house corpus with a direct impact on FYW teachers and learners. Thus, we chose a different approach. As a learning organization, we recognized the importance of concentrating on issues directly relevant to the faculty/TAs involved in teaching FYW courses. We conducted multiple meetings with FYW teachers/TAs to gather their valuable feedback on the challenges their undergraduate ESL students faced and questions they wanted addressed that corpus-based methods could answer. Together, we collaboratively developed a list of areas worth exploring, prioritizing aspects that could have the most immediate impact on classroom instruction. We collectively developed our research agenda tailored to the local context. This approach allowed us to explore relevant research questions as a group, engaging teachers/TAs in various corpus-based research projects and the development of classroom materials. By focusing on the local context and involving all stakeholders in shaping the research agenda, we fostered a collaborative environment that led to meaningful research and practical outcomes for both students and teachers.

CONCLUSION

This article presented research conducted using a locally sourced corpus to drive purposeful changes to the local setting, while also discussing challenges and considerations in building a program-level corpus. As a learning organization, Middlewood and Abbott (2015) advocate for “regularly reassessing internal practices” to improve teaching (p. 23). Creating a locally-sourced corpus allowed us to reevaluate classroom approaches and better meet the needs of AGC learners and teachers. While building a program-level corpus is challenging, I highly recommend other writing programs to develop and use their in-house corpus (or corpora) to explore issues directly relevant to their distinctive context, leading to informed and purposeful curriculum and classroom teaching improvements. This endeavor, however, must be nonevaluative, collaborative, and exploratory; all stakeholders to varying degrees must be involved in this effort. By seeking answers to locally relevant questions together, writing programs can effectively use locally sourced corpora to address the learning and teaching needs of their specific setting.
THE AUTHOR

Joseph J. Lee, PhD, is a senior lecturer in the Department of English at Dalarna University. His research and teaching interests include ESP/EAP, academic writing, and applied corpus linguistics. His publications have appeared in English for Specific Purposes, Journal of English for Academic Purposes, and Journal of Second Language Writing.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

FUNDING STATEMENT

No funding was received for this work.

REFERENCES


