Research Note

Curating a Corpus: A Three-Phase Model

Siân Alsop, *Centre for Global Learning, Coventry University*

Michael Laudenbach, *New Jersey Institute of Technology*

William Marcellino, *RAND, Pardee RAND Graduate School, Carnegie Mellon University*

Scopus Abstract

We describe in this Research Note processes and protocols for curating a corpus of texts for analysis, from collection of data to readiness for analysis. We offer example case studies working with texts in different genres and languages, and using different tools, to illustrate general principles for corpus curation. Rather than a comprehensive guide for researchers interested in corpus linguistics methods, we offer a conversational starting point, supplementing our overview of three phases (*collection*, *cleaning*, and *pre-processing*) with authentic experiences from our own diverse research. Further, we reflect on the pedagogical implications associated with corpus linguistics, as well as the challenges and next steps in corpus curation and analysis in the age of generative AI. Our experiences show how common curation phases can be applied to different studies and contexts, and the considerations that arise when doing so.

Structured Abstract

- **Aim**: We aim to support corpus analysts in the practical preparation of messy semi-structured data for scalable machine analysis while raising conceptual issues regarding the potential and challenges of generative AI and discussing pedagogical implications. As a Research Note, this paper explores illustrative examples of curation that draw on disparate datasets and use cases. We hope that the three phases of corpus curation that we observed—collection, cleaning, and pre-processing—lead to more explicit considerations of these phases in future research articles and associated materials in the field.
- **Problem Formation**: Our model and cases demonstrate that the main driver for corpus linguistics work presents an inherent conflict: corpus curators need machines to do the work at a scale that humans cannot handle, but machines do not have human-like mechanisms, for example, for dealing with the semi-structured language data at the heart of this work. Detailed guides for conducting corpus research are plentiful, and while they address the brittleness of machine reading, their focus tends to be technical discussions of study design, or the theoretical foundations for and applications of quantitative methods of analysis. Alternatively, we probe our own research experiences for the complicated decisions made during the project-specific phases of corpus curation—the messy, middle stages that rarely appear in published research articles.
- **Methods and/or Discussion**: We illustrate the considerations corpus curators face when preparing data for analysis by sharing three distinct accounts of our own curation experiences, and we hope our readers notice parallels between our narratives and their own processes. Our corpora span genres (student writing, children's storybooks, and argumentative text) and languages (English and Mandarin). Our combined cases show that despite different purposes, processes, and datasets, we pursued common phases of curation work, albeit to different extents, using different methods. We then present these common phases in a model: Phase 1, collection, addresses issues of data location, the form the data takes, ownership, and ethical concerns. Phase 2, cleaning, focuses on removing noise to make text fully machine- and/or package-readable. Phase 3, pre-processing, considers decisions around structural units, standardisation, markup and metadata standards, and the relationship between such decisions and the insights later analysis can provide. It looks forward to necessary data structures in light of the tools selected for carrying out analysis.
- **Conclusions**: While the purposes and questions that underpin the nature of corpora differ, we contend that these curation phases enable questions to be asked and answered through analysis similarly. Accordingly, we draw attention to some ways in which researchers attend to corpus-specific concerns to ensure that findings from analysis are valid and meaningful, especially for pedagogical application.

- **Directions for Further Action**: We suggest that addressing the identified gap between machine capabilities and machine constraints poses an ongoing challenge for the human researchers who curate corpora and the pedagogic insight those corpora can yield. We acknowledge the need for further research (across disciplines, genres, languages, and datasets) to refine the curation phases we present. In parallel, we highlight the need to standardize reporting in every curation phase, with particular concern for issues of transparency, reliability, and replicability in light of the evolution of generative AI.

Keywords: corpus curation, corpus research, generative artificial intelligence, writing analytics

1.0 Introduction

Corpus linguistics and associated methodologies have been applied in a wide variety of studies, including, for instance, sociolinguistic analyses of language change throughout history, real-time tracking of speech in social media spaces, and pedagogical inquiries into disciplinary discourses. Regardless of the context, the usefulness and reliability of corpus analysis does, however, depend on the quality and integrity of the underlying dataset, which is a question of curation. In this Research Note, we present disparate datasets and use cases that speak to the core reason for, but also tension in, corpus analysis. Although we do not formally pursue research questions in the present writing, we nevertheless address and problematize the following questions:

- What do we gain in large-scale quantitative analyses of language, and what do we lose?
- How do our decisions in corpus curation shape our analytical possibilities, as well as the reliability of results?

Our goal is not to provide readers with a decision-making algorithm or blueprint for each granular step involved in corpus analysis; there are a variety of resources that successfully serve that function (e.g., Biber & Conrad, 2019; Brezina, 2018; Brown & Wetzel, 2023; Egbert et al., 2022; Friginal & Hardy, 2020; Geisler & Swarts, 2019; Itchuaqiyaq et al., 2021; Lang et al., 2023; McEnery & Brezina, 2022; Upton & Cohen, 2009; Wallis, 2021). Instead, we wish to offer our own experiences with corpus curation as examples and inroads for a discussion of overlapping concerns in our decision-making processes as corpus researchers. Each example we provide underscores particularities depending on the linguistic artifacts we collect. Our cases show that while machines offer affordances for linguistic and rhetorical analysis, we must also address the brittleness of machine reading to fully realise pedagogical benefit. We aim to illustrate how the rich and reliable analysis of language data at scale requires a process of curation that prepares text datasets for machine reading and brings to the surface the many features that humans use to make sense of that language data. With writing analytics in mind, we hope to initiate a broader discussion of what it means to collect and curate textual artifacts as data, especially in the context of increasingly digital, imagistic, and artificially generated streams of communication.

Here, we would like to clarify what we refer to as the *curation* stage of a corpus study: we see it encompassing the collecting, cleaning, and pre-processing of authentic linguistic data prior to analysis. In this stage of research, the curator makes decisions, each with particular epistemological assumptions, that can unintentionally diminish various aspects of the study.

Curators decide the boundaries for acceptable collection and process the data in ways that carry implications for the results—this pre-processing further abstracts the data from its real-world context. This marks a difference in human- and machine-reading: the accompanying factors necessary for human comprehension are lost in order for a machine to make anything meaningful out of it. Generative AI (GenAI) and large language models (LLMs) address some of the brittleness of machine reading, but they reproduce the core tensions of human- versus machine-reading in new ways. Incorporating the affordances of LLMs into corpus analysis while mitigating and accounting for their constraints is an important next step for corpus analysis as a discipline; Marcellino's curation example described below, collecting Mandarin text for the RAND Corporation, describes how LLM-based encoding was incorporated into the cleaning and pre-processing pipeline.

Curators of a corpus must first consider what types of language acts are represented in the artifacts comprising a linguistic dataset. For instance, researchers interested in how language changes over time must collect longitudinal data and design a diachronic analysis accordingly. For other research questions, a synchronic study design—a snapshot of writing or speech in a single timeframe—might suffice. In any case, linguistic data can be messy, and some researchers may find themselves constrained by real-world challenges for collecting data and metadata. Additionally, curators must consider the limits of their quantitative corpus studies by identifying which research questions might require additional qualitative data, like associated writing prompts or the thought processes or intentions of a writer. A corpus study can describe a single group of texts, or it might compare two or more corpora to identify key differences and similarities (e.g., between student work and published expert writing, between different demographics, before and after a learning intervention).

Our aim is to lay out key phases in such corpus curation from the start of collection to readiness for analysis, in sequence and in general terms, and to explore some considerations that accompany these phases. We illustrate general principles through examples drawn from our own experience, and we expound on them in light of these experiences. Our corpora are very different in nature, as are the experiences of gathering and transforming the texts we describe. Our texts span genres and languages, as well as public and private domains. Some texts originated in digital format and some required conversion from physical form. We highlight commonality across our processes and protocols despite these differences. In doing so, we hope to provide readers with a starting point for future conversations.

1.1 Corpus-Based Approaches to Writing Pedagogy

This Research Note explores questions about how decisions in the early stages of research affect downstream analyses. Recognizing that many readers are instructors themselves, we briefly note the significant pedagogical contributions of corpus linguistics in higher education to underscore the implications that early-stage curation choices have on the late-stage transformations of research findings into educational materials. Corpus methods grant researchers the ability to zoom out and identify tacit linguistic patterns that might otherwise go unnoticed; such patterns are particularly important for novice writers to know when they enter a specific discourse community (i.e., their chosen discipline, industry, and academia more broadly). A corpus-linguistics-informed pedagogy calls explicit attention to the linguistic frameworks typified in specific genres that new learners can use as writing strategies (Slagle, 2025); CROW's repository of teaching materials exemplifies this type of corpus-informed praxis (Staples et al., 2024). While the researchers cited in this section focus mostly on student writing in higher education contexts, we wish to note that the affordances of corpus linguistics carry over to many workplace situations involving professional and technical writing (e.g., Conrad, 2017; Friginal & Hardy, 2020; Vine, 2020).

Using studies from corpus linguistics to design instruction is especially useful for helping both student and professional writers build metalinguistic, or textual, awareness (e.g., Helberg et al., 2018; Slagle, 2025; Staples et al., 2024). Such an approach motivates writers to "become more responsible agents of their text," fostering a metacognitive awareness of their own communication decisions (Kaufer & Wetzel, 2017, p. 718). Additionally, corpus studies have the potential to equip writing instructors and practitioners alike with knowledge of disciplinary writing characteristics that are valuable for both workplace writing and writing across the curriculum/writing in the disciplines (WAC/WID) programs. For instance, scholars in WAC/WID have used corpus-based study designs to analyse student writing and published writing in engineering and propose pedagogical directions based on the results (Boettger, 2014; Boettger & Palmer, 2010; Boettger & Wulff, 2014; Conrad 2017, 2018, 2019; Cotos, 2017). Corpus studies that reveal differences between the advice of instructional texts and actual instances of authentic writing illustrate the potential for these methods to align pedagogy with practice (Boettger & Wulff, 2014; Conrad, 2018; Lancaster, 2016; Wolfe, 2009). Moreover, such studies underscore the impact of turning to authentic writing (or spoken data) and taking an explicitly descriptive approach.

Drawing from the Saussurean conceptions of linguistic competence and performance, Amy Devitt (2015) calls for a genre-based writing pedagogy that attends to both the "communicative event and individual language-users" (p. 44). We reflect more on pedagogical implications below, after presenting our specific narratives of corpus curation, but we find the distinction between genre competence and genre performance particularly useful for corpus-based research: each text we collect as researchers captures a single instantiation, or *performance,* of the given genre at that particular time, in that particular context. Phenomenologically, rhetorical situations never truly recur but are instead typified for readers (and writers) who recognize when certain features are used for certain purposes, in certain exigent circumstances, for certain audiences, and with certain formal constraints (Devitt, 2015; Miller, 1984). Devitt's (2015) conceptualization of genre performance, though originally aimed at studies of student writing, captures the concerns that researchers ought to continuously return to during the process of corpus curation. That is, authentic instances of real-world human language acts must be carefully collected, cleaned, and pre-processed if we are to learn anything useful about the genre performance itself.

1.2 Matters of Corpus Curation

Corpora can address and invite questions that underpin writing analytics research. They are composed of particular and authentic language varieties or domains which, when analysed, can reveal characteristic features or generalisable patterns; founding principles of representativeness, size, balance, and sampling (e.g., Biber, 1993; Faigley & Witte, 1981; Leech, 1991; Sinclair, 1991) remain fundamental. Asking and answering these research questions meaningfully requires the right kind of language data and metadata, in sufficient proportions, along with appropriate permissions and ethical processes around data usage.

We anticipate that some readers may find that suitable existing corpora are not available and choose to make their own. A range of comprehensive guides are available to support this work. Questions of theoretical design and access are laid out in detail in the substantial scholarship of corpus building, from fundamental principles and general considerations (e.g., Reppen, 2024; Xiao, 2010) to reflections on corpora and LLMs (e.g., Crosthwaite & Baisa, 2023; Curry et al., 2024). The practicalities and particularities of curating—or, purposefully compiling—corpora matched to aims are considered in light of overall type (e.g., general, specialised, historical) (Murphy & Riordan, 2016), as well as characteristics like mode (e.g., written, spoken, multimodal), genre, and language. Options for selecting analysis software, from commercial to open source, are well documented (e.g., Berberich & Kleiber, 2023; CLARIN, 2025). Available resources tend to be either broad enough to provide general guidance or aimed at highly technical or specialized studies. Practical, example-based options for addressing curation issues across different fields are less common.

Running in parallel to the many design and process questions curators face are those of ethics. The tenets of research ethics in applied linguistics (e.g., consent, privacy, anonymization, legality, transparency) are inextricable from decision making about data collection and handling. Ethical considerations prioritize protecting the human participants who create the natural language data used in corpora, which is drawn from a range of real-world contexts and modes (e.g., Brookes & McEnery, 2024; Sterling & De Costa, 2018). These principles of corpus design and ethics remain important, and there is increasing advocacy for case-by-case ethical decision making (e.g., Brookes & McEnery, 2024). Rudniy (2018), for example, focuses on legal regulation in relation to de-identification and pseudonymization, highlighting how barriers to the automation of these processes in a corpus of STEM student writing necessitated manual post-editing. In the same vein, Leedham et al.'s (2021) work on building a corpus of highly sensitive texts takes questions of the ethics of de-identification further by deeming some texts "not for sharing," distinguishing researcher-only and public corpus versions. Particular ethical considerations may hold different weight, or nuance, in particular projects, due to the vast range of material and contexts with and in which corpus linguists work.

2.0 Our Corpora

We draw on three very different curation experiences and sets of expertise to illustrate the three general phases outlined: Michael Laudenbach worked with student assignments from American institutions of higher education, Siân Alsop worked with English-language children's storybooks published in hardcopy form, and William Marcellino worked with Mandarin-language policy documents. Our different purposes and processes required movement through the same phases with different emphases. Alsop lingers more on questions of collection, for example, and Laudenbach on issues of cleaning. Marcellino offers more in terms of pre-processing. We summarize the focus of our accounts across curation phases in Table 1. All of this is meant to generate questions and considerations for future discussions of corpus curation, especially for newer practitioners.

Table 1

Comparison of Curation Phases Across Three Accounts

		2.1. Student Writing	**2.2 Children's Storybooks**	**2.3. Policy Documents**
Overall purpose		Comparing academic genres and registers	Building a vocabulary resource	Identifying argumentation structures
Curation phase focus	Collecting	Processes of site-specific data extraction	Pragmatics of accessing and transforming hardcopy texts	Using an LLM-powered pipeline: ingest text; chunk, tokenize, and embed text; and pass to an LLM via API with specific prompt guidance to apply an analytical heuristic
	Cleaning	Issues of anonymization, ethics, and risks of re-identification	Messiness in *noisy* text and limited LLM use	
	Pre-processing	Tagging rhetorical features using corpus software (DocuScope) and generating parallel corpora using LLMs	PoS and word difficulty tagging using corpus software (SketchEngine)	

2.1 Student Writing (Laudenbach)

In higher education, administrators and researchers collect student writing for placement or other assessment purposes, but these corpora also serve as the basis for investigations into the registers and genres of specific academic disciplines. Corpus linguists have used authentic collections of student writing to pinpoint which linguistic and rhetorical strategies students need to hone to effectively participate in their discourse communities (Swales, 1990).

The research questions I (Laudenbach) have pursued in my research into student writing amount to a large-scale rhetorical task analysis: in a specific discipline, what types of genres and registers do experts use (i.e., academics and practitioners), and how do students develop and employ these practices? This approach seeks to assist writing pedagogy across the disciplines by identifying the specific rhetorical and linguistic features to prioritize in classroom practice and feedback.

2.1.1 Collection and Anonymization of Student Writing

Student writing data, however, must be collected, maintained, and analysed with particular care, abiding by the U.S. Family Educational Rights and Privacy Act (or FERPA) and the ethical standards for human subject research set by the researcher's institutional review board (IRB). If any data is collected in the European Union (EU) or European Economic Area (EAA), then the curator must follow the more stringent guidelines outlined by the General Data Protection Regulation (GDPR).

Perhaps one of the more important steps in collecting student writing is anonymizing or de-identifying the sources to maintain the privacy of students under relevant legislation, which includes the GDPR in the EU/EEA and FERPA in the US. In computational linguistics, researchers have used named entity recognition (NER) in machine learning to automate and augment text anonymization (e.g., Pilán et al., 2022; Rudniy, 2018). This includes both direct and indirect identifiers, or "quasi-identifiers," such as geographic, demographic, biographical, or temporal markers (Pilán et al., 2022, pp. 1057–1058). Should corpus data include, for instance, legal texts with identifying biographical information strewn throughout, then more advanced computational techniques for anonymization should be considered, including metrics for benchmarking the process (Pilán et al., 2022). For student data, though, we targeted course assignments wherein no biographical information is shared by the student. This differs from, say, admission letters, statements of purpose, or literacy narratives, genres in which students are expected to share personal details.

At Carnegie Mellon University, we collected student writing from statistics courses, where students submitted papers as either RMarkdown files or through a digital learning tool that stored submitted writing in HTML files (Nugent et al., 2019). This lucky set of circumstances allowed us to write a custom R script to extract text from papers without capturing any identifying information. In a similar effort at the New Jersey Institute of Technology, we obtained IRB approval to collect student texts through Canvas, our learning management system, allowing us to easily download files from our target courses. However, this method required repeated and extensive passes of manual anonymization to ensure that no identifying information remained in the cleaned data. In the context of our research questions in higher education, our metadata included the course section, project type, year, semester, and a randomized identification number for each document, usually embedded in the file name (i.e., [RandomID]_36-200_F2020_01). Lastly, file order was randomized to avoid accidentally maintaining the texts in alphabetical order, a risk of re-identification. For our purposes, this satisfied the data protection protocols set forth by our IRB; the full scope of data anonymization, de-identification, and pseudonymization in corpus linguistics scholarship is outside the purview of this Research Note (see Rudniy, 2018).

In a study with enough technical oversight, curators could control the file types they receive by having instructors restrict the file submission type; for instance, Word Documents (.doc, .docx) might be preferable to PDFs, especially since non-embedded PDFs require the use of optical character recognition (OCR), which can lead to issues during data cleaning that Alsop explores in detail below. If the target corpus includes writing from more technical disciplines, instructors might ask students to submit files in HTML, Quarto, or RMarkdown, which lend themselves to more efficient cleaning techniques since the curator can pinpoint specific sections of the text to extract.

2.1.2 Cleaning

These texts were then cleaned using custom scripts in Python and R. In this context, curators can use regular expressions to capture and remove any remaining identifiers like names, school IDs, course instructor, etc., or they might use automated techniques for NER (see Rudniy, 2018). Depending on the study design, researchers should also consider removing and noting textual features, like section headers, figures, or tables in their metadata for a greater ease of organization in future analyses. Recent research has been published using the student corpora described here (DeLuca et al., 2025; Laudenbach, *forthcoming*; Markey et al., 2024). The metadata used in these studies includes the course section, semester, and project number, which are each embedded in the file names (Laudenbach, *forthcoming*). As we state elsewhere, the cleaning phase of corpus curation should aim to transform the data into a state from which all future analyses can stem. This way, researchers can return to the exact dataset used in other analyses.

2.1.3 Pre-Processing for Genre Analysis

Corpus-based studies of style, register, or genre (Biber & Conrad, 2019) use various methods and methodologies, and for a large-scale descriptive study, text-taggers and advanced statistical techniques are common. We chose to tag the Carnegie Mellon University corpus of student writing in statistics for part-of-speech—using CLAWS7 (Leech et al., 1994)—and the 63 lexicogrammatical features developed by Biber (1992). For the latter, we used pseudobibeR, an R package that approximates Biber's tagset and returns raw and normalized counts (Brown, 2024).

Perhaps the more interesting part of this study involved the use of DocuScope, a dictionary-based text-tagger, to tag rhetorical functions. Part of a broader research project that began at Carnegie Mellon University in 1998 (Brown & Wetzel, 2023; Kaufer & Ishizaki, 2023), DocuScope is somewhat unique in its approach to text-tagging: whereas several corpus-based studies tag texts for lexicogrammatical structures or parts of speech, DocuScope uses a robust dictionary to categorize tokens according to their rhetorical function, thereby connecting micro-level linguistic choices to macro-level rhetorical moves. It can be used in either an interactive application (useful for exploratory analysis) or by simply processing all files within a directory. Recent scholarship has demonstrated the effectiveness of DocuScope for investigating genre and register in written texts from a rhetorical perspective (e.g., Brown & Wetzel, 2023; Kaufer et al., 2004; Laudenbach, *forthcoming*; Marcellino, 2014; Taguchi et al., 2017; Zhao & Kaufer, 2013).

Additionally, a newer tool developed by David West Brown, DocuScope Corpus Analysis (CA; Brown, 2025), has many pre-processed corpora built-in, enabling users to compare, for example, student writing from the Michigan Corpus of Upper-Level Student Papers to published scientific writing hosted by Elsevier. It also includes the British Academic Written English Corpus (BAWE) and the Human-AI Parallel Corpus, discussed below. DocuScope CA is free to use and makes standard corpus linguistics methods (e.g., n-gram analysis, collocations, keyness analysis, principal component analysis) easily accessible for experienced researchers and instructors as well as those wanting to familiarize themselves with corpus analysis. Indeed, students in first-year writing courses at Carnegie Mellon University are currently using the tool. Under the hood, it functions as a text tagger, processing corpora with the CLAWS7 part-of-speech tagset and a spaCy model trained on the DocuScope dictionary, docuscospaCy, which can be found on huggingface or accessed directly via its Python library with the same name. This makes DocuScope CA distinct from previous versions of DocuScope: instead of dictionary lookups, it now uses a pre-trained machine learning model to tag tokens, which greatly reduces the computation time while providing easy access to its uniquely rhetorical approach to corpus analysis. Using the tool, users can also download the data as CSV files containing normalized tag frequencies for further analysis.

Similar to Biber's (1992) multidimensional analysis, DocuScope's many feature categories (numbering anywhere from 9 to over 30, depending on the user's choice of dictionary) are typically transformed using statistical dimension-reduction methods like factor analysis, principal component analysis, or linear discriminant analysis (see Brown & Wetzel, 2023, for an extensive discussion of DocuScope-related research). To prepare the tagged data for analysis, the curator simply needs to normalize and output the tag counts into a file, likely a CSV, where each row is a document from the corpus, and each column is one of the DocuScope categories. This data can then be read into R or Python for further processing and analysis. These pre-processing decisions were informed by previous decisions made during early phases of study design.

For my own research, I chose to apply DocuScope's uniquely rhetorical tagset because of my focus on genre-based pedagogy. I take an approach to corpus-based genre analysis exemplified by Aull's (2015) discussion of the disciplinary separations of English for Academic Purpose (EAP) and rhetorical genre studies (RGS)—noting, specifically, their shared emphasis on genre awareness. Where critiques of RGS center on the "erasure of the sentence," those of EAP mention a disregard for social contexts and rhetorical exigence (Aull, 2015, p. 18). Aull (2015) attempts to reconcile these differences in her analysis, which "highlights rhetorical cues of [student] essay prompts (often absent in EAP corpus linguistic research) alongside shared linguistic patterns (often absent in RGS studies)" (p. 1). Scholars have employed corpus methods to explore linguistic patterns that become norms in certain contexts, sometimes refuting prescriptivist views of genre and language by proving them to be unsubstantiated by actual evidence (Lancaster, 2016; Perales-Escudero, 2011; Swales et al., 1998; Wulff et al., 2012). Central to the descriptivist approach is a shift away from *right or wrong* and *good or bad* writing to *appropriate* writing for a given rhetorical situation (e.g., Bitzer, 1968; Gere et al., 2021).

Following Aull (2015, 2020), my approach to corpus studies aims to make "tacit linguistic expectations and language-level patterns more a part of [student] writers' explicit genre knowledge" (2015, p. 174), and multivariate analysis with text-taggers like DocuScope helps me to accomplish that since the variables are intuitive and related to rhetorical moves discussed in writing curricula. Additionally, EAP research suggests that text naming and genre naming help ease the cognitive burden of L2 learners who may encounter different obstacles than those L1 students face upon entering English-speaking university settings (Cortes, 2004; Johns, 2011). More specifically, both L1 and L2 novice writers benefit greatly from the explicit teaching of "lexical bundles," especially those that recur in specific disciplinary contexts (Cortes, 2004).

With this in mind, we in TeachStat wanted our pre-processing to be extensive enough to answer potential future research questions. However, curators should note that pre-processing can be iterative: researchers might return to the raw or cleaned text data and perform various transformations to investigate different linguistic or rhetorical dimensions that can emerge during analysis. For our purposes, DocuScope, part-of-speech, and lexicogrammatical tagging were more than sufficient for a large-scale genre analysis of student writing tasks.

2.1.4 A Note on Generating Parallel Corpora Using LLMs

In a separate study, the TeachStat team at Carnegie Mellon University created the Human-AI Parallel Corpus in English (HAP-E) for the purpose of comparing human writing to texts generated by six different LLMs (Reinhart et al., 2025). While corpus linguists have been experimenting with LLM-generated texts in comparative analyses, HAP-E is distinct in its one-shot prompting. The researchers took the first 500 words of a human-written text sampled from the Corpus of Contemporary American English (COCA), then asked a given LLM to generate the next 500 words with the following prompt:

> In the same style, tone, and diction of the following text, complete the next 500 words, generate exactly 500 words, and note that the text does not necessarily end after the generated words.

It should be noted that on more than one occasion, different models generated nonsensical outputs that were discarded during collection. The corpus therefore allows a direct linguistic comparison between human and machine output, organized by genre according to the corresponding human-authored COCA texts. HAP-E is publicly available on huggingface, where readers can also find documentation on the corpus contents and instructions for pre-processing (see Reinhart et al., 2025).

2.2 Children’s Storybooks (Alsop)

At Coventry University, our primary research need was to curate a corpus from which to develop a new way of testing the extent to which children aged 3–6 acquire storybook vocabulary. As we could find no viable existing dataset of such words, we built the Early Years StorYbook (EYSY) corpus. Specifically, our goal was to identify 48 words that characterise the storytelling genre through analysis of our large corpus of 1000 books (including over 280,000 words, around 18,000 of which are unique). The words we identified underpinned a test rolled out in schools across England as part of the ‘Story Choices Trial’ within the ‘Teacher Choices’ initiative, funded by the Education Endowment Foundation (2025). To make our genre-specific test, to measure child vocabulary, to gauge intervention effectiveness, to ultimately guide classroom practice, we first needed words. We needed authentic storybook words, and an awful lot of them. The task demanded the speed and scalability afforded by machine methods (despite limitations), as the depth of human reading would have been too slow at such scale.

2.2.1 Collection

With our very specific need weighing heavily as a project lynchpin, our corpus design was critical, particularly in terms of ensuring representativeness. We needed to know which books children across the UK were likely to encounter, and sample accordingly. We consulted library usage statistics, authentic school provision, home provision, and expert guidance. Our end list of 1000 books represented common usage, which looked like many books written by many authors, published by many different publishing houses, across time.

Due to our purpose and design, emphasis quickly shifted from the highly conceptual to the highly practical. Collecting electronic versions of these texts (including old favourites on school bookshelves that were not necessarily digitally available) emerged as a logistically impossible task within our timeframe. We instead decided to purchase physical copies and make them digital. This was a consequential decision. We faced significant issues in acquiring the books, despite our highly supportive bookseller (who was equally patient about massive institutional delays in paying the invoice). Once finally delivered, transforming these beautiful physical artefacts into beautiful digital pictures posed particular challenges. Paying experts to do it was prohibitive and would have meant destroying the books, which we planned to pass on to our schools after the project concluded. We did have access to a high-quality scanner that could produce high quality images, but logistical constraints, including scanning time per book and patchy building access, made this unfeasible.

We opted instead for convenient technology—a tablet and/or phone, and various creative ways of holding pages in place. We explored various software options, and Adobe Scan proved most effective for our purposes. Our institution offered sufficient space to safely store the resultant PDF files. Along the way, we encountered numerous, and often comical, unforeseen practical aspects of taking pictures of a book collection that, when stacked, climbed halfway up a bedroom wall.

To scan 1000 physical books to make 1000 PDFs, we exhaustively used the Adobe Scan mobile app, then saved each PDF with a consistent naming convention. Here, we met numerous challenges related to unconventional book design, both physical elements (e.g., fold/pop-outs, shiny features) and layout features (e.g., text merged with illustration).

To extract the text (1000 PDFs to 1000 raw TXT files), we used Google Cloud Vision in document_text_detection mode, which suited our varied layouts. Using an automated Python script, we converted each PDF to an image (using the PyMuPDF library (fitz)), then applied OCR to each image (Vision API), extracted plain text, and saved separate TXT files. Here our challenges (and learning) came in many flavours. Some text was part of illustrations, not the main body of the story (e.g., a character reading a newspaper with words in the picture). The OCR struggled to recognise some highly stylised fonts and was confused by physically textured pages (a fun feature of children's books). At scale, processing thousands of pages involved tens of thousands of application programming interface (API) calls, which needed to be managed. While not perfect, OCR produced sufficiently accurate raw text versions of our main body story words, along with plenty of extraneous information.

2.2.2 Cleaning

In our cleaning phase, the very physical met the very artificial. We needed to turn 1000 very noisy plain text files into 1000 very clean plain text stories. Attempts to write scripts to undertake data cleaning proved painful—we simply had too many variables and too little time to do so successfully. We looked to GenAI, and input from more experienced colleagues helped here (thanks, Bill Marcellino). As our data is for research purposes only and, for copyright reasons, cannot be fully reproduced or made public, we needed an offline and private system. To isolate only the main body stories, we used a locally hosted language model (mistralai/mistral-7b-instruct-v0.2). We ran the model through LM studio, accessed through a local API, which allowed us to engage with the LLM privately on a local computer. This also let us control the design of the instructions, or prompts, we input for cleaning.

After some trial and error, our prompt focused on the following objectives: 1) maintaining the story content, 2) removing non-story text (e.g., author bios, page numbers, special symbols, ISBNs), and 3) avoiding rewriting the story. The process involved various challenges, largely related to the consistency of formatting, noise from illustrations, paraphrasing and/or gap filling, and capacity constraints due to local hosting. This approach was far more successful than scripting for us and resulted in clean versions of only our stories. We linked these stories to corresponding metadata (e.g., title, author info, and publication info) held in a separate spreadsheet provided by our helpful bookseller. To match story text and metadata, we largely relied on ISBNs, which we extracted from the initial noisy versions of our plain texts using regex and validation logic. We defaulted to title matching in the few cases ISBNs were missing. A few small challenges here related to errors in the metadata and similarity in titles.

The result of these processes was the production of a corpus of 1000 machine-readable story texts and associated metadata that represents the books young children across the UK encounter daily at school. Because of the messiness of our process (digitising, extracting, and cleaning the text), we wanted to ensure we had achieved above acceptable reliability beyond our ongoing manual checking processes (which gave us confidence in > 80% reliability across the corpus). This was acceptable for our initial purposes given time pressures. We are currently comparing a gold standard sample (i.e., 100 books / 10% of fully human-checked texts) to the produced corpus, and early indications point to high reliability (> 90%).

2.2.3 Pre-Processing

We chose to use the corpus tool SketchEngine (Kilgarriff et al., 2014), into which we could upload our clean story body texts and automatically import matched metadata to form the header, then commence pre-processing. SketchEngine also allowed us to create a private and safe corpus (i.e., ISO 27001 certified), add the extra information we needed, and later use pre-existing in-built corpora (in our case, a subset of CHILDES) as reference data to perform keyness analysis. It also offered a user-friendly experience across the research team, not all of whom were familiar with corpus linguistic approaches. The extra information we added included part of speech (PoS) tags (via the English TreeTagger PoS tagset with SketchEngine modifications). We also tagged the difficulty level of each word for children, matched from Kuperman's (2012) Age of Acquisition (AoA) wordlists. This information allowed us to meet our inclusion and exclusion criteria to identify characteristic story words.

2.2.4 Takeaways

A practical complication arose post-curation in the form of the expiry of our institutional access to SketchEngine and limited local resources to renew the license. We maintain access for now, but this unforeseen narrative twist may influence how we conduct further analyses. Regardless, by curating a clean corpus associated with full metadata, we are well positioned to look beyond the initial research need to address broader questions of language and genre using various computational means.

Acquiring and digitising our texts was the least skilled yet pragmatically most challenging part of our process. Several times, we made unforeseen *art of the possible* decisions to balance adhering to important design principles while meeting project milestones. To this end, the use of GenAI was critically enabling; we accepted with reservations a lack of transparency during part of our cleaning phase, accordingly. Limiting the scope of black-box work and implementing regular reliability checks made this trade-off more palatable at the time, and a good decision in retrospect. Use of off-the-shelf corpus analysis software was valuable for giving access to built-in reference corpora and for enabling a team with mixed experience to work together easily, but it did raise concerns about the longer-term risks of relying on third-party software. However, the curation work we did will allow us to flexibly carry out further analyses using any software. Throughout this work, buy-in and goodwill among our established research team was probably the most important factor in overcoming both technical and logistical challenges.

2.3 Chinese Language Data and LLMs (Marcellino)

2.3.1 Collection

At RAND, we have developed an AI application to analyse argumentative texts using a Toulmin model framework—identifying claims, grounds, and unspoken warrants (Toulmin, 2003) in argument. Our goal was to extract possible points of conflict in contested public sphere issues to better inform public policy. As part of testing, we wanted to apply this approach to non-English data to consider whether the Toulmin model framework would work across languages and to evaluate how well multilingual LLMs worked on other languages. As a pilot test, we chose Chinese language (Mandarin), and a small corpus of three People's Liberation Army policy documents on artificial intelligence was curated by a native Chinese (Taiwanese) speaker with decades of experience in China military affairs. The small document count reflects the nascency of GenAI as a technology and our need to pilot with human qualitative inspection of each stage of the pipeline. Because claims tend to be located at the paragraph level, each of the three doctrinal publications contained 20+ claims, sufficient for this pilot stage.

Our Toulmin model workflow employed an LLM-powered pipeline: text documents were first split into paragraphs, tokenized, and then encoded as dense numerical vectors using OpenAI's ADA-2 embedding model—a neural network that maps texts into a high-dimensional space where semantically similar documents appear closer together. These embedded paragraphs were then passed to an LLM via API calls to our internal model service. We used Microsoft's Azure service, and GPT-4-nano as our specific model (relatively cheap and fast for enterprise use). However, the pipeline itself—ingest text, chunk/tokenize/embed text, pass to an LLM via API with specific prompt guidance—was agnostic to the specific model used or whether the process employed a commercial cloud vendor or local AI model deployment (local deployment may be key for data privacy concerns). Further, while this case involved GenAI rather than deterministic natural language processing (NLP) methods or statistical corpus methods, the text pre-processing considerations are generalizable.

2.3.2 Cleaning

A key benefit of using LLMs is that modern frontier LLMs do not require much in the way of corpus cleaning; the global attention mechanism that lets LLMs learn language relationships broadly from trillions of tokens makes them robust readers. Contemporary LLMs can handle orthography variation, typographic errors, language variations and nonstandard forms, punctuation presence/absence, and so on. As a result, our human attention could be more valuably focused on the analytic output of our pipeline, letting us ask questions such as was the model extracting warrants in ways that made sense to an expert native-language analyst and were results sufficiently similar over multiple iterations to be trustworthy.

2.3.3 Pre-Processing

In early trials, our system ingested Rich Text Format (RTF) documents that had been converted from Word files. When tokenized, these produced nonsensical or corrupted tokens, which in turn led to spurious LLM outputs. LLMs are stochastic, and while powerful, can be highly unreliable, producing plausible but misleading output. In this case, the model did what it was supposed to do—follow instructions—and with no meaningful input, cheerfully complied and made-up plausible output based on text file names, not file contents. This scenario highlights a serious limitation with using LLMs for NLP, one that requires researchers to carefully inspect and vet inputs and verify outputs.

Ultimately, the cause of our nonsensical output was a mismatch between text encoding formats—the RTF file contained hidden markup and non-UTF-8-character sequences that distorted Mandarin characters. The fix, converting all source material to UTF-8 encoded plain text, ensured accurate character representation and consistent token boundaries. Once the data pipeline used standardized UTF-8 text, tokenization and argument extraction performed as expected, correctly identifying implicit logical relationships in Chinese prose, as shown in Table 2.

Table 2

Model Pipeline

Passage Text	Claim	Grounds	Unspoken Warrant
The People’s Liberation Army (PLA) has developed incrementally, historically proving adept at modernizing and mastering quality developments.	The PLA is historically adept at modernizing and mastering quality developments.	The PLA has shown a capacity for incremental development and adaptation through history.	A history of successful adaptation indicates future success in new developments.

2.3.4 Broader Lessons for Preparing Mandarin Corpora

This case illustrates the importance of methodical corpus preparation for Mandarin NLP and GenAI analysis. Key best practices drawn from recent reviews include:

1. Encoding consistency: Always normalize source text to Unicode (UTF-8). This preserves both simplified and traditional characters and prevents corruption when passing through different NLP systems (Aliero et al., 2023).
2. Segmentation accuracy: Mandarin lacks whitespace word delimiters. Apply hybrid segmentation methods—combining rule-based, statistical, and transformer-based models—to ensure tokens align with meaningful linguistic units before embedding or downstream analysis (Zhang et al., 2022).
3. Punctuation handling: Distinguish between Chinese and Western punctuation forms. Proper punctuation normalization is essential for sentence boundary detection and for preserving rhetorical structure in argument extraction tasks (Aliero et al., 2023; Zhang et al., 2022).
4. Normalization of nonstandard forms: Use multi-level normalization to handle dialects, colloquial expressions, and variant orthography. Combining dictionary-based and neural sequence-to-sequence approaches improves robustness in mixed or user-generated content (Aliero et al., 2023).
5. Corpus documentation and maintenance: Maintain annotated corpora and lexicons tuned to Mandarin. Document pre-processing steps, encoding standards, and segmentation parameters to ensure reproducibility and future reuse (Aliero et al., 2023).

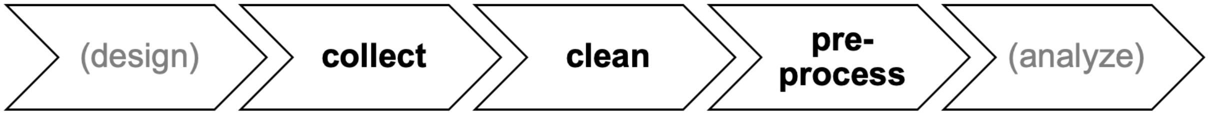
2.3.5 Takeaways

The *unspoken warrant* extraction experiment demonstrated that technical pre-processing choices can directly determine analytic validity. Encoding errors obscure meaning at the most basic level, while high-quality, linguistically informed pre-processing enables LLMs to reason accurately about argument structure. For corpus linguists and AI practitioners, robust Mandarin text preparation—rooted in encoding integrity, segmentation precision, and normalization discipline—is foundational to meaningful computational analysis.

3.0 Corpus Curation: Three-Phase Model

Although corpus curators seek data purposefully, our experiences reveal a practicality to questions of making this data ready for analysis that may be underestimated in design phases. We used the three-phase model shown in Figure 1 to help each of us organize and articulate the narratives presented above, and here, we expound on those phases with the hopes of highlighting and prompting a similar discussion in the field. It is important to note that the phases we list here are iterative; curators may choose to return to their cleaning or pre-processing phase to transform the raw data in different ways for various research questions, refining the data through repeated assessment of its quality and reliability. Our present discussion broadly covers the work done after a study has been designed and before analysis is undertaken.

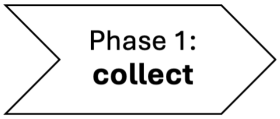
Figure 1
Three Corpus Curation Phases: Collecting, Cleaning, and Pre-Processing for Analysis



Note: These phases are affected by upstream design choices and affect downstream analysis of the cleaned and processed data.

Some genres may already be available in digital form, like student assignments or policy documents, and some may need to be transformed, such as lectures or hardcopies of manuscripts. *Collection* involves first locating where the raw target data lives and how to access it. Although most data can be retrieved digitally with relative ease, non-natively digital data may require the researcher to step away from their computer—to capture spoken events not available online or to visit physical archives, for example.

Figure 2
Phase 1: Collect



- In what form does (or will) the raw collected target data exist? (e.g., modality, language, file format)
- Where is target data and metadata held and/or encoded?
- Who owns the data? Who can and cannot access it?
- Does the raw data need to be changed to create a written, machine-readable version?

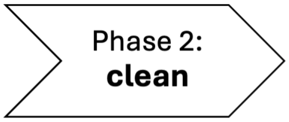
All types of corpus curation meet in the *cleaning* phase. It is here that digitally held data is wrangled into a clean form—in other words, stripped of all noise, fully machine-readable. This phase is about achieving accuracy and consistency while preserving authenticity and preventing harm. Questions of anonymization may be particularly complex in relation to the ownership of certain text types; removing direct identifiers (e.g., names, reference codes) may be insufficient when combinations of indirect identifiers (e.g., reference to a place, a particular course, a rare experience, or a timepoint) can be cross-referenced with external information to identify the author. Cleaning involves careful decision making about what data should be kept and in what form, and how these decisions are implemented at scale.

The influence of GenAI complicates questions of ethics and makes them significantly more important. At this point, we acknowledge that the use of what might be considered traditional AI is not new to corpus linguistics. Under the broad umbrella of work done by machines with human-like intelligence (e.g., understanding language, translation, learning, problem solving), most corpus linguists' toolkits have an AI influence. Indeed, Curry et al. (2024) argue that "any study published under the banner of corpus linguistics over the past 30 years is indebted, to some degree, to advances in AI" (p. 2). By nature, corpus linguistics methods are related to language models, machine learning, and NLP methods; these areas share methodological aims related to processing large amounts of text. Where corpus linguistics approaches diverge from the goals (and role) of AI is in a commitment to interpretation beyond automation (e.g., McEnery & Hardie, 2012; Teubert, 2005), in privileging discourse meaning over algorithmic modelling. Despite tensions, a mutually enabling relationship has generally existed, where corpus linguistics offers descriptive frameworks, language data, and linguistic insight, and NLP provides the computational tools to carry out automated tasks.

GenAI obscures the processes and decisions behind this work, however, from criteria for data gathering to phases of data processing. Beyond following rules to understand or categorise existing data, it creates new data (e.g., Kalota, 2024). Increasing recent engagement with GenAI (e.g., through user-friendly chatbots and platforms underpinned by powerful LLMs) complicates possibilities in terms of what *can* be done and amplifies questions around what *should* be done with language data. What current discussion across the field brings to the fore is the very immediate and practical nature of considerations stemming from ethics, particularly those that relate to how language data is acquired, handled, and processed. Understanding the risks and rewards of GenAI use is far past theoretical territory. The resultant need for researchers to engage in informed decision making about where and how data collection takes place and the way in which data is handled is tied to the growing availability of both this data and the methods that are used to prepare and process it.

Figure 3

Phase 2: Clean



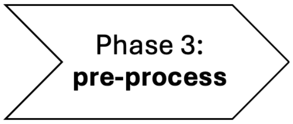
- Does non-textual material need to be removed? (e.g., metadata, page numbers, tables)
- Does textual data need to be removed? (e.g., references)
- Does any data need to be anonymized?
- Is formatting consistent? (e.g., removing duplicate or inconsistent spacing)
- What character encoding standard will be used to represent text (e.g., UTF-8)?
- Is language encoding consistent? (e.g., converting to a single standard, then ensuring special characters display properly)

Once clean, raw text will need to be pre-processed in preparation for analysis. The choices made in this step are not merely technical; rather they shape what sort of linguistic insights corpora can later reveal. For example, tokenization may split text into smaller, more flexible pieces (like a part of a word, or punctuation), which affects the way that text is computed. These decisions about standardization require the researcher to also question what, as a result, is lost. When considering lemmatization options, the quality of the information that groups inflected word forms under their base will determine the accuracy of analysis of word behaviours beyond surface form (e.g., morphological variance). Likewise, tagging text with functional roles guides how lexicogrammatical patterns can be revealed, thus impacting the accuracy of how comparisons of salience (keyword analysis) are made. Broader decisions about the metadata standard employed—the system for marking up the various data about data, such as author and publication information, modality, language, or recording—determine the context in which later analysis and interpretation of linguistic patterns take place.

So choices pertaining to the rules inputted (or tagsets applied) and metadata included, and the purposes driving these choices, all matter—and relate to overarching research aims. Practical guides break down these choices (e.g., Newman & Cox, 2020), and Berberich and Kleiber (2023) helpfully maintain a list of corpus linguistic tools and their uses, including those used for tagging. Keeping track of these decisions is important to principles of transparency and replicability.

Figure 4

Phase 3: Pre-Process



Phase 3:
pre-process

- What counts as a word? How will the text be divided into countable units, or tokenized?
- What other structural units matter (e.g., sentences, paragraphs, speaker turns)?
- Is standardisation necessary (e.g., case, spelling)?
- What linguistic markup is required (e.g., lemmatization, parts of speech)?
- Which system for adding linguistic markup is fit for purpose?
- Which standard and/or taxonomy is most suitable for metadata?
- How will all decisions be recorded?

Pre-processing texts for analysis assumes that some form of script will be used to run analytical tools, either directly or within corpus software. Considerations here include determining the time-to-completion for running those scripts (which affects the size of each invocation of the analysis), as well as further processing requirements (such as randomizing data). Student submissions, for example, may or may not be considered sufficiently random for IRB purposes, so for those corpora there may well be additional requirements for the organization of the corpus. For very large corpora of all types, size constraints may need to be applied so that program invocations terminate in a reasonable time.

Considerations in this last phase are, in essence, about determining how best to set the corpus up for the computations that will answer the questions being asked. The choice of computing tool brings with it the usual risks of focusing on quantitative solutions without considering the implications for what exactly those numbers represent. There are several powerful language analysis tools available: some free, some with nominal-to-moderate costs, and some best available by contacting those who have written the software. Some tools are more transparent than others. Each brings different affordances and constraints. Broader considerations will have influence here, such as aligning tool selection to research questions/aims, as well as the value of multidisciplinary perspectives in the analysis work ahead. In these choices, questions of design will likely (once more) be weighed against more pragmatic considerations of, for example, preference, ease, time, and cost. We aim to illustrate such decision making in our real-world examples.

4.0 Conclusions

4.1 Tensions in the Process

Our immediate purpose, and the intended takeaway for our paper, is to help corpus analysts with practical considerations for preparing rich but messy, semi-structured language data for scalable analysis by machines. As our examples show, the process of corpus curation itself is a pedagogical exercise in research design, data collection and maintenance, as well as methodological interrogation. Corpus curation, as an intentional practice, forces the researcher to reflect on their epistemological assumptions as they collect data in the wild, so to speak, and transform that data for specific research questions and analyses. For instance, in each of our examples above, logistical constraints influenced our methods, and, subsequently, the data we were able to collect bounded our analyses.

Ultimately, our disparate datasets and challenges speak to the core reason for, but also tension in, corpus analysis: machines can read at scale in a way that complements human high-context reading, but machines lack the robust reading mechanisms that allow humans to deal with the semi-structured nature of language data. If we photograph or photocopy text, for example, we may need to use OCR (or GenAI, as Alsop notes) to make the text machine-readable. If the language is character-based, or logographic (e.g., Mandarin), we may need a segmenter to make individual words readable. If we want to detect and make sense of style and stance features that are cumulative over a text and deeply implicated in constructs like genre, we may need to use tagging and statistical analysis software like DocuScope.

This tension between the slowness and limited scale of robust human reading versus the high-speed and scalable, but brittle, reading done by machines remains a core technical challenge for corpus analysis. The three-phase process we have laid out, combined with example technical challenges in a diverse set of corpus curation projects, has been meant to bring this tension to life and offer practical guidance for future corpora. Additionally, we have included Marcellino's GenAI example, where an LLM is used to bypass much of the labor-intensive work of pre-processing text (e.g., segmentation, punctuation handling, normalization). The integration of LLMs into existing NLP pipelines appears to be a convincing use case currently being explored.

LLMs do, however, replicate the core tension between human and machine reading in a new way. Humans can make robust sense of language data because we have rich, long-tailed conceptual knowledge and high-level context knowledge. While humans do use information-compression techniques (e.g., higher organizing schema: *birds* containing members like sparrows, robins, starlings, eagles, ostriches, & penguins), we still maintain extremely rich and nuanced information, even though this comes at the cost of storing and accessing extremely rare information (Shani et al., 2025). We can see this in language data, where word frequencies follow Zipf's Law: very few items make up the vast majority of usage, with an extremely long tail of very infrequent but extremely useful language, which gives humans tremendous conceptual flexibility and power. Machines, meanwhile, at least the LLMs being used now for a variety of tasks, trade off conceptual richness and reasoning in favour of scale and speed (Morris et al., 2025).

As our brief accounts of experience indicate, those who curate and analyse corpora face a growing number of decisions about the acceptability of such trade-offs, which will inform next steps in corpus analysis as a discipline. On the one hand, we think GenAI is too potentially powerful and useful to ignore; on the other hand, it also presents a new set of challenges, and we cannot naïvely apply LLMs to any given task. Just as corpus analysis has over time built up rigorous processes and tools to compensate for the brittleness of machine reading, we will have to do the same in the age of GenAI.

Our work towards a model of curation points to a need, and opportunity, for researchers to report standards and checklists in each phase of their work, especially when GenAI is involved. Development of such documentation will help to ensure that researchers can fully account for how data is collected, cleaned, and pre-processed, for validation purposes. Further, with future GenAI developments in mind, quantifying the reliability of automated tools seems to be a weighty concern; it is not one that our accounts sufficiently address. We believe that establishing clear verification metrics to audit the accuracy of familiar steps that are increasingly drawing on GenAI to overcome traditional limitations (e.g., OCR and text segmentation) is a priority. As the nature of these steps evolves, establishing standardized documentation will surely strengthen the field's position in terms of fostering transparency and replicability.

4.2 Pedagogical Implications

By taking actual, authentic instances of discourse-use instead of institutionalized prescriptivism as a starting point (Kaufer & Ishizaki, 2006), corpus-based studies can help distinguish specific linguistic features of genre to target in teaching, in practice, and in future scholarship. Researchers have used numerous methods, both quantitative and qualitative, to explore academic and non-academic corpora for that very purpose. Additionally, concordancing software packages like DocuScope CA, discussed above, allow researchers to conduct complex analyses in simple and (mostly) intuitive ways, making corpus-based methods more accessible to instructors (see also Charles & Frankenberg-Garcia, 2021).

As noted above, some studies target specific linguistic features to evaluate hypotheses based on preexisting composition theories—for example, analysing hedging and boosting to substantiate claims about stance in academic writing (Aull & Lancaster, 2014; Hyland, 2005). Other studies are more exploratory, taking a descriptive and inductive approach to corpora of academic writing (e.g., Aull, 2020; Brown & Aull, 2017; Brown & Wetzel, 2023; Conrad, 2017). Further, many researchers use comparative methods, like keyness and correlation analysis, to explore variation between groups like expert and novice writers, L1 and L2 learners, or between academic and professional contexts (Granger & Paquot, 2009; Römer & Wulff, 2010). Lessons drawn explicitly from corpus studies—such as epistemic stance (hedging/boosting) or the appropriate contexts to use imperatives in an academic register (e.g., Aull & Lancaster, 2014; Reppen, 2010; Swales & Feak, 2012)—are undoubtedly valuable in higher education classrooms, especially in the context of GenAI.

While corpus-based studies reveal teachable linguistic patterns and offer insights into tacit linguistic conventions in different academic registers, they also uncover assumptions surrounding the preferences of readers. With more diverse studies of corpora from a variety of sociolinguistic contexts, educators can expand the range of linguistic registers to incorporate into lesson plans for students and resources for professionals.

Finally, just as understanding the difference between statistical significance and effect size is essential for interpreting basic findings, so, too, should those findings be evaluated alongside scrutiny of methods and methodologies, especially in data collection, cleaning, and pre-processing, as we have outlined in this Research Note. In short, just because a linguistic feature is statistically significant does not mean it warrants attention in course curricula. In any case, corpus-based studies are useful for building genre and disciplinary awareness, but the extent of their incorporation into curricula requires continual, iterative research to remain useful and up to date (e.g., Conrad, 2017).

4.3 Limitations and Future Directions

The three-phase model we have loosely suggested here results from three different (and quite niche) practical curation cases, which limits the comparisons and generalizations we can draw. Limitations are compounded by differences in the third-party tools we used, the constraints on data we faced, and the extent of our engagement with black-boxed steps. We present our model as a way of seeing commonality across phases of work, supplementing existing guides with real-world examples in writing analytics. Further cases that span a much wider variety of disciplines, genres, languages, and diverse datasets are, of course, needed to refine the phases we outlined here along with their accompanying considerations.

We believe that establishing a curation heuristic accompanied by real-world examples is particularly useful for those who are new to building corpora (or building new corpora) and that there is scope to incorporate such a model into training, perhaps in new researcher or graduate programs. Research into writing studies may also benefit from lessons on corpus curation and applications of GenAI, which may in turn support the development of curricula in the field. For instance, the use of the HAP-E corpus outlined by Laudenbach highlights opportunities for research into human-AI text comparison, which, with recent publications, appears likely to be a focus of writing analytics inquiry in the immediate future.

Acknowledgments

We gratefully acknowledge the generosity of our colleagues and the support of the writing analytics community. The editors-in-chief of this journal made this collaboration possible. We received feedback from Dr. Benet Vincent and Dr. Megan Kane, which helped shape early drafts, and feedback from two expert anonymous reviewers, which greatly refined our final draft. The three individual accounts reflect work undertaken by three valued research teams at Carnegie Mellon University (TeachStat), the RAND Corporation, and Coventry University, where special mention is owed to Mrs. Idara Bassey for work on the EYSY Corpus. Despite the wealth of expertise we have benefitted from, all errors are, of course, ours alone.

Author Biographies

Siân Alsop is an assistant professor (research) in the Centre for Global Learning at Coventry University. Her work uses corpus linguistic approaches to understand issues of language and literacy in and beyond academic contexts. ORCID: https://orcid.org/0000-0002-9639-0298

Michael Laudenbach is an assistant professor in the Department of Humanities and Social Sciences at New Jersey Institute of Technology. His research examines rhetorical features of academic writing, as well as the differences between human writing and AI-generated text. ORCID: https://orcid.org/0000-0003-1691-0562

William Marcellino is a senior behavioral scientist at the RAND Corporation, professor of text analytics at the Pardee RAND Graduate School and lecturer at Carnegie Mellon University. He was trained as a sociolinguist and corpus linguist, and at RAND he conducts AI policy research and oversees RAND's AI_Tools initiative.

References

Aliero, A. A., Adebayo, B. S., Aliyu, H. O., Tafida, A. G., Kangiwa, B. U., & Dankolo, N. M. (2023). Systematic review on text normalization techniques and its approach to non-standard words. *International Journal of Computer Applications, 185*(33), 44–55. https://doi.org/10.5120/ijca2023923106

Aull, L. L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy.* Palgrave Macmillan.

Aull, L. L. (2020). *How students write: A linguistic analysis*. Modern Language Association.

Aull, L. L., & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Written Communication*, *31*(2), 151–183.

Berberich, K., & Kleiber, B. (2023). *Tools for corpus linguists*. https://corpus-analysis.com

Biber, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities, 26*(5/6), 331–345. www.jstor.org/stable/30204629

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Biber, D., & Conrad, S. (2019). *Register, genre and style* (2nd ed.). Cambridge University Press.

Bitzer, L. F. (1968). The rhetorical situation. *Philosophy & Rhetoric, 1*(1), 1–14.

Boettger, R. K. (2014). Explicitly teaching five technical genres to English first-language adults in a multi-major technical writing course. *Journal of Writing Research, 6*(1), 29–59. https://dx.doi.org/10.17239/jowr-2014.06.01.2

Boettger, R., & Palmer, L. (2010) Quantitative content analysis: Its use in technical communication. *IEEE Transactions on Professional Communication, 53*(4), 346–357. https://doi.org/10.1109/TPC.2010.2077450

Boettger, R., & Wulff, S. (2014). The naked truth about the naked *this*: Investigating grammatical prescriptivism in technical communication. *Technical Communication Quarterly, 23*(2), 115–140. https://doi.org/10.1080/10572252.2013.803919

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

Brookes, G., & McEnery, T. (2024). Corpus linguistics and ethics. In P. I. De Costa, C. Cinaglia, & A. Rabie-Ahmed (Eds.), *Ethical issues in applied linguistics scholarship* (pp. 28–44). John Benjamins.

Brown, D. W. (2024). *pseudobibeR: Aggregate counts of linguistic features* (Version 1.1) [R package]. Github. https://github.com/browndw/pseudobibeR

Brown, D. W. (2025). *DocuScope corpus analysis* (Version 0.4.0) [Software]. Github. https://github.com/browndw/docuscope-ca-online

Brown, D. W., & Aull, L. L. (2017). Elaborated specificity versus emphatic generality: A corpus-based comparison of higher-and lower-scoring Advanced Placement exams in English. *Research in the Teaching of English*, *51*(4), 394–417.

Brown, D. W., & Wetzel, D. Z. (Eds.). (2023). *Corpora and rhetorically informed text analysis: The diverse applications of DocuScope* (Vol. 109). John Benjamins.

Charles, M., & Frankenberg-Garcia, A. (2021). *Corpora in ESP/EAP writing instruction*. Routledge.

CLARIN. (2025). *Corpus query tools*. https://www.clarin.eu/resource-families/corpus-query-tools

Conrad, S. (2017). A comparison of practitioner and student writing in civil engineering. *Journal of Engineering Education, 106*(1), 191–217. https://doi.org/10.1002/jee.20161

Conrad, S. (2018). The use of passives and impersonal style in civil engineering writing. *Journal of Business and Technical Communication, 32*(1), 38–76. https://doi.org/10.1177/1050651917729864

Conrad, S. (2019). Register in English for academic purposes and English for specific purposes. *Register Studies, 1*(1), 171–201. https://doi.org/10.1075/rs.18008.con

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423.

Cotos, E. (2017). Language for specific purposes and corpus-based pedagogy. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 248–264). John Wiley & Sons. https://doi.org/10.1002/9781118914069.ch17

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

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics, 4*, 1–9.

DeLuca, L. S., Reinhart, A., Weinberg, G., Laudenbach, M., Miller, S., & Brown, D. W. (2025). Developing students' statistical expertise through writing in the age of AI. *Journal of Statistics and Data Science Education, 33*(3), 1–13. https://doi.org/10.1080/26939169.2025.2497547

Devitt, A. (2015). Genre performances: John Swales' genre analysis and rhetorical-linguistic genre studies. *Journal of English for Academic Purposes, 19*, 44–51. https://doi.org/10.1016/j.jeap.2015.05.008

Education Endowment Foundation. (2025). *Story choices – Teacher choices trial*. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/story-choices-2024-25-ey-teacher-choices-trial

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

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414.

Friginal, E., & Hardy, J. A. (Eds.). (2020). *The Routledge handbook of corpus approaches to discourse analysis* (1st ed.). Routledge. https://doi.org/10.4324/9780429259982

Geisler, C., & Swarts, J. (2019). *Coding streams of language: Techniques for the systematic coding of text, talk, and other verbal data*. WAC Clearinghouse.

Gere, A. R., Curzan, A., Hammond, J. W., Hughes, S., Li, R., Moos, A., Smith, K., Zanen, K. V., Wheeler, K. L., & Zanders, C. J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication, 72*(3), 384–412.

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 193–214). Continuum.

Helberg, A., Poznahovska, M., Ishizaki, S., Kaufer, D., Werner, N., & Wetzel, D. (2018). Teaching textual awareness with DocuScope: Using corpus-driven tools and reflection to support students' written decision-making. *Assessing Writing, 38,* 40–45. https://doi.org/10.1016/j.asw.2018.06.003

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies, 7*, 173–192.

Itchuaqiyaq, C. U., Ranade, N., & Walton, R. (2021). Theory-to-query: Developing a corpus-analysis method using computer programming and human analysis. *Technical Communication*, *68*(3), 7–28.

Johns, A. (2011). The future of genre in L2 writing: Fundamental, but contested, instructional decisions. *Journal of Second Language Writing, 20,* 56–68. https://doi.org/10.1016/j.jslw.2010.12.003

Kalota, F. (2024). A primer on generative artificial intelligence. *Education Sciences, 14*(2), 172. https://doi.org/10.3390/educsci14020172

Kaufer, D. S., Ishizaki, S., Butler, B. S., & Collins, J. (2004). *The power of words: Unveiling the speaker and writer's hidden craft*. Routledge.

Kaufer, D., & Ishizaki, S. (2006). A corpus study of canned letters: Mining the latent rhetorical proficiencies marketed to writers-in-a-hurry and non-writers. *IEEE Transactions on Professional Communication*, *49*(3), 254–266.

Kaufer, D., & Ishizaki, S. (2023). The DocuScope project: History, theory, and future directions. In D. W. Brown & D. Z. Wetzel (Eds.), *Corpora and rhetorically informed text analysis* (pp. 2–24). John Benjamins.

Kaufer, D., & Wetzel, D. (2017). Rhetoric, composition, design. In M. J. MacDonald (Ed.), *The Oxford handbook of rhetorical studies* (pp. 709–722). https://doi.org/10.1093/oxfordhb/9780199731596.013.054

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

Lancaster, Z. (2016). Do academics really write this way? A corpus investigation of moves and templates in "They Say/I Say." *College Composition & Communication*, *67*(3), 437–464.

Lang, S., Buell, D., & Elliot, N. (2023). Computer-assisted corpus analysis: An introduction to concepts, processes, and decisions. *IEEE Transactions on Professional Communication, 66*(1), 94–113.

Laudenbach, M. (*Forthcoming*). Student writing at scale: A corpus rhetorical analysis of client-facing and expert-facing genres in statistics & data science. *IEEE Transactions on Professional Communication.*

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 8–29). Longman.

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)* (pp. 622–628). https://aclanthology.org/C94-1103.pdf

Leedham, M., Lillis, T., & Twiner, A. (2021). Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP corpus. *Applied Corpus Linguistics, 1*(3). https://doi.org/10.1016/j.acorp.2021.100011

Marcellino, W. M. (2014). Talk like a Marine: USMC linguistic acculturation and civil–military argument. *Discourse Studies, 16*(3), 385–405.

Markey, B., Brown, D. W., Laudenbach, M., & Kohler, A. (2024). Dense and disconnected: Analyzing the sedimented style of ChatGPT-generated text at scale. *Written Communication*, *41*(4), 571–600.

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

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Miller, C. (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2), 151–167. https://doi.org/10.1080/00335638409383686

Morris, J. X., Sitawarin, C., Guo, C., Kokhlikyan, N., Suh, G. E., Rush, A. M., Chaudhuri, K., & Mahloujifar, S. (2025). How much do language models memorize? *arXiv preprint*. https://arxiv.org/abs/2505.24832

Murphy, B., & Riordan, E. (2016). Corpus types and uses. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 388–403). Routledge.

Nugent, R., Yurko, R., Burckhardt, P., & Kovacs, F. (2019, May). "Many students, one dataset": Using ISLE to teach reproducibility and the impact of data analysis decisions on conclusions [Breakout session]. *U.S. Conference on Teaching Statistics (USCOTS), 2019*, State College, PA, United States. https://www.causeweb.org/cause/uscots/uscots19/breakout/1D

Newman, J., & Cox, C. (2020). Corpus annotation. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 25–48). Springer.

Perales-Escudero, M. D. (2011). To split or to not split: The split infinitive past and present. *Journal of English Linguistics*, *39*(4), 313–334.

Pilán, I., Lison, P., Øvrelid, L., Papadopoulou, A., Sánchez, D., & Batet, M. (2022). The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, *48*(4), 1053–1101.

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, *122*(8), e2422455122. https://doi.org/10.1073/pnas.2422455122

Reppen, R. (2010). *Using corpora in the language classroom* (Vol. 6). Cambridge University Press.

Reppen, R. (2024). Exploring individual longitudinal development in a corpus of 'natural' disciplinary writing: What could it mean for teaching? *International Review of Applied Linguistics in Language Teaching*, *62*(1), 61–78.

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research, 2*(2), 99–127. https://doi.org/10.17239/jowr-2010.02.02.2

Rudniy, A. (2018). De-identification of laboratory reports in STEM. *The Journal of Writing Analytics, 2*, 176–202.

Shani, C., Jurafsky, D., LeCun, Y., & Shwartz-Ziv, R. (2025). From tokens to thoughts: How LLMs and humans trade compression for meaning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2505.17117

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Slagle, T. (2025). Developing writers' engagement in academic genres: Insights from linguistically informed instruction. *Journal of Writing Analytics*, *8*, 90–132. https://doi.org/10.37514/JWA-J.2025.8.1.03

Staples, S., Conrad, N., Dang, A., & Wang, H. (2024). Building language and genre awareness through learner corpus data in a second language writing course. In S. Götz & S. Granger (Eds.), *Learner corpus research for pedagogical purposes* (Vol. 10, pp. 146–182). John Benjamins. https://doi.org/10.1075/ijlcr.00043.sta

Sterling, S., & De Costa, P. (2018). Ethical applied linguistics research. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 163–182). Palgrave Macmillan.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M., Ahmad, U. K., Chang, Y. Y., Chavez, D., Dressen, D. F., & Seymour, R. (1998). Consider this: The role of imperatives in scholarly writing. *Applied linguistics*, *19*(1), 97–121.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills.* University of Michigan Press.

Taguchi, N., Kaufer, D., Gomez-Laich, P., & Zhao, H. (2017). A corpus linguistics analysis of on-line peer commentary. *Pragmatics and Language Learning, 14*, 357–370.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1–13. https://doi.org/10.1075/ijcl.10.1.01teu

Toulmin, S. (2003). *The uses of argument* (Updated ed.). Cambridge University Press.

Upton, T. A., & Cohen, M. A. (2009). An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies*, *11*(5), 585–605.

Vine, B. (2020). *Introducing language in the workplace*. Cambridge University Press.

Wallis, S. (2021). *Statistics in corpus linguistics.* Routledge.

Wolfe, J. (2009). How technical communication textbooks fail engineering students. *Technical Communication Quarterly, 18*(4), 351–375. https://doi.org/10.1080/10572250903149662

Wulff, S., Römer, U., & Swales, J. (2012). Attended/unattended this in academic student writing: Quantitative and qualitative perspectives. *Corpus Linguistics & Linguistic Theory*, *8*(1), 129−157. https://doi.org/10.1515/cllt-2012-0006

Xiao, R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35.

Zhang, X., Mao, R., & Cambria, E. (2022). A survey on syntactic processing techniques. *Artificial Intelligence Review, 56*(6), 5645–5728. https://doi.org/10.1007/s10462-022-10300-7

Zhao, H., & Kaufer, D. (2013). DocuScope for genre analysis: Potential for assessing pragmatic functions. In N. Taguchi & J. M. Sykes (Eds.), *Technology in Interlanguage Pragmatics Research and Teaching* (pp. 235–259). John Benjamins. https://doi.org/10.1075/lllt.36.12zha