Research Article

Tracking Citation Practices Across Disciplines: A SourceMapper Analysis of MICUSP

Megan Kane, Seton Hall University

Scopus Abstract

This study describes the development and validation of three machine-classification models for analyzing citation rhetorics in student writing at scale. Using a sample of Biology and English papers from the Michigan Corpus of Upper-Level Student Papers (MICUSP) as a test case, I compare a TF–IDF classification model with two embeddings-based models (SciBERT and Qwen3) in identifying three rhetorical actions students take when citing sources—Reporting, Transforming, and Evaluating. Model evaluation showed that the SciBERT classifier performed most reliably, achieving 86.80% agreement with human coding for rhetorical function identification, with the TF–IDF and Qwen3 models also performing with moderate levels of agreement (69.90%– and 73.20% agreement, respectively). To demonstrate the explanatory power of this approach to computational-rhetorical analysis, I applied the strongest classifier across all MICUSP English and Biology papers. I supplemented this automated tagging with manual identification of whether each citation involved a primary or secondary source. The findings reflect clear and discipline-consistent contrasts: English papers cited more primary sources, were substantially more citation-dense, and relied more heavily on Transforming moves (paraphrasing, synthesis, recontextualization), while Biology papers cited more secondary sources, contained fewer citations overall, and more often used the Reporting function. Together, these results demonstrate that the computational-rhetorical approach to citation function analysis produces patterns consistent with established disciplinary writing research, offering both a scalable model for citation-function analysis and a practical tool for instruction. I conclude by outlining how this classification approach, implemented in a user-friendly interface I call SourceMapper, can help instructors and students visualize citation practices and support writing pedagogy across the disciplines.

Structured Abstract

- **Background**: The purpose of study is to develop and validate machine classification models for examining students' citation practices at scale that remains sensitive to rhetorical nuance. Using 67 Biology and 98 English papers from the Michigan Corpus of Upper-Level Student Papers (MICUSP) as a test case, the project evaluates three machine-classifiers—a TF–IDF model and two embeddings-based models (SciBERT and Qwen3)—to determine how effectively they identify three rhetorical citation functions: Reporting, Transforming, and Evaluating. After assessing model performance, I apply the most accurate classifier to explore how citation practices vary across the two disciplines and how they shift when students engage primary versus secondary sources. These findings illustrate the kinds of insights such classification methods can reveal about how writers position themselves in disciplinary conversations and how their citation behaviors reflect broader norms of knowledge-making. They also highlight the pedagogical potential of the machine-classification approach as implemented in the user-friendly SourceMapper tool, which can help instructors make citation practices visible and support more rhetorically aware source use.
- **Research Questions**:
    1. RQ1. How well do three different machine-learning approaches—TF-IDF features, SciBERT embeddings, and Qwen3 embeddings—perform in identifying rhetorical citation functions across a multidisciplinary student-writing corpus, and which approach offers the most reliable and valid method for large-scale analysis?
    2. RQ2. How do the rhetorics of citation differ between Biology and English student writing within the MICUSP corpus?
        - In what contexts and with what frequency do students employ reporting, transforming, and evaluating functions when citing sources in Biology versus English?
        - To what extent do Biology and English disciplines privilege certain rhetorical citation functions over others?
    3. RQ3: How do differences in engagement with primary versus secondary sources influence the rhetorical citation functions used by students within these disciplines?
- **Methodology**: The study analyzed a 165-paper MICUSP subcorpus (67 Biology, 98 English) comprising 15,474 sentences. Sentences containing citations were first identified using a Python-based "theory-to-query" approach. A representative subset was then hand-coded for four categories—Reporting, Transforming, Evaluating, and No Citation—using a previously validated codebook. To automate large-scale tagging, three machine classifiers were trained on a combined dataset of 12,148 manually coded sentences drawn from MICUSP and an additional first-year writing corpus. The trained machine classifiers were (1) a TF–IDF ensemble baseline, (2) a SciBERT model fine-tuned for rhetorical classification, and (3) a Qwen3-Embedding-8B model using hierarchical embeddings. Model performance was evaluated using accuracy, precision, recall, and F1 metrics. Close readings of representative passages contextualized the computational results and validated the automated classifications. The highest-performing model (SciBERT) was then applied to the full MICUSP sample to generate large-scale rhetorical

function tags; this tagging was supplemented by manual coding of each sentence containing a citation, notating whether the source cited was a primary or secondary source. Chi-square and two-proportion z-tests were used to assess differences in citation functions across disciplines and source types.

- **Results**: Among the three machine-learning approaches, SciBERT achieved the strongest overall performance (86.80% accuracy) confirming the value of domain-specific embeddings for rhetorical function classification. Applying the SciBERT classifier to the full MICUSP subcorpus revealed significant disciplinary contrasts. English papers were far more citation-dense (87% of sentences) than Biology papers (30%). English students predominantly used Transforming functions (≈62%), followed by Reporting (≈22%) and rare Evaluating functions (≈4%). Biology papers contained a majority of No Citation sentences (≈70%), with far lower frequencies of Reporting (≈16%), Transforming (≈12%), and Evaluating (≈2%). Source type further shaped these patterns: primary sources prompted more Transforming and Evaluating, while secondary sources were linked mainly to Reporting. English students transformed and evaluated primary texts at high rates, reflecting humanities conventions of synthesis and critique. Biology students relied chiefly on secondary sources, where concise reporting dominated and evaluation was minimal. Most disciplinary and source-type differences were statistically significant, underscoring how machine classification can reveal the way disciplinary norms govern students' rhetorical engagement with sources.
- **Conclusions**: This study demonstrates that machine classification is a viable method for examining citation functions at scale. The strong performance of the SciBERT classifier provides evidence that rhetorical citation functions in student writing can be identified accurately and reliably through automated methods. The findings also show, through the MICUSP use case, how machine classification can illuminate disciplinary differences in students' citation practices. English writers more often transformed—and occasionally evaluated—sources, positioning themselves as interpreters within ongoing conversations. Biology writers, by contrast, relied heavily on reporting or writing without citation, reflecting norms that emphasize concise presentation of prior research. Pedagogically, integrating this machine-classification approach into the SourceMapper tool shows promise for helping instructors and students visualize citation patterns, support revision, and strengthen cross-disciplinary writing instruction. Future work should expand the training data to more diverse student populations and test SourceMapper in classroom settings to assess its impact on citation awareness and formative assessment.

Keywords: citation, first-year writing, machine learning, writing across the curriculum, writing analytics

1.0 Background

Why do students cite, and how can we study their citation practices at scale in a way that remains valid, accessible, and rhetorically grounded? This study takes up those questions by developing and testing machine classification models for analyzing the rhetorical functions students use when engaging with sources. To do so, I examined the Michigan Corpus of Upper-Level Student Papers (MICUSP) (O'Donnell & Römer, 2012; Römer & O'Donnell, 2011), a collection of high-scoring undergraduate and graduate writing from the University of Michigan. I constructed a subcorpus of 165 papers from MICUSP—67 Biology papers and 98 English papers—and used three machine classification approaches (trained on TF-IDF features, SciBERT embeddings, and Qwen3 embeddings) to identify the citation rhetorics in these texts. Here, "citation rhetorics" refers to the functional actions students perform when citing sources, such as 1) reporting information from sources, 2) transforming or synthesizing ideas from sources, and 3) evaluating a source's merit, strength, or effectiveness. While the study compares citation patterns across English and Biology and examines how students cite primary versus secondary sources, these disciplinary findings serve primarily as a showcase for the methodological finding: the automated tagging pipeline produces results consistent with well-established understandings of disciplinary writing, suggesting that computational classification of citation functions can be both scalable and pedagogically relevant. In this way, the study contributes both a computational-rhetorical model for large-scale citation analysis and a concrete demonstration of how such machine classification methods, when incorporated into a user-friendly tool I call SourceMapper, can help instructors better understand how students position themselves in disciplinary conversations.

Much existing cross-disciplinary citation research relies on detailed manual coding or small, locally developed schemes that are difficult to scale and seldom generalize across contexts (Jamieson, 2013; Petrić, 2007; Lee, Hitchcock, & Casal, 2018; Kastman Breuch & Larson, 2017). While such work has generated important insights, it remains limited in its ability to capture citation practices across large and diverse sets of student texts, leaving open the question of how citation functions might be examined systematically and at scale. To address this gap, the present study advances a computational approach for tagging rhetorical citation functions. As part of its development, three automated classification methods were tested to determine which most reliably identified citation functions in student writing.

Once validated, this methodological framework was applied to a disciplinary test case, building on longstanding research showing that citation practices vary across fields and reflect deeper differences in how knowledge is constructed (Hyland, 2004, 2011). Using MICUSP—a corpus frequently used to investigate disciplinary variation in linguistic and rhetorical features (Hardy & Römer, 2013; Hardy & Friginal, 2016; Wang, 2022)—the study examines sentence-level citation functions, or "the intentions writers realize by using citations" (Petrić, 2007), and how these functions shift across disciplines and manually coded source types. The results not only align with known disciplinary writing patterns but also demonstrate the explanatory potential of computational tagging for analyzing citation rhetorics at scale. These findings suggest concrete WAC/WID applications; making citation functions visible through a platform like SourceMapper, through which the machine classification system can be integrated "under the hood" for user-friendly citation tagging, can help instructors better articulate disciplinary expectations and support students in transferring rhetorical strategies across writing contexts.

The article explores the following questions:

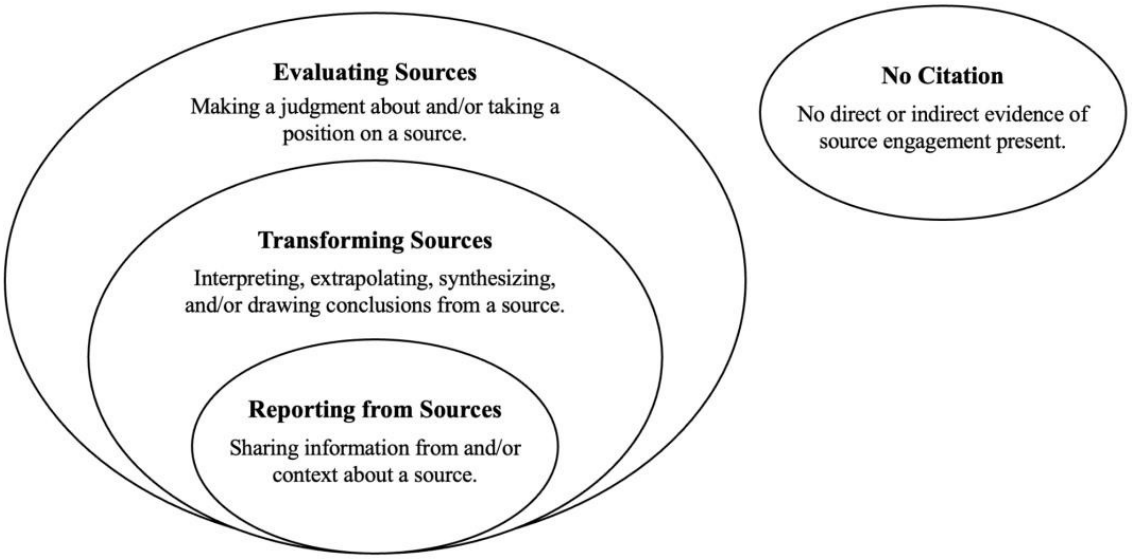
1. RQ1. How well do three different machine-learning approaches—TF-IDF features, SciBERT embeddings, and Qwen3 embeddings—perform in identifying rhetorical citation functions across a multidisciplinary student-writing corpus, and which approach offers the most reliable and valid method for large-scale analysis?

2. RQ2. How do the rhetorics of citation differ between Biology and English student writing within the MICUSP corpus?

- In what contexts and with what frequency do students employ reporting, transforming, and evaluating functions when citing sources in Biology versus English?
- To what extent do Biology and English disciplines privilege certain rhetorical citation functions over others?

Figure 1

Rhetorics of Citation Construct



As the figure shows, the citation rhetoric codebook which I adopt in this study (Kane, 2024) includes three types of citation functions—Reporting, Transforming, and Evaluating. These functions operate hierarchically, increasing in complexity from Reporting as the most basic form of engagement to Evaluating as the most advanced. Reporting involves conveying information from a source with little modification, often by summarizing or restating. Transforming reflects a more complex function, where writers interpret, synthesize, or extrapolate from source material in support of their own argument.

Evaluating represents the highest level of rhetorical engagement, requiring writers to make judgments about the credibility, significance, or limitations of a source. In addition to these rhetorical functions, I also tracked sentences without citation. Any sentence lacking a direct citation in MLA or APA style, or an indirect reference to source material (e.g., "The author suggests…" or "This idea leads to…"), was coded as No Citation—essentially, instances where the writer is articulating their own claims or analysis.

Although not a central focus of this study, the No Citation category supports proportionality analysis (i.e., the extent to which citations comprise the overall paper) and provides a comparative corpus for future investigation.

3. RQ3: How do differences in engagement with primary versus secondary sources influence the rhetorical citation functions used by students within these disciplines?

In addition to automating coding of rhetorical citation functions as prevalent in English and Biology papers, I manually coded which citations referenced primary versus secondary sources in the MICUSP subcorpus. This distinction clarifies not only why sources are being cited but also what kinds of sources students cite in each way—for example, primary sources such as empirical research articles in the sciences or literary texts in the humanities, and secondary sources such as review articles, empirical syntheses, scholarly criticism, or interpretive theoretical texts. Making this distinction allows the later analysis to consider how much variation in rhetorical citation functions is attributable to source type itself and to further discuss what broader disciplinary conventions might determine which sources students are expected to draw on.

2.0 Literature Review

2.1 Approaches to Citation Classification

John Swales (1986) was among the first discourse analysts to argue for a systematic citation categorization scheme that attended not only to the form of citations, but also to their context, quality, and rhetorical weight. He proposed this work as a necessary bridge between two strands of citation research: studies in the social sciences, which focused primarily on typologies of form and content, and studies in applied linguistics, which emphasized syntactic patterns and rhetorical effects but remained, in his words, "zealously non-interpretive" in their classifications (p. 44). To that end, Swales (1986) introduced three dimensions for categorizing citations: whether a citation is short (one sentence or less) or extensive (more than one sentence); whether a citation is evolutionary (used for the new paper), juxtapositional (presented as an alternative viewpoint), or neither; and whether a citation is confirmative (affirmed as containing accurate information), negational (subject to critique by the writer) or neither (Swales, 1986). In tracing how a single source was cited across a decade of publications, he demonstrated how citation analysis could illuminate how knowledge is taken up, evaluated, and integrated into a disciplinary community. He further positions the work of citation analysis in the context of the rhetorical efforts made in academic research papers, proposing that article introductions are a particular site of citation as readers seek to give background about a field of study before making their own contribution (1986). Swales' discussion of the rhetorical "moves" made in academic writing and to what extent they involve source engagement was further crystallized a few years later (Swales, 1990).Taken together, Swales's (1986, 1990) contributions ground the study of citation as a matter of context, rhetorical purpose, and disciplinary form. He articulates several reasons writers invoke other authors: to provide background, to frame a critique, and to support or extend new claims. The rhetorical classification used in the present study draws heavily on this tradition. Although I adopt a smaller unit of analysis—the sentence-level "function" (Petrić, 2007) rather than multi-sentence moves—the underlying understanding of rhetorical purpose is shaped by Swales's work.

Swales' (1986) citation classification system is only one of many that have been developed and taken up over the years. Student citation, in particular, has been a sustained focus of study: a major literature review identified 69 empirical investigations of how students cited sources through 2016 (Cumming, Lai, & Cho, 2016), and scholarship has continued to grow over the past eight years (Lee, Hitchcock, & Casal, 2018; Scheidt & Middleton, 2021; Sun & Soden, 2022; Doolan, 2023). Much of this research highlights the extent to which citation is context-dependent—shaped by discipline, genre, assignment, and local expectations—and, as a result, a wide range of coding schemes has been proposed to capture the complex rhetorical facets of citation. Below, I will discuss several influential typologies that focus specifically on rhetorical classification. These models form the basis for the current study and inform the classification system I have developed for this study.

Bojana Petrić's (2007) study was one of the first to examine student citation through the lens of rhetorical function. Acknowledging that many typologies of citation had been proposed from Swales (1986) onwards, Petrić (2007) made the case that most are meant to study published academic writing, rather than attuned to the goals, audiences, and genres of student citers. Building on the work of Thompson and Tribble (2001), who proposed a multilevel typology to examine the functions of non-integral and integral citations in agriculture dissertations, Petrić (2007) proposes a typology of nine rhetorical functions students utilize in academic writing. Here, Petrić (2007) defines "rhetorical function" as "the intentions writers realise by using citations" (pg. 241). Her categories of citation function, which were developed through qualitative coding, include 1) Attribution, 2) Exemplification, 3) Further reference, 4) Statement of use, 5) Application, 6) Evaluation, 7) Establishing Links between sources, 8) Comparison of writer's findings and sources', and 9) Other. She employs these categories in studying masters' theses in gender studies, noting how usage of different citation functions corresponds to different sections of the theses and correlates to paper scores. Interestingly, most citations fell into a single category, Attribution, in which students simply credited information to a source without making additional interpretive or analytical moves. This pattern was especially pronounced in the lowest-scoring papers. Petrić (2007) argues that instructors should work to make such forms and functions more explicit in the classroom, whether by showcasing examples of different rhetorical citation functions or by designing activities in which students practice crafting citations for varied purposes.

While emphasizing the value of her study—particularly for improving pedagogical support—Petrić cautions against broadly applying her coding scheme to other genres or learner groups, given that it was developed for a specific population writing within a narrowly defined genre. The scheme also lacks interrater reliability in the sense articulated by Geisler and Swarts (2019), as it was created and validated by a single coder. Enthusiasm for Petrić's (2007) approach to rhetorical citation analysis, as well as persistent challenges in adapting such schemes, is evident in more recent work. For instance, Lee, Hitchcock, and Casal (2018) adapted about half of Petrić's categories to analyze the citing practices of L2 first-year research writers. Like Petrić (2007), Lee, Hitchcock, and Casal (2018) coded most student citations (87.43 percent) as "attribution." The next most frequent type of citation used was evaluation (6.69 percent), followed by exemplification (4.58 percent) and an extremely low count of links between sources (1.23 percent). Likewise, Yan and Ma's (2024) development of a four-macro-function, eleven-micro-function framework reflects ongoing attempts to refine and recalibrate Petrić's (2007) categories, so they better accommodate the diversity and complexity of contemporary academic writing contexts.

Their system organizes citations into four overarching rhetorical functions—**A**ttribution, Links Between Sources, Evaluation, and Supporting—each of which is subdivided into more granular micro-functions (e.g., synthesizing, contrasting, positive evaluation, negative evaluation, application). While Yan and Ma's (2024) framework offers rhetorical nuance and clear pedagogical value—supported by high interrater agreement (88.3%)—its locally developed, many-category structure makes it difficult to scale beyond its original context. Such research yields important insight into what students do with rhetorical functions, but it also underscores a persistent methodological limitation: functional categories are difficult to translate across contexts, and few attempts have succeeded in producing broadly applicable coding schemes. This challenge extends well beyond Petrić's study. For example, Kastman Breuch and Larson (2017), drawing on Swales' (1990) rhetorical move framework, observed that students most often used citations for a single purpose—supporting a topic or idea—rather than introducing new lines of inquiry or establishing comparative frames. Their adaptation likewise failed to map neatly onto student writing, forcing them to exclude half of their initial corpus from analysis. In a similar vein, Scheidt and Middleton (2021), in their First-Year Writing program study The Upward Project, developed a five-category "source engagement" scheme—Inform, Explain, React, Develop, and Connect—to evaluate synthesis papers.

They found that students who drew more heavily on the Explain, Develop, and Connect categories tended to earn higher scores. Yet they, too, cautioned against replication, describing their categories as "stabilized-for-now" and subject to change as curricular contexts evolved. They also reported only moderate interrater reliability (68%), further illustrating how difficult it is to create a reliable coding scheme for student citation practices. While local coding frameworks hold tremendous pedagogical value, I also propose the need for a reliable, dynamic framework developed from a multi-disciplinary corpus: one that can be widely employed and refined through use of computational tools. This is the approach the present study seeks to employ and test using a multidisciplinary subcorpus from MICUSP. Thus, a brief look at prior cross-disciplinary citation research is essential for understanding the disciplinary conventions that shape citation practices and for clarifying what large-scale analysis of citation rhetorics must account for and potentially enhance.

2.2 Citation Across the Disciplines

A substantial body of research demonstrates that citation is not merely a matter of style or convention, but also a practice deeply ingrained in disciplinary communities. Ken Hyland's Disciplinary Discourses (2004) documents how scholars construct "manifest intertextuality" through parenthetical citations, author names, reporting structures, and other syntactic signals. Although citation forms remain relatively stable due to style guidelines, Hyland argues that writers' rhetorical motives and disciplinary contexts "complicate this picture considerably" (p. 22). His framework differentiates among integral and non-integral citations, among citations accompanying quotation, summary, and synthesis, and across the range of reporting verbs used to present others' work. Across a comparative corpus of research articles, Hyland (2004) found clear disciplinary contrasts. Humanities writers used citations more frequently and favored integral structures that foregrounded authors as participants in ongoing interpretive debates.

Conversely, writers in the hard sciences tended to rely on non-integral citations with more neutral reporting verbs, signaling a positivist stance in which sources function primarily as repositories of information rather than interlocutors. These differences, Hyland argues, reflect underlying epistemological orientations: where humanities knowledge is seen as negotiated and dialogic, scientific knowledge is often presented as stable and cumulative.

Other scholars have extended this disciplinary view. Chang (2008) showed how citation practices differ not only across fields but also across genres within fields, demonstrating that applied linguistics research articles contain more hidden references and clustered citations than practitioner-oriented design articles. Hyland and Jiang (2017) examined leading journals in four disciplines to test the widespread claim that academic writing has become increasingly informal. Analyzing a 2.2-million-word corpus sampled over three time periods, they found only a modest rise in informal features—and primarily in the hard sciences, not the social sciences. Their results suggest that academic discourse continues to be shaped by disciplinary norms and genre expectations, with changes emerging unevenly across fields.

Students, too, internalize disciplinary differences in their writing—though not always consistently or effectively. Research indicates that expert writers overwhelmingly favor summary and synthesis over direct quotation, a pattern that Swales (2014) confirmed in student Biology papers, where only 11 of 327 citations were tied to quotations. Both Swales (2014) and Hyland (2004) interpret such choices as strategic: writers arrange, summarize, and synthesize sources in ways that position their claims within an evolving scholarly landscape. When looking beyond quotation practices to broader citation forms, similar disciplinary patterns emerge. Using data from the British Academic Written English (BAWE) corpus, Hilary Nesi shows that non-integral citations dominate across disciplines, echoing trends identified by Samraj (2008) and Swales (2014). Integral citation becomes more common at higher academic levels, except in the physical sciences, where the "name (date) verb" pattern remains rare. Importantly, Nesi argues that these formal patterns are

metadiscoursal: students use reporting verbs, stance markers, and citation placement to guide readers' interpretations, thereby linking citation form to rhetorical function.

This functional lens is elaborated further in corpus-assisted studies of novice researchers. Finally, Wette (2017) analyzes 27 Arts and Social Sciences assignments and finds that the majority of student citations served only to attribute information to sources, reflecting a view of sources primarily as repositories of content rather than as participants in a disciplinary conversation. This stands in stark contrast to the wider range of rhetorical citation purposes observed in published scholarship, where writers position themselves as interlocutors within the field. Wette concludes that explicit attention to disciplinary conventions of citation should be a greater focus of instruction.

Taken together, these studies underscore both the value and the complexity of examining citation across the disciplines at scale and in student writing. Hyland's (2004) work demonstrates how citation choices index disciplinary epistemologies, while research on student writing (e.g., Nesi, 2013; Samraj, 2008; Swales, 2014) illustrates how novices navigate these expectations with uneven rhetorical awareness and limited exposure to effective models. Further contributions from Chang (2008) and Hyland and Jiang (2017) reveal how citation practices vary not only across fields but also across genres and historical periods, underscoring their dynamic and situated nature. More recent pedagogically oriented research, such as Wette (2017), shows that explicit scaffolding is often necessary for students to develop the more complex evaluative and synthesis-oriented citation functions used by experts. The present study builds on this body of scholarship by proposing methodological interventions that can support both students' understanding of their citation practices and instructors' efforts to teach them. To do so, the study analyzes a student-specific corpus that spans two rhetorically distinct disciplines, using this comparison as a representative test case.

For its corpus of multidisciplinary student writing, the present study uses the Michigan Corpus of Upper-Level Student Papers (MICUSP), a well-established corpus for analyzing disciplinary writing across genres and fields (O'Donnell & Römer, 2012; Römer & O'Donnell, 2011). Previous MICUSP research has included Römer and Wulff's (2010) analysis of function words and demonstratives, Hardy and Friginal's (2016) exploration of genre variation, Wang's (2022) study of hedging, Aull, Bandarage, and Miller's (2017) investigation of generalization in student versus expert writing, and Hardy, Römer, and Roberson's (2015) work on using MICUSP pedagogically to expose students to discipline-specific citation practices. Additional studies have examined phrase-frames (Walcott, 2021), and grammatical patterns (Kim, 2018; Larsson, 2018). Collectively, these studies highlight MICUSP's robustness as a resource for linguistic analysis, offering detailed insights into students' use of fine-grained textual features across genres and disciplines. Its limitations—particularly its restricted linguistic diversity beyond one institutional context and its age, given that it was compiled from student writing in the early 2000s, prior to major shifts such as the COVID-19 pandemic and the rise of AI—should be acknowledged and are discussed further below. Even so, MICUSP remains a valuable site for exploration, especially when paired with frameworks that illuminate disciplinary differences in rhetorical function.

2.3 Leveraging Computational Tools to Advance Citation Analysis

Prior research indicates how computational citation analysis is of value to this study. Knight et al. (2020) conduct a survey of computational tools which have been leveraged to annotate rhetorical markers, separating these into rule-based and machine learning approaches. Rule-based approaches generate annotations based on pre-set constructs of writing. For example, the AcaWriter (Knight et al, 2020) uses a rule-based parser that identifies text features that communicate rhetorical intentions, such as summarizing sources, describing background, and claiming centrality. The AcaWriter is framed as a formative feedback tool, enabling students to submit their drafts and receive automatically generated reports through its user-friendly web interface. In contrast, machine-learning classifiers, like Cotos, Huffman, and Link's (2020) Research Writing Tutor (RWT), has been trained on a corpus of 900 manually annotated research papers

from 30 disciplines, from which it makes predictions about the rhetorical moves present within a given text. The RWT tool, too, has been leveraged in the service of providing feedback for advanced academic writers, including a color-coded display of move types and a concordance that allows students to search for example moves, as well as microlevel feedback to help clarify purpose and focus of individual sentences. Though these computational move classifier projects have typically been taken up in individual contexts, without much overlap between researchers or corpora, they yield valuable insights within their local contexts and, I suggest, provide an avenue for researchers to better understand broad patterns of rhetorical actions across the corpus and for students to gain insights into their own rhetorical practices.

This article involves rhetorical function analysis conducted through machine classification models I developed and trained on a coding scheme denoting common citation functions in student writing. The development of this identification, validation, and computational training pipeline follows the methodological model established by Omizo et al. (2019). Their scheme, grounded in rhetorical genre theory and Swalesian move theory (1990), identified three comprehensive functions made by online users in learning forums: Inviting, Staging, and Evoking. After achieving acceptable levels of interrater reliability within their research team, Omizo et al. (2019) applied the scheme to training and testing datasets for use with a machine classifier, which reached an accuracy rate of .70—sufficient for use as a tagging mechanism. They subsequently developed an interface, the Faciloscope, which allowed users to input texts for machine tagging and analysis, and demonstrated its application through a sample webpage.

In employing a machine classification approach in this study, and in building the resulting user-friendly tagging interface, SourceMapper, which I describe in the Discussion section, I adopted a similar methodological pipeline to Omizo et al (2019). This research also draws from the work of my dissertation (Kane, 2024), where I first constructed and validated a coding scheme for citation functions. I established the validity of the coding scheme in that earlier project through iterative testing and interrater reliability, reaching simple agreement of 0.85 and Cohen's Kappa of 0.77. The manual coding provided a stable foundation for computational tagging. For the present study, I retrained models on a new corpus of student writing and experimented with different embedding approaches to improve tagging accuracy and scalability. This process mirrors the sequence established by Omizo et al. (2019)—human coding, validation, computational training—while extending it to the domain of citation rhetorics in student writing. Ultimately, this project functions both as a computational model for analyzing rhetorical citation functions in student writing and as a demonstration of how large-scale citation analysis can be meaningfully applied to disciplinary contexts. Using MICUSP as a showcase, the study affirms the value of examining citation at scale and offers insights into how Biology and English students engage in disciplinary citation practices, including how those practices vary by source type.

3.0 Research Methodology

3.1 Research Design and Procedures

This study analyzes a subcorpus of 165 papers drawn from the Michigan Corpus of Upper-Level Student Papers (MICUSP), consisting of 67 Biology papers and 98 English papers. Because MICUSP contains uneven numbers of papers across disciplines, the resulting subcorpus is likewise imbalanced. I controlled for this imbalance in my analysis by calculating relative frequency values on a per-paper basis, as detailed in the Results section. The Michigan Corpus of Upper-Level Student Papers (MICUSP) was selected because it offers a substantial, publicly accessible collection of high-scoring upper-level student writing across a wide range of disciplines. According to the Fair Use Statement on the MICUSP home page, the files are freely available for research and teaching purposes, and no additional permissions were required to access or analyze this corpus. All papers were segmented into sentences

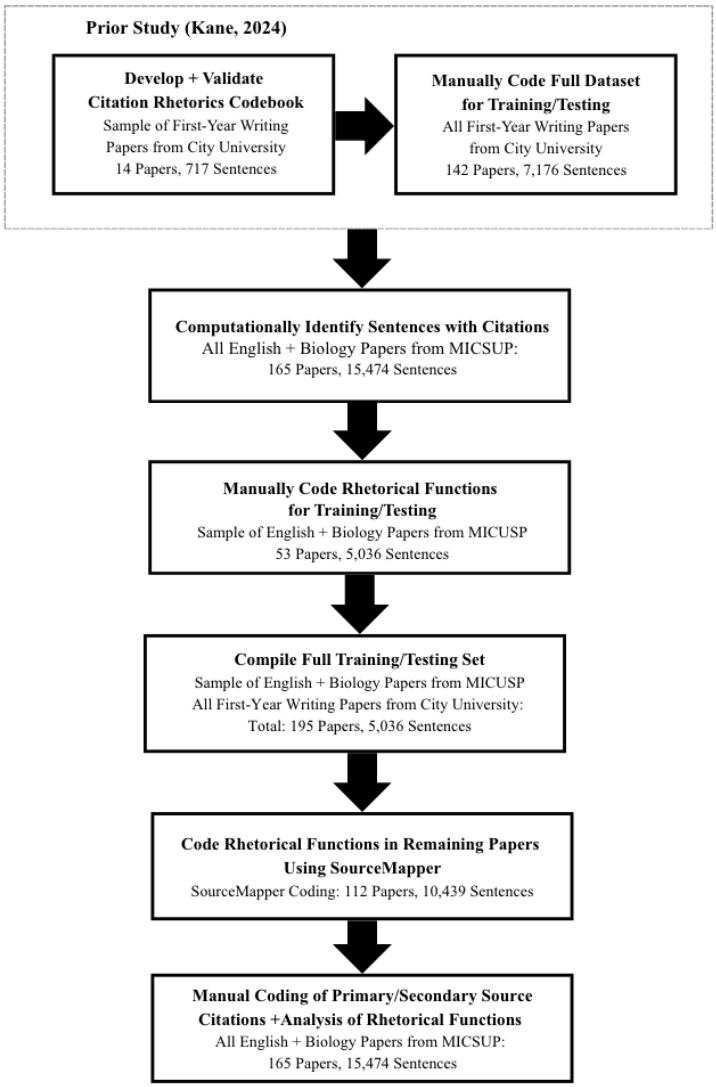
(15,474 in total), and those containing citations were identified using a computationally assisted "theory-to-query" approach (Itchuaqiyaq, Ranade, & Walton, 2021). This framework involves four steps: defining project parameters, gathering initial data, creating a custom computer program, and testing program outputs. I also followed Itchuaqiyaq, Ranade, and Walton (2021) in using Python to write code for identifying citation types. Python is an open-source coding platform which has been established as a reliable tool to support natural language processing in industry and the academy. My own familiarity with the tool also makes it an ideal platform, and a result of this project is a set of Python notebooks which any researcher can use to replicate my coding processes with their own corpora.

Once all sentences containing citations were identified, I selected a representative subset (30 percent) of the MICUSP subcorpus for manual coding. These manually coded sentences were then used to refine and strengthen SourceMapper, the machine classification tool I have been developing. Figure 2 illustrates the paper intake and analysis workflow. In using a manually coded subset to iteratively train and adjust the model prior to full deployment, I follow a validation process similar to those outlined by Omizo et al. (2019).

The rhetorical coding scheme used in this manual phase was originally developed as part of my IRB-approved analysis of 48 first-year writing portfolios (142 papers; 7,176 sentences) from my university's writing program (Kane, 2024). The framework draws heavily on the rhetorical function coding schemes proposed by Petrić (2007) and Scheidt and Middleton (2021), adopting the same unit of analysis they employ—the sentence—and adapting their more extensive taxonomies to the specific needs of the first-year writing corpus I analyzed. In that earlier study, I collapsed their broader sets of functions into three core categories that emerged as most central to students' citation practices: Reporting, Transforming, and Evaluating (see Figure 1). Sentences that were not identified as containing citation, using the "theory-to-query" approach (Itchuaqiyaq, Ranade, & Walton, 2021), were automatically labeled as a fourth functional category, "No Citation."

Figure 2

Data Intake and Processing Pipeline



To validate the coding scheme as part of my dissertation research, I recruited the help of two coders who worked within my university's first-year writing program. Following Geisler and Swarts (2019), I employed a multi-part validation process in which I first worked with one coder to establish acceptable levels of agreement, then affirmed validation of the coding scheme with a second coder. I used samples of approximately 10% of the data (14 papers, 717 sentences) representative of the range of genres and scores in the dataset, to validate my codes with the first data. This involved four rounds of coding, through which my coding scheme underwent several adjustments, and culminated in reaching simple agreement of 0.868 and Cohen's Kappa of 0.796 (Geisler and Swarts, 2019). Coding an additional 10 percent of the data with a second coder resulted in a simple agreement of 0.85 and Cohen's Kappa of 0.77. These rates fell into acceptable ranges of agreement (Geisler and Swarts, 2019) and thus affirmed the reliability of my coding scheme.

I applied the same validated coding scheme—containing three categories: Reporting, Transforming, and Evaluating—to manually classify the citation function of each citation-present sentence in a 30 percent subset of the 165 MICUSP papers analyzed in this study (31 English papers, 22 Biology papers; 5,036 total sentences). I conducted a self-validation of this coding, following procedures similar to Petrić (2007), and achieved an accuracy rate of approximately 92 percent. Once the subset was coded, these sentences were incorporated into the machine classification pipeline as training data.

They were combined with the training data generated in my dissertation study, which consisted of a fully hand-coded dataset of 142 papers (7,176 total sentences: 4,668 containing citation and 2,508 without). The full training set for SourceMapper therefore comprised both the dissertation corpus and the MICUSP subset (5,036 sentences: 3,145 with citations and 1,891 without), resulting in a combined dataset of 12,212 sentences (7,813 with citations and 4,399 without). This combined dataset was used to test three machine classification strategies—an ensemble classifier with TF–IDF vectors, SciBERT, and Qwen3—in order to determine the most effective model for rhetorical citation-function tagging.

Following the evaluation of machine-classification strategies and selection of the best-performing model, I used SourceMapper to tag the remaining papers in the MICUSP English and Biology subcorpora (112 papers, 10,439 sentences). To supplement this analysis, I manually coded whether each sentence containing evidence of citation referenced a primary source (such as an original research study, historical document, or literary text) or a secondary source (such as interpretive or critical scholarship). Although identifying citations by primary or secondary status is a planned feature of the SourceMapper platform, it is currently beyond its capacities. Doing this step manually thus reflects the human–machine augmented nature of the project.

After all sentences in the English and Biology papers were tagged, I conducted relative frequency analyses in Python to examine two dimensions of variation, in response to RQ2 and RQ3:

1. differences across disciplines (Biology vs. English), and
2. differences within disciplines by source type (primary vs. secondary).

Tests of significant difference—including chi-square tests and two-proportion z-tests—were used to assess whether observed variation by discipline and by source type exceeded what would be expected by chance. This statistical framework provided the basis for interpreting patterns in students' citation rhetorics.

3.3 Instruments

Two main instruments were employed in this study. First, standard NLP packages in Python were used for tokenizing and cleaning the data, consistent with the methods outlined in Itchuaqiyaq, Ranade, and Walton (2021) for identifying citations and in Omizo et al. (2019) for preparing data for machine learning classification. Once citation rhetorics had been annotated, Python packages were again used to conduct statistical analyses, including chi-square tests and two-proportion z-tests. Second, machine classification was conducted using Python packages that facilitated both traditional feature-based approaches and embedding-based models. TF–IDF representations were generated programmatically, and pre-trained Qwen and SciBERT models were used to produce sentence-level embeddings, which were then classified using standard supervised learning procedures. Finally, the classifier determined to be most accurate during testing was integrated into SourceMapper, an online application that automates rhetorical citation tagging. The application was developed in Python using Flask as the web framework, with a Bootstrap-based frontend that provides an interface for users to input academic text and receive detailed analysis results. The application employs multiple machine learning approaches for rhetorical function classification, and it is trained using the same general approach as Omizo et al. 's

(2019) Facilioscope but adapted for citation rhetorics. The tool is openly available for review and use on GitHub (Kane, 2025).

3.4 Data Analysis

The analysis proceeded in several stages shown in Figure 3. To investigate RQ1, I examined three approaches to machine classification for use in SourceMapper to tag rhetorical functions. First, an ensemble classifier was trained on TF–IDF vectors, a long-established approach that represents texts by word frequency patterns (Salton & Buckley, 1988) and has been widely applied in text classification tasks (Joachims, 1998). In this study, TF–IDF offers a long-established baseline for text classification against which newer methods can be measured. The ensemble classifier approach is grounded on the recognition that "no single classifier is the best performer for classifying all rhetorical categories of sentences" (Widyantoro et al., 2013); as such, leveraging an ensemble model combines the output of multiple classifiers, weighting their responses and assigning the level generated with the greatest confidence.

Next, a fine-tuned machine classifier was trained on SciBERT embeddings shown in Figure 4, which draw on domain-specific pretraining in large collections of scientific papers, making them well-suited for analyzing disciplinary writing (Beltagy, Lo, & Cohan, 2019). Though SciBERT has demonstrated strong performance in scholarly discourse tasks such as citation intent classification (Cohan, Ammar, van Zuylen, & Cady, 2019) and citation context classification (Maheshwari, Singh, & Varma, 2021), it has rarely been applied to student writing. Its inclusion therefore tests whether a model optimized for professional research articles can also capture the rhetorical functions of citation in novice texts.

Finally, a machine classifier was trained on Qwen3-Embedding-8B embeddings, a recently released model which has been shown to provide context-aware representations with strong performance on semantic tasks (Zhang et al., 2025). Although new to writing studies, Qwen3 offers an opportunity to explore whether the latest generation of embedding models can advance rhetorical function classification.

Figure 3

TF-IDF Vectorization and Classification Workflow

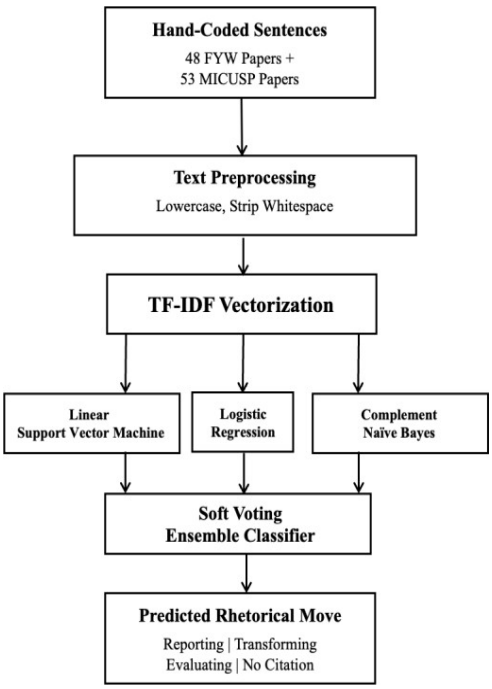


Figure 4

SciBERT Classification Workflow

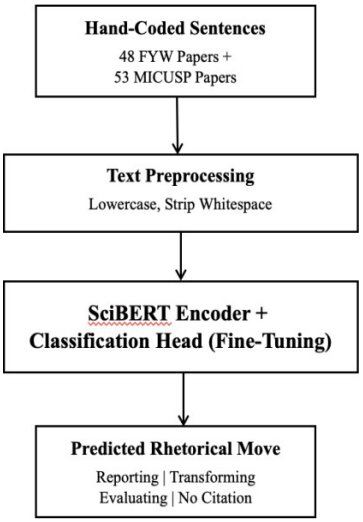
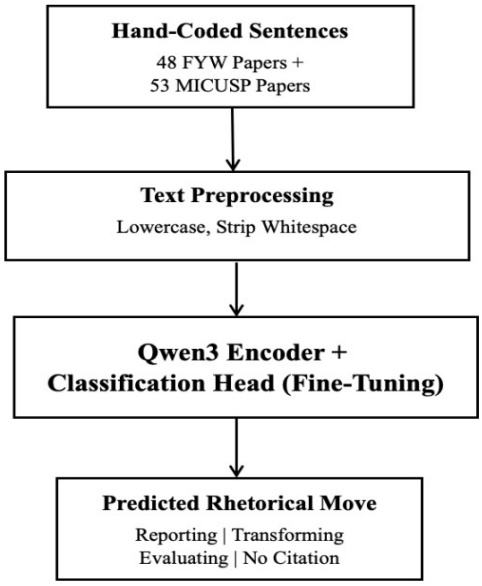


Figure 5

Qwen3 Classification Workflow



As shown in Figure 5, each machine classifier was trained on a stratified sample that combined manually annotated MICUSP papers as well as the set of hand-coded first-year writing papers from my dissertation study (Kane, 2024).

Incorporating both datasets enabled a feature set of 12,148 sentences which had been hand-coded with citation function labels (Reporting, Transforming, Evaluating, and No Citation) and allowed the classifiers to capture rhetorical patterns of citation across levels of student writing. Ultimately, this resulted in stronger performances than when training exclusively on MICUSP papers.

Taken together, these models represent three points of comparison: a traditional baseline (TF–IDF), a proven domain-specific model with demonstrated success in scholarly discourse analysis but limited application to student writing (SciBERT), and a cutting-edge embedding model that extends the frontier of context-aware text representation (Qwen3). This combination allows the study to assess both established and novel pathways for computational analysis of student writing. My dissertation work (Kane, 2024), which posited the value of feature-based models and BERT embeddings in a first-year writing corpus, serves as a benchmark for evaluating SourceMapper's generalizability to upper-level student writing. SourceMapper's reliability was measured as consistency across repeated trials, while validity was framed as the extent to which model outputs captured meaningful rhetorical constructs (Huot, 1990; Kane, 2006).

To address RQ2 and RQ3, the complete MICUSP subcorpus of English and Biology papers were tagged using the most accurate machine classification model, SciBERT, as integrated into the SourceMapper platform. Manual coding of all sentences contained citations was also performed to distinguish which sentences cited primary sources (e.g. empirical research articles, literary texts) versus secondary sources (e.g. critiques, interpretive reviews). Then, relative frequency analyses were conducted for each rhetorical function across the Biology and English subcorpora, and rhetorical function usage was compared between sentences using primary and secondary sources in each discipline and across the full subcorpus. Tests of significant difference (e.g., chi-square tests, two-proportion z-tests) were performed to assess variation across discipline and source type (primary vs. secondary). In sum, these combined methods—statistical, computational, and rhetorical—support a multi-dimensional understanding of how academic writers engage

sources across disciplines, and of how tools like SourceMapper might support instruction and assessment aligned with disciplinary conventions.

4.0 Results

4.1 RQ1: Training and Evaluating Machine Learning Models for Rhetorical Function Classification

To evaluate SourceMapper's effectiveness in classifying rhetorical functions across disciplines, I trained and tested three machine classifiers on hand-coded data, including approximately 30% of the papers from the MICUSP sub-corpus I created from the English and Biology papers. Combined with the sentences originally used to train this classifier, a collection of 48 first-year writing portfolios used in my prior dissertation study (Kane, 2024), the total sample of hand-coded data used to train the machine classifiers consisted of a total of 12,148 sentences. Each sentence in the corpus had been manually labeled with one of four rhetorical categories: Reporting, Transforming, Evaluating, or No Citation (for sentences without citation functions). Each classifier was trained on at least 75% of this sample and tested on the remaining 25%, allowing for comparison of performance metrics such as accuracy, precision, recall (Leacock et al, 2014), and F1-score across different approaches.

The ensemble classifier trained on the TF-IDF feature set achieved an accuracy level of 69.90%. Table 1 below highlights the breakdown in classification of each rhetorical function of citation, capturing precision (proportion of sentences classified into a category that were actually correct), recall (proportion of all true instances of a category that were successfully identified by the model) and the F1 score (mean of precision and recall). As evidenced here, the TF-IDF model's performance was somewhat uneven between the four categories. Reporting sentences were identified with moderate success (precision = .55, recall = .59, F1 = .56), while Transforming functions achieved higher precision (.70) but somewhat lower recall (.58), suggesting the model was effective when it predicted a Transforming sentence but often failed to detect them. Evaluating functions showed a similar pattern (precision = .53, recall = .60, F1 = .56). Other than the accuracy of the No Citation category (precision = .80, recall = .86, F1 = .83), the model was most successful at predicting Transforming sentences—interestingly, one of the most challenging categories for human coders to predict, as I note further down in this section. Overall, the TF-IDF model offers a usable baseline for classroom applications and tool integration, even if more advanced embedding models outperform it in raw accuracy.

Table 1

Results of Ensemble Classifier Trained on TFIDF Features

Category	Precision	Recall	F1-Score
Reporting	0.55	0.59	0.56
Transforming	0.70	0.58	0.64
Evaluating	0.53	0.60	0.56
No Citation	0.80	0.86	0.83

The second classifier, trained using SciBERT, achieved an overall accuracy of 86.80%, a substantially improved performance across all categories compared to TF–IDF. As evidenced by Table 2, Reporting and Transforming functions both achieved strong results (precision = .81, recall = .81, F1 =.81; precision = .82, recall = .82, F1 = .82, respectively), indicating a balanced ability to correctly identify and capture most instances. Evaluating functions also showed robust performance (F1 = .77), though slightly lower than

Reporting and Transforming, reflecting the challenge of modeling evaluative citation functions. The No Citation category achieved the highest accuracy overall (precision = .96, recall = .95, F1 = .95), showing that SciBERT embeddings provided clear separation between cited and uncited sentences. These results demonstrate the value of domain-specific embeddings; SciBERT's pre-training on scientific articles appears particularly effective in capturing the rhetorical patterns of academic writing.

Table 2

Results of Classifier Trained on SciBERT Embeddings

Category	Precision	Recall	F1-Score
Reporting	0.81	0.81	0.81
Transforming	0.82	0.82	0.82
Evaluating	0.76	0.77	0.77
No Citation	0.96	0.95	0.95

Shown in Table 3. the third model was trained using Qwen3-Embedding-8B, a high-capacity transformer designed to generate dense sentence, paragraph, and chunk-level embeddings, which were then combined hierarchically to support rhetorical function classification. As part of the Tongyi Qianwen family of large-scale language models, Qwen3 has performed strongly in recent benchmarks for contextual reasoning and semantic nuance (Bai et al., 2023; Yang et al., 2025). On the MICUSP corpus, however, this model achieved an overall accuracy of 73.2%, with a macro F1-score of 0.662 and a weighted F1-score of 0.730—slightly outperforming the retrained TF-IDF model but falling short of SciBERT's performance. The "No Citation" category saw the highest precision (0.866), recall (0.885), and F1-score (0.876), suggesting that Qwen was particularly effective at identifying non-citation sentences. For rhetorical functions, Qwen performed best on Transforming (F1 = 0.683) and showed moderate success on Reporting (F1 = 0.555), but it struggled with Evaluating, which had the lowest scores across all metrics (F1 = 0.534). Although Qwen's multi-level embeddings offer a training model pulling from robust semantic context, this added complexity comes with trade-offs. The model is computationally expensive, with significantly longer training and inference times than either SciBERT or TF-IDF. Given the relatively modest gains in classification accuracy—especially in the categories most relevant for rhetorical analysis—its practicality for instructor-facing tools like SourceMapper may be limited.

Table 3

Results of Ensemble Classifier Trained on Qwen3 Embeddings

Category	Precision	Recall	F1-Score
Reporting	0.60	0.52	0.56
Transforming	0.67	0.70	0.68
Evaluating	0.54	0.53	0.53
No Citation	0.87	0.89	0.88

After experimenting with multiple machine learning models and embedding strategies, SciBERT-based embeddings emerged as offering the most effective balance of performance, efficiency, and construct validity for rhetorical function classification. While SciBERT was originally designed for analyzing large

collections of published research articles (Beltagy, Lo, & Cohan, 2019), this study demonstrates that its embeddings can also be applied effectively to student writing, capturing rhetorical functions with a high degree of accuracy. The ranking is shown in Table 4.

Table 4

Rank of Three Machine Classifiers Trained on MICUSP Subcorpus

Rank	Model	Overall Accuracy	Macro F1	Weighted F1
1	SciBERT	86.80%	0.84	0.87
2	Qwen3-Embedding-8B	73.20%	0.66	0.73
3	TF-IDF	69.90%	0.65	0.70

In this way, the study extends SciBERT's utility beyond expert discourse into the less stable, more heterogeneous rhetorical landscape of undergraduate and graduate student texts. By contrast, Qwen3-Embedding-8B, though state-of-the-art in semantic similarity tasks (Zhang et al., 2025), underperformed in this specific classification setting and required greater computational resources. Shown in Table 5, these findings suggest that more complex, general-purpose embeddings do not necessarily yield better results when the task involves nuanced rhetorical distinctions rather than broad semantic similarity. SciBERT thus is a pragmatic choice for scaling rhetorical function classification as well as a theoretically significant one, showing that domain-specific models trained on expert writing can be productively adapted for pedagogical research on student writing.

Table 5

Comparison (by Category) of Three Machine Classifiers Trained on MICUSP Subcorpus

Category	SciBERT	Qwen3	TF-IDF
Reporting	0.81	0.56	0.56
Transforming	0.82	0.68	0.64
Evaluating	0.77	0.53	0.60
No Citation	0.95	0.88	0.83

From a construct validity perspective, SciBERT embeddings also appear to best capture the distinctions among Reporting, Transforming, and Evaluating functions as defined in (Kane, 2024). Table 6 below illustrates how each citation function is defined in my validated codebook. It also highlights to what extent SciBERT accurately captures sub-constructs of each citation function, as well as where the model (and all models) falter.

Table 6

Citation Rhetoric Constructs Identified by SciBERT

Citation Function	Sub-Construct of Citation Function	Examples of Sub-Construct Detected by SciBERT
Reporting	Report facts, statistics, anecdotes or other information from sources	*The number of introductions caused by international commerce is enormous (Mooney and Cleland, 2001).*
	Report on the context surrounding sources	*The Book of Margery Kempe, considered to be the first autobiography written in English, presents the story of a woman whose life is guided by visions of, and conversations with, God.*
Transforming	Analyze the source's content or arguments	*Polanski's take on MacBeth thus highlights the depravity of this overly ambitious man, in that his black and deep desires remain to rot within himself.*
	Use source material to support (or challenge) the writer's claims	*Efforts to promote this interaction, such as the conference arranged by Chrisopher Brochu (2004), have already begun, and must continue.*
	Interpret the source's rhetorical effects	*Shylock is sympathetic here, especially as the importance of wedding rings becomes integral in Act V when Portia and Nerissa berate their husbands for giving their rings away.*
	Extrapolate and/or draw inferences from source's content	*Effective management of fragmented populations thus depends on an understanding of the processes and parameters of dispersal (MacDonald & Johnson 2001).*
	Synthesize information from two or more sources	*Although various aspects affecting the infectivity of Diplostomum sp. in species of fish has been shown in a variety of locations (Marcogliese 1962, Mamer 1978, Valtonen and Gibson 1997), it has never been sufficiently examined in fish of Douglas Lake, Michigan.*
Evaluating	Show emotion in response to a source's argument	*Among the most striking and least well understood patterns in influenza are those of subtype replacement and coexistence (Earn et al. 2002).*

Citation Function	Sub-Construct of Citation Function	Examples of Sub-Construct Detected by SciBERT
	Evaluate a source's merit, strength or effectiveness	*By doing this, he overshadowed any goodness that Shakespeare may have originally intended for the play.*

It is also useful to examine the most frequently misclassified rhetorical functions, as these patterns illuminate both the limits of automated classification and the strengths of human coding. Across all three machine classifiers, the Evaluating category was misclassified most often, with models frequently predicting Evaluating in sentences that were actually performing Reporting or Transforming. Close reading explains this pattern: the classifiers tended to over-rely on the presence of evaluative or sentiment-bearing words, even when those words did not function as evaluations in context. For instance, the sentence "Thomson (1993) considered the whole pool-drying scenario to be logically inadequate, instead claiming that ecological conditions had to be the driving factor" (BIO.G1.04.1) was labeled Evaluating by all classifiers—almost certainly triggered by the word "inadequate"—yet under the manual coding scheme it is categorized as Reporting, since it simply conveys the source's stance rather than offering the writer's evaluative judgment. While a larger training sample or expanded context window may reduce some of these errors, the pattern underscores that automated systems struggle to distinguish lexical evaluation from rhetorical evaluation, whereas human readers readily interpret such distinctions through contextual reasoning. In contrast, in this study and in Kane (2024), manual coders most often struggled with the Transforming category, noting the conceptual breadth of the move and sometimes over-identifying transformation in sentences that were principally reporting. Although distinct, these two patterns together highlight an important methodological point: human and machine approaches excel at different aspects of rhetorical classification. As such, a combined workflow—one in which manual insight informs and checks automated tagging—would offer a more robust pathway than relying on either approach alone. The pedagogical uses of such a model are highlighted in the Discussion section.

In sum, SciBERT provides the strongest foundation for machine classification of citation functions. This model was used to analyze the full MICUSP sample, supplemented by manual annotations identifying whether each citation referenced a primary or secondary source. The next section presents the findings from this combined analysis.

4. 2 Descriptive Statistics within the Michigan Corpus of Upper-Level Papers Subcorpus

The MICUSP sample used in this study, shown in Table 7, comprises 165 upper-level academic papers drawn from two disciplines: Biology (n = 67) and English (n = 98). As shown in Table 1, these papers are similar in overall length and sentence structure. English papers include 9,437 sentences and 250,764 words, while Biology papers contain 6,037 sentences and 147,276 words. On average, English papers contain slightly longer sentences (M = 26.57 words) than Biology papers (M = 24.40 words) and have a marginally higher number of sentences per paper (M = 96.30 vs. 90.10). These patterns suggest relatively comparable document lengths across disciplines, though minor differences in sentence complexity may reflect disciplinary writing norms—particularly the tendency for English papers to engage in extended analysis and interpretation.

Table 7

Total & Average Counts in MICUSP Corpus (165 Papers)

DISCIPLINE	Total Sentence Count	Total Word Count	Total Paper Count	Avg Sentence Length	Avg Sentences per Paper
Biology	6037	147276	67	24.40	90.10
English	9437	250764	98	26.57	96.30
BOTH DISCIPLINES	**15474**	**398040**	**165**	**25.72**	**93.78**

When examining the frequency of citations across the corpus, however, disciplinary differences become more pronounced. Table 8 shows that English papers are far more citation-dense, with 87.23% of sentences containing at least one citation, compared to only 29.80% of Biology sentences. On a per-paper basis, English papers average over 75 cited sentences, while Biology papers average around 20. This suggests that English writing at the upper-division level is more intertextual, possibly reflecting the field's emphasis on literary analysis and critical synthesis of secondary sources.

These differences in citation density point to distinct rhetorical expectations across disciplines, which are further reflected in the functions that citations serve within texts. Table 9 summarizes the normalized frequency of each rhetorical function type identified in the full MICUSP sample.

Table 8

Citation Counts in MICUSP Corpus (165 Papers)

DISCIPLINE	Total Sentences	Sentences with Citations	% Sentences with Citations	Avg # of Sentences with Citations per Paper
Biology	6037	1799	29.80	20.45
English	9435	8230	87.23	75.61
BOTH DISCIPLINES	**15474**	**10029**	**64.82**	**55.19**

Table 9

Rhetorical Functions Counts in MICUSP Corpus (165 Papers)

Rhetorical Function	Raw Count	Frequency (Normalized)
Transforming	6550	0.42
No Citation	5458	0.35
Reporting	2986	0.19
Evaluating	478	0.03

The most common rhetorical function across the corpus is Transforming (42.33%), which includes paraphrasing, synthesis, and recontextualization of source material. This is closely followed by No Citation sentences (35.28%), which represent original contributions or general claims that do not rely on sources. Reporting functions—sentences that summarize or otherwise relay information from sources without interpretation—make up 19.30% of the corpus, while Evaluating functions are the least frequent (3.09%), indicating that critical commentary on sources is relatively rare overall. To assess whether rhetorical function usage is associated with discipline, a chi-square test for independence was conducted. The test yielded a highly significant result, χ^2 (3, N = 15,474) = 5,662.33, $p < .001$, indicating a strong association between rhetorical function and disciplinary context. In other words, Biology and English papers differ systematically in how they employ Reporting, Transforming, Evaluating, and No Citation sentences. These findings confirm that disciplinary conventions not only shape how frequently sources are used but also influence the rhetorical purposes they serve.

Before proceeding with further interdisciplinary analysis of citation, it is important to note the genres represented in this study. MICUSP categorizes student writing into several "Paper Types," including Argumentative Essay, Creative Writing, Critique/Evaluation, Proposal, Report, Research Paper, and Response Paper, and the distribution of these genres differs markedly across English and Biology.

The largest contrast appears between Argumentative Essays (65 English papers, 3 Biology papers) and Research Papers (26 Biology papers, 0 English papers). Biology also includes a substantial number of Response Papers (28, compared to 2 in English), while both disciplines contribute a similar number of Reports (31 Biology; 22 English). Although the relatively small group sizes discouraged genre-specific analysis of rhetorical functions, it is nonetheless important to recognize that disciplinary assignments shape the kinds of citation actions students are likely to perform: Argumentative Essays—more commonly assigned in the English—may encourage Transforming or other interpretive moves, whereas Research Papers and Response Papers—which more Biology students write—may more frequently elicit Reporting. This pattern underscores that genre expectations are themselves disciplinary constructs. Thus, the genre differences observed here reinforce the logic of treating this as a disciplinary comparison, since disciplinary conventions determine the genres students write and, in turn, the citation functions they employ.

4.3 RQ2: Differences in Rhetorical Function Usage Across Disciplines

To better understand how rhetorical citation functions vary across disciplines, I applied the finalized SciBERT-based classifier to 165 upper-level student papers from the MICUSP corpus—67 from Biology and 98 from English, respectively. This classifier, which showed the highest overall accuracy across all tested models (86.8%), was used to assign each sentence to one of four rhetorical categories: Transforming, Reporting, Evaluating, or No Citation. The following analysis reflects automated tagging using SciBERT-generated sentence embeddings.

As shown in Table 10, Biology papers were overwhelmingly composed of sentences without reference to sources, with No Citation accounting for 70.20% of the corpus. These sentences, which often consist of original analysis, description, or procedure, occurred an average of 63.25 times per paper (SD = 37.72), though this number varied widely across papers. Among cited sentences, Reporting functions were the most common (15.79%), followed closely by Transforming (12.19%) and with Evaluating (1.82%) trailing far behind. The low use of Evaluating functions in particular is consistent with discipline-specific conventions in scientific writing, which tend to minimize overt critique or commentary.

Table 10

Citation Function Counts in Biology Corpus (n = 67 papers, SciBERT embeddings)

Rhetorical Function	Raw Count	Normalized Count	Avg Use Per Paper (SD)

Transforming	736	0.12	13.63 (15.61)
No Citation	4238	0.70	63.25 (37.72)
Reporting	953	0.16	15.88 (16.14)
Evaluating	110	0.02	2.89 (2.40)

As shown in Table 11, English papers were dominated by Transforming functions, which accounted for 61.62% of all sentences. On average, students employed 59.33 Transforming sentences per paper (SD = 39.78), though usage varied considerably across the corpus. No Citation represented 12.93% of sentences, averaging 19.06 per paper (SD = 33.26), with substantial variability in how often students developed ideas without explicit source use. Reporting functions comprised 21.55% of sentences, with students using an average of 29.96 per paper (SD = 19.83). Finally, Evaluating functions made up 3.90% of the corpus, occurring about 4.04 times per paper (SD = 3.59). Compared to Biology papers, these results show English students rely more heavily on Transforming and Evaluating functions, reflecting humanities conventions that emphasize synthesis and critical engagement with sources.

Table 11

Citation Function Counts in English Corpus (n = 98 papers, SciBERT embeddings)

Rhetorical Function	Raw Count	Normalized Count	Average Use Per Paper (SD)
Transforming	5814	0.62	59.33 (39.78)
No Citation	1220	0.13	19.06 (33.26)
Reporting	2033	0.22	20.96 (19.83)
Evaluating	368	0.04	4.04 (3.59)

Finally, a chi-square test revealed significant disciplinary differences in how students incorporated sources overall ($\chi^2(3, N = 165) = 5662.327$, $p < .001$). Post hoc Z-tests (see Table 12) showed that Biology papers were far more likely than English papers to include sentences without any citation ($z = 72.72$, $p < .001$). In contrast, English papers were significantly more likely to use Reporting functions ($z = –8.86$, $p < .001$), Evaluating functions ($Z = –7.29$, $p < .001$), and especially Transforming functions ($z = –60.70$, $p < .001$). Taken together, these results shown in Table 12 confirm that English writing is characterized by far more synthesis and critical engagement with sources, while Biology writing favors concise reporting.

Table 12

Z-Test Results for Discipline + Rhetorical Function (165 Papers)

Function	Bio Proportion	Eng Proportion	*z* - Statistic	*p* - Value	Interpretation
No Citation	0.70	0.13	72.72	< .001	Bio papers include many more sentences without citations
Reporting	0.16	0.22	-8.86	< .001	English papers use significantly more sentences with reporting citations
Transforming	0.12	0.04	-60.70	< .001	English papers use significantly more sentences with transforming citations
Evaluating	0.02	0.62	-7.29	< .001	English papers use significantly more sentences with

While these disciplinary contrasts are significant, some differences may be shaped by how students engage primary and secondary sources in their writing. For this study, primary sources are defined as works that serve as the subject of analysis—such as novels, short stories, dramatic works, or poems. Secondary sources, by contrast, often include scholarly articles, book chapters, or journalistic texts that students draw on to support or extend their arguments. Each citation in the dataset was manually labeled as referencing either a primary or a secondary source; in cases where both were referenced within the same sentence, the citation was counted in both categories. This distinction provides a more nuanced lens for interpreting rhetorical function patterns across disciplines.

4.4 RQ3: Differences in Rhetorical Function Usage Between Primary and Secondary Sources

Across the corpus, students engaged more extensively with primary sources than with secondary sources. As shown in Table 13, primary source citations accounted for 6,683 instances, compared to 2,894 for secondary sources. Within both categories, Transforming functions were by far the most frequent, making up 71.64% of all primary-source sentences and 10.05% of all secondary-source sentences. Reporting and Evaluating functions followed in frequency, though their relative distribution differed. When students cited primary sources, Reporting (23.40%) and Evaluating (4.95%) occurred at nearly comparable levels. By contrast, when citing secondary sources, Reporting (8.00%) was much more common than Evaluating (0.66%).

These patterns align with disciplinary expectations for how primary and secondary sources function in student writing. Primary sources, particularly in English papers, serve as the objects of analysis, prompting students to summarize, interpret, and critique them in relatively equal measure. Secondary sources, more characteristic of Biology papers, are used primarily to provide factual support or background, where concise reporting dominates and explicit evaluation is minimized.

Table 13

Primary/Secondary Citation Function Counts in MICSUP Corpus (165 Papers)

Rhetorical Function	Primary Source Raw Count	Primary Source Normalized Count	Secondary Source Raw Count	Secondary Source Normalized Count
Transforming	4788	0.72	1555	0.10
Reporting	1564	0.23	1237	0.08
Evaluating	331	0.05	102	0.01
Total	6683	1	2894	1

As shown in Table 14, disciplinary differences in source use become clear when primary and secondary citations are examined separately. In Biology papers, primary source citations were extremely rare, appearing in only three research papers. The types of papers prevalent in the Biology subcorpus (Reports, Response Papers, and Research Papers) typically require students to summarize and assess the validity of published articles in the field. Within this small set, Reporting functions were most common (used in around 12 sentences per paper), followed closely by Transforming (approximately 11 sentences per paper) and Evaluating (approximately 8 sentences per paper). Although the averages per paper appear relatively high,

these values reflect the concentration of citations in only a handful of papers rather than consistent use across the subcorpus.

Table 14

Primary Source Citation Function Counts in Biology Corpus (67 Papers)

Rhetorical Function	Raw Count	Normalized Count	Average Use Per Paper (SD)*
Transforming	37	0.01	11.33 (6.03)
Reporting	34	0.01	12.33 (11.50)
Evaluating	25	< 0.001	8.33 (4.04)

**Based on papers where these functions are used*

By contrast, English papers in this corpus relied heavily on primary sources, which were cited in 88 out of 98 papers. As Table 15 illustrates, transforming dominated, averaging 54.02 sentences per paper (SD = 30.19). Reporting occurred next most common (around 18 sentences per paper), with Evaluating sentences appearing in only around four sentences per paper when counting primary source citations.

Table 15

Primary Source Citation Function Counts in English Corpus (98 Papers)

Rhetorical Function	Raw Count	Normalized Count	Average Use Per Paper (SD)
Transforming	4754	0.50	54.02 (30.19)
Reporting	1527	0.16	17.76 (17.33)
Evaluating	306	0.03	3.97 (3.47)

As shown in Table 16, secondary sources were far more prevalent in Biology papers than in English papers. In Biology, secondary citations appeared in 59 of 67 papers and accounted for the bulk of source use. Within this category, Transforming functions were the most frequent, averaging 13.34 per paper (SD = 15.92), though their distribution varied widely. Reporting functions followed with a similarly high degree of variation across papers (16.53). Evaluating functions were rare, making up only around two sentences per paper (SD = 1.75). This pattern reflects disciplinary conventions in scientific writing, where students primarily summarize and synthesize texts while offering limited explicit critique.

Table 16

Secondary Source Citation Function Counts in Biology Corpus (67 Papers)

Rhetorical Function	Raw Count	Normalized Count	Average Use Per Paper (SD)*
Transforming	667	0.11	13.34(15.92)
Reporting	852	0.14	15.21 (16.53)
Evaluating	62	0.01	2.38 (1.75)

Comparatively, English papers drew on secondary sources less extensively, with citations occurring in 60 of 98 papers and at lower normalized frequencies overall. As Table 17 shown, Transforming again was the most frequent rhetorical function used, averaging 17.08 per paper (SD = 30.93), though the high standard deviation indicates uneven distribution across the subcorpus. Reporting was less common, averaging about 8.37 sentences per paper (SD = 11.84). Evaluating remained the least frequent, occurring about two times per paper (SD = 1.55).

Table 17

Secondary Source Citation Function Counts in English Corpus (98 Papers)

Rhetorical Function	Raw Count	Normalized Count	Average Use Per Paper (SD)
Transforming	888	0.09	17.08 (30.93)
Reporting	385	0.04	8.37 (11.84)
Evaluating	40	< 0.001	1.90 (1.55)

Z-tests shown in Table 18 further clarified the distinctions between rhetorical functions when citing primary versus secondary sources (Table 18). Reporting functions were far more common with secondary sources than with primary sources, a difference confirmed by a strong negative z-statistic (−19.11, $p < .001$). In contrast, Transforming functions occurred more frequently in primary source contexts than in secondary, a significant difference in the opposite direction ($z = 17.02$, $p < .001$). Similarly, Evaluating functions were more prevalent with primary sources than with secondary, again supported by a highly significant z-statistic (3.09, $p < .001$). Together, these results suggest that students use secondary sources primarily for reporting established information, while primary sources prompt greater synthesis and evaluation.

Table 18

Z-Test for Primary vs. Secondary Source Use Per Rhetorical Function

Function	Primary Source Use	Secondary Source Use	z - Statistic	p - Value	Interpretation
Reporting	0.23	0.43	-19.11	< .001	Use of Reporting is significantly higher in secondary source use than primary source use
Transforming	0.72	0.54	17.02	< .001	Use of Transforming is significantly higher in primary source use than secondary source use
Evaluating	0.05	0.04	3.09	< .001	Use of Evaluating is significantly higher in primary source use than secondary source use

A chi-square test revealed significant disciplinary differences in how students used rhetorical functions with primary sources. Th post hoc Z-tests shown in Table 19 indicated that Biology papers were significantly more likely than English papers to use Reporting functions when citing primary sources ($Z = –7.93$, $p < .001$). In contrast, English papers were significantly more likely than Biology papers to use Transforming functions ($Z = 3.53$, $p < .001$) and Evaluating functions ($Z = 9.59$, $p < .001$). These findings suggest that Biology writers tend to prioritize conveying source content directly, while English writers engage more in interpretation and evaluation when integrating primary sources.

Table 19

Z-Test Results for Discipline + Rhetorical Function (Primary Sources, 91 Papers)

Function	Bio	English	*z* - Statistic	*p* - Value	Interpretation
Reporting	0.39	0.23	-7.93	< .001	Significantly more use of Reporting in Biology sentences citing primary sources than English sentences citing primary sources
Transforming	0.35	0.72	3.53	< .001	Significantly more use of Transforming in English sentences citing primary sources than Bio sentences citing primary sources
Evaluating	0.26	0.05	9.59	< .001	Significantly more use of Evaluating in English sentences citing primary sources than Biology sentences citing primary sources

Similarly, significant differences appeared when students cited secondary sources ($\chi^2 = 189.26$, $df = 2$, $p < .001$). As shown in Table 20, Biology papers were significantly more likely than English papers to use Reporting functions when citing secondary sources ($z = 13.30$, $p < .001$). Conversely, English papers were significantly more likely to use Transforming functions ($z = –13.67$, $p < .001$). No significant disciplinary difference emerged for Evaluating functions ($z = 1.27$, p = .20).

Table 20

Z-Test Results for Discipline + Rhetorical Function (Secondary Sources, 119 Papers)

Function	Bio	English	*z* - Statistic	*p* - Value	Interpretation
Reporting	0.54	0.29	13.30	< .001	Reporting is used significantly more frequently in reference to secondary sources in Bio papers
Transforming	0.42	0.68	-13.67	< .001	Transforming is used significantly more frequently in reference to secondary sources in English papers
Evaluating	0.04	0.03	1.27	0.20	There is no significant difference in the use of Evaluating between English and Bio papers (both use the functions least often of the three citation functions)

5.0 Discussion

In response to RQ1, the findings show that among the three machine-learning approaches tested—TF-IDF, Qwen3 embeddings, and SciBERT embeddings—SciBERT offers the most reliable and valid method for large-scale identification of rhetorical citation functions in student writing. Whereas TF-IDF provided a workable baseline and Qwen3 captured broad semantic relations but struggled with finer rhetorical distinctions, SciBERT consistently achieved the highest accuracy, precision, recall, and F1 scores across all categories, closely approximating human-coded judgments. This performance indicates that domain-specific embeddings pre-trained on academic discourse are particularly well suited to the task of distinguishing functions such as Reporting, Transforming, and Evaluating. Importantly, this methodological finding extends prior rhetorical-function frameworks—such as Petrić's (2007) and the more elaborated schemes proposed by Yan and Ma (2024)—by showing that their underlying conceptual distinctions can be scaled beyond small, manually coded datasets. In other words, SciBERT does not supersede these frameworks but operationalizes them, enabling the kind of corpus-level, cross-disciplinary analyses that have long been called for but were previously infeasible due to the labor-intensive nature of manual coding.

In response to RQ2, the MICUSP case study demonstrates clear and systematic disciplinary differences in how students use and rhetorically frame citations. Although English and Biology papers were similar in overall length and sentence structure, their patterns of source use diverged sharply.

English papers were substantially more citation-dense, with nearly nine in ten sentences containing at least one citation, compared to fewer than one-third of sentences in Biology—a contrast that reflects the

humanities' reliance on sustained textual engagement. More importantly, the rhetorical functions of those citations differed markedly across fields. English writers overwhelmingly favored Transforming moves, drawing on synthesis, paraphrase, and recontextualization to construct interpretive claims, often around primary texts. Biology writers, by contrast, were less likely to reference sources in general and, when they did cite, used Reporting far more frequently, consistent with disciplinary expectations that emphasize succinct presentation of prior research rather than extended analysis of it. Evaluative moves were infrequent in both fields but occurred more often in English, aligning with humanities norms of critique and interpretive positioning. Together, these patterns mirror long-established disciplinary distinctions identified by Hyland (2004), who argues that citation practices mediate the tension between demonstrating originality and signaling humility to the community. In this sense, humanities writers position themselves as interlocutors in an ongoing scholarly conversation, whereas scientific writers tend to treat many sources as background information that can be reported efficiently or omitted altogether.

The MICUSP analysis thus illustrates not only that citation rhetorics differ across disciplines, but also that those differences manifest consistently in the frequency, distribution, and rhetorical purpose of Reporting, Transforming, and Evaluating functions in student writing.

In response to RQ3, the analysis shows that students' rhetorical citation functions shift meaningfully depending on whether they are engaging primary or secondary sources—and these shifts are mediated by disciplinary norms. English writers drew heavily on Transforming moves when working with primary texts, using synthesis, paraphrase, and interpretive recontextualization to advance argumentative claims. Primary texts also prompted the highest rates of Evaluating in English papers, reflecting humanities conventions that encourage critique, interpretive positioning, and the articulation of competing readings. In contrast, secondary sources in English were more often associated with Reporting, used to situate interpretations within existing scholarship but not extensively analyzed in their own right. Biology writers displayed the opposite pattern: when citing empirical literature—treated here as secondary research—they overwhelmingly relied on Reporting functions, summarizing prior findings with minimal transformation and rarely offering evaluative commentary. Primary sources, which in Biology often take the form of data sets, methodological descriptions, or scientific phenomena rather than textual artifacts, appeared far less frequently and elicited little rhetorical elaboration beyond basic reporting. These contrasts demonstrate that "primary" and "secondary" are not simply bibliographic categories, but disciplinary constructs that shape what counts as evidence and how students are expected to use it. Collectively, these findings reinforce that citation is not a uniform or generic skill; it is deeply tied to disciplinary "ways of knowing and doing" (Carter, 2007), manifesting differently depending on the epistemological role a source plays in a field.

5.1 Pedagogical Implications

The findings from RQ1–RQ3 have clear implications for writing pedagogy across disciplines, particularly in supporting students' metacognitive awareness of citation as a rhetorical practice. Because these results demonstrate that citation functions vary systematically across disciplines and source types, they highlight the need for instructional approaches that make such expectations visible to students rather than allowing them to remain tacit. SourceMapper—the tool developed as a result of machine classification training and testing in this study—offers one means of doing so. By providing sentence-level tags of rhetorical citation functions, SourceMapper gives students and instructors a shared metalanguage for discussing how sources are being used and whether those uses align with disciplinary norms.

Integrated into writing pedagogy, the tool can serve as a scaffold that helps students understand not only when they are citing, but why they are doing so, and how those rhetorical purposes shift across contexts.

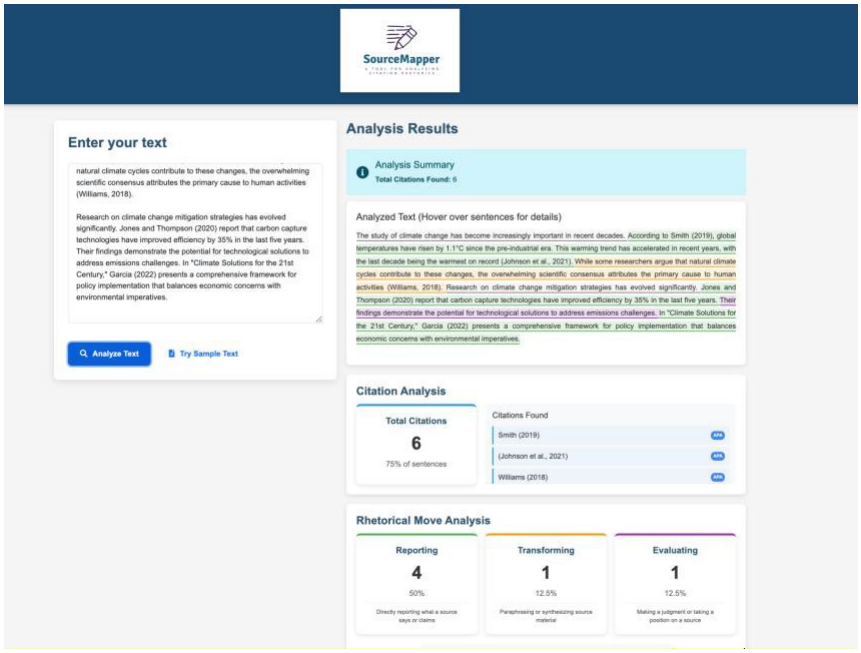
Several tools precede SourceMapper in this arena. Notably, DocuScope's classroom platforms, integrated into writing courses at Carnegie Mellon University, have been shown to promote student metacognition by enabling them to visualize their rhetorical and linguistic choices and compare them to peers (Helberg et al., 2018). Reflection exercises built around DocuScope Classroom reports help students see how their language patterns relate to genre conventions, thereby cultivating what Helberg et al. (2018) call "textual awareness"—a metacognitive recognition of how individual composing choices aggregate into genre-appropriate rhetorical effects. Such tools not only inform course-level instruction but also allow programs to make broader curricular claims about writing development and align assessment with rhetorical and genre-based goals across the curriculum (Wetzel et al., 2021). Similarly, Elena Cotos's Research Writing Tutor has long been used to support student understanding of rhetorical function structures in academic genres such as introductions, methods, and conclusions (Cotos, Huffman, and Link, 2020). By making disciplinary patterns explicit and offering feedback on rhetorical function, RWT helps students recognize how writing conventions vary across fields and invites them to adapt their rhetorical strategies accordingly. In this sense, SourceMapper can serve as a complementary analysis tool alongside platforms like DocuScope or the Research Writing Tutor, with its distinctive focus on citation type and rhetorical function. Like these tools, it allows researchers to analyze individual texts or entire corpora while attending to metadata such as discipline, genre, or source type, thereby deepening interpretive possibilities. The result would be an even more layered analysis—one that supports a more nuanced understanding of textual awareness.

SourceMapper integrates the trained and validated SciBERT-based classifier for rhetorical citation functions into a user-friendly interface for sentence-level tagging and analysis. The tool provides both students and instructors with an accessible platform for examining how and why sources are used within a piece of writing and for comparing those patterns to disciplinary expectations. In the sections that follow, I describe the interface and outline several pedagogical applications that leverage these capabilities.

The SourceMapper interface shown in Figure 6 allows users to paste or upload text for citation analysis. During the analysis process, the system first detects whether each sentence contains a citation by combining regular expressions, rules-based heuristics, and a named entity recognition (NER) pipeline. It then matches the detected references to major citation styles such as MLA, APA, Chicago, or others. If a sentence contains an author's name or other clear markers of citation consistent with major academic styles, it is classified as an "attributed" citation, following the distinctions used by Citation Project researchers such as Jamieson (2013). Conversely, if a sentence signals engagement with a source through phrases like "this research shows" or "they argue" but contains no explicit author name or formal parenthetical reference, it is classified as "non-attributed." This distinction does not necessarily imply plagiarism; rather, it highlights a common pattern in student writing where ideas are referenced indirectly (often, in a sentence following one in which a source has been referenced explicitly). Identifying these patterns can offer students and instructors a useful diagnostic tool for understanding when source use is being made explicit and when it is left implicit, and can help guide revisions toward clearer, more intentional citation practices.

Figure 6

Overview of SourceMapper Interface



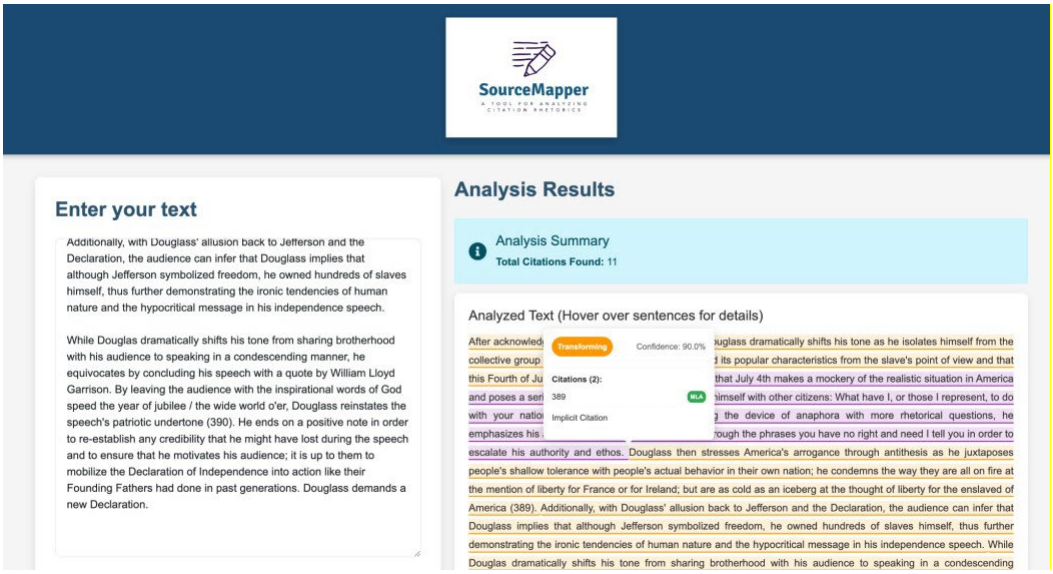
Once the presence and format of a citation are identified, the SciBERT-based classifier determines the rhetorical function of that citation—whether it is primarily reporting, transforming, or evaluating a source. The results are displayed in two complementary ways, as pictured above. First, the text itself is returned in an annotated view, with color-coding or highlighting to indicate rhetorical functions sentence by sentence. Second, a summary dashboard provides an overview of the analysis, including the proportion of sentences with and without citations, the balance of rhetorical function types, and the distribution of citation styles across the text. Instructors and students can thus see whether sources are present and formally correct and how they function rhetorically within the paper. In addition to the interface, users can also download a CSV of tagged sentences in an individual text or corpus.

5.2. Proposed Classroom Implementation: Using MICUSP As a Teaching Tool

SourceMapper could be of use in the classroom would as a teaching tool for the investigation of a corpus like MICUSP and for comparison across disciplines. This builds on the long tradition of using MICUSP and other corpora as a pedagogical resource for writing instruction, especially in WAC/WID contexts where disciplinary expectations need to be made explicit (O'Donnell and Römer, 2011; Hardy and Römer, 2013; Yan and Ma, 2024). The goal in such activities is to help students see not only what counts as strong writing in their own field but also how practices may differ across disciplines. For example, as shown in Figure 7, an English instructor might project SourceMapper results from the analysis of a single MICUSP paper, such as ENG.G0.01.2.

Figure 7

Results of SourceMapper Analysis of Paper ENG.G0.01.2



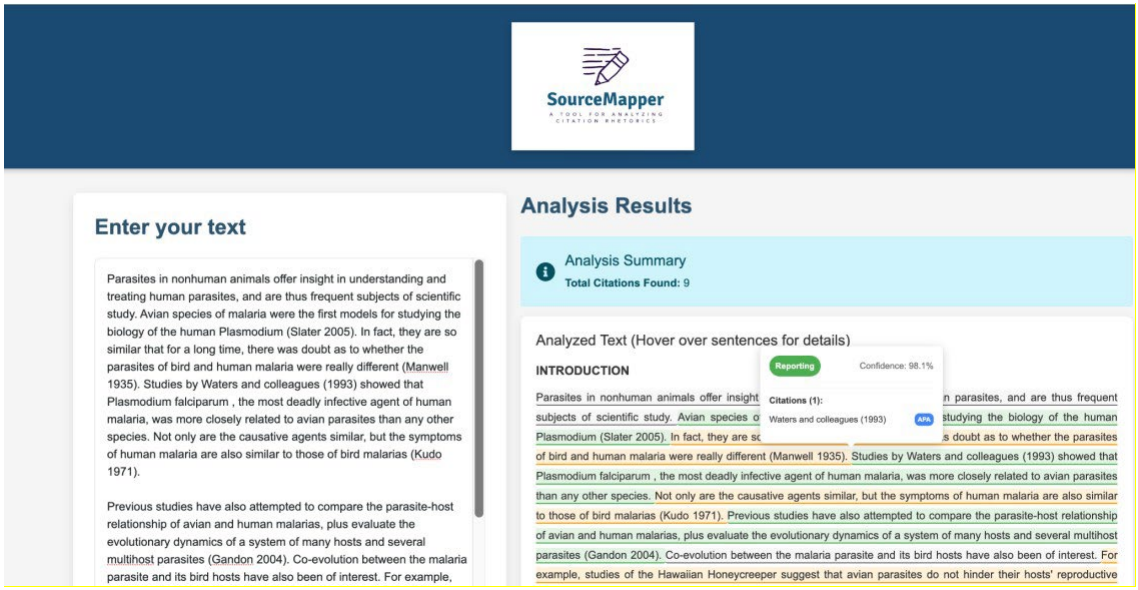
(https://micusp.elicorpora.info/view?pid=ENG.G0.01.2)

They could then lead a discussion on two fronts: first, the type of citation the student author is employing (MLA vs. APA, attributed vs. non-attributed), and second, the extent to which the rhetorical balance of citation functions (e.g., reporting vs. transforming vs. evaluating) reflects disciplinary expectations. In this paper, for example, the use of citation is significant; indeed, every sentence includes either an attributed MLA citation or non-attributed but implicit referential to the source in question (Frederick Douglass' declaration). Further, the patterns of citation rhetoric in this paragraph closely mirror the large-scale findings of this study. The primary citation function is Transforming, as the student takes up and analyzes the literary characteristics of Douglass's speech. In fact, all but two sentences Transform the source; the others Evaluate its rhetorical impact, another common function made in English papers. Such a lesson helps demystify disciplinary writing conventions while also inviting students to critique and reflect on authentic examples of student writing, an approach that resonates with MICUSP's original pedagogical mission.

Similarly, an instructor could ask students to compare papers from two different disciplines using SourceMapper, such as Biology and English, as shown in Figure 8. Students would examine each paper in turn through the tool and then discuss the rhetorical and citation patterns they observe. For instance, compared to the English paper analyzed above, a Biology paper might feature significantly fewer citations per paragraph overall, reflecting disciplinary conventions of concentrating references in the introduction and methods sections.

Figure 8

Results of SourceMapper Analysis of Paper BIO.G0.21.1



(https://micusp.elicorpora.info/view?pid=BIO.G0.21.1)

Moreover, students may notice that Biology papers display a much higher prevalence of Reporting functions—directly attributing findings to previous studies—while only occasionally Transforming to synthesize or reinterpret sources and rarely making evaluative judgments. The Biology paper excerpt shown here illustrates this trend: most citations appear as concise reporting of prior research. Such a comparative exercise could help students see concretely how citation practices differ across fields, fostering disciplinary awareness and metacognitive reflection on their own use of sources.

5.3 Proposed Classroom Implementation: Using SourceMapper as a Revision Aide

A second implementation for the classroom would be to encourage students to upload their own papers to SourceMapper for analysis. This activity would allow students to visualize their citing habits in real time. They would be able to explore 1) whether their work adheres to correct citation standards (MLA/APA/Chicago), 2) to what extent they are attributing their sources properly, and 3) the rhetorical balance of their citations across a text. Instructors could scaffold this practice through a sequence of activities: first, introducing the rhetorical functions of citation (reporting, transforming, evaluating) and having students hand-code examples in a model text; then, using SourceMapper to analyze a model rhetorical analysis paper so students can see how typical citation patterns play out in practice; and finally, inviting students to analyze their own drafts and compare their patterns to the conventions they have observed. In this way, students would not be given prescriptive disciplinary rules but would instead encounter typical conventions and expectations against which to reflect upon their own practice. Although not yet tested formally in the classroom, I propose that such activities would yield effects like those reported for DocuScope Classroom. Wetzel et al. (2021) found that automated rhetorical tagging tools create “spaces for student agency” by supporting habits of rhetorical reasoning (p. 294), while Helberg et al. (2018) demonstrated that DocuScope Classroom enhanced revision and

heightened students' metacognitive awareness of genre. SourceMapper has the potential to offer similar benefits in the understudied area of citation rhetorics—not only by helping students reflect on and revise how they use sources, but also by making expectations around citation clearer across a range of writing contexts, whether disciplinary (as evidenced in this study), genre-based, or assignment-specific.

6.0 Conclusions

This article used machine-learning classification to examine how student writers in Biology and English engage sources through three citation functions: Reporting, Transforming, and Evaluating. As a whole, the study highlights several implications for how we understand and teach citation as a rhetorical practice shaped by discipline and source type. First (RQ1), the comparison of three machine-learning approaches demonstrated that SciBERT offers the most reliable and valid method for large-scale rhetorical-function tagging, establishing a scalable foundation for analyzing student citation practices.

Second (RQ2), the full MICUSP analysis revealed clear disciplinary contrasts: English writers relied heavily on Transforming and Evaluating functions, while Biology writers predominantly used Reporting or produced No Citation sentences, underscoring longstanding disciplinary differences in how knowledge is constructed rhetorically. Third (RQ3), source type shaped citation function use in discipline-specific ways: English students transformed and evaluated primary texts at high rates, whereas Biology writers primarily reported information from secondary research. Together, these findings show that citation is not merely a formal requirement, but a rhetorical act shaped by disciplinary "ways of knowing and doing," and they demonstrate how a machine-learning–driven tool like SourceMapper—when complemented by descriptive statistics and close reading—can surface such patterns at scale with both analytical and pedagogical value.

While the potential of machine classification for writing pedagogy is significant, there are also important limitations and risks. As noted in the 2014 CCCC Position Statement on Assessment, the best forms of writing assessment are contextual and reflect the value of a range of rhetorical skills and literacies. Automated systems risk reinforcing superficial lexical patterns unless they are continually validated against human-coded data and interpreted through rhetorical frameworks. Relatedly, there is the danger of over-reliance. While SourceMapper can provide valuable diagnostic insights, it should not replace human judgment or the rich, dialogic processes of peer review and instructor feedback. Instead, it should be positioned as one resource among many in a balanced ecology of formative assessment.

In addition, issues of fairness and representativeness must also be considered. Given that the classifiers were trained on two corpora—a smaller, more recent first-year writing dataset collected in 2023 and the larger but nearly twenty-year-old MICUSP corpus—this combination presents its own challenges. The FYW corpus offers contemporary writing but is limited in size and institutional scope; MICUSP provides breadth across disciplines and genres but reflects older writing practices and one U.S.-based academic context. Together, they form a hybrid training set that is richer than either corpus alone, yet still not fully representative of the range of student writers, institutions, and linguistic backgrounds present in higher education today. Because writing and citation practices are always "stabilized for now," it is likely that they have shifted considerably in the two decades separating these corpora—especially given changes in digital research habits, the increasing globalization of academic Englishes, and the emergence of AI-assisted writing tools. These constraints underscore the need for more diverse, contemporary datasets if large-scale citation-function analysis is to develop responsibly. Comparable corpora such as BAWE offer useful points of expansion, being larger and more balanced across disciplines, though they share similar timepoint limitations. Building and sharing new, representative corpora that include multilingual writers, first-year coursework, multiple varieties of English, and a broader range of institutional contexts would substantially strengthen both the fairness and the generalizability of computational approaches like SourceMapper. Given

the tool's ability to handle large datasets, developing such corpora is not only feasible but essential for future work.

As noted above, this research has several applications for the classroom. In WAC and WID contexts, SourceMapper can help students navigate shifting expectations as they function across courses in the humanities and the sciences, where citation norms and rhetorical functions often differ. By analyzing model texts or incorporating SourceMapper as a revision step, instructors can use the tool to make disciplinary citation patterns more visible and to guide students in reflecting on their own practices. While the platform does not yet distinguish between primary and secondary sources, the citation counts and rhetorical function distributions it produces already offer a comparative baseline for heightened awareness. Developing analytic methods to identify and track primary versus secondary source use represents an important future research direction. More broadly, classroom-based research is essential for determining how SourceMapper can be most effectively implemented. Pilot studies could, for example, integrate the tool into scaffolded assignments where students first analyze model texts, then apply the tool to their own drafts during revision, and finally reflect on how their citation rhetorics compare to disciplinary expectations. Such research would provide a critical avenue for exploring the extent to which these practices positively influence students' awareness and use of citation functions in ways that align with genre- and discipline-based conventions. Beyond formative classroom use, I also envision future applications for summative assessment. For writing program administrators, for example, aggregated SourceMapper analyses could serve as a valuable form of evidence for accreditation and curricular assessment, provided results are interpreted in light of local outcomes and student demographics. Finally, future directions must also include continued model training, since the landscape of machine learning evolves rapidly, with new models and benchmarks for text classification emerging monthly (and sometimes even daily).

In sum, this study validates a computational approach to rhetorical function analysis, extends citation scholarship to student writing across disciplines, and highlights pedagogical applications that foreground fairness and accessibility. Returning to the opening question—why do students cite?—the analyses enabled by this tool suggest that students cite to report, to transform, and to evaluate knowledge in ways that align with disciplinary expectations. By making these rhetorical functions visible and enabling their large-scale classification, this work contributes to our body of knowledge about how disciplinary conventions shape the ways students engage in knowledge making. Further, the SourceMapper application offers a promising way for students to see their own work in relation to broader academic practices, and to reflect on how their rhetorical decisions position them within a field. Ultimately, by developing a machine-classification methodology with clear classroom applications, this work advances research on citation rhetorics while creating new opportunities to connect classroom writing more directly to disciplinary discourse.

Author Biography

Megan Kane is a Visiting Assistant Professor of English at Seton Hall University. Her research centers on the computational analysis of student writing, with particular attention to the rhetorical choices students make when citing sources. She also explores how AI is reshaping writing classrooms and influencing students' composing practices. She is the 2025–2026 WAC Clearinghouse Associate Publishers New Scholar Fellow, a role in which she serves as the 2025-206 Assistant Editor of *The Journal of Writing Analytics*. ORCID: https://orcid.org/0000-0003-1817-2751

References

Aull, L. L., Bandarage, D., & Miller, M. R. (2017). Generality in student and expert epistemic stance: A corpus analysis of first-year, upper-level, and published academic writing. *Journal of English for Academic Purposes, 26,* 29–41. https://doi.org/10.1016/j.jeap.2017.01.005

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3613–3618). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1903.10676

Carter, M. (2007). Ways of knowing, doing, and writing in the disciplines. *College Composition & Communication, 58*(3), 385-418. https://doi.org/10.58680/ccc20075912

Chang, Y. Y. (2008). Citation and disciplinary knowledge: An investigation of citation practices in Applied Linguistics versus Computer-Aided Architectural Design. *Concentric: Studies in Linguistics, 34*(2), 121-142.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2019). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.207

Cotos, E., Huffman, S., & Link, S. (2020). Understanding graduate writers' interaction with and impact of the Research Writing Tutor during revision. *Journal of Writing Research, 12*(1), Article 1. https://doi.org/10.17239/jowr-2020.12.01.07

Hardy, J. A., & Friginal, E. (2016). Genre variation in student writing: A multi-dimensional analysis. *Journal of English for Academic Purposes, 22*, 119–131. https://doi.org/10.1016/j.jeap.2016.03.002

Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-Level Student Papers (MICUSP). *Corpora, 8*(2), 183–207. https://doi.org/10.3366/cor.2013.0041

Hardy, J. A., Römer, U., & Roberson, A. (2015). The power of relevant models: Using a corpus of student writing to introduce disciplinary practices in a first-year composition course. *Across the Disciplines, 12*(1). https://wac.colostate.edu/docs/atd/articles/hardyetal2015.pdf

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, 1–15. https://doi.org/10.1016/j.asw.2018.06.003

Hyland, K. (2011). Disciplines and discourses: Social interactions in the construction of knowledge. In D. Starke-Meyerring, A. Paré, N. Artemeva, M. Horne, & L. Yousoubova (Eds.), *Writing in Knowledge Societies* (pp. 193-214). Parlor Press & the WAC Clearinghouse.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing* (Michigan Classics Edition). University of Michigan Press.

Hyland, K. (2002). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics, 34*, 1091-1112. http://dx.doi.org/10.1016/S0378-2166(02)00035-8

Hyland, K., & Jiang, F. (2017). Is academic writing becoming more informal? *English for Specific Purposes, 45*, 40–51. https://doi.org/10.1016/j.esp.2016.09.001

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.

Jamieson, S. (2013). Reading and Engaging Sources: What students' use of sources reveals about advanced reading skills. *Across the Disciplines*, *10*(4), 1–20. https://doi.org/10.37514/ATD-J.2013.10.4.15

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), Machine learning: ECML-98 (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Kane, Megan, (2024). *Assessing citation practices in first-year writing: A computational-rhetorical approach* [Doctoral dissertation)] Philadelphia, Temple University. http://dx.doi.org/10.34944/dspace/10643

Kane, M. (2025). Citation analysis tool [Computer software]. GitHub. https://github.com/mkane968/citation-analysis-tool

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Kastman Breuch, L.-A., & Larson, B. N. (2017). Research and rhetorical purpose: Using genre analysis to understand source use in technical and professional writing. In T. Serviss & S. Jamieson (Eds.), *Points of Departure: Rethinking Student Source Use and Writing Studies Research Methods* (pp. 182–208). Utah State University Press. https://doi.org/10.7330/9781607326250.c006

Kim, S. (2018). 'Two rules are at play when it comes to none': A corpus-based analysis of singular versus plural none. *English Today, 34*(3), 50-56. https://doi.org/10.1017/S0266078417000554

Knight, S., et al. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. Journal of Writing Research, 12(1), 141-186. https://doi.org/10.17239/jowr-2020.12.01.06

Knight, S., et al. (2020). Are you being rhetorical? A description of rhetorical move annotation tools and open corpus of sample machine-annotated rhetorical moves. *Journal of Learning Analytics, 7*(3), Article 3. https://doi.org/10.18608/jla.2020.73.10

Larsson, T. (2018). Is there a correlation between form and function? A syntactic and functional investigation of the introductory it pattern in student writing. *ICAME Journal, 42*(1), 13- 40. https://doi.org/10.1515/icame-2018-0003

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). Springer

Lee, J. J., Hitchcock, C., & Elliott Casal, J. (2018). Citation practices of L2 university students in first-year writing: Form, function, and stance. *Journal of English for Academic Purposes, 33,* 1–11. https://doi.org/10.1016/j.jeap.2018.01.001

Li, Y., & Casanave, C. P. (2012). Two first-year students' strategies for writing from sources: Patchwriting or plagiarism? *Journal of Second Language Writing, 21*(2), 165–180. https://doi.org/10.1016/j.jslw.2012.03.002

Nesi, Hilary (2021). Sources for courses: Metadiscourse and the role of citation in student writing. *Lingua, 253,* 1-17. https://doi.org/10.1016/j.lingua.2021.103040.

O'Donnell, M. B., & Römer, U. (2012). From student hard drive to web corpus (Part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *7(*1), 1–18. https://doi.org/10.3366/corp.2012.0015

Omizo, R., Clark, I, Nguyen, M. & Hart-Davidson, W. (2019). Inventing Rhetorical Machines: On Facilitating Learning and Public Participation in Science. In J. Ridolfo & W. Hart-Davidson (Eds.), *Rhetorical Machines: Writing, Code, and Computational Ethics* (pp. 110-136), University of Alabama Press.

Petrić, B. (2007). Rhetorical functions of citations in high- and low-rated master's theses.

Journal of English for Academic Purposes, 6(3), 238–253. https://doi.org/10.1016/j.jeap.2007.09.002

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

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.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval.

Information Processing & Management, 24(5), 513–523. https://doi.org/10.1016/0306-

4573(88)90021-0

Scheidt, D., & Middleton, H. (2021). The tacit values of sourced writing: A study of source "engagement" and the FYW program as community of practice. *Writing Program Administration, 45*(1), 91–110. https://link.gale.com/apps/doc/A687753894/AONE?u=anon~922d575a&sid=googleScho

lar&xid=93bd369c

Swales, J. (1986). Citation Analysis and Discourse Analysis. *Applied Linguistics, 7*(1), 39–56. https://doi.org/10.1093/applin/7.1.39

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge, UK: Cambridge University Press.

Swales, J. M. (2014). Variation in Citational Practice in a Corpus of Student Biology Papers: From Parenthetical Plonking to Intertextual Storytelling. Written Communication, 31(1), 118–141. https://doi.org/10.1177/0741088313515166

Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for Academic Purposes. *Language, Learning & Technology, 5*(3), 91–91. https://doi.org/10.64152/10125/44568

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. https://doi.org/10.1177/1461445609341006

Wang, X. (2022). Hedging in academic writing: Cross-disciplinary comparisons in the Michigan Corpus of Upper-Level Student Papers (MICUSP). *EdArXiv*. https://doi.org/10.35542/osf.io/5xj27

Wette, R. (2017). Source text use by undergraduate post-novice L2 writers in disciplinary assignments: Progress and ongoing challenges. *Journal of Second Language Writing, 37,* 46-58. https://doi.org/10.1016/j.jslw.2017.05.015

Wetzel, D., Kaufer, D., & Ishizaki, S. (2021). Toward a rhetorical approach to digital writing assessment. *Journal of Writing Assessment, 14*(1), 289–298. https://wac.colostate.edu/docs/jwa/vol14/wetzel.pdf

Widyantoro, D. H., Khodra, M. L., Trilaksono, B. R., & Aziz, E. A. (2013). A multiclass-based classification strategy for rhetorical sentence categorization from scientific papers.

Journal of ICT Research and Applications, 7(3), 235–249. https://doi.org/10.5614/itbj.ict.res.appl.2013.7.3.5

Yan, J. and Ma, Q (2024). Developing advanced citation skills: A mixed-methods approach to corpus technology training for novice researchers. *Journal of English for Academic Purposes, 72*, 1-14. https://doi.org/10.1016/j.jeap.2024.101451

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models (Version 3). *arXiv*. https://doi.org/10.48550/arXiv.2506.05176