

RESEARCH ARTICLE

JASIST WILEY

Methodological challenges in content-based citation analysis: Expertise, reliability, and the primacy of citance identification

Zehra Taşkın^{1,2} 

¹Department of Information Studies,
Institute of Informatics, Hacettepe
University, Ankara, Turkey

²Scholarly Communication Research
Group, Adam Mickiewicz University,
Poznań, Poland

Correspondence

Zehra Taşkın, Department of Information
Studies, Institute of Informatics,
Hacettepe University, Ankara, Turkey;
Scholarly Communication Research
Group, Adam Mickiewicz University,
Poznań, Poland.
Email: ztaskin@hacettepe.edu.tr

Funding information

Polish National Agency for Academic
Exchange, NAWA Poland, Grant/Award
Number: PPN/ULM/2019/1/00062

Abstract

Content-based citation analysis seeks to capture the meaning and functions of citations but continues to face unresolved methodological challenges. This study analyzes a stratified sample of library and information science publications to examine how citance segmentation and annotator expertise influence the consistency of classification. Using two annotators with different professional backgrounds, the findings show that agreement is high when citances are defined identically, but reliability decreases sharply once text boundaries diverge. Citance length, rather than subject category or citation density, emerges as the strongest predictor of disagreement. These results identify segmentation as a methodological rather than a purely technical issue, shaping both human and automated tagging outcomes. By highlighting the interplay between expertise effects and boundary definitions, the study underscores the need for clearer operational frameworks in citation analysis. The contribution lies in demonstrating that methodological refinements in citance identification are essential for improving reproducibility, enhancing hybrid human-machine approaches, and strengthening the validity of citation-based indicators in research evaluation.

1 | INTRODUCTION

In recent years, content-based citation analysis has gained increasing attention as a methodological approach to better understand the functions and contexts of citations within scholarly communication (Anderson & Lemken, 2023; Donner et al., 2025; Iqbal et al., 2020; Ma et al., 2025; Nishikawa & Koshiba, 2024). Originally, this study aimed to develop a new classification scheme for citation contexts in English-language publications in the library and information science (LIS) field, as a follow-up to earlier work on the Turkish LIS field (Taşkın &

Al, 2018). However, rapid advancements in artificial intelligence (AI) and natural language processing (NLP) have made the technical task of categorizing citation sentences increasingly efficient (Nicholson et al., 2021; scite.ai, 2020), raising important questions about what truly constitutes success in supervised classification systems. For example, we found that content-based citation analysis faces persistent challenges in accurately understanding the positive and negative meanings of citations, reliably extracting citances (citation sentences) from texts, and classifying multiple citations within the same context. All of which limit the consistency and validity of

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2026 The Author(s). *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

both human and automated tagging (Taşkın et al., 2023). These limitations are further compounded in automated systems, where most features used for classification have low predictive value, suffer from reproducibility issues, and rely on small or unbalanced datasets that limit generalizability (Pride & Knoth, 2017).

As automated tools promise higher accuracy and scalability, it becomes critical to examine the nuanced human factors that influence classification outcomes (Li et al., 2019). One such factor is the level of subject expertise of those who annotate or tag citances. While it might be assumed that deeper expertise naturally leads to more accurate tagging, studies have shown a more complex reality: annotator reliability can vary by instance and task, and aggregated non-expert judgments can sometimes rival or even exceed the consistency of expert annotations (Li et al., 2019; Snow et al., 2008). This expertise effect shows that experts, by reading between the lines, may introduce greater variability into the tagging process. Differences in raters' subjective interpretations and thresholds can therefore increase random error and bias (Narimanzadeh et al., 2023). At the same time, disagreement in content-based citation analysis does not emerge solely from classification decisions themselves. A more fundamental source of variability lies in the identification of citance boundaries (Abu-Jbara & Radev, 2012; Nakov et al., 2004). Citation contexts are rarely self-evident textual units, and annotators must decide where a citance begins and ends, whether surrounding sentences belong to the same rhetorical context, and how much contextual information is necessary for interpretation. These decisions are inherently interpretive and may vary according to disciplinary reading practices, professional experience, and assumptions about rhetorical coherence.

Such variability has important implications for reproducibility. Differences in citance segmentation influence how meaning, purpose, and evaluative tone are interpreted, which in turn affects both human annotation and automated classification outcomes. Despite major advances in NLP and large language models, the methodological instability introduced during citance identification remains insufficiently examined.

This study investigates how expertise and citance segmentation shape annotation consistency in content-based citation analysis. To do so, I draw on a large dataset of citances in the LIS domain (Taşkın, 2025) independently tagged by two annotators with different professional backgrounds: one an academic researcher specializing in scholarly communication and bibliometrics, the other a practitioner working with citation indexing systems. The comparison makes it possible to examine how interpretive decisions emerge during citance identification and how these decisions influence downstream classification.

The contribution of the study lies in repositioning citance identification as a methodological and epistemic problem rather than a purely technical preprocessing step. The findings demonstrate that disagreement frequently emerges before classification itself, at the stage where annotators define the textual boundaries of citation contexts. By foregrounding this issue, the study contributes to ongoing discussions about reliability, reproducibility, and interpretive validity in content-based citation analysis.

2 | CONTENT-BASED CITATION ANALYSIS

Content-based citation analysis has developed steadily from early rhetorical and functional classifications (Garfield, 1965; Moravcsik & Murugesan, 1975) to the modern integration of semantic and syntactic methods powered by NLP and deep learning (Athar, 2011; Pride & Knoth, 2017). Early scholarship highlighted the limitations of simple citation counts by showing that citations differ in purpose, polarity, and rhetorical role (Teufel, 1999). Over the past decade, the emergence of large full-text corpora and advances in machine learning have made it possible to scale citation contents, addressing questions about citation sentiment (Abu-Jbara & Radev, 2012; Small, 2011), citation function (Teufel et al., 2006), and more recently, geopolitics and epistemic context (Taşkın et al., 2023).

Recent studies have expanded these perspectives in several directions. Surveys of deep learning applications in bibliometrics emphasize how transformer-based architectures and large datasets are reshaping both methodological approaches and research questions in the field (Zhang et al., 2025). Semi-supervised models such as GAN-based frameworks demonstrate how generative methods can support citation function analysis where annotated data are limited (Chatrinan et al., 2025). Parallel to these advances, argumentative zoning has been incorporated into citation recommendation systems, showing that aligning rhetorical structure with user queries improves both retrieval and contextual relevance (Ma et al., 2025).

Alongside function and recommendation, the problem of correctly identifying the cited source within complex citation contexts has been tackled with novel disambiguation strategies that aim to increase precision in large-scale analyses (Yang et al., 2025). Other work has addressed classification challenges by developing multi-task and ensemble learning frameworks that improve the accuracy of citation intent recognition (Jiang, 2025; Xu et al., 2025). These approaches illustrate the technical

maturation of the field, where feature engineering is gradually supplanted by deep learning and hybrid architectures.

Citation contexts are also increasingly applied to questions of research quality and integrity. Several studies have examined how retracted publications continue to be cited, often without acknowledgment of their retracted status. For example, some analyses reveal that retracted articles can still attract substantial attention, suggesting systemic weaknesses in how corrections to the scientific record are disseminated (Schmidt, 2024). Others have compared patterns across different classes of retracted papers, showing how citation behavior varies depending on the reason for retraction (da Silva & Nazarovets, 2023). Further work has explored how retraction is perceived through citation contexts, highlighting that many citances remain neutral or even positive in tone despite the retraction (Liu et al., 2022). These findings illustrate that content-based approaches are not limited to function classification but can also illuminate broader evaluative and ethical dimensions of research practice.

The application of large language models to citation context analysis has also opened a new research frontier, where the interpretive and generative capacities of such models are tested against established citation analysis tasks (Nishikawa & Koshiba, 2024). This strand builds directly on earlier NLP-driven work but raises new challenges about reproducibility, transparency, and the interpretive authority of AI in bibliometrics.

Recent research has also increasingly emphasized document-level and context-aware approaches to citation function classification. Studies on multi-label and multi-sentence citation analysis demonstrate that citation functions frequently extend beyond isolated sentence boundaries and may involve overlapping rhetorical roles distributed across broader argumentative structures (Lauscher et al., 2022; Zhang et al., 2021). Other work highlights the growing role of native contextual cues, discourse structure, and large language models in improving citation interpretation and contextual sensitivity within automated systems (Zhang et al., 2022, 2023). Collectively, these developments reinforce the importance of citance segmentation and contextual scope as central methodological concerns in citation analysis.

In recent years, content-based citation analysis has become a rapidly expanding area of research, with new studies published every year that push beyond the limits of traditional citation analysis. The growing interest reflects both methodological advances and the recognition that conventional citation counts cannot capture the full complexity of scholarly communication. The field now draws on deep learning, semi-supervised learning,

and large language models to address increasingly nuanced questions about citation function, sentiment, and epistemic context. Several reviews have already synthesized earlier phases of this literature (Ding et al., 2014; Iqbal et al., 2020), yet the pace of change and the arrival of new generation techniques underscore the need for an updated review that summarizes recent developments and evaluates their implications for research evaluation and the study of scientific communication.

Overall, the field has progressed from conceptual and rhetorical classifications to complex machine learning approaches that address function, sentiment, recommendation, and evaluation. At the same time, it continues to grapple with questions of validity, interpretive consistency, and the broader epistemic consequences of how citations are classified and used.

3 | METHODOLOGY

The methodology of this study is organized into several interrelated stages. First, I describe the process of identifying the dataset, including the sources from which citances were retrieved and the criteria applied to ensure relevance and quality. Second, I outline the sample selection procedure, clarifying the scope and representativeness of the data. Third, I present the classification scheme that guided the tagging process. This is followed by an explanation of expert tagging, including the background and expertise levels of the annotators. Finally, I detail the analysis and visualization process of this paper.

3.1 | Identifying dataset

This study builds upon a preliminary effort to map the publication and citation patterns of the LIS field (Taşkın, 2021). That earlier study assembled a century-long dataset of 123,742 articles indexed in the Web of Science (WoS) core indexes (SCI, SSCI, and A&HCI), covering the period 1921–2018. This corpus provided one of the most comprehensive baselines available for LIS and offered an empirically grounded classification of the field into distinct sub-fields, as well as detailed descriptive insights into its publication and citation dynamics.

In the present research, the same dataset of 123,742 articles was used as the foundation. The classification and descriptive findings established in the preliminary study serve here as a validated framework for selecting, organizing, and analyzing citation contexts. By relying on this established groundwork, the current study ensures both continuity and comparability, while also extending

TABLE 1 Descriptive characteristics of the stratified sample.

Subject category	<i>n</i> of pub
Librarianship and law librarianship (LIBR)	324
Scientometrics and information retrieval (SCIE)	146
Management and information systems (MIS)	119
Health information in LIS (HEALT)	74
Publication years	<i>n</i> of pub
1921–1950	11
1951–1980	97
1981–1990	85
1991–2000	115
2001–2010	157
2011–2018	198
Publication country by region	<i>n</i> of pub
North America	290
Europe and Central Asia	146
No affiliation data	116
East Asia and Pacific	73
Middle East and North Africa	12
Sub Saharan Africa	10
South Asia	8
Latin America and Caribbean	6
2 or more regions' collaboration	2
Publication country by income	<i>n</i> of pub
High income country	487
No affiliation data	116
Upper middle income country	45
Lower middle income country	13
Upper and high income country collaboration	2
Number of collaborating authors	<i>n</i> of pub
Single author	325
Two authors	163
Three authors	90
Four or more authors	85
Number of citations	<i>n</i> of pub
1–4 citations	233
5–13 citations	157
14 or more citations	159
No citation	114
Number of references	<i>n</i> of pub
1–9 references	174
10–22 references	164
23–42 references	162
43 or more references	163

the analysis to the interpretive and methodological challenges of tagging citances.

It should be noted that while WoS offers structured metadata and robust citation linking, its subject classification introduces well-known limitations (Milojević, 2020). The LIS category includes journals from neighboring domains (e.g., management information systems, health informatics) and omits many regional or practice-oriented journals central to the profession (Ni & Sugimoto, 2011; Taşkın, 2021; Tseng & Tsay, 2013). As a result, the dataset must be understood as a structured and analyzable proxy for LIS rather than a definitive representation of the entire global field.

3.2 | Sample selection

Given the size and heterogeneity of the LIS corpus, it was neither feasible nor necessary to analyze all 123,742 articles. Instead, a stratified sampling approach was employed to ensure that the selected sample represented the diversity of the field. Seven strata were defined based on established descriptive dimensions of LIS scholarship: (1) the four main sub-field categories identified in the preliminary study (librarianship, scientometrics/information retrieval, management and information systems, and health information in LIS); (2) number of authors as a proxy for collaboration patterns; (3) publication period; (4) and (5) affiliated countries grouped by both income level and geographical region; (6) citation counts; and (7) reference list length.

From this framework, a total of 663 articles were selected. This sample size was calculated to provide representation of the population at the 99% confidence level, ensuring that the findings could be generalized while remaining analytically manageable. The distribution of the sample across each stratum was aligned with the distribution of the full population, so that the share of the universe and the share of the sample run in parallel. In this way, the use of stratified sampling minimized the risk of overrepresenting highly productive sub-fields (e.g., scientometrics and information retrieval) or dominant regions (e.g., North America and Western Europe), while also guaranteeing the inclusion of publications from less represented contexts, such as lower-middle income countries and regions outside the Global North.

The descriptive characteristics of the sample are presented in Table 1. This multidimensional stratification ensured that the sample captures the core of the LIS literature together with its disciplinary, temporal, geographic, and bibliometric diversity. The resulting 663-article corpus thus provides a robust foundation for the subsequent analyses of citation contexts and tagging practices.

Eng site Hi Zehra Logout | Home | Cite | Tagged citations | Reference raw list |

Select Year: 2018 | Select an Article: LIBR7428 | Article ID: LIBR7428 PDF HTML

libraries 827

is offering new potential for researchers to harness the new technologies and offer limitless digital information content through a single package (Garcia *et al.*, 2011). The Semantic Web is useful for enhancing access to information and knowledge. There is a strong nexus between the mission and aims of libraries and the objectives of the Semantic Web. The Semantic Web is helpful to increase usability of digital data. It offers Semantic Web representation, semantic outreach and Semantic Web services and resources (Greenberg, 2009).

The new form of digital library will be the semantic digital library (Alotaibi, 2010). The use of Semantic Web and resource description and framework (RDF) technology in digital libraries will aid in managing metadata and digital contents, as well as offer new techniques for the discovery of information (Han, 2006). The Semantic Web will increase the usability of library material on a large scale. It is an important source to access information available in digital form. There is a need to investigate the uses of Semantic Web and web ontology language to develop a richer conceptual model to be shared on the web (Ferguson, 2007).

Semantic Web technology

The development of the web is divided into four major eras: that is, Web 1.0, Web 2.0, Web 3.0 and Web 4.0. The Web 1.0 era encompasses 1990-2000, Web 2.0 2001-2010, and the Web 3.0 and Web 4.0 eras are envisioned from 2011 to 2020 and from 2021 to 2040, respectively (Aghaei *et al.*, 2012). It is believed that by using mashup technologies and artificial intelligence on the web, Web 3.0 will bring information together in a more meaningful way (Das, 2012).

Berners-Lee *et al.* (2001) explained that the Semantic Web brings structure to the meaningful content of web pages by allowing software agents to roam from page to page to carry out sophisticated work for the users. The Semantic Web defines the information in a more meaningful way for computers and people to work in cooperation. By exploiting the value of the Semantic Web, the computer will better understand and process the data not merely to display it but also to retrieve it in an effective way.

The basic function of the Semantic Web is to use the extensible markup language (XML) which allows the use of tags to describe digital objects. XML developers relationship between

Citation

The development of the web is divided into four major eras: that is, Web 1.0, Web 2.0, Web 3.0 and Web 4.0. The Web 1.0 era encompasses 1990-2000, Web 2.0 2001-2010, and the Web

Press to copy the highlighted text Clear

Ref: None selected 30 / 30 ref.

Meaning: Neutral Posit[+] Negat[-] Pos & Neg[+/-]

Purpose: Literature Definition Method Comparison
Data Validation

Section: Introduction Methodology Research Dis & Conc
Not IMRAD Other

Shape: Mention multiple Quotation N/A

SAVE

FIGURE 1 Tagging interface.

3.3 | Classification scheme

For the content-based analysis of citances, I relied on the classification scheme originally developed by Taşkın and Al (2018). That scheme was designed to capture both semantic and syntactic dimensions of citations and introduced four overarching categories: meaning, purpose, shape, and array. In its initial formulation, each of these categories was further subdivided into fine-grained classes. For example, meaning distinguished between positive, negative, and neutral citations; purpose was coded as literature, definition, data, method, or validation; shape covered author mentions, quotations, and multiple citations; and array referred to section, frequency of occurrence, and distribution across sections.

In the present study, I concentrated on two categories, meaning and purpose, as these dimensions are central to understanding the evaluative tone and functional role of citations. The other two categories (shape and array) were coded in the dataset but were not included in the statistical analysis presented here. This approach allowed me to focus more directly on interpretive challenges while still producing an openly shared dataset that covers all four dimensions. By making the complete dataset available, I aim to support continuity with prior research and enable future studies to explore shape and array classifications alongside meaning and purpose.

3.4 | Designing web interfaces

To facilitate the systematic tagging of citances, a dedicated web-based interface that supported both data management and expert annotation was developed. The interface was designed with two main screens. On the left side, the full text of the article was displayed, while the right side contained the tagging panel (Figure 1).

All articles in the sample were first downloaded in PDF format and converted into rich text format to enable processing. In cases where PDFs lacked text recognition, optical character recognition (OCR) was applied to ensure accurate text extraction. Once prepared, the files were uploaded to the system and linked to a relational database that stored all metadata, including bibliographic information and full reference lists.

The tagging interface allowed annotators to select citances directly within the text by highlighting them with the cursor. The selected text was then automatically copied into the tagging panel. At this stage, the tagger classified the citance according to the four main categories of the classification scheme (meaning, purpose, shape, and array) using drop-down menus. A key feature of the system was the integration of reference lists: taggers were required to assign each citance to the correct reference from a structured drop-down menu populated by the database. This ensured that every citance was linked to its cited source and that tagging was consistent across annotators.

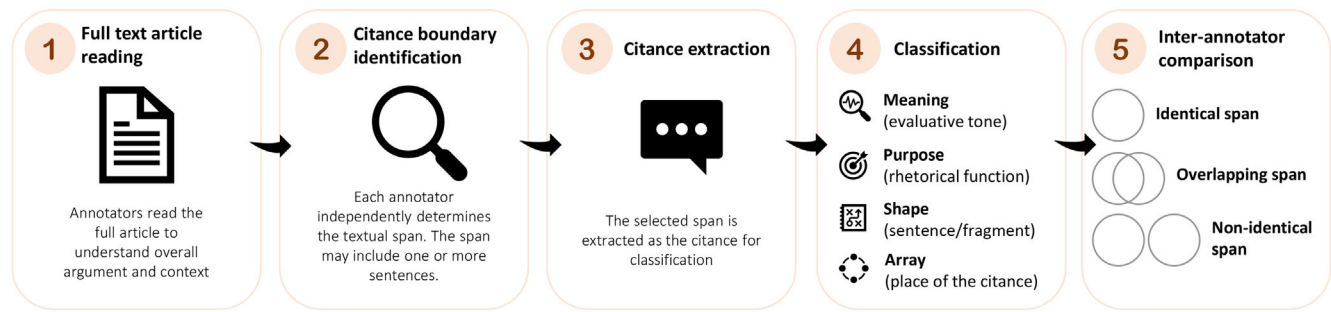


FIGURE 2 Workflow of citance identification and annotation.

By combining automated preprocessing (PDF conversion, OCR, and database integration) with structured manual tagging, the interface reduced annotation errors and streamlined the workflow for expert raters. The design also facilitated inter-annotator agreement testing, as each tagged citance was stored with user identifiers and version control in the backend database.

3.5 | Annotators

The tagging process was carried out by two annotators with different types of expertise in the LIS field. The first annotator is an associate professor with extensive academic experience in LIS (the author). She has published widely in the field and has been actively engaged in research on scholarly communication, citation analysis, and bibliometrics. Her long-standing involvement in the field ensured a deep understanding of both the disciplinary context and the interpretive nuances of citances.

The second annotator is a graduate of the same LIS department who currently works for the Turkish national citation index (TR Dizin). His professional role has provided him with substantial familiarity with citation data, indexing practices, and the technical aspects of scholarly publishing. While he does not have research experience comparable to the first annotator, his daily exposure to the infrastructure of citation indexing systems equipped him with practice-oriented expertise and a solid conceptual grasp of how citations function in scholarly communication.

By including annotators with these complementary profiles, the study was able to compare tagging practices across academic research expertise and indexing-based professional expertise. This design made it possible to evaluate how scholarly research experience versus technical familiarity with citation systems shaped the interpretation and classification of citances.

Figure 2 summarizes the annotation workflow used in the study. The process involved independent

identification of citance boundaries, extraction of the relevant textual span, and subsequent classification according to meaning, purpose, shape, and array dimensions. The workflow also illustrates how interpretive variability may emerge during citance identification prior to classification itself.

3.6 | Analysis

The statistical analyses were conducted using SPSS for inferential testing and reliability assessment, while R Commander with the *kmggplot* plug-in was employed to generate visualizations and descriptive plots. The analysis followed a stepwise procedure to ensure both the robustness of statistical inference and the clarity of graphical presentation.

First, descriptive statistics were computed to summarize the main characteristics of the sample, including frequencies, distributions, and measures of central tendency. These provided a baseline understanding of the dataset before conducting inferential tests. Second, inferential analyses were performed to examine potential differences and relationships between variables. Depending on the type and distribution of the data, appropriate non-parametric tests were selected. For comparisons between groups, chi-square tests and Mann–Whitney U tests were applied to categorical and ordinal data, while the Kruskal–Wallis test was employed to evaluate differences across multiple groups. In each case, significance levels were set at $p < 0.05$.

To complement statistical significance testing, effect sizes were calculated using Cohen's formulas (Cohen, 2013), providing an indication of the magnitude of observed effects beyond p-values. Following conventional thresholds, effect sizes were interpreted as small, medium, or large, thereby offering a more nuanced understanding of the substantive importance of the results. In addition, inter-annotator reliability was assessed using Cohen's kappa coefficient (Landis & Koch, 1977), which measures the extent of

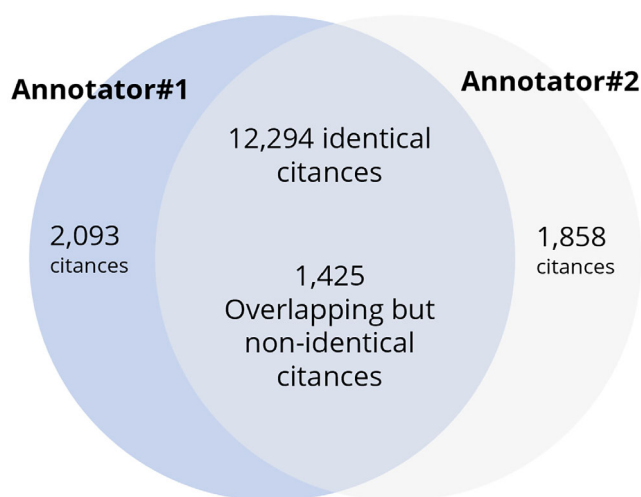


FIGURE 3 Overlap of citances identified by Annotator#1 and Annotator#2.

agreement between annotators beyond chance. Kappa values were interpreted according to conventional benchmarks, ranging from slight to almost perfect agreement, thereby enabling a more robust evaluation of consistency across coding categories.

Finally, results were visualized using R Commander with the *kmggplot* package, which allowed for the production of distribution plots and boxplots. These visualizations were designed to illustrate statistical outcomes and highlight patterns of agreement, divergence, and variability across annotators. The combination of statistical testing, effect size estimation, and visualization ensured that the analysis was both methodologically rigorous and easily interpretable.

4 | FINDINGS

4.1 | Descriptives

Figure 3 presents the overlap and divergence between the two annotators in tagging citances. Examples of identical, overlapping, and non-identical citance selections are provided in Appendix A (Table A1). While the majority of citances (12,294) were identically tagged, the distribution of unique citances is particularly revealing: Annotator#1 identified 2093 citances that were not captured by Annotator#2, while Annotator#2 found 1858 unique citances. These unique cases suggest that even with close reading, human annotators may overlook certain citances depending on their familiarity with the structure of scientific texts. In this respect, Annotator#1's deeper research experience may have increased sensitivity to contextual and rhetorical cues that shaped citance identification.

This variability underscores the potential of AI-based or automated citance detection techniques. Unlike human annotators, automated systems are less constrained by fatigue and can potentially provide more exhaustive coverage of citation contexts. The fact that each annotator independently captured several thousand citances missed by the other illustrates the limitations of relying solely on manual identification, reinforcing the importance of automation for scalability and reproducibility.

At the same time, the 1425 overlapping but non-identical citances draw attention to another methodological challenge: the definition of citance length. Annotators agreed on the presence of a citation but differed on the precise textual boundaries, revealing the “text span problem.” These boundary discrepancies highlight a fundamental limitation that affects both manual and automated approaches, emphasizing the need for consistent operational rules for citance segmentation.

4.2 | Length of citances

Determining the precise boundaries of a citation context is one of the most persistent challenges in content-based citation analysis. Nakov et al. (2004) introduced the term *citance* to denote the sentence or sentences surrounding a citation, arguing that citation contexts often extend beyond a single sentence and may instead encompass multiple sentences or fragments. They identified the “text span problem” as a central issue: deciding whether a citance should be treated as a phrase, a clause, a single sentence, or several sentences together. Later work confirmed these difficulties. Abu-Jbara and Radev (2012) demonstrated that identifying the scope of a reference is especially complex when citations are grouped or spread across sentences. Similarly, Taşkın and Al (2018) noted that what is often coded as a “citation sentence” may in fact consist of several sentences, complicating manual and automated tagging. More recently, Taşkın et al. (2023) observed that entire paragraphs sometimes needed to be treated as citances, since relying on heuristics such as linking words (however, although) proved insufficient for determining where a citation begins and ends. Collectively, these studies highlight that the citance, rather than the citation sentence, is the more appropriate analytical unit, making it important to examine the actual distribution of citance length in empirical research.

As Figure 4 illustrates, most identical citances consist of a single sentence, with longer cases appearing only as rare outliers. In contrast, both non-identical and unique citances show systematically longer spans, reflected in higher medians and wider distributions. This confirms

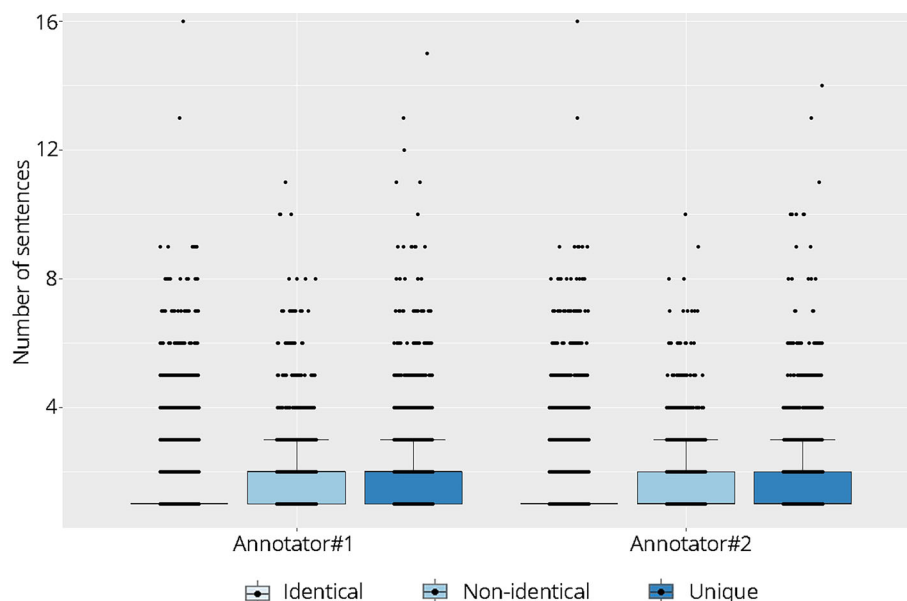


FIGURE 4 Distribution of citance length across identical, non-identical, and unique cases.

that the segmentation problem becomes more pronounced as citances extend beyond a single sentence, creating greater potential for annotator disagreement and interpretive divergence. The fact that annotators did not always capture the same sentences may therefore be explained by citance length itself: Once a reference extended over multiple sentences, annotators were more likely to identify different contextual boundaries and select partially overlapping citances. The Kruskal–Wallis test showed highly significant differences in length across identical, non-identical, and unique citances ($\chi^2(2) = 4297, p < 0.001$).

4.3 | Inter-annotator agreement in identical citances

Assessing inter-annotator agreement on identically tagged citances provides an essential benchmark for evaluating the reliability of the classification scheme and the consistency of annotators' judgments. Because identical citances refer to cases where both annotators recognized and tagged the same text span, the resulting agreement rates are less affected by segmentation issues and more directly capture concordance in applying meaning and purpose categories. By calculating measures such as Cohen's kappa, it is possible to quantify the extent to which the observed agreement exceeds chance, providing a more robust indication of reliability than simple percentage agreement.

The confusion matrix for the meaning class shows that the majority of citances were coded identically, with 11,724 jointly identified as neutral (Table 2). This dominance of neutral cases contributes to the overall high

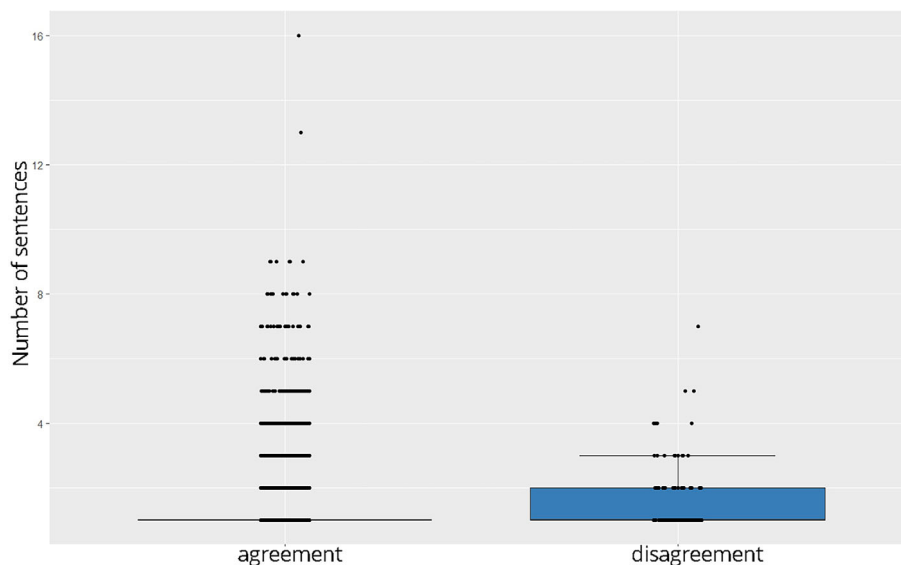
agreement, but the strength of alignment also extends to other categories. The Cohen's kappa coefficient of 0.894 ($p < 0.001$) falls into the “almost perfect” range, and this result reflects not only the large neutral base but also substantial consistency in positive, negative, and mixed classifications. These results indicate that annotators share a strong baseline in recognizing neutral citations and also achieve notable consistency in identifying evaluative tones. Although divergences are more likely in positive, negative, and mixed categories, the overall pattern confirms that judgments were largely stable across meaning classes, providing confidence in the reliability of the classification scheme when citances are defined identically.

To explore potential factors that might influence the classification of citances, I conducted additional tests. First, I examined whether the subject category of the citation (the four main LIS domains) created significant differences in annotator agreement using a chi-square test. Although the result was statistically significant ($\chi^2(3) = 8.158, p = 0.043$), the effect size was negligible (Cramer's $V = 0.026$), indicating no meaningful difference across domains. I then tested whether the number of references cited within a single sentence affected annotator disagreement by applying the Mann–Whitney U test. The result was not significant ($U = 638,910, Z = -1.407, p = 0.159$), and the effect size was minimal ($r_G = 0.013$). In contrast, the length of the citance had a significant impact ($U = 544,178.5, Z = -5.964, p < 0.001$), with a small effect size ($r_G = 0.053$). These findings suggest that annotator differences are not shaped by subject category or the number of references cited, but rather by sentence length, which increases the likelihood of divergent interpretations as citances become longer. As

TABLE 2 Confusion matrix showing agreement on meaning class (identical citances).

		Annotator#2			
		Neutral	Positive	Negative	Positive and negative
Annotator#1	Neutral	11,724	30	13	3
	Positive	44	396	0	1
	Negative	9	0	47	0
	Positive and negative	2	3	1	21

Note: Darker red cells indicate agreement between annotators (diagonal cells), while lighter shades indicate lower levels of agreement. The color gradient is used to facilitate visual interpretation of the confusion matrix and to highlight patterns of inter-annotator agreement.

FIGURE 5 Distribution of citance length in sentences across agreement and disagreement cases.**TABLE 3** Confusion matrix showing agreement on purpose class (identical citances).

		Annotator#2					
		Comparison	Data	Definition	Literature	Methodology	Validation
Annotator#1	Comparison	798	41	15	493	29	12
	Data	25	318	1	96	10	2
	Definition	13	0	413	278	12	2
	Literature	249	25	137	7561	129	31
	Methodology	16	5	12	250	911	10
	Validation	19	12	4	117	12	236

Figure 5 shows, the vast majority of agreements occur in one-sentence citances, while the dots in the agreement group represent occasional longer cases that are statistical outliers. In disagreements, however, longer citances are more common, indicating that once a citance extends beyond a single sentence, annotators are more likely to diverge in their interpretations.

To examine how annotators converged or diverged when classifying the purpose of identical citances, I calculated Cohen's kappa alongside the confusion matrix.

The kappa coefficient was 0.667 ($p < 0.001$), which indicates substantial agreement but is notably lower than the almost perfect agreement observed for the meaning class. This difference shows that annotator perceptions diverged more often when judging the rhetorical or functional role of citations than when judging their evaluative tone.

The confusion matrix (Table 3) demonstrates that most citances were consistently coded as literature (7561 cases), forming the backbone of the observed agreement.

However, a considerable portion of the disagreements involved spillovers into the literature category: for example, 493 citances coded as comparison by Annotator#1 were tagged as literature by Annotator#2, and 278 citances coded as definition by Annotator#1 were also categorized as literature by Annotator#2. Additional inconsistencies appeared between methodology and literature (250 cases). These patterns suggest that while annotators shared a broad understanding of what counts as literature-driven citations, they differed in how narrowly or broadly they interpreted comparison, definition, or methodology as distinct functions.

In short, the purpose class reveals how expertise and interpretive stance influence tagging: agreement was strong overall, but the boundaries between literature-related functions were more porous and therefore more susceptible to subjective judgment.

To explore the factors that might account for annotator differences in the purpose classification, I conducted supplementary statistical analyses. First, I tested whether the subject category of the citing article (four LIS domains) influenced inter-annotator agreement. The chi-square test was statistically significant ($\chi^2(3) = 14.450$, $p = 0.002$), but the effect size was very small (Cramer's $V = 0.034$). This suggests that although minor variations exist across domains, subject category does not meaningfully explain the observed differences. Second, I examined whether the number of references cited in a single sentence contributed to disagreements by applying the Mann-Whitney U test. The result was not significant ($U = 10,936,368$, $Z = -0.182$, $p = 0.856$), with a negligible effect size ($r_G \approx 0.002$). Thus, the number of citations within a sentence did not increase the likelihood of divergent judgments. Finally, I tested whether citance length (measured in number of sentences) had an effect. Here, the Mann-Whitney U test was significant ($U = 10,464,266$, $Z = -5.365$, $p < 0.001$). The effect size was small but meaningful ($r_G \approx 0.048$), showing that disagreements were more frequent in longer citances. This indicates that, similar to the findings for meaning, longer textual contexts increase interpretive variation between annotators in purpose classification.

The agreement observed in identical citances suggests that differences in expertise had limited influence when annotators were working with the same textual unit. Once the boundaries of a citance were fixed, both the experienced researcher and the indexer drew on comparable cues in the text to classify meaning and purpose, leading to convergence in their judgments. This indicates that the role of expertise becomes less decisive under conditions of shared segmentation, where interpretive discretion is minimized and the classification task is more straightforward. In such cases, the consistency of results

is driven more by the clarity of the textual evidence than by differences in professional background.

4.4 | Inter-annotator agreement in overlapping but non-identical citances

When turning to the overlapping but non-identical citances, that is, cases where annotators agreed on the presence of a citance but marked different text spans, the results reveal a markedly lower level of agreement. The Cohen's kappa for the meaning category dropped to 0.261 ($p < 0.001$), which corresponds only to fair agreement. This sharp decline compared to identical citances underscores the impact of the "text span problem": once annotators selected different sentence boundaries, their subsequent judgments about meaning diverged much more often.

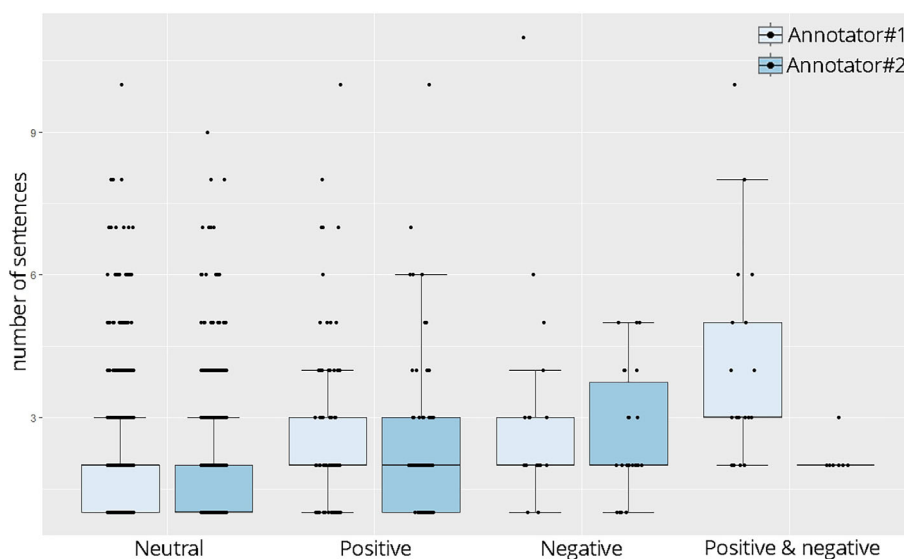
The confusion matrix (Table 4) illustrates this pattern. Although 1236 citances were jointly coded as neutral, disagreements were distributed across evaluative categories. For example, 54 citances labeled as positive by Annotator#1 were classified as neutral by Annotator#2, while 17 citances marked as neutral by Annotator#1 were tagged as negative by Annotator#2. Even within the small set of mixed cases, inconsistencies occurred, such as three citances tagged as positive by Annotator#2 but as mixed by Annotator#1. These discrepancies indicate that without having the full citance in view, annotators could not always capture the context, which led to differences in how they interpreted the evaluative tone.

The disagreements in non-identical citances can be grouped into several systematic patterns. The most central factor is citance length. When annotators did not select the same sentence span, the full evaluative context of the author was not always visible. In such cases, one annotator often coded the citance as neutral, while the other, who included an additional sentence, identified it as positive or negative. This shows that when the full stretch of text is not captured, the underlying tone or emotion of the author is easily overlooked, and this omission becomes the primary source of disagreement. This tendency is confirmed in Figure 6, which shows that longer citances are more often associated with disagreement. Especially in positive, negative, and mixed categories, annotators' distributions diverge as citance length increases, underlining how segmentation decisions shape the perceived evaluative tone.

In the case of non-identical citances, differences between annotators became more apparent, particularly in evaluative categories. While both annotators produced nearly identical lengths for neutral citances, the distributions for positive, negative, and mixed

TABLE 4 Confusion matrix showing agreement on meaning class (non-identical citances).

		Annotator#2			
		Neutral	Positive	Negative	Positive and negative
Annotator#1	Neutral	1236	46	17	5
	Positive	54	27	2	2
	Negative	14	1	3	0
	Positive and negative	14	3	0	1

FIGURE 6 Distribution of citance length across annotator labels for non-identical citances.

cases diverged. Here, the first annotator's disciplinary expertise led to systematically longer citances, suggesting that greater familiarity with scholarly argumentation encouraged the inclusion of broader contextual spans. By contrast, the second annotator tended to delimit citances more narrowly, likely reflecting a procedural and indexing-oriented approach. This pattern indicates that expertise exerts its strongest influence when evaluative or ambivalent tones are present, as the academic annotator is more inclined to read beyond minimal sentence boundaries to capture nuance, whereas the indexer emphasizes concise, technically bounded units.

A second cluster of disagreements is linked to contrastive constructions. When citances contained expressions such as *although*, *however*, or similar connectors, the annotators diverged in their interpretation. One annotator tended to see these phrases as neutral signals that simply contextualize prior studies, whereas the other understood them as carrying an implicit critique and therefore coded the citance as negative. This indicates that the presence of contrastive markers leaves greater room for subjective interpretation of whether a statement should be read as evaluative.

A third cluster emerges from citances with ambivalent or dual tones. In these cases, the sentence combined a form of recognition with a limitation or critique. One annotator consistently acknowledged both aspects by assigning a mixed (positive-negative) label, while the other simplified the classification into a single polarity, usually positive. These differences highlight how annotators varied in their sensitivity to subtle shifts in tone within the same textual unit.

Finally, disagreements were also observed in general descriptive formulations that referred to groups of studies or approaches. One annotator treated such phrasing as neutral background information, whereas the other considered it to imply endorsement and therefore coded the citance as positive. This shows that even in seemingly descriptive sentences, interpretive differences can arise about whether the author's intention was to merely report or to evaluate.

For the purpose classification of non-identical citances, the level of inter-annotator agreement decreased substantially compared to identical cases. The Cohen's kappa coefficient was 0.412 ($p < 0.001$), which corresponds to moderate agreement. This drop illustrates once again that differences in sentence

TABLE 5 Confusion matrix showing agreement on purpose class (non-identical citances).

		Annotator#2					
		Comparison	Data	Definition	Literature	Methodology	Validation
Annotator#1	Comparison	83	4	2	109	13	5
	Data	3	22	0	8	3	2
	Definition	3	0	19	42	11	1
	Literature	75	4	12	681	48	11
	Methodology	9	3	6	62	113	5
	Validation	6	1	0	32	3	24

segmentation amplify interpretive variation, making it more difficult for annotators to converge on the functional role of a citation when they did not highlight the exact same span.

The confusion matrix (Table 5) shows that most agreements clustered around the literature category (681 cases), which dominates the distribution. However, a significant portion of the disagreements involved spills over between literature and other functions. For example, 109 citances coded as comparison by Annotator#1 were categorized as literature by Annotator#2, while 42 citances marked as definition by Annotator#1 were also assigned to literature by Annotator#2. Additional discrepancies were observed between methodology and literature (62 cases). These patterns suggest that, in longer or partially overlapping citances, annotators struggled to consistently separate functions such as comparison, definition, and methodology from the broader frame of literature review.

When compared to the meaning classification in non-identical citances ($\kappa = 0.261$), the purpose classification ($\kappa = 0.412$) shows a slightly higher but still only moderate level of agreement. This contrast underlines that while both dimensions are strongly affected by segmentation issues, purpose labels allow somewhat more common ground between annotators than evaluative meaning, which is more sensitive to missing contextual information.

5 | DISCUSSION

This study demonstrates that while automated tools offer scalability in citation context analysis, human annotation continues to be shaped by interpretive complexity. The comparison between two annotators with different professional profiles provides insights into how subject expertise influences tagging practices and how these effects interact with methodological challenges such as citance segmentation.

5.1 | Expertise and reliability

The results show that when annotators worked with identical citances, agreement was consistently high, particularly in the meaning category. This indicates that once textual boundaries were fixed, expertise exerted limited influence on reliability. Both the academic researcher and the indexer relied on similar textual cues, suggesting that the clarity of the citance itself outweighed individual differences in interpretive stance. In other words, agreement was driven less by expertise than by the reduced discretion involved in classifying a shared text span.

By contrast, in non-identical citances, expertise played a more visible role. The academic annotator systematically included longer spans, particularly in evaluative categories, reflecting a tendency to capture broader argumentative context. The Annotator#2, in turn, adopted a more procedural approach, delimiting citances more narrowly. These differences underscore what I describe as the expertise effect: subject knowledge deepens interpretive engagement but may also expand the unit of analysis in ways that complicate consistency. This pattern suggests that expertise does not always lead to convergence in classification outcomes but can, under certain conditions, amplify divergence.

5.2 | The segmentation problem

The “text span problem” emerges as a central methodological barrier. Agreement levels dropped sharply when annotators diverged in defining citance boundaries, highlighting that segmentation decisions shape subsequent judgments of meaning and purpose. This finding echoes prior studies noting that citances extend variably across phrases, sentences, or even paragraphs (Abu-Jbara & Radev, 2012; Nakov et al., 2004; Taşkın & Al, 2018). The strong effect of citance length in the statistical tests reinforces this point: the longer the span, the

higher the likelihood of disagreement. Automated systems face similar challenges, which suggests that improving operational rules for citance definition is critical for both human and machine-driven analyses.

The findings also suggest that disagreement in citance segmentation should not be interpreted solely as annotator inconsistency or cognitive bias. Citation contexts do not exist as fully objective textual units waiting to be extracted. Instead, annotators construct citances by interpreting rhetorical continuity, evaluative meaning, and contextual relevance within the surrounding discourse. As a result, segmentation variability reflects an inherent characteristic of citation context interpretation rather than a simple annotation failure.

The findings further indicate that citation meaning and rhetorical function cannot always be reliably interpreted within isolated sentence boundaries. In many cases, evaluative tone, comparison, or methodological positioning emerges across multiple sentences and depends on broader argumentative continuity. This helps explain why longer citances produced greater disagreement in the study and why fixed-window approaches may fragment rhetorical meaning by excluding adjacent contextual cues.

This perspective has important implications for reproducibility in content-based citation analysis. High inter-annotator agreement may depend as much on shared operational assumptions about citance boundaries as on agreement in classification categories themselves. Consequently, methodological transparency in how citances are identified becomes essential for interpreting and comparing findings across studies, datasets, and computational frameworks.

5.3 | Meaning and purpose classifications

Another important finding concerns the difference between meaning and purpose classifications. While annotators reached almost perfect agreement on evaluative tone in identical citances, agreement levels were notably lower in the purpose dimension. Disagreements clustered around literature-related categories such as comparison, definition, and methodology, revealing blurred boundaries between functional roles. This aligns with earlier critiques that citation function categories often overlap in practice and require continuous refinement. It also suggests that expertise shapes interpretation differently across dimensions: whereas evaluative judgments benefit from clearer cues, functional judgments are more susceptible to variation.

The findings also point to the broader difficulty of imposing discrete functional categories on citation

contexts that may simultaneously perform multiple rhetorical roles. In practice, citances frequently combine comparison, methodological reference, evaluation, and background framing within the same argumentative unit. Such overlaps increase interpretive variability and complicate efforts to establish fully standardized annotation frameworks across datasets, domains, and research traditions. The results therefore suggest that differences in segmentation and operational definitions may influence inter-annotator agreement as much as the classification categories themselves.

The findings do not suggest that expert knowledge eliminates interpretive ambiguity or guarantees universally consistent classification. Instead, expertise appears to increase sensitivity to contextual nuance, rhetorical structure, and implicit scholarly positioning within citation contexts. From this perspective, the value of expert interpretation for automated systems lies less in producing perfectly stable labels and more in revealing the contextual signals and ambiguities that simplified annotation frameworks tend to overlook. These observations support the development of computational approaches that incorporate contextual uncertainty, document-level interpretation, and more transparent operational definitions of citation contexts.

6 | CONCLUSIONS

6.1 | Implications for automated and human tagging

These findings carry two main implications. First, they caution against assuming that more expertise automatically improves reliability in content-based citation analysis. Instead, expertise appears to operate as a double-edged factor. It can reduce bias in some contexts while increasing variability in others. Second, the study highlights that inter-annotator agreement cannot be disentangled from segmentation practices. Any attempt to benchmark annotation quality, whether human or automated, must therefore account for the fundamental role of citance definition.

The findings also carry several implications for computational citation analysis research. First, expert-labeled data may be most valuable not as a source of universally stable ground truth, but as a means of identifying contextual ambiguity and rhetorical complexity within citation contexts. Second, annotation inconsistency should be treated as an informative characteristic of citation interpretation rather than solely as an annotation error. Finally, automated segmentation and classification systems may benefit from approaches that incorporate

contextual uncertainty, document-level interpretation, and transparent operational definitions of citance boundaries. Such strategies may help bridge human interpretive sensitivity with the scalability of computational methods.

6.2 | Contribution to the field

By situating human annotation within the broader landscape of automated classification, this study contributes to ongoing debates about validity, reproducibility, and the interpretive limits of content-based citation analysis. The findings suggest that the future of the field lies not in replacing human expertise with automation, but in designing systems that integrate the strengths of both: the scalability of machine learning and the contextual sensitivity of expert interpretation. For research evaluation and scholarly communication studies, this means that advances in AI should be accompanied by critical awareness of the interpretive challenges embedded in citation practices.

The study therefore does not aim to propose a scalable annotation infrastructure for large bibliometric databases or automated citation indexing systems. Instead, it focuses on the epistemic and methodological challenges that emerge during citance identification and interpretation. By foregrounding these issues, the findings provide a basis for more critically evaluating how automated citation analysis systems operationalize citation contexts at scale.

Future research may also benefit from incorporating annotators with little or no disciplinary expertise in order to further examine how professional background, domain familiarity, and interpretive experience shape citance identification and classification practices.

ACKNOWLEDGMENTS

I would like to thank NAWA Poland and the AMU Scholarly Communication Research Group for the great 2 years full of research and collaboration. I am especially grateful to Abdülkadir Taşkın for serving as the second annotator in this study and for designing the databases and interfaces that made the tagging process possible.

FUNDING INFORMATION

This research was supported by a research grant from the Polish National Agency for Academic Exchange, NAWA Poland (PPN/ULM/2019/1/00062).

CONFLICT OF INTEREST STATEMENT

The author has declared no conflict of interest.

DATA AVAILABILITY STATEMENT

All data supporting the findings of this study is openly available in an online repository (<https://osf.io/a6m93/>).

ORCID

Zehra Taşkın  <https://orcid.org/0000-0001-7102-493X>

REFERENCES

- Abu-Jbara, A., & Radev, D. (2012). Reference scope identification in citing sentences. In *Proceedings of the 2012 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 80–90). Association for Computational Linguistics. <https://aclanthology.org/N12-1009>
- Anderson, M. H., & Lemken, R. K. (2023). Citation context analysis as a method for conducting rigorous and impactful literature reviews. *Organizational Research Methods*, 26(1), 77–106. <https://doi.org/10.1177/1094428120969905>
- Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session* (pp. 81–87). Association for Computational Linguistics. <https://aclanthology.org/P11-3015>
- Chatrinan, K., Noraset, T., & Tuarob, S. (2025). GAN-CITE: Leveraging semi-supervised generative adversarial networks for citation function classification with limited data. *Scientometrics*, 130(2), 679–703. <https://doi.org/10.1007/s11192-025-05233-1>
- Cohen, B. H. (2013). Statistical power and effect size. In *Explaining psychological statistics* (pp. 237–270). John Wiley & Sons Ltd.. <https://doi.org/10.1002/9781394260638.ch8>
- da Silva, J. A. T., & Nazarovets, S. (2023). Partial citation analysis of five classes of retracted papers, and devising a new four-tier citation classification system for retracted (and other) papers. *Scientometrics*, 128(8), 4887–4894. <https://doi.org/10.1007/s11192-023-04769-4>
- Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, 65(9), 1820–1833. <https://doi.org/10.1002/asi.23256>
- Donner, P., Stahlschmidt, S., Haunschild, R., & Bornmann, L. (2025). Does citation context information enhance the validity of citation analysis for measuring research quality? An empirical comparison of peer assessments and enriched citations. *Quantitative Science Studies*, 6, 967–987. <https://doi.org/10.1162/qss.a.15>
- Garfield, E. (1965). Can citation indexing be automated. In *Statistical Association Methods for Mechanized Documentation Symposium Proceedings, 1964*, (NBS26) (189–+). National Bureau of Standards Miscellaneous Publication.
- Iqbal, S., Hassan, S.-U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2020). A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. *arXiv:2008.13020 [Cs]*. <http://arxiv.org/abs/2008.13020>
- Jiang, X. (2025). Ensembling approaches to citation function classification and important citation screening. *Scientometrics*, 130(3), 1371–1419. <https://doi.org/10.1007/s11192-025-05265-7>

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Cohan, A., Jurgens, D., & Lo, K. (2022). MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In M. Carpuat, M.-C. de Marneffe, & I. V. M. Ruiz (Eds.), *Proceedings of the 2022 conference of the north American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 1875–1889). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.137>
- Li, M., Fahlström Myrman, A., Mu, T., & Ananiadou, S. (2019). Modelling instance-level annotator reliability for natural language labelling tasks. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Volume 1 (long and short papers)* (pp. 2873–2883). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1295>
- Liu, X., Wang, C., Chen, D.-Z., & Huang, M.-H. (2022). Exploring perception of retraction based on mentioned status in post-retraction citations. *Journal of Informetrics*, 16(3), 101304. <https://doi.org/10.1016/j.joi.2022.101304>
- Ma, S., Zhang, C., Zhang, H., & Gao, Z. (2025). Citation recommendation based on argumentative zoning of user queries. *Journal of Informetrics*, 19(1), 101607. <https://doi.org/10.1016/j.joi.2024.101607>
- Milojević, S. (2020). Practical method to reclassify web of science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, 1(1), 183–206. https://doi.org/10.1162/qss_a_00014
- Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, 5(1), 86–92.
- Nakov, P. I., Schwartz, A. S., & Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics*. Association for Computing Machinery (ACM).
- Narimanzadeh, H., Badie-Modiri, A., Smirnova, I. G., & Chen, T. H. Y. (2023). Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 334:1–334:29. <https://doi.org/10.1145/3610183>
- Ni, C., & Sugimoto, C. R. (2011). Four-facets study of scholarly communities: Artifact, producer, concept, and gatekeeper. In *Proceedings of the 2011 ASIS&T Annual Meeting*. Wiley. <https://doi.org/10.1002/meet.2011.14504801343>
- Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., & Rife, S. C. (2021). Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3), 882–898. https://doi.org/10.1162/qss_a_00146
- Nishikawa, K., & Koshiba, H. (2024). Exploring the applicability of large language models to citation context analysis. *Scientometrics*, 129(11), 6751–6777. <https://doi.org/10.1007/s11192-024-05142-9>
- Pride, D., & Knott, P. (2017). *Incidental or Influential?—Challenges in Automatically Detecting Citation Importance Using Publication Full Texts* (p. 578). Springer. https://doi.org/10.1007/978-3-319-67008-9_48
- Schmidt, M. (2024). Why do some retracted articles continue to get cited? *Scientometrics*, 129(12), 7535–7563. <https://doi.org/10.1007/s11192-024-05147-4>
- scite.ai. (2020). Journals: What is the scite index? <https://scite.ai/journals>
- Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, 87(2), 373–388. <https://doi.org/10.1007/s11192-011-0349-2>
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In M. Lapata & H. T. Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Association for Computational Linguistics. <https://aclanthology.org/D08-1027/>
- Taşkın, Z. (2021). Forecasting the future of library and information science and its sub-fields. *Scientometrics*, 126(2), 1527–1551. <https://doi.org/10.1007/s11192-020-03800-2>
- Taşkın, Z. (2025). Methodological Challenges in Content-Based Citation Analysis: Expertise, Reliability, and the Primacy of Citance Identification [dataset]. <https://osf.io/a6m93/overview>
- Taşkın, Z., & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics*, 114(1), 335–357. <https://doi.org/10.1007/s11192-017-2560-2>
- Taşkın, Z., Krawczyk, F., & Kulczycki, E. (2023). Are papers published in predatory journals worthless? A geopolitical dimension revealed by content-based analysis of citations. *Quantitative Science Studies*, 4(1), 44–67. https://doi.org/10.1162/qss_a_00242
- Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Text [PhD, University of Edinburgh]. <https://www.cl.cam.ac.uk/~sht25/thesis/t1.pdf>
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing, EMNLP '06* (pp. 103–110). Association for Computational Linguistics.
- Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, 95, 503–528. <https://doi.org/10.1007/s11192-013-0964-1>
- Xu, X., Xie, Y., Zhao, X., & Liu, Y. (2025). Mf-cite: Citation intent classification in scientific papers based on multi-feature fusion. *Scientometrics*, 130(8), 4465–4493. <https://doi.org/10.1007/s11192-025-05365-4>
- Yang, S., Wu, L., Wu, B., & Song, Y. (2025). Which is the cited source? A new perspective on article evaluation based on semantic similarity—Citation contribution attribution. *Journal of the Association for Information Science and Technology*, 76(11), 1532–1552. <https://doi.org/10.1002/asi.70012>
- Zhang, Y., Wang, Y., Sheng, Q. Z., Mahmood, A., Emma Zhang, W., & Zhao, R. (2021). TDM-CFC: Towards document-level multi-label citation function classification. In W. Zhang, L. Zou, Z. Maamar, & L. Chen (Eds.), *Web*

Information Systems Engineering—WISE 2021 (pp. 363–376). Springer International Publishing. https://doi.org/10.1007/978-3-030-91560-5_26

Zhang, Y., Wang, Y., Sheng, Q. Z., Yao, L., Chen, H., Wang, K., Mahmood, A., Zhang, W. E., Zaib, M., Sagar, S., & Zhao, R. (2025). Deep learning meets bibliometrics: A survey of citation function classification. *Journal of Informetrics*, 19(1), 101608. <https://doi.org/10.1016/j.joi.2024.101608>

Zhang, Y., Wang, Y., Wang, K., Sheng, Q. Z., Yao, L., Mahmood, A., Zhang, W. E., & Zhao, R. (2023). When Large Language Models Meet Citation: A Survey. <https://doi.org/10.48550/ARXIV.2309.09727>

Zhang, Y., Zhao, R., Wang, Y., Chen, H., Mahmood, A., Zaib, M., Zhang, W. E., & Sheng, Q. Z. (2022). Towards employing native

information in citation function classification. *Scientometrics*, 127(11), 6557–6577.

How to cite this article: Taşkın, Z. (2026). Methodological challenges in content-based citation analysis: Expertise, reliability, and the primacy of citance identification. *Journal of the Association for Information Science and Technology*, 1–16. <https://doi.org/10.1002/asi.70102>

APPENDIX A

TABLE A1 Examples of identical, overlapping, and unique citance selections identified by the two annotators.

Type	Annotator 1	Annotator 2	Description
Identical	“... a fundamental difference exists between information retrieval on one hand and certain other language processing tasks on the other. In retrieval one needs to render a document retrievable, rather than to convey the exact meaning of the text. Thus two items dealing with the same subject matter but coming to different conclusions are treated identically in retrieval, that is, either they are both retrieved or they are both rejected. In a question-answering system or language translation system, these documents would of course be treated differently” (Salton and McGill, 1983, p. 267)		Both annotators selected identical textual boundaries
Overlapping	A block of indicators should load higher on the construct that they are intended to measure as compared to their loadings on other constructs (Chin, 1998). The results indicated that our measures satisfied this condition, thus providing support for discriminant validity (Table 3)	A block of indicators should load higher on the construct that they are intended to measure as compared to their loadings on other constructs (Chin, 1998)	Citances partially overlap but differ in contextual scope
Non-identical	Annotators selected distinct citation contexts		

Note: The examples illustrate how differences in contextual scope and citance boundary selection contributed to interpretive variability during annotation.