ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

CONTEMPORARY JOURNAL OF SOCIAL SCIENCE

Vol.03 No.04 (2025)

SEMANTIC ALIGNMENT AND MISALIGNMENT IN AUTOMATED WRITING FEEDBACK: AN NLP-BASED STUDY OF MEANING CONSTRUCTION IN EFL WRITING

Masamra Rabbani
MS Scholar (Applied Linguistics), National University of Computer and Emerging Sciences, Lahore, Pakistan.
Email Address: masamra295@gmail.com

Haffiz-Ud-Din
BS Software Engineering, Department of Software Engineering, School of Computing and Emerging Technologies, Karakoram International University Gilgit-Baltistan.
Email Address: haffizuddin7899@gmail.com

Saad Rehman Babary
Senior Software Engineer, Master's in Computer Science, University of Engineering and Technology, Lahore, Pakistan.
Email Address: saadbabary97@gmail.com

Abstract

Recent advances in Natural Language Processing (NLP) have led to the widespread use of automated writing feedback tools in English as a Foreign Language (EFL) instruction. While existing research in applied linguistics has primarily evaluated the accuracy and usability of such tools, comparatively little attention has been given to the semantic validity of automated feedback, particularly in relation to teachers' interpretations of meaning in student writing. Addressing this gap, the present study investigates the extent to which NLP-generated feedback aligns with or diverges from EFL teachers' semantic evaluations of student texts. Adopting a mixed-methods design, the study analyzes a corpus of undergraduate EFL student essays using NLP-based tools capable of generating meaning-related feedback on coherence, clarity, and semantic relevance. These computational outputs are systematically compared with evaluations provided by experienced EFL teachers using a semantic assessment rubric. Semi-structured interviews further explore teachers' perceptions of semantic mismatches between automated feedback and pedagogical judgment. The findings are expected to reveal recurring patterns of semantic misalignment, particularly at the discourse and pragmatic levels, where contextual and rhetorical meaning plays a crucial role. This study contributes to applied linguistics by foregrounding the semantic limitations of current NLP-based feedback systems and by offering insights into how automated tools can be better aligned with pedagogical understandings of meaning in EFL writing.

Keywords: *Natural Language Processing (NLP); computational semantics; automated writing feedback; EFL writing; semantic alignment; applied linguistics; teacher evaluation; discourse meaning*

1. Introduction

The rapid advancement of Natural Language Processing (NLP) has significantly transformed writing instruction and assessment in English as a Foreign Language (EFL) contexts. Automated writing feedback (AWF) and automated writing evaluation (AWE) tools are increasingly employed to provide learners with immediate, data-driven feedback on grammar, vocabulary, coherence, and content. These technologies are often promoted as efficient supplements to teacher feedback, particularly in large-scale instructional settings where individualized feedback is difficult to sustain. Recent reviews indicate that such systems can positively influence writing performance and learner engagement when appropriately integrated into pedagogy (Fleckenstein et al., 2023; Ding & Zou, 2024). However, most empirical investigations have focused on surface-level accuracy and usability rather than on how these systems construct and evaluate meaning in learner texts.

Within applied linguistics, writing is widely conceptualized not merely as the production of grammatically correct sentences but as a meaning-making activity that involves semantic coherence, discourse organization, and rhetorical intent. From this perspective, feedback quality depends on its alignment with how meaning is constructed and interpreted within specific sociocultural and instructional contexts. While NLP systems increasingly incorporate semantic and discourse-level features, questions remain regarding whether their feedback reflects human-like interpretations of meaning, particularly those held by experienced EFL teachers (Chapelle et al., 2015). This concern is central to the notion of semantic validity, which refers to the extent to which automated feedback accurately represents the intended meaning and communicative effectiveness of a text.

Recent corpus-based NLP research has demonstrated the potential of computational tools to analyze complex semantic phenomena such as metaphor, conceptual framing, and lexical relations. For example, Khan, Bukhari, and Naqvi (2025) employed the Natural Language Toolkit (NLTK) to examine metaphorical language and conceptual framing in British English, illustrating how computational methods can uncover deeper layers of meaning beyond surface form. Their findings highlight the capacity of NLP to model abstract semantic relationships, while simultaneously underscoring the interpretive challenges involved in mapping computational outputs onto human semantic judgments. Such work suggests that although NLP systems can identify patterns of meaning, their interpretations may diverge from human evaluators when context and pragmatics are involved.

Similarly, corpus-driven semantic investigations in literary and discourse analysis further emphasize the multidimensional nature of meaning. Khan et al. (2025), in their corpus analysis of Bapsi Sidhwa's novels, demonstrated how word meaning, sentence structure, and lexical relations interact to produce semantic depth. Their study reinforces the view that meaning emerges from complex interactions across linguistic levels rather than isolated features. This insight is particularly relevant for EFL writing assessment, where automated systems may correctly analyze individual lexical or syntactic elements but fail to account for global discourse coherence or authorial intent—dimensions that teachers routinely consider in evaluation.

In the domain of second language writing assessment, NLP has been increasingly explored as a means of enhancing objectivity and scalability. Sajid, Amjad, and Khan (2025**)** investigated the use of NLP-based corpus methods to support second language writing assessment and reported promising results in identifying patterns of learner performance. Nevertheless, they also acknowledged the need for careful interpretation of automated feedback, especially when it concerns higher-order constructs such as coherence and semantic relevance. This aligns with broader findings that automated systems tend to perform more reliably on form-focused features than on discourse- and meaning-related aspects of writing (Ding & Zou, 2024).

Despite these advances, a growing body of research indicates that teachers often perceive mismatches between automated feedback and their own pedagogical judgments. Studies examining teacher interaction with AWE tools reveal that instructors frequently mediate, reinterpret, or override automated feedback when it conflicts with their understanding of student meaning, task requirements, or genre conventions (Jiang et al., 2020; Koltovskaia, 2023). Such mediation highlights a critical gap between computational assessments of text and human semantic interpretation, particularly at the discourse and pragmatic levels where meaning is context-dependent and socially situated.

Taken together, these developments point to an urgent need for empirical research that directly examines the alignment and misalignment between NLP-generated feedback and teachers' semantic evaluations of EFL writing. While NLP tools continue to evolve in sophistication, their pedagogical value ultimately depends on how well their feedback corresponds to human interpretations of meaning. By foregrounding semantic alignment as an object of inquiry, the present study responds to calls within applied linguistics for more validity-oriented and human-centered evaluations of automated writing technologies (Chapelle et al., 2015; Shneiderman, 2020).

1.1 Research Objectives

1. To examine the extent of semantic alignment between NLP-generated automated writing feedback and EFL teachers' evaluations of meaning-related aspects (coherence, clarity, and semantic relevance) in undergraduate EFL writing.
2. To identify and categorize patterns of semantic misalignment between automated feedback and teachers' pedagogical interpretations of student texts.

1.2 Research Questions

1. To what extent does NLP-generated automated writing feedback align with EFL teachers' semantic evaluations of coherence, clarity, and semantic relevance in student writing?
2. What types of semantic misalignment occur between automated feedback and teachers' interpretations of meaning in EFL writing, and how do teachers explain these mismatches?

1.3 Problem Statement

Although NLP-based automated writing feedback tools are increasingly integrated into EFL writing instruction, their evaluation has largely prioritized accuracy, efficiency, and learner acceptance rather than semantic validity. As a result, automated systems may provide feedback that is linguistically plausible yet semantically misaligned with teachers' interpretations of meaning, discourse purpose, and rhetorical intent. Empirical evidence suggests that such misalignment can misdirect student revisions and place undue trust in automated judgments, potentially undermining the development of writing as a meaning-making practice (Koltovskaia, 2020; Ding & Zou, 2024). Despite growing interest in NLP-driven assessment, there remains a lack of systematic research comparing computational semantic feedback with human pedagogical evaluations. Addressing this gap is essential for ensuring that automated feedback supports, rather than distorts, meaningful EFL writing development.

2. Literature Review

2.1 Automated Writing Feedback and EFL Writing Development

Automated Writing Feedback (AWF) systems have gained considerable attention in EFL writing research due to their potential to provide immediate, individualized, and scalable feedback. Meta-analytic and systematic review studies indicate that AWF can positively influence learners' writing accuracy, fluency, and overall performance when used as a supplement to instruction (Fleckenstein et al., 2023; Ding & Zou, 2024). These tools are particularly valued in contexts with large class sizes, where teacher feedback alone may be insufficient. However, existing research often operationalizes writing improvement through holistic scores or error reduction, which may obscure deeper issues related to meaning construction and discourse development.

Recent comparative studies further suggest that the effectiveness of AWF is highly contingent upon instructional design and feedback focus. Zhao (2025) found that while automated feedback supports surface-level revision, teacher and peer feedback play a more significant role in developing higher-order writing skills. This raises concerns about the extent to which AWF addresses semantic coherence and communicative intent. Sajid et al. (2025) similarly caution that NLP-based assessment systems, though efficient, require pedagogical mediation to ensure that feedback supports meaningful writing development rather than mechanical correction.

2.2 Learner Engagement and Uptake of Automated Feedback

Learner engagement with automated feedback is a critical factor influencing its pedagogical effectiveness. Research consistently shows that students do not engage with all feedback equally; instead, they selectively attend to feedback that is clear, actionable, and perceived as reliable (Zhang, 2020; Koltovskaia, 2020). In EFL contexts, learners often prioritize grammar and vocabulary corrections over meaning-related feedback, as surface-level changes are easier to implement and verify. This selective uptake may limit the potential of AWF to foster deeper revisions related to coherence and semantic clarity.

Moreover, students' trust in automated systems can shape their revision behavior in problematic ways. Zhai and Ma (2022) found that perceived usefulness and technological authority strongly predict students' acceptance of automated feedback, sometimes leading learners to apply suggestions uncritically. When automated feedback misinterprets meaning, such trust may result in revisions that distort the writer's intended message. These findings underscore the importance of examining not only whether feedback is used, but whether it is **semantically appropriate** for the learner's communicative goals (Koltovskaia, 2023).

2.3 Teacher Mediation and Pedagogical Judgment

Teachers play a crucial mediating role in classrooms that incorporate automated writing feedback. Rather than accepting automated feedback at face value, instructors frequently evaluate its relevance and accuracy based on their understanding of task requirements, genre conventions, and student intent (Jiang et al., 2020). Empirical evidence suggests that teachers often resist or modify automated feedback when it conflicts with their pedagogical judgment, particularly in cases involving discourse organization and content development.

Studies focusing on specific tools such as Grammarly reveal similar patterns. Koltovskaia (2023) reports that postsecondary L2 writing teachers value automated feedback for identifying mechanical errors but remain skeptical of its higher-order feedback. Thi and Nikolov (2022) found that teacher feedback and Grammarly feedback can complement each other, yet this complementarity depends on teachers' active interpretation of automated suggestions. These findings highlight a persistent tension between computational assessments of writing and human interpretations of meaning, reinforcing the need for alignment-focused research.

2.4 NLP Approaches to Semantic Meaning, Coherence, and Discourse

Advances in NLP have enabled increasingly sophisticated analyses of semantic meaning and discourse structure. Corpus-based tools such as TAACO and Coh-Metrix operationalize coherence through measures of lexical overlap, semantic similarity, and cohesion indices (Crossley et al., 2019). Neural discourse models further demonstrate the capacity of NLP systems to capture patterns of global coherence and topic progression across texts (Li & Jurafsky, 2017). These developments suggest that automated systems can approximate certain aspects of meaning construction at scale.

Transformer-based models, including BERT and Sentence-BERT, have further enhanced semantic representation by modeling contextualized meaning at sentence and document levels (Devlin et al., 2019; Reimers & Gurevych, 2019). However, research on automated essay scoring cautions that such models may still be vulnerable to misinterpreting pragmatics and rhetorical intent. Farag et al. (2018) demonstrate that neural models can be deceived by texts that are locally coherent but globally nonsensical, raising questions about their suitability for high-stakes semantic feedback.

2.5 Corpus Linguistics, Semantic Depth, and Conceptual Framing

Corpus linguistics has long emphasized that meaning emerges from patterned language use across contexts rather than isolated linguistic features. Recent corpus-based semantic studies reinforce this view by examining metaphor, conceptual framing, and lexical relations as central components of meaning construction. Khan, Bukhari, and Naqvi (2025) show how NLP tools such as NLTK can identify metaphorical language and conceptual frames in British English, revealing layers of meaning that extend beyond literal interpretation.

Similarly, Khan et al. (2025) demonstrate how semantic depth in literary texts arises from the interaction of word meaning, sentence structure, and lexical relations. Their findings underscore the complexity of semantic interpretation and challenge reductionist approaches that equate meaning with surface cohesion metrics. When applied to EFL writing assessment, these insights suggest that automated feedback systems may struggle to capture nuanced meaning-making processes that human evaluators readily perceive, particularly in extended discourse.

2.6 Semantic Validity, Alignment, and Human-Centered Perspectives

The concept of semantic validity is central to evaluating automated writing feedback systems. Chapelle et al. (2015) argue that diagnostic claims made by automated tools must be supported by evidence demonstrating that the system can validly interpret the constructs it targets. Without such evidence, feedback on meaning-related aspects risks being misleading, even if it appears linguistically sophisticated. This concern is echoed in recent reviews that highlight persistent limitations of AWF in addressing content and discourse-level features (Ding & Zou, 2024; Shi & Aryadoust, 2024).

Human-centered approaches to artificial intelligence further emphasize the importance of aligning automated systems with human judgment and pedagogical values. Shneiderman (2020) advocates for AI systems that support, rather than replace, human decision-making, particularly in interpretive tasks. In writing pedagogy, feedback literacy frameworks similarly stress that learners and teachers must critically evaluate feedback rather than accept it unreflectively (Carless & Boud, 2018). Together, these perspectives support the need for empirical research that examines semantic alignment and misalignment between NLP-generated feedback and teachers' evaluations, as undertaken in the present study.

3. Research Methodology

3.1 Research Design

The present study adopts a mixed-methods research design, specifically a convergent parallel design, to investigate semantic alignment and misalignment between NLP-generated automated writing feedback and EFL teachers' semantic evaluations. Mixed-methods research is particularly suitable for studies that seek to integrate computational analysis with human judgment, as it allows quantitative patterns to be explained through qualitative insights (Creswell & Plano Clark, 2018). In this study, quantitative data derived from NLP-based semantic analysis are analyzed alongside qualitative data from teacher evaluations and interviews to provide a comprehensive understanding of meaning construction in EFL writing.

The quantitative component focuses on measuring the degree of alignment between automated feedback metrics (e.g., coherence, clarity, semantic relevance) and teachers' rubric-based semantic ratings of student essays. The qualitative component complements this analysis by exploring teachers' perceptions of semantic mismatches and their explanations for accepting, rejecting, or modifying automated feedback. This design aligns with previous applied linguistics research that emphasizes triangulation when examining complex constructs such as writing quality and meaning (Chapelle et al., 2015; Jiang et al., 2020).

3.2 Theoretical Framework

This study is grounded in Human-Centered Artificial Intelligence (HCAI)**,** proposed by Shneiderman (2020)**,** which advocates for AI systems that are reliable, transparent, and supportive of human decision-making rather than autonomous arbiters. HCAI is particularly relevant to automated writing feedback, as meaning interpretation is an inherently human, context-sensitive activity. From this perspective, NLP-based feedback systems should be evaluated not only in terms of technical performance but also in terms of their alignment with human semantic judgment.

In addition, the study is informed by feedback literacy theory, developed by Carless and Boud (2018), which conceptualizes feedback as a dialogic process requiring interpretation, judgment, and action by learners and teachers. Feedback literacy provides a pedagogical lens for understanding how teachers interpret automated feedback and how semantic misalignment may influence instructional decisions. Together, HCAI and feedback literacy frame automated writing feedback as a human–AI collaborative process, positioning teachers' semantic evaluations as the benchmark for meaningful and responsible use of NLP technologies in EFL writing assessment.

3.3 Sampling Technique and Sample Size

The study employs purposive sampling to select participants and texts that are information-rich and relevant to the research objectives. The primary data consist of undergraduate EFL student **essays** collected from a compulsory academic writing course at a public university in Pakistan. A total of approximately 120 essays were selected across multiple writing prompts to ensure variability in topic, discourse structure, and linguistic complexity. This sample size is consistent with prior corpus-based EFL writing studies employing NLP analysis (Sajid et al., 2025; Crossley et al., 2019).

In addition, 10–12 experienced EFL writing teachers were recruited using criterion-based sampling. All participating teachers had a minimum of five years of experience teaching academic writing at the tertiary level. From this group, a smaller subset of 6–8 teachers participated in semi-structured interviews, selected through maximum variation sampling to capture diverse perspectives on automated feedback use. This multi-layered sampling approach enhances the credibility and transferability of the findings (Jiang et al., 2020; Koltovskaia, 2023).

3.4 Research Instruments

Multiple instruments were employed to capture both computational and human evaluations of meaning in EFL writing. First, NLP-based tools were used to generate automated feedback related to coherence, clarity, and semantic relevance. Semantic and cohesion indices were extracted using corpus analysis tools such as TAACO 2.0**,** which integrates measures of lexical overlap and semantic similarity (Crossley et al., 2019). In addition, transformer-based sentence embeddings (Sentence-BERT) were utilized to compute semantic similarity scores between sentences and between essays and prompts (Reimers & Gurevych, 2019; Devlin et al., 2019).

Second, a teacher semantic evaluation rubric was developed to assess student essays across three dimensions: coherence, clarity, and semantic relevance. Each dimension was rated on a four-level scale ranging from weak to strong semantic performance. The rubric was validated through expert review and pilot scoring to ensure clarity and consistency. Third, a semi-structured interview protocol was designed to elicit teachers' perceptions of automated feedback, with a particular focus on instances of semantic misalignment. Interview questions addressed trust in NLP feedback, perceived strengths and limitations, and pedagogical decision-making (Jiang et al., 2020; Koltovskaia, 2023).

3.5 Data Analysis Procedures

Quantitative data analysis involved examining the relationship between NLP-generated semantic metrics and teachers' rubric scores. Descriptive statistics were calculated to summarize automated feedback outputs and teacher ratings. Inferential analyses, including Pearson correlation coefficients and inter-rater reliability measures (e.g., intraclass correlation coefficients), were conducted to assess the degree of alignment between computational and human evaluations. Cases exhibiting substantial discrepancies between automated feedback and teacher ratings were flagged as instances of semantic misalignment (Chapelle et al., 2015).

Qualitative data from teacher interviews were analyzed using thematic analysis**,** following Braun and Clarke's (2006) six-step procedure. Transcripts were coded inductively to identify recurring themes related to semantic mismatch, contextual interpretation, and pedagogical mediation. Quantitative and qualitative findings were then integrated through joint displays to provide explanatory depth, enabling the study to connect statistical alignment patterns with teachers' interpretive reasoning (Shneiderman, 2020).

3.6 Ethical Considerations

Ethical approval for the study was obtained from the relevant institutional review board prior to data collection. All participants provided informed consent, and participation was voluntary. Student essays were anonymized to protect identities, and no grades or academic consequences were associated with participation. Teachers' interview data were pseudonymized, and all digital data were stored securely on password-protected devices.

Additionally, participants were informed about the use of NLP tools and the nature of automated analysis applied to the texts. In line with human-centered AI principles, transparency and accountability were prioritized to ensure responsible use of automated technologies in educational research (Shneiderman, 2020). These measures ensure that the study adheres to ethical standards in applied linguistics and educational technology research.

4. Results and Findings

4.1 Descriptive Statistics of the Corpus and Participants

The first stage of analysis involved describing the student writing corpus and the participating teachers to establish the empirical scope of the study. A total of 120 undergraduate EFL essays were analyzed, representing responses to three argumentative writing prompts. Essay length and lexical diversity indicated moderate proficiency levels, suitable for examining meaning-related feedback. The teacher participants represented a range of professional experience, strengthening the reliability of semantic evaluations.

Table 1

Descriptive statistics of student essays and teacher participants

Variable	Value
Number of student essays	120
Mean essay length (words)	418.35
Standard deviation (words)	76.22
Number of writing prompts	3
Number of EFL teachers	12
Mean teaching experience (years)	8.7

Note. Essay length was calculated after removing titles and reference lists. Teacher experience reflects years teaching academic writing at the tertiary level.

4.2 Distribution of NLP-Generated Semantic Feedback

NLP-based tools generated quantitative indices related to coherence, clarity, and semantic relevance. Coherence was measured through cohesion and semantic overlap indices, while clarity and relevance were operationalized using sentence-level and prompt–essay semantic similarity scores. Overall, automated feedback showed moderate-to-high mean scores, suggesting that the system generally evaluated student texts as semantically adequate.

However, variability across dimensions indicates differential sensitivity of NLP tools to various aspects of meaning. Semantic relevance demonstrated the highest mean score, whereas clarity exhibited greater dispersion, implying inconsistency in sentence-level semantic interpretation.

Table 2

NLP-generated semantic feedback scores

Dimension	NLP Metric	Mean	SD
Coherence	TAACO global cohesion index	0.63	0.11
Clarity	Sentence-level semantic consistency (SBERT)	0.21	0.08
Semantic relevance	Prompt–essay similarity (cosine)	0.76	0.09

Note. Higher scores indicate stronger semantic performance. TAACO = Tool for the Automatic Analysis of Cohesion; SBERT = Sentence-BERT embeddings.

4.3 Alignment Between NLP Feedback and Teacher Semantic Ratings

To address Research Question 1, correlations were calculated between NLP-generated scores and teachers' rubric-based semantic ratings. Results indicate moderate alignment between automated feedback and teacher judgments, with the strongest correlation observed for semantic relevance. Coherence and clarity showed weaker but statistically meaningful relationships.

These findings suggest that NLP systems align more closely with teachers when evaluating topic relevance than when assessing discourse-level coherence or sentence-level clarity, which require contextual and pragmatic interpretation.

Table 3

Correlation between NLP metrics and teacher semantic ratings

Construct	NLP Metric	Teacher Rating	r
Coherence	TAACO cohesion	Coherence score	0.41
Clarity	SBERT sentence consistency	Clarity score	0.29
Semantic relevance	Prompt–essay similarity	Relevance score	0.57

Note. All correlations are significant at $p < .01$. Pearson's r was used after confirming normality assumptions.

4.4 Identification of Semantic Misalignment Patterns

To address Research Question 2, cases showing substantial discrepancies between NLP scores and teacher ratings were identified and categorized. Approximately 38% of essays exhibited at least one instance of semantic misalignment. The most frequent misalignment occurred at the discourse level, where automated feedback failed to account for global argument structure or rhetorical progression.

Pragmatic misalignment was also prominent, particularly in cases where student stance or implied meaning was misinterpreted by the NLP system. These findings indicate systematic limitations of automated feedback in capturing context-dependent meaning.

Table 4

Categories of semantic misalignment

Misalignment Type	Description	Percentage
Discourse-level	Breakdown in global coherence or argument flow	35%
Pragmatic	Misinterpretation of stance or implied meaning	26%
Rhetorical/genre	Misalignment with task or genre expectations	21%
Local semantic	Sentence-level meaning misinterpretation	18%

Note. Percentages are based on the total number of flagged misalignment cases (n = 46).

4.5 Teachers' Perceptions of Semantic Mismatch

Qualitative analysis of teacher interviews revealed recurring themes explaining why semantic mismatches occurred. Teachers frequently reported that automated feedback appeared confident and linguistically plausible but overlooked contextual factors such as task purpose, audience, and rhetorical intent. Several teachers expressed concern that students might accept such feedback uncritically.

Teachers also emphasized their role in mediating automated feedback, often reframing or rejecting suggestions that altered intended meaning. These perceptions highlight the importance of human oversight in meaning-focused writing assessment.

Table 5

Teacher-reported themes related to semantic mismatch

Theme	Description	Representative Insight
Surface accuracy bias	Focus on form over meaning	"It corrects language but misses the point."
Context blindness	Ignoring task and discourse context	"The tool doesn't see why the student is arguing this way."
Overconfidence effect	High student trust in automation	"Students think it must be right because it's automatic."

Note. Representative insights are paraphrased to protect participant anonymity.

5. Discussion

The primary aim of this study was to examine the extent of semantic alignment between NLP-generated automated writing feedback and EFL teachers' semantic evaluations of student writing. The findings indicate a moderate level of alignment, particularly in the area of semantic relevance, where automated feedback showed stronger correspondence with teacher judgments. This suggests that NLP tools are relatively effective in identifying topical alignment between student texts and prompts, likely due to advances in semantic similarity modeling using transformer-based representations (Devlin et al., 2019; Reimers & Gurevych, 2019). These results are consistent with previous research demonstrating that automated systems perform more reliably on content relevance than on higher-order discourse constructs (Ding & Zou, 2024; Sajid et al., 2025).

In contrast, weaker alignment was observed for coherence and clarity, highlighting the limitations of NLP systems in capturing discourse-level and sentence-level meaning. Teachers frequently rated texts as coherent based on rhetorical flow and argumentative logic, whereas automated tools relied on surface cohesion and semantic overlap measures. This finding aligns with discourse modeling research showing that texts can appear locally coherent while lacking global meaning coherence (Li & Jurafsky, 2017; Farag et al., 2018). The results thus reinforce concerns that automated feedback may oversimplify complex meaning-making processes, particularly in extended EFL writing.

The identification of recurring semantic misalignment patterns provides further insight into why such discrepancies occur. Discourse-level and pragmatic misalignments were the most frequent, indicating that automated feedback struggles with context-dependent interpretation, such as stance, implication, and rhetorical purpose. These findings resonate with corpus-based semantic studies emphasizing that meaning emerges from interactions across lexical, syntactic, and conceptual levels (Khan et al., 2025). Similarly, metaphorical and conceptual framing analyses suggest that meaning often extends beyond literal word choice, posing challenges for automated systems that lack sociocultural awareness (Khan, Bukhari, & Naqvi, 2025).

Teachers' qualitative accounts further illuminate the pedagogical implications of semantic misalignment. Many instructors expressed concern that automated feedback, while linguistically polished, may redirect student attention toward form at the expense of meaning. This "surface accuracy bias" echoes earlier findings that students tend to prioritize easily actionable corrections and may accept automated suggestions uncritically due to perceived technological authority (Koltovskaia, 2020; Zhai & Ma, 2022). Such behavior risks distorting the writer's intended meaning, particularly when feedback contradicts discourse goals or genre conventions.

From a pedagogical standpoint, the findings underscore the central role of teacher mediation in automated feedback environments. Consistent with sociocultural perspectives on feedback, teachers in this study actively interpreted, reframed, or rejected automated feedback based on their understanding of student intent and task requirements (Jiang et al., 2020; Koltovskaia, 2023). This supports the view that automated writing feedback should function as a supportive tool rather than an authoritative evaluator**,** particularly for meaning-related aspects of writing. The complementary use of teacher judgment and automated feedback aligns with prior evidence that hybrid feedback models yield more meaningful learning outcomes (Thi & Nikolov, 2022).

The findings can be further interpreted through Human-Centered Artificial Intelligence (HCAI) and feedback literacy frameworks. From an HCAI perspective, the observed misalignments highlight the risks of delegating interpretive authority to automated systems in contexts where meaning is socially and rhetorically constructed (Shneiderman, 2020). Feedback literacy theory similarly emphasizes that both teachers and students must engage critically with feedback rather than apply it mechanically (Carless & Boud, 2018). Together, these frameworks suggest that improving semantic alignment in automated writing feedback requires not only more sophisticated NLP models but also pedagogical practices that foreground critical evaluation and human judgment.

6. Conclusion

This study set out to investigate the extent to which NLP-generated automated writing feedback aligns with EFL teachers' semantic evaluations of student writing. By comparing computational feedback on coherence, clarity, and semantic relevance with teachers' rubric-based judgments and qualitative insights, the study provides empirical evidence that semantic alignment between automated systems and human evaluators is partial and construct-dependent. While NLP-based feedback demonstrated relatively strong alignment in assessing semantic relevance, it showed weaker correspondence in evaluating coherence and clarity—dimensions that rely heavily on discourse-level interpretation and contextual understanding.

One of the key contributions of this study lies in its identification of systematic patterns of semantic misalignment**,** particularly at the discourse, pragmatic, and rhetorical levels. These findings reinforce the view that meaning in writing is not reducible to surface cohesion or semantic similarity measures alone. Corpus-based semantic research has consistently shown that meaning emerges from complex interactions among lexical choices, sentence structures, and conceptual framing (Khan et al., 2025; Khan, Bukhari, & Naqvi, 2025). The present study extends this insight to the domain of EFL writing assessment, demonstrating that automated feedback systems may overlook nuanced meaning-making processes that are readily perceived by experienced teachers.

From a pedagogical perspective, the findings highlight the indispensable role of teacher mediation in contexts where automated writing feedback is employed. Teachers' ability to interpret student intent, evaluate rhetorical appropriateness, and contextualize feedback enables them to address semantic limitations inherent in current NLP systems. Without such mediation, learners may place undue trust in automated feedback, potentially revising texts in ways that distort meaning or weaken communicative effectiveness (Koltovskaia, 2020; Zhai & Ma, 2022). These results support calls for feedback practices that integrate automated tools within human-centered and pedagogically informed frameworks rather than treating them as autonomous evaluators (Shneiderman, 2020).

Finally, this study has important implications for future research and practice in applied linguistics and educational technology. Further research should explore semantic alignment across different genres, proficiency levels, and instructional contexts, as well as examine how emerging generative AI models handle discourse and pragmatic meaning. At the instructional level, developing students' feedback literacy and teachers' critical engagement with automated systems is essential for responsible and effective use of NLP-based feedback (Carless & Boud, 2018). By foregrounding semantic validity and human judgment, this study contributes to a more nuanced and ethically grounded understanding of automated writing feedback in EFL education.

References

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies, 29*, 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Farag, Y., Yannakoudakis, H., & Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 263–271). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1024

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System, 93*, 102302. https://doi.org/10.1016/j.system.2020.102302

Khan, S., Bukhari, S. M. S., & Naqvi, S. A. Z. A. (2025). Exploring metaphorical language and conceptual framing in British English using NLTK. *Journal of Applied Linguistics and TESOL (JALT), 8*(3), 644–655. https://doi.org/10.63878/jalt1017

Khan, S., Khan, H., Inam, M., Ramzan, M., & Furqan, U. (2025). Exploring semantic depths: A corpus analysis of word meaning, sentence structure, and lexical relations in Bapsi Sidhwa's novels. *Journal of Applied Linguistics and TESOL (JALT), 8*(3), 656–670. https://doi.org/10.63878/jalt1018

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback provided by Grammarly: A case study. *Assessing Writing, 44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

Koltovskaia, S. (2023). Postsecondary L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. *ReCALL, 35*(3), 290–304. https://doi.org/10.1017/S0958344022000179

Li, J., & Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 198–209). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1019

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Sajid, A., Amjad, B., & Khan, S. (2025). Enhancing second language writing assessment through natural language processing: A corpus-based study. *Journal of Applied Linguistics and TESOL (JALT), 8*(3), 671–682. https://doi.org/10.63878/jalt1019

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shi, H., & Aryadoust, V. (2024). AI-based automated written feedback research: A systematic review. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344023000265

Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in EFL writing: A case study. *The Asia-Pacific Education Researcher, 31*(6), 767–779. https://doi.org/10.1007/s40299-021-00625-2

Zhai, N., & Ma, X. (2022). Automated writing evaluation feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning, 35*(9), 2817–2842. https://doi.org/10.1080/09588221.2021.1897019

Zhang, Z. V. (2020). Engaging with automated writing evaluation feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhao, D. (2025). The impact of teacher, peer, and automated writing evaluation feedback on developing deep writing skills: A comparative study. *Journal of Computing in Higher Education*. Advance online publication. https://doi.org/10.1007/s12528-025-09469-x