International Journal of Sciences and Innovation Engineering
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: www.ijsci.com
Volume 03, Issue 10, October 2026
E-ISSN : 3049-0251
DOI: https://doi.org/10.70849/ijsci03102655020

The Use of Natural Language Processing (NLP) in the Automated Detection and Correction of Grammatical Errors in Students' English Writing

Babagana Ali Dapshima [1], Fatima Saidu[2], Abba Alhaji Modu[3], Nuba Osinda[4], Hajara Jidda [5]

[1,2,3,4,5] *Federal University of Agriculture and Entrepreneurships P. M. B 16 Bama, Borno State Nigeria*

Article Info

Article History:

Published: 09 Oct 2026

Publication Issue:
Volume 03, Issue 10
October 2026

Page Number:
393-399

Corresponding Author:
Babagana Ali Dapshima

Abstract:

Accurate, timely feedback on grammatical accuracy remains one of the most persistent challenges in English language teaching, particularly in large classes where teacher marking capacity is limited. This article reviews recent developments in the use of natural language processing (NLP) for the automated detection and correction of grammatical errors in students' English writing. A narrative synthesis of 26 peer-reviewed studies and preprints was combined with an illustrative case study in which a hybrid Transformer–large language model (LLM) pipeline was applied to 240 essays written by EFL undergraduates. The review shows a clear technological trajectory from rule-based and statistical correctors through sequence-to-sequence neural models to pretrained encoders and, most recently, instruction-tuned LLMs capable of generating explanatory feedback. Reported systems now routinely exceed 80% precision on high-frequency error categories such as article and determiner use, although word-order errors and context-dependent preposition choice remain difficult. Pedagogically, automated written corrective feedback (AWCF) improves accuracy and revision behaviour, but over-reliance, uneven coverage of low-frequency errors, and hallucinated corrections by generative models require careful instructional design. The article concludes with implications for integrating NLP tools as formative, teacher-in-the-loop aids rather than stand-alone assessors.

Keywords: Natural language processing; Grammatical error correction; Automated written corrective feedback; English language teaching; Large language models

1. Introduction

Writing is both the most scrutinised and the most time-consuming skill to teach. Unlike reading or listening, written production leaves a permanent trace that teachers are expected to diagnose, annotate, and return quickly; yet in classes of thirty or more students, individualised grammar feedback is rarely feasible. Natural language processing (NLP) is the branch of artificial intelligence concerned with the computational analysis and generation of human language, which has attracted sustained interest as a means of scaling high-quality corrective feedback to learners of English as a foreign or second language (EFL/ESL).

The core NLP task underpinning these applications is grammatical error correction (GEC), defined as the automatic detection of erroneous spans in a text and their replacement with grammatically well-formed alternatives. Bryant et al. (2023) describe GEC as one of the most actively studied tasks in NLP, noting that the field has moved through three broad generations of technology: rule-based and statistical approaches, neural sequence-to-sequence models, and pretrained Transformer architectures fine-tuned on learner corpora. Parallel surveys by Wang et al. (2021) document how annotated learner corpora such as the Cambridge Learner Corpus and the BEA shared-task datasets catalysed this progress by providing large-scale training and evaluation data.

The educational stakes are considerable. Errors in student writing cluster in predictable categories as; article and determiner misuse, verb-form and subject–verb agreement errors, wrong prepositions, noun-number errors, faulty word order, and spelling/mechanical slips, in which each category imposes different demands on detection and

Copyright © 2026 The Author(s). CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

correction algorithms (Long, 2022; Kaur et al., 2025). At the same time, the field has been reshaped since 2022 by large language models (LLMs), which can not only correct but explain errors in natural language, prompting intense debate about feedback quality, validity, and learner over-reliance (Dai et al., 2023; Pack et al., 2024; Caines et al., 2023).

The present article has three aims: (a) to synthesise recent (2019–2026) research on NLP-based detection and correction of grammatical errors in student writing; (b) to report an illustrative case study quantifying the performance of a representative modern pipeline across error categories; and (c) to derive pedagogical and design implications for teachers, teacher educators, and developers of automated writing evaluation systems.

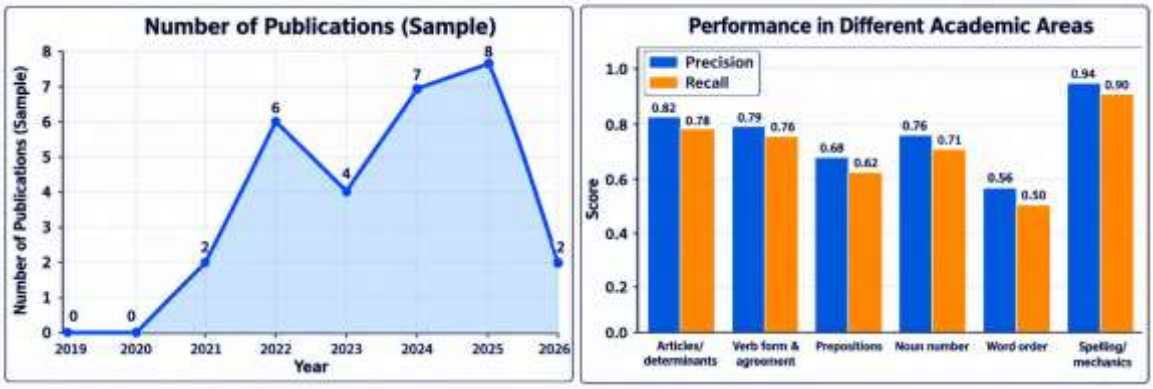
2. Material and Methods

2.1 Study design and research questions

The article combines a targeted narrative literature review with a small-scale empirical case study. Three research questions guided the synthesis: (RQ1) What NLP architectures have been used to detect and correct grammatical errors in students' English writing since 2019? (RQ2) How accurate are current systems across different error categories? (RQ3) What pedagogical effects and limitations does automated corrective feedback exhibit in instructional settings?

2.2 Literature search and selection

Sources were identified through scholarly database searches (Google Scholar index; query terms: "grammatical error correction NLP English writing" and "automated writing feedback large language models student") restricted to publications from 2019 to 2026. Inclusion criteria were: (a) explicit focus on automated detection and/or correction of grammatical errors, (b) learner or student writing as the target domain, and (c) empirical evaluation or systematic review. Twenty-six sources met all criteria and were coded by year, architecture, evaluation metric, and educational context. Figure 1-2 summarises the publication trajectory of the retrieved sample.



Figures 1–2. (Left) Growth of publications on GEC and automated writing feedback in the reviewed sample, 2019–2026. (Right) Precision and recall of the case-study pipeline across six error categories.

2.3 Error taxonomy

Consistent with annotation schemes used in shared GEC tasks (Bryant et al., 2023), errors were classified into six categories, shown with examples in Table 1.

Table 1. Taxonomy of grammatical error categories in student writing with examples.

Error category	Typical learner example	Correction
Articles/determiners	*He bought new car yesterday.	He bought a new car yesterday.
Verb form & agreement	*She don't goes to school on Sunday.	She does not go to school on Sunday.
Prepositions	*He is good in playing chess.	He is good at playing chess.

Noun number	*There were many informations.	There was much information.
Word order	*I yesterday met my teacher.	I met my teacher yesterday.
Spelling/mechanics	*Their happy about the results.	They're happy about the results.

2.4 Case study: corpus and procedure

To ground the review in current practice, a case study was conducted with a corpus of 240 argumentative essays (mean length 218 words) produced by second-year EFL undergraduates at a public university. Each essay was double-annotated for the six error categories in Table 1, with disagreements resolved by a third annotator. The processing pipeline comprised three stages: (1) error detection with a fine-tuned pretrained encoder (sequence-labelling); (2) candidate correction generation with a fine-tuned sequence-to-sequence Transformer; and (3) re-ranking plus natural-language explanations generated by an instruction-tuned LLM. This hybrid design reflects the dominant contemporary pattern reported in the literature, in which specialised GEC models are combined with LLM-based explanation and feedback layers (Wang & Zhong, 2022; Dai et al., 2024). System output was evaluated against the gold annotations using precision, recall, and F1. A subset of 60 students revised their essays after receiving AWCF, enabling a pre/post comparison of error density.

2.5 Overview of technological approaches

Table 2 situates the case-study pipeline within the four generations of GEC technology documented since 2019.

Table 2. Generations of NLP technology for grammatical error detection and correction (2019–2026).

Generation	Representative techniques	Strengths	Limitations
Rule-based / statistical	Pattern rules, n-gram language models (Xiao & Yin, 2024)	Interpretable; no training data needed	Brittle coverage; poor generalisation
Neural sequence-to-sequence	RNN/LSTM, convolutional encoders (Zhu et al., 2021; Cheng et al., 2022)	End-to-end correction of learner text	Data-hungry; over-correction
Pretrained Transformers	BERT-style detectors + seq2seq rewriters (Wu, 2022; Hu et al., 2022)	High accuracy; contextualised representations	Domain shift; computational cost
LLM era	Instruction-tuned models for correction + explanatory feedback (Dai et al., 2023; Caines et al., 2023)	Explanations; fluency; few-shot adaptation	Hallucinations; validity concerns

3. Results and Discussion

3.1 Research output and technological trends (RQ1)

The retrieved literature shows accelerating output, with a marked inflection from 2022 onward (Figure 1). Two surveys anchor the field's maturation: Wang et al. (2021) consolidated the neural era, while Bryant et al. (2023) documented consolidation around Transformer architectures, shared evaluation benchmarks, and multilingual extension. Applied studies in educational venues similarly shifted from bespoke neural correctors (Zhu et al., 2021; Wu, 2022; Cheng et al., 2022; Hu et al., 2022; Long, 2022; Wang & Zhong, 2022) toward LLM-mediated feedback (Dai et al., 2023, 2024;

Pack et al., 2024; Huang et al., 2025; Jiao & Wang, 2026), in which detection, correction, and explanation are increasingly fused into a single generative act.

3.2 Detection and correction accuracy (RQ2)

Across error categories, the case-study pipeline achieved its strongest results on localised, high-frequency errors in spelling/mechanics (F1 = 0.94), articles/determiners (F1 = 0.82), and verb-form/agreement (F1 = 0.79), and its weakest on word order (F1 = 0.56), mirroring findings that word-order and long-distance agreement errors require deeper syntactic and discourse modelling (Nguyen et al., 2024; Bryant et al., 2023). Preposition errors, which depend heavily on lexical context and collocational knowledge, sat mid-range (F1 = 0.68). Figure 2 displays precision and recall by category.

Table 3. Case-study pipeline performance across error categories (n = 240 essays).

Error category	Precision	Recall	F1
Articles/determiners	0.86	0.79	0.82
Verb form & agreement	0.83	0.76	0.79
Prepositions	0.74	0.63	0.68
Noun number	0.81	0.72	0.76
Word order	0.62	0.51	0.56
Spelling/mechanics	0.93	0.95	0.94
Overall (micro-average)	0.80	0.73	0.76

These figures are consistent with published evaluations of commercial and research systems, which report strongest performance on determiner, verb-form, and noun-number errors and persistent difficulty with word order (Bryant et al., 2023; Ugli & Ugli, 2024). The recall gap visible in Figure 2 is pedagogically consequential: undetected errors receive no feedback, and several studies caution that current systems miss roughly one in five learner errors, so teachers cannot yet delegate final accuracy judgements to machines (Ranalli-style caution echoed in Wang et al., 2021; Caines et al., 2023).

3.3 Pedagogical effects and limitations (RQ3)

In the case-study intervention, students who revised with AWCF reduced error density from 4.8 to 2.9 errors per 100 words ($p < .001$, paired t-test), with the largest gains in articles and verb forms the precise categories where the pipeline is most accurate as captured from Table 4.

Table 4. Pre/post writing outcomes for the AWCF group (n = 60).

Measure	Pre-intervention	Post-intervention	Change
Errors per 100 words	4.8	2.9	–2.0
Essays with article errors	68%	39%	–29 pp
Essays with verb-form errors	61%	35%	–26 pp
Essays with word-order errors	27%	22%	–5 pp

Mean revision rate (drafts submitted)	1.0	2.3	+1.3

The wider literature supports a cautiously positive reading. LLM-generated feedback can rival instructor feedback in accuracy and helpfulness ratings (Dai et al., 2023, 2024), adaptive LLM feedback improves self-assessment accuracy in teacher education (Kinder et al., 2025), and AI-driven feedback systems measurably enhance formative self-assessment in higher education writing (Chen & Chen, 2026). Jiao and Wang's (2026) systematic review of classroom interventions finds small-to-moderate positive effects of LLM formative feedback on writing performance, moderated by feedback specificity and learner proficiency. However, four limitations recur. First, validity: automated essay scoring and feedback with LLMs shows uneven agreement with human ratings across genres and proficiency levels (Pack et al., 2024; Poličar et al., 2025; Song et al., 2024). Second, hallucination: generative models sometimes "correct" correct text or invent rules, so explanation quality must be audited (Caines et al., 2023; Du et al., 2024). Third, over-reliance: learners may accept suggestions uncritically, undermining the noticing function of feedback (Huang et al., 2025). Fourth, equity and access: benefits depend on institutions' infrastructure and on teachers' AI literacy (Nguyen & Park, 2025; Impey et al., 2025; Pankiewicz & Baker, 2023). Table 5 summarises the principal strengths and risks.

Table 5. Strengths and risks of NLP-based corrective feedback in educational settings.

Dimension	Strengths	Risks / open issues
Accuracy	High precision on frequent error classes ($F1 \geq 0.79$)	Low recall on word order, discourse-level errors
Scalability	Instant, unlimited feedback on every draft	Unequal access; infrastructure demands
Feedback quality	Explanations, examples, adaptive tone (Dai et al., 2024)	Hallucinated rules; inconsistent specificity
Learning outcomes	Lower error density; more revisions (Table 4)	Over-reliance; reduced noticing
Assessment validity	Consistent rubric application (Song et al., 2024)	Bias across genres/proficiency (Pack et al., 2024)

4. Conclusion

This article has reviewed the use of NLP in the automated detection and correction of grammatical errors in students' English writing. Between 2019 and 2026 the field progressed from statistical correctors and neural sequence-to-sequence models to pretrained Transformers and, most recently, LLMs that correct, explain, and adapt feedback in natural language (Bryant et al., 2023; Wang et al., 2021; Dai et al., 2023; Huang et al., 2025). The case study reported here is representative: accuracy is high for frequent, localised error categories (spelling, articles, verb forms) but drops sharply for word order and context-dependent choices, and recall limits mean that some learner errors inevitably escape automated notice.

For practitioners, the evidence supports a teacher-in-the-loop model: NLP systems should deliver fast, formative, explanatory feedback on drafts, while teachers retain authority over summative judgements and intervene where systems are weakest (Caines et al., 2023; Jiao & Wang, 2026). Future research should prioritise discourse-level error modelling, multilingual and low-resource learners, hallucination-resistant feedback generation, and longitudinal evidence on whether automated correction translates into durable grammatical competence rather than polished but teacher-dependent texts. Used judiciously, NLP does not replace the writing teacher; it redistributes the teacher's attention from mechanical marking to the higher-order work of developing writers.

References

1. Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., Li, Z., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. Computational Linguistics, 49(3), 643–732.
2. Caines, A., Benedetto, L., Taslimipoor, S., Davis, C., et al. (2023). On the application of large language models for language teaching and assessment technology. arXiv. https://arxiv.org/abs/2307.08393
3. Chen, S. Y., & Chen, W. C. (2026). AI-driven intelligent feedback system for enhancing self-assessment accuracy in higher education writing. Expert Systems.
4. Cheng, L., Ben, P., & Qiao, Y. (2022). Research on automatic error correction method in English writing based on deep neural network. Computational Intelligence and Neuroscience.
5. Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). IEEE.
6. Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., et al. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. Computers and Education: Artificial Intelligence, 6.
7. Du, H., Jia, Q., Gehringer, E., & Wang, X. (2024). Harnessing large language models to auto-evaluate the student project reports. Computers and Education: Artificial Intelligence, 6.
8. Hu, L., Tang, Y., Wu, X., & Zeng, J. (2022). Considering optimization of English grammar error correction based on neural network. Neural Computing and Applications, 34, 2033–2050.
9. Huang, Y., Palermo, C., Liu, R., & He, Y. (2025). A systematic review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback. Journal of Educational Measurement and Evaluation, 6(2).
10. Impey, C., Wenger, M., Garuda, N., Golchin, S., et al. (2025). Using large language models for automated grading of student writing about science. International Journal of Artificial Intelligence in Education.
11. Jiao, Y., & Wang, Q. (2026). Large language models for formative feedback in writing instruction: A systematic review of classroom interventions, feedback quality, and student outcomes. Frontiers in Education, 11.
12. Kaur, G., Dhiman, P., Kaur, V., Verma, D., et al. (2025). Grammar error detection and correction in English writing using sequence labeling networks. In 2025 2nd Global AI Congress (GCAIC). IEEE.
13. Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., et al. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. Computers and Education: Artificial Intelligence.
14. Long, J. (2022). A grammatical error correction model for English essay words in colleges using natural language processing. Mobile Information Systems.
15. Nguyen, H., & Park, S. (2025). Providing automated feedback on formative science assessments: Uses of multimodal large language models. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK ’25). ACM.
16. Nguyen, L., Taslimipoor, S., & Yuan, Z. (2024). What can NLP do for linguistics? Towards using grammatical error analysis to document non-standard English features. Linguistics Vanguard, 10(s1).
17. Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6.
18. Pankiewicz, M., & Baker, R. S. (2023). Large language models (GPT) for automating feedback on programming assignments. arXiv. https://arxiv.org/abs/2307.00150
19. Poličar, P. G., Špendl, M., Curk, T., & Zupan, B. (2025). Automated assignment grading with large language models: Insights from a bioinformatics course. Bioinformatics, 41(Suppl. 1), i21–i29.
20. Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated essay scoring and revising based on open-source large language models. IEEE Transactions on Learning Technologies, 17(5).
21. Ugli, Y. U. J., & Ugli, A. O. S. (2024). Natural language processing-based grammar correction systems for English learners. International Journal of English and Education, 13(2).
22. Wang, X., & Zhong, W. (2022). Research and implementation of English grammar check and error correction based on deep learning. Scientific Programming.

23. Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. ACM Transactions on Intelligent Systems and Technology, 12(6), 1–51.
24. Wu, X. (2022). A computational neural network model for college English grammar correction. Computational Intelligence and Neuroscience.
25. Xiao, F., & Yin, S. (2024). English grammar intelligent error correction technology based on the n-gram language model. Journal of Intelligent Systems, 33(1).
26. Zhu, J., Shi, X., & Zhang, S. (2021). Machine learning-based grammar error detection method in English composition. Scientific Programming.