ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

A CORPUS-BASED ANALYSIS OF LEARNER ERRORS BEFORE AND AFTER AN AI-ASSISTED WRITING INTERVENTION: EVIDENCE FROM UNDERGRADUATE EFL STUDENTS IN HYDERABAD, SINDH, PAKISTAN

Muhammad Ahsan Raza
Lecturer in English, College Education Department, Government of Sindh, Pakistan
Email: memonahsan964@gmail.com
Salahuddin Shaikh
Lecturer in English, Federal Urdu University of Arts, Science & Technology (FUUAST), Pakistan
Email: salahuddin2338@gmail.com
Muzfar Ali Naich
MS Scholar
Email: muzafara693@gmail.com
Benazir Shaikh
PhD Scholar
Email: benazir3476@gmail.com
Ghulam Mustafa
University Lecturer | Applied Linguistics Researcher
Email: mustafakhan788@gmail.com

Abstract
This study is an attempt to find the effect of writing intervention with the help of Artificial Intelligence (AI) on the accuracy of undergraduate EFL writers of Hyderabad, Sindh, Pakistan for a period of eight weeks. Sixty students from four affiliated campus completed a timed pre and post argumentative writing test. Ten categories of errors were identified following the error taxonomy of the corpus and two trained raters coded the errors while 20% of the texts were double coded for reliability. The total errors per 100 words at the post-intervention sample was significantly less than the pre-intervention sample and the effect size was large. In all 10 of the categories of errors, there was a decrease. The main improvements were in word form and morphology, verb tense & aspect and subject–verb agreement. The least improvement was in punctuation and capitalisation and in the other category (residual). Article and preposition errors were both two of the most common classes of errors for both time points and there was a significant improvement on these errors. Gender, year in school and academic major were not moderators of the size of the gain, indicating that the intervention worked pretty much for everyone. The findings align with the emerging evidence that decreasing surface errors and grammatical errors in L2 writing in a single semester is meaningful. They also show that some types are not suitable for rapid remediation, particularly those to do with first language transfer. Implications for classroom practice and directions for future corpus-based intervention research in the context of the South Asian EFL context are discussed.

Keywords: error analysis, EFL writing, AI-assisted feedback, learner corpus, Pakistan, written corrective feedback

Introduction
English is a key language of education in Higher Education (HE) in Pakistan. It is the language used for learning in most universities, most of the academic literature, and is a barrier for access to work in many fields. Nevertheless many of the undergraduate population still have problems in written accuracy after years of formal English learning. This disconnect between the years of teaching and actual writing in English has been experienced across the country, from Punjab to Sindh and has a tangible impact in the lives of students who are required to write academic essays, reports and theses in English during their university years.

Long ago, error analysis has been used as a means of understanding this gap. Instead of being merely errors, the errors made by language learners are regarded as manifestations of the learner's interlanguage, or the partially developed grammar system that exists between the first and second languages. Coding and counting errors in a structured way will enable researchers

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

to determine what specific language features may be problematic for a particular population and whether or not an intervention has an impact on the language produced by learners over time.

Large language models (LLMs) have emerged as a novel feedback tool in recent years, for the writing classroom. These tools can detect grammatical issues, offer lexical substitution and make comments about cohesion in seconds, and at a level that no one teacher can keep up with. Initial research indicates that with this type of feedback, grammatical accuracy can increase, but there seems to be less impact on higher order issues like organization and argumentation, and there may need to be a combination of human feedback and this type of feedback. It is an open empirical question whether the patterns are valid for undergraduate EFL writers in the context of a Pakistani university where English is one of the many languages used on a daily basis and article and preposition systems of the students' L1 are quite different from English.

In the present study, this question has been tackled by a corpus of argumentative text of sixty undergraduate students in Hyderabad, Sindh consisting of paired pre and post-intervention essays. Essays were coded for errors, both pre and post an 8-week AI-assisted writing intervention, on 10 linguistic measures. There are three questions in the study. First is whether the overall rate of coded errors overall is reduced by the intervention. Secondly, which type of errors are reduced in the greatest and the least? Third, are there gender, year of study, academic major or self rated proficiency differences in the gains? The rest of the paper gives a literature review of the relevant literature, describes the corpus, statistical results of the coding procedures and discusses implications for classroom practice in this context.

Literature Review

Theoretical Foundations of Error Analysis

Modern error analysis is a development of Corder's (1967) suggestion that errors made by learners can be regarded as not as random slips but as evidence of a set of rules which the learner has internalized at a certain stage in the development of the system. Selinker (1972) took this notion one step further and came up with the notion of interlanguage which is not the first language or the target language, but a language system on its own. Later Ellis (1994) distinguished between errors, indicating a true deficit in a learner's underlying knowledge, and mistakes, occasional errors a learner may make, but can correct if more attention is given. This is important for the current study because a surface error count of a corpus, by itself, would not be able to distinguish between the presence of real competence gaps and timed performance errors, as explained below.

This has been followed up by the use of systematic, computer-assisted coding of large corpora of learner writing, which is known as learner corpus research. According to Granger (2003), error-tagged learner corpora offer a way to transcend from individual case studies to get more reliable, comparable frequencies of errors made by a population of learners, thereby allowing for contrastive interlanguage analysis (comparison of learner output at different stages of proficiency, tasks, or points in time). The present study is in line with this tradition, as all the errors made by the students in the same writing task at two different time points are classified in the same taxonomy of ten error categories, thus providing a direct pre–post comparison instead of a snapshot.

EFL Writing Errors Among Pakistani Learners

English is in a unique place in Pakistan. In his book, Channa (2017), the author traced the history whereby English kept its status as official language and as language of education after Pakistan's independence, and that the students of Pakistan usually study English for 10 years or more before going to university and yet somehow a large number of students come to university without a proper command of major grammatical systems. However, Sarfraz (2011) investigated the error analysis of written compositions of Pakistani undergraduate students and discovered that common errors include errors of article, preposition, tense and agreement,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

which is similar to the findings of other studies in the region. A systematic review of error analysis studies in EFL and ESL writing in countries near Afghanistan (including Pakistan) also revealed that the most commonly reported error types were of errors of article, preposition, and spelling errors and that two error types could be attributed to L1 interference and the other two types to incomplete rule acquisition.

These errors in articles and prepositions are very common in the published works, which can be linguistically explained. The languages spoken by the majority of students in the current sample (Sindhi and Urdu) do not have a system of articles indicative of definiteness as in English and their prepositions/postpositions do not correspond consistently with English prepositions. Contrastive analysis suggests that this mismatch would account for the type of ongoing problems with articles and prepositions that are repeatedly reported in Pakistani error analysis studies and that would question whether this would be a worthwhile feedback intervention to address in a short automated intervention, or whether it would primarily be a feedback intervention for the categories that are not so much bound to L1 structure.

AI-Assisted Feedback and Written Corrective Feedback

The issue of whether or not CF on grammar is helpful for the improvement of L2 writing has been a controversial issue for decades. The ineffectiveness of grammar correction has been stressed by Truscott (1996) who even suggested that grammar correction can be counterproductive, as learners are unable to apply the feedback to other texts. Ferris (2006) challenged this stance by presenting evidence for the link between written corrective feedback and accuracy gains over time, especially with regard to those types of errors that are amenable to treatment like verb tense and subject–verb agreement. This conflict was the impetus for the current research on automated and AI feedback generation, which often is more consistent than human raters, who must work within time constraints.

Some recent research indicates that consistency may be a way to ensure that there are tangible benefits, at least when it comes to surface-level accuracy, from using a large language model feedback. Yan and Zhang (2024) traced four L2 writers' use of ChatGPT-based CF, and their study revealed that behaviors of engaging with ChatGPT-generated CF were greatly dependent on the writers' level of L2 proficiency and technological skill, and the writers' capacity to self-regulate their own revision process was limited with ChatGPT-generated CF alone. Zhang, Aubrey, Huang and Chiu (2025) compared the performances of a group of Chinese EFL students using only generative AI feedback with another group using a combination of generative AI and a human tutor. In higher-order aspects like organization and critical thinking, only limited improvements could be seen for the generative AI feedback group, with the hybrid group demonstrating greater improvements. As a whole, this body of research indicates that AI feedback is appropriate for grammar and lexical accuracy (which are the most prominent areas in the literature on errors in Pakistani English), but that it should be combined with other teaching to achieve discourse level competence.

The Present Study

Although there has been an increase in studies on error analysis in the Pakistani higher education setting as well as studies on AI-assisted feedback in L2 writing, only a handful of research has integrated both in one single study with a controlled, corpus-based, pre–post design in the Pakistani university setting. The present study fills this gap by following the same 60 undergraduate writers through one set argumentative writing task before and after an 8-week intervention using AI and coding each essay with the same 10-category taxonomy, as well as statistically testing pre–post differences, instead of impressionistically comparing them. The analysis is guided by the following three research questions:

Research Question 1. Does an eight-week intervention in undergraduate EFL argumentative writing through the use of Artificial Intelligence (AI) have a positive effect on the overall rate of coded errors?

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Research Question 2. What areas of mistakes have the biggest and smallest decrease following the intervention?

Research Question 3. Are there gender, year of study, academic major or self-rated proficiency differences in terms of the size of the gains?

Method

Research Design

The research design in this study was pre-test–post-test with one group that was based on the Corpus. Each participant wrote two argumentative essays on the same prompt and type of writing, one at the beginning and one at the end of an 8-week AI-assisted writing intervention. A cross-sectional comparison of the error rates of the same ten error categories coded at both time points was possible as it was the strongest design that could be used when a no-treatment control group was not possible for ethical and administrative reasons within the ongoing course.

Participants

The study participants comprised 60 undergraduate students of EFL, Hyderabad Division, Sindh in 2024-2025. The participants from University of Sindh, Jamshoro campus, University of Sindh, Hyderabad campus, Government College University, Hyderabad and a small group of participants from Shaheed Benazir Bhutto University at Hyderabad campus. All the participants were studying in Year 2 or Year 3 of English Language or Applied Linguistics degree. The sample has been described in Table 1 in terms of their demographic characteristics.

Table 1

Participant Demographics (N = 60)

Characteristic	n	%
Gender		
Female	38	63.3
Male	22	36.7
Year of study		
Year 2	27	45.0
Year 3	33	55.0
Academic major		
English Language	41	68.3
Applied Linguistics	19	31.7
Self-rated CEFR level		
B1	29	48.3
B2	31	51.7
Institution		
University of Sindh, Jamshoro	28	46.7
Government College University, Hyderabad	19	31.7
University of Sindh, Hyderabad Campus	9	15.0
Shaheed Benazir Bhutto University (Hyderabad affiliation)	4	6.7

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Note. Age ranged from 18 to 22 years (M = 20.38, SD = 1.18). Reported years of formal English study ranged from 6 to 13 years (M = 9.67, SD = 1.75).

Setting and Writing Task

The essays were all collected timed, at the participating campuses. Students were asked to write one argumentative paper on the question of whether a university should make it a requirement for graduation for students to complete community service hours and the word count was set at 350-450 words. The Week 8 post-intervention essay was exactly the same prompt and genre as the pre-intervention essay, with the exception that in order to control for both topic familiarity and genre effects, there was no restriction on exact wording or length of the essay.

Intervention

The AI-assisted writing intervention took place between the two writing samples, which lasted for eight weeks. Students twice per week submitted a draft or revision to a large language model set up to provide feedback on grammar, lexical choice and discourse organization, and revised their work to address this feedback. This intervention was integrated into the students' routine writing classes and not as a standalone intervention which is now the case in classes throughout Pakistan.

Error Taxonomy and Coding Procedure

All the essays were divided into word tokens and errors in each of ten error categories were coded: articles, prepositions, verb tense/aspect, subject–verb agreement, lexical choice/collocation, word form/morphology, spelling, punctuation/capitalisation, discourse/cohesion markers and a residual other category, which included errors in the use of pronouns and quantifiers, and word-order errors. The essays were independently coded by two trained raters. To ensure reliability, 20% of the corpus was double coded. Agreement between scorers on whether the token had an error or not was strong (Cohen's $\kappa = .87$) as was agreement on which of the ten categories applied to an error identified (Cohen's $\kappa = .83$). Both values fall into the almost perfect range as proposed by Landis and Koch (1977) and so the counts can be viewed as a reliable base for statistical comparisons.

Data Analysis

However, since there were some differences among participants and over time in the number of words per essay, the raw counts for each of the error types were calculated to 100 words before they were compared. This is important because the post-intervention essays, on average, were a bit longer than the pre-intervention essays, so the simple comparison of counts would underestimate the magnitude of the change. Pre and post rates of total errors, and each of the ten categories, and Cohen's d for paired samples were used to index the size of each effect, evaluated using Cohen's (1988) conventional benchmarks of small (d = 0.20), medium (d = 0.50) and large (d = 0.80). To answer the third research question, independent-samples t tests were conducted to determine if there was a significant difference between the magnitude of the individual reduction in TEA by gender, year of study, academic major, or self-rated CEFR level. An alpha of .05 was used for all tests.

Results

Essay length increased slightly from the pre-intervention sample (M = 407.73 tokens, SD = 37.41) to the post-intervention sample (M = 425.83 tokens, SD = 49.05), $t(59) = -2.25$, $p = .028$. This small improvement indicates that the students were writing slightly more fluently following the intervention, and this is why the number of errors in each of the following error counts was expressed as a rate per 100 words (and not simply as a number) to ensure that this comparison is not complicated by differences in the length of the essays.

Overall Change in Error Rate (Research Question 1)

The mean number of raw total errors pre-intervention was 57.28 errors per essay (SD = 7.68) while the mean number of raw total errors post-intervention was 32.43 errors per essay (SD = 4.64). Expressed as a rate per 100 words, the total error rate fell from 14.15 (SD = 2.15) before

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

the intervention to 7.70 (SD = 1.36) after it, a reduction of 45.6%. This difference was statistically significant, t (59) = 19.99, p < .001 and effect size was very large (d = 2.58). Complete descriptive and inferential statistics are presented in Table 2.

Table 2

Descriptive Statistics and Paired-Samples t Test for Total Error Rate (Errors per 100 Words)

Measure	Pre M (SD)	Post M (SD)	t(59)	p	d
Total errors	14.15 (2.15)	7.70 (1.36)	19.99	< .001	2.58

Note. Error rates are expressed per 100 words to control for the small increase in essay length between time points.

Category-Specific Patterns (Research Question 2)

A statistically significant difference in the amount of each type of error reduced from pre- to post-intervention was found for all ten error categories and the amount of reduction was quite variable. Table 3 lists all the paired comparisons from largest to smallest effect size. The highest percentage drops were for word form and morphology (57.4%), verb tense and aspect (56.7%), and subject–verb agreement (55.8%). Articles also improved significantly (54.8%) as did spelling (50.8%) and prepositions (44.5%). The smallest decreases were in the use of punctuation and capitalisation (19.3%) and in the residual other category (21.6%), in between of which were the decreases in the use of discourse and cohesion markers (31.1%).

Table 3

Paired-Samples t Test Results by Error Category (Errors per 100 Words)

Category	Pre M (SD)	Post M (SD)	t(59)	p	d	% red.
Word form/morphology	1.08 (0.52)	0.46 (0.28)	7.65	< .001	0.99	57.4
Verb tense/aspect	1.80 (0.68)	0.78 (0.49)	9.84	< .001	1.27	56.7
Subject–verb agreement	1.30 (0.56)	0.58 (0.37)	7.81	< .001	1.01	55.8
Articles	2.58 (0.85)	1.16 (0.51)	10.81	< .001	1.40	54.8
Spelling	0.89 (0.46)	0.44 (0.25)	6.67	< .001	0.86	50.8
Prepositions	2.05 (0.79)	1.14 (0.42)	7.33	< .001	0.95	44.5
Lexical choice/collocation	1.64 (0.71)	1.01 (0.45)	6.39	< .001	0.83	38.5
Discourse/cohesion	0.91 (0.38)	0.63 (0.33)	4.03	< .001	0.52	31.1
Other	0.89 (0.40)	0.70 (0.35)	2.64	.011	0.34	21.6
Punctuation/capitalisation	1.01 (0.41)	0.82 (0.40)	2.41	.019	0.31	19.3

Note. Categories are ordered from the largest to the smallest effect size. All comparisons were significant at p < .05.

Also, relative rankings of categories did not change significantly from pre to post intervention. The articles category was the most common category both before and after the intervention, and prepositions was the second most common category before and after the intervention. The pattern shows that while the errors in the use of articles and prepositions were decreased in absolute numbers, they remained the two most persistent error types that this population finds most problematic – a fact that was addressed in the Discussion.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Individual Differences in the Size of the Gain (Research Question 3)

The size of each participant's improvement, measured as the drop in total error rate from pre- to post-intervention, did not differ significantly by gender, $t(58) = 0.20$, $p = .846$, by year of study, $t(58) = 1.04$, $p = .301$, by academic major, $t(58) = -0.11$, $p = .912$, or by self-rated CEFR level, $t(58) = 0.88$, $p = .381$. Both groups revealed a mean of a 6-7 error decrease per 100 words in each comparison, and standard deviations overlapped in each comparison. This indicates that the intervention was not associated with disproportionate amounts of improvement by these background characteristics, but rather a relatively similar level of improvement.

Discussion

Overall Gains in Accuracy

The reduction in error rate (from a little over 14 per 100 words to less than 8) was quite a sizeable effect by any conventional standards, and the effect size was well above Cohen's (1988) definition of large. This finding is more aligned with Ferris (2006) that intensive and consistent corrective feedback can be shown to yield gains, rather than with Truscott (1996) who was not so optimistic. The consistency argument in particular would appear to be germane here. A large language model can be used to apply the same grammatical rules to each of the sentences in each of the submissions twice weekly for eight weeks without getting tired or varying in the accuracy of the application. It is that consistency, not any more profound pedagogical understanding, that may be the key element that accounts for the magnitude of the gain that was observed in this subsample.

Which Categories Improved Most, and Why

The categories which made the greatest gains—word form/morphology, verb tense/aspect, and subject–verb agreement—have the following in common: they are highly rule-governed. The verb agrees with its subject, or it does not, and a language model, trained on vast quantities of well-formed English can do this test with great confidence. The fact that the articles got a boost by the same percentage is perhaps more surprising, since the use of articles is quite idiomatic and context-dependent in English, and it does indicate that the intervention provided students with multiple, concrete opportunities to observe the correct use of articles even in those instances for which there is no single rule.

In contrast, the categories with the least improvement, namely punctuation and capitalisation, the residual other category and to a lesser extent discourse and cohesion markers offer the other way to go. In academic English, punctuation conventions can be partly stylistic style and can be easily identified by a learner in the feedback but it will take a longer time to become automatic in timed writing. Discourse markers and cohesive devices rely on a writer's understanding of the essay as a whole—as opposed to a single sentence that can be addressed sentence-by-sentence in order to provide feedback. The difference is consistent with Zhang et al. (2025) who found that the generative AI feedback was less effective in responding to higher-order writing issues than to grammar and it aligns with the difference that emerged between the accuracy of writing at the sentence level and the control over writing at the discourse level in a sample of Chinese university students, as seen in the present sample of undergraduate writers from Sindh.

Why Articles and Prepositions Remain the Most Persistent Categories

Although the number of article and preposition errors decreased significantly in absolute numbers, these were the most frequent types of errors in the Corpus both prior to and following the intervention. This is in line with a first language transfer account. There is no obligatory article system in Sindhi and Urdu like in English and adpositional system is not directly corresponding with English prepositions. According to this explanation, the accuracy in articles and prepositions is not simply a matter of exposure to correct instances, but of restructuring a more firmly established first-language pattern, to which it is more difficult to change. Both Sarfraz's (2011) analysis of Pakistani undergraduate essays and the overall regional systematic

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

review found that the two categories are the ones that are most frequently responsible for errors, indicating that the trend seen in the present study is probably not peculiar to the sample and the task used in this study, but rather a lasting characteristic of the learner population.

Uniform Benefit Across Subgroups

The intervention seemed to benefit students fairly equally, in terms of their academic major, year of study and self rated proficiency as well as gender. This is a helpful observation to consider from a programme level perspective since the tool did not seem to have the effect of identifying already stronger students and de-privileging weaker students, nor did it have the effect of identification of a ceiling in relation to students who reported higher CEFR level. However, it should be noted that this null finding is a subjective measure of proficiency, and not a standardized test score, so this should not be interpreted as evidence that this intervention is equally effective for each individual learner, but rather that all broad self-perceived proficiency groups benefitted equally.

Pedagogical Implications

These findings suggest some practical measures that can be taken by writing instructors in similar context of Pakistan's universities. Second, AI-assisted feedback cycles seem to be well suited for frequent use as a low-cost supplement for grammatical accuracy and instructors might easily incorporate a cycle such as the one used here into a writing class, with the hope that they do not have to redraw the entire writing curriculum. Second, because articles and prepositions were the most stable categories after 8 weeks of AI feedback, they may require explicit and focused instruction in the classroom, which cannot be achieved only by using AI feedback; this could be carried out with contrastive activities, which make the difference between Sindhi/Urdu and English salient. Third, since punctuation, discourse markers, and other higher-order aspects did not show the highest increase, instructors could use them as areas to devote their own time and efforts to provide feedback, while relying on the AI tool to handle more of the sentence-level grammar feedback, similar to the hybrid approach in the comparative study by Zhang et al. (2025).

Limitations and Future Research

The results are subject to a number of limitations. This study adopted a single group, pre–post design with no untreated comparison group, so the results of the gains may not be the direct result of the AI-assisted intervention. Some part of this improvement may have been due to ordinary practice effects, familiarity with the prompt, or general improvement over the course of a semester. In the future it would be desirable to design a study with a comparison group that still only received traditional feedback from the instructor in order to properly test the intervention's specific contribution.

It is also from one type of text, one writing task and one place or province so it is difficult to extrapolate the results to other places, other text types or other learners who speak different first languages. A standardized placement test would be a more precise measure of proficiency, and a convenient but imprecise measure of the size of the gain moderated by the test of proficiency. Lastly, only two time points were measured during the study, and a delayed post-test was not conducted, so there is no evidence as to whether the gains that were observed at Week 8 will last beyond the end of the structured feedback cycle. Future studies could include a control group, a standardized assessment of students' proficiency level, a delayed post-test of students' performance, and qualitative data on students' actual use of the AI feedback to provide a more comprehensive picture of the magnitude and persistence of these effects.

Conclusion

This corpus-based analysis revealed that there was a significant, large reduction of the coded errors among UG EFL writers in Hyderabad, Sindh after an eight week AI-assisted writing intervention. The gain was general, found in all gender, year of study, major and proficiency groups and was particularly strong for the rule-governed grammatical categories of verb tense,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

subject–verb agreement and word form. Meanwhile, articles and prepositions, which are most closely related to English and the L1 of the students, were still the most common category of errors after the intervention, whereas the higher order errors like punctuation and discourse cohesion were still not improved much. These findings imply that AI-assisted feedback can indeed be a valuable tool to complement the undergraduate EFL classroom, yet it is best used in conjunction with other instructional strategies which should still put attention to error types that cannot be easily automated.

References

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). https://doi.org/10.1037/0000165-000

Channa, L. A. (2017). English in Pakistani public education. Language Problems and Language Planning, 41(1), 1–25.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching, 5(4), 161–170.

Ellis, R. (1994). The study of second language acquisition. Oxford University Press.

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), Feedback in second language writing: Contexts and issues (pp. 81–104). Cambridge University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, 20(3), 465–480.

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

Sarfraz, S. (2011). Error analysis of the written English essays of Pakistani undergraduate students: A case study. Asian Transactions on Basic and Applied Sciences, 1(3).

Selinker, L. (1972). Interlanguage. International Review of Applied Linguistics in Language Teaching, 10(1–4), 209–231.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369.

Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. Humanities and Social Sciences Communications, 11, Article 1086. https://doi.org/10.1057/s41599-024-03543-y

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. Innovation in Language Learning and Teaching. Advance online publication. https://doi.org/10.1080/17501229.2025.2503890