Vol. 4 No. 1 (2026)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

CORPUS-BASED EVALUATION OF AI WRITING FEEDBACK TOOLS' ACCURACY AGAINST HUMAN-ANNOTATED LEARNER CORPORA

Muhammad Khalil[*1], Shah Nawaz khan[2], Iqbal Ghani Anas[3]

*[*1]M. Phil English Linguistics, Department of English, Ghazi University Dera Ghazi Khan, Punjab, Pakistan*
[2]Lecturer, Department of English, Govt. Postgraduate College Dargai, Malakand, Khyber Pakhtunkhwa, Pakistan.
[3]BS Computer Science, Federal Urdu University of Arts, Science & Technology, Pakistan

[*1]mkhalil.jami@gmail.com, [2]shankhanicup@gmail.com, [3]ma6749429@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

DOI: http://doi.org/10.5281/zenodo.21335208

Abstract

The use of automated writing evaluation (AWE) tools and generative artificial intelligence (AI) writing assistants like Grammarly, Microsoft Editor, and large language model (LLM) based systems of L2 writing feedback has been widely adopted in second and foreign language (L2) writing pedagogy. Though these tools are popular, doubts linger over their ability to detect and correct the error types human raters can identify in authentic learner texts using the help of existing annotation schemes. The aim of this study is to perform a corpus-based assessment of the accuracy of the three most popular AI writing feedback tools against two human annotated learner corpora, namely the Cambridge Learner Corpus First Certificate in English (FCE) and the NUS Corpus of Learner English (NUCLE). The study uses quantitative, comparative research design based on Error Analysis and Automated Writing Evaluation validity framework, which calculate precision, recall, and F0.5 scores for each tool for every of seven error categories, namely error category of article, determiner, preposition, verbs' tense and agreement, spelling, punctuation, word choice and word order. Three hundred error-annotated sentences were sampled and AI-made corrections were compared systematically against the gold standard human corrections by the MaxMatch scoring method. Results show that AI tools can accurately detect mechanical errors (spelling and punctuation) and provide high precision and recall on these types of error, but their performance is significantly lower on the context-dependent types of errors (word choice and usage of articles, word order). This suggests that there is an important difference between the ability to detect errors at the surface level and the ability to make judgements about errors at the level of discourse, which is what expert human feedback is able to do. The study ends with pedagogical and developmental suggestions for effective incorporation of AI feedback tools in L2 writing classrooms.

Keywords: *AI, L2 writing, Grammarly, Microsoft Editor, large language model, MaxMatch*

1. Introduction

1.1. Background of the Study

The use of computers to analyze student writing is not novel; it was first pioneered by Page (1966), who showed that a number of statistical properties of a text could be used to predict human holistic scores. This early vision was developed into a broad family of automated writing evaluation (AWE) systems, such as Educational Testing Service's e-rater engine (Attali & Burstein, 2006), the Criterion Online Writing Service (COWS) based on the e-rater engine (Burstein, Chodorow, & Leacock, 2004), and Vantage Learning's IntelliMetric, which combined natural language processing techniques with statistical or rule-based models to score essays, and later, to generate diagnostic feedback on grammar, mechanics, and style (Dikli, 2006; Shermis & Burstein, 2013).

These have since evolved even more in the last 10 years, leading to the mainstreaming of consumer-facing grammar and style checkers like Grammarly and Microsoft Editor, and more recently, the rise of large language model (LLM) based writing assistants that can generate fluent, contextually relevant feedback on-demand. Today, these tools are integrated into word processors, browsers and learning management systems and are a fact of life for the L2 writing process for millions of learners worldwide (Barrot, 2021; Koltovskaia, 2020). The attractions of their approach are that they provide access to immediate, individual, and ever patient feedback that can be used to supplement or, in some cases, replace the more limited feedback capacity of a human teacher (Warschauer & Ware, 2006).

Alongside these technological developments, there has also been a vast infrastructure of learner corpora, or collections of authentic learner texts, often with error annotations, systematically gathered and used

to describe the interlanguage patterns and to train and validate natural language processing systems (Granger, 2002). The Cambridge Learner Corpus (Nicholls, 2003), the FCE error-coded corpus (Yannakoudakis, Briscoe, & Medlock, 2011), and NUS Corpus of Learner English (NUCLE) (Dahlmeier, Ng, & Wu, 2013) are famous error-coded and hand-tagged corpora that provide very detailed error tags with which the accuracy of any automated error-detection system could, in principle, be assessed. The shared task tradition in grammatical error correction (GEC) has also enriched the formalization of rigorous, corpus-based evaluation metrics, with the recently introduced MaxMatch (M2) scorer (Dahlmeier & Ng, 2012) and the ERRANT toolkit (Bryant, Felice, & Briscoe, 2017). This is what motivates the present study: the combination of the rich sources of mature learner corpora, the well-developed evaluation frameworks developed in computational linguistics, and the explosive growth of AI writing feedback tools in classrooms.

1.2. Research Problem

While the empirical literature on the topic of AI writing feedback tools has grown in the last few years and has been used with great frequency by L2 writers, it has been dominated by two lines of research that are useful but fail to fill a crucial gap. The first strand involves the perceptions of learners and teachers of AWE tools, usually conducted via surveys or interviews, and has yielded a mixed bag of positive and sometimes mixed perceptions of usefulness and trust (Koltovskaia, 2020; O'Neill & Russell, 2019). The second strand focuses on the impact of the AWE-supported teaching on writing achievement within a semester/academic term, typically using quasi-experimental studies of classrooms (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Ranalli, 2018). However, few studies directly and systematically assess the linguistic accuracy of AI feedback tools by taking a rigorous approach to precision-recall-F-scores as computational linguistics routinely do when evaluating grammatical error correction systems (Bryant et al., 2017; Dahlmeier & Ng, 2012), by benchmarking against hand-annotated error tags embedded in a well-constructed learner corpus. In cases where accuracy has been tested, it has been almost exclusively on a single instrument, a limited type of errors, and self-constructed, relatively small samples of sentences, which reduces the generalizability and comparability of the findings (Ranalli, Link, & Chukharev-Hudilainen, 2017). This renders it difficult for language teachers, materials developers, and tool designers to obtain a clear, corpus-supported view of precisely where the current AI writing feedback tools are sound, and where they are systematically deficient in terms of the variety of types of errors made by L2 learners.

1.3. Research Questions

1. To what extent do AI writing feedback tools align with human annotations in the FCE and NUCLE corpora, measured through precision, recall, and F0.5 scores?
2. Does tool accuracy vary across error types such as articles, prepositions, verb forms, spelling, punctuation, word choice, and word order?
3. How do variations in accuracy across tools and error types inform their pedagogical use in L2 writing instruction?

1.4. Research Objectives

1. To measure the agreement between three AI writing tools and human-annotated learner corpora using standard evaluation metrics.
2. To compare the accuracy of different tool types (rule-based, statistical, and LLM-based) to identify performance differences.
3. To determine which error categories are reliably detected and which require deeper contextual understanding.

1.5. Significance of the Study

It is important to emphasize that this study benefits three audiences, which are related to each other. It provides a scaffolding for pedagogical use of AI-generated feedback that can inform classroom practitioners of which types of feedback will benefit learners after the least amount of teacher check and which will be more useful with more teacher mediation.It provides an empirically based roadmap for pedagogical use of AI feedback that can guide classroom practitioners in determining which categories will benefit students with less teacher involvement and which will require more mediation. The category-specific accuracy profile derived here reveals the concrete language phenomena, especially in terms of article usage, word choice, and word order, that need to be further improved in the models for developers of AI writing feedback systems, complementing benchmarking work that is mostly focused on computational linguistics platforms (Napeles, Sakaguchi, &

Tetreault, 2017; Ng et al., 2014). The study also illustrates the potential of transferring some of the methodological tools developed in the field of benchmarking grammatical error correction systems in natural language processing (NLP), such as the M2 scoring and ERRANT error typology, to the field of applied corpus linguistics, in the classroom, offering a methodological bridge between these two linguistic strands that have often operated in parallel (Leacock, Chodorow, Gamon, & Tetreault, 2010).

2. Literature Review

One of the first attempts at automated writing evaluation took place when Page (1966) showed that certain text characteristics that could be computed (e.g., essay length, word variety) could be used to statistically approximate the holistic scores assigned by trained human evaluators. This early proof of concept was the basis for increasingly sophisticated systems that have been built since the 1990s, such as the e-rater system developed by Educational Testing Service (ETS) that integrates natural language processing algorithms for grammar, usage, mechanics, and style, along with statistical models trained on a large number of human-scored essays (Attali & Burstein, 2006). E-rater was then incorporated into the Criterion Online Writing Service, a writing service that was explicitly designed to provide teaching feedback, not just limited to summative scoring, which marked a step toward formative, diagnostic AWE (Burstein et al., 2004). The extensive work of Dikli (2006) and later, that of Shermis and Burstein (2013), both provide a thorough record of the use of automated essay-scoring systems in the commercial and educational environments, and continuously emphasize that while the scores may be accurate overall, the feedback may not be accurate or actionable at the level of the individual grammatical error, as will be evident throughout the present study.

There has been significant research on the pedagogical effects of AWE tools in the classroom, once they have been implemented in an actual writing course. Warschauer and Ware (2006) eloquently reviewed the literature and contended that the AWE research program for the classroom should shift from efficacy to more complex questions of classroom practices of teachers and students dealing with automated feedback in social and institutional settings. The line of inquiry was continued by Chen and Cheng (2008) in three EFL writing classes who found that the effectiveness perceived by the students was not only attributed to the technical capability of AWE, but also to how the teachers incorporated the feedback into their overall pedagogical design. However, in a multi-site case study, Grimes and Warschauer (2010) also concluded that AWE tools were perceived as not always perfect, but still helpful, and that additional teacher scaffolding was required for success. More recently, Nunes, Cordeiro, Limpo, and Castro (2022) reviewed the literature on AWE in school-based studies from 2000 to 2020 and found that, overall, AWE use is correlated with improvements to students' writing quality, but the evidence is not uniform in terms of methodological rigor and very few studies independently validate the linguistic accuracy of the automated feedback students receive, rather than only measure the positive outcomes at the end of the writing process.

All of this pedagogical research is embedded in a rich theoretical literature on written corrective feedback (WCF) in SLA, which already offers the conceptual vocabulary for what constitutes accurate and useful feedback, in the first place. Ellis (2009) posited a useful typology that classified feedback as either direct or indirect, and focused or unfocused, which has since been adopted to describe the feedback behaviour of automated systems as well as human teachers. Ferris (2006) examined research on the benefits of error feedback for student writers, and found that the effectiveness of error feedback depends upon its accuracy and specificity, which is very relevant to any tool that has a varying degree of accuracy and specificity when detecting errors. In their detailed discussion of written corrective feedback in L2 writing, Bitchener and Ferris (2012) also highlighted the fact that some types of errors, for instance those that have complex, meaning-dependent rules, e.g., article usage, are more difficult to correct reliably, regardless of whether by human or machine, than others are, e.g., subject-verb agreement, spelling.

The method followed in the present study is directly inspired by the methodology of learner corpus research and grammatical error correction (GEC) benchmarking in computational linguistics. In his seminal overview, Granger (2002) has identified learner corpora as a valuable resource in the study of L2 acquisition and in natural language processing applications, and this is due to its systematic error annotation. Error-tagged learner corpora have also been seen as having special potential for use in computer-assisted language learning by Granger (2003), who was one of the first to predict that much of the future research on AWE and GEC technology will be application of this kind. The tradition of annotating errors has been implemented at large scale in resources like the NUS Corpus of Learner English (NUCLE, Dahlmeier et al., 2013), which has been

developed specifically for GEC system development and evaluation, the Cambridge Learner Corpus (Nicholls, 2003), and the FCE dataset released by Yannakoudakis et al. (2011) as part of their research into automatic essay-grading. Despite their slight differences in genre, level of proficiency, and level of granularity in the annotations, all of these corpora feature human-validated sentence-level error tags as gold standard which allow for the accurate evaluation of rigorously the corpora.

The community of computational linguistics built upon these corpora, and produced a series of shared tasks that was designed to provide a framework for assessing automated error correction. Helping Our Own is a pilot task that was one of the first to carry out benchmarking of automated EC systems against human-edited learner text, using structured metrics (Dale & Kilgarriff, 2011). The MaxMatch scorer of Dahlmeier and Ng (2012) was subsequently adopted for the evaluation of the shared task of CoNLL-2014 (Ng et al., 2014), which matches system-generated edits with the human gold-standard edits at the phrase level and computes precision, recall, and a score F0.5 that gives more weight to precision than recall, because spurious or incorrect automated corrections are pedagogically more harmful than missed errors. Subsequent, Naples et al. (2017) presented the JFLEG corpus and pointed to the shortcomings of an edit-based approach to fluency errors, while Bryant et al. (2017) introduced their rule-based method, ERRANT (Error Recovery and Revision Analysis of Natural Transcriptions), for automatically categorizing edits into linguistically motivated error types, thus providing the basis for reporting accuracy scores per category as done in the present study. Leacock et al. (2010) synthesized the research on automated grammatical error detection by language learners and explicitly concluded that article, preposition, and collocation errors are persistent categories of errors that are hard to correct for automated systems, since they require world knowledge, discourse context, or subtle semantic judgement, none of which solely syntactic or statistical models can capture.

The last strand of literature to have been assessed is writing feedback tools for use by consumers, mostly for student engagement or perception, rather than for corpus-validated accuracy, as in the previous strand. In a study of university students' attitudes towards Grammarly, O'Neill and Russell (2019) reported that attitudes were largely positive, although it seemed that students sometimes found that the tool failed to detect errors or suggested errors that were not warranted in a specific context. In Koltovskaia (2020), the researcher used multiple case studies to examine the students' engagement with the automated written corrective feedback provided with Grammarly. The results indicated that there was a wide range of student attitudes towards the tool's suggestions, with some students simply accepting incorrect suggestions. In his study of the effectiveness of automated written corrective feedback in facilitating student learning, Ranalli (2018) reported that students were not always able to leverage the AWCF, that is, to understand and apply the automated suggestions, especially when they were ambiguous or incorrect, and that in his study, explicit accuracy evidence was required for automated suggestions to be validated for formative assessment purposes, as argued by Ranalli, Link and Chukharev-Hudilainen (2017). In the classroom context, Barrot (2021) examined the impact of AWCF on L2 writing accuracy and concluded that while there was evidence of improvements in some aspects, others did not improve, indicating an uneven accuracy profile across linguistic phenomena. As a whole, this literature points to an important gap: Although the corpus-based field is well-developed with a number of good resources available, the domain of GEC has received a number of evaluations, and the number of AI writing feedback tools that have been used in classrooms is increasing, the combination of these three elements is remarkably scarce in the literature and is the focus of the present study.

3. Research Methodology

3.1. Nature of the Study

The type of study is quantitative and comparative-evaluative study. It is not an experimentation of the learner's behavior or classroom factors, but it employs a corpus-based benchmarking methodology that draws on traditions from computational linguistics evaluation in order to address an applied linguistics research question related to the classroom-relevant accuracy of AI writing feedback tools. The study is non-interventionist and analytic, with a focus on the existing corpus of learner annotations that are publicly documented as a gold standard for evaluation, and not a reliance on new subjective judgments by human subjects.

3.2. Research Design

The evaluation design used in this research is a cross section comparative corpus based. The AI writing feedback tool, operationalized in the design at the three levels, is a rule-based & statistics-based commercial grammar checker (Tool A), an integrated word-processor grammar checker (Tool B), and a large language

model-based grammar feedback system (Tool C). It measures the standard information-retrieval accuracy measures of precision, recall, and F0.5, for seven linguistically defined error categories. Both a between-tool comparison (which tool is best) and a within-tool, between-category comparison (which type of errors is the system better/worse at addressing) are addressed by this design, jointly answering RQ1 and RQ2. The design is explicitly based on the shared-task paradigm in GEC research (Dahlmeier & Ng, 2012; Ng et al., 2014), but is not used for system development purposes here; instead it is oriented towards an applied, classroom-oriented research question.

3.3. Data Collection Procedure

The gold-standard data used in this study came from two publicly documented error-annotated learner corpora: the FCE dataset (Yannakoudakis et al., 2011) – a collection of upper-intermediate level ESOL examination scripts which are error-annotated and corrected to twenty-five categories of error, and NUS Corpus of Learner English, or NUCLE (Dahlmeier et al., 2013) – a set of essays written by university students, annotated with errors and classified in accordance with a twenty-seven-category error taxonomy. All original error tagging labels were first translated to seven broad, linguistically coherent error categories that exist in both taxonomies and are found in the second language writing literature, where they are well documented and are pedagogically relevant to errors (Bitchener & Ferris, 2012; Ellis, 2009): article and determiner errors, preposition errors, verb tense and agreement errors, spelling errors, punctuation errors, word choice errors, word order errors.

Three hundred error annotated sentences were then sampled from the combined corpora in a stratified manner so that roughly 40-45 sentences were selected from each of the seven error categories, ensuring that there was enough statistical power for comparisons to be made between the different categories, while at the same time being a manageable sample for detailed manual error analysis. Each sampled sentence was kept in its original, uncorrected learner form, along with the gold standard human correction as reported in the source corpus annotation. Those three hundred uncorrected sentences were then presented to each of the three AI writing feedback tools, all in the same format (no added context beyond the sentence itself) and under the same conditions as used in the previous step, using each tool's standard grammar- and style-checking function; in the case of the large language model tool, a fixed standardized prompt was used for every sentence, asking it to identify and correct grammar, word usage, and mechanical errors, so that elicitation conditions would be held constant across the sample. Suggested edits for each tool were recorded verbatim, and then aligned and scored with the relevant gold standard edits.

3.4. Data Analysis Procedure

The alignment process used was the one established in GEC evaluation research, in which the proposed edit for each sentence was first automatically aligned with the gold edit for the sentence following the scoring approach described in Dahlmeier & Ng (2012) – called MaxMatch, or M2, – for which both the system and the gold edit for a sentence are represented as sets of token-span corrections and the optimal alignment between the two is computed using a lattice-based matching algorithm. An accuracy score was computed for each type of error as well as the overall accuracy by applying the classification procedure described by Bryant et al. (2017), which was based on a linguistically motivated error classification system (ERRANT). Edits were then classified into the seven categories of errors described above.

A true positive was assigned to an error category and to each tool if an edit suggested by the AI was equal in scope and content to an edit made by the gold standard annotation, a false positive was assigned if the proposed edit was not the same as the gold standard edit, or if the proposed edit corrected the error incorrectly, and a false negative was assigned if the AI did not detect an error that was identified in the gold standard annotation. Precision was calculated as the ratio of the number of edits that the AI system proposed which were considered correct in the gold standard to the number of edits in the gold standard, whereas recall was defined as the ratio of the number of gold standard errors correctly recognized by the AI system and corrected, over the number of all gold standard errors. The F0.5 score was calculated as the weighted harmonic mean of precision and recall, with a weighting of 0.5 to give precision more weight than recall due to the pedagogical cost of incorrect automatic corrections being greater than the cost of missed corrections (Ng et al., 2014). To complement the descriptive profile of accuracy (precision-recall-F0.5), an inferential measure was used, namely a chi-square test of independence, which compared the distribution of correct versus incorrect edits with the combination of tool and error category, thus directly addressing RQ2.

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

3.5. Theoretical Framework

This study is theoretically informed by two complimentary frameworks. The first is Error Analysis (Corder, 1967), which was originally developed to state that errors made by learners are not random but systematic, reflecting the learners' developing interlanguage system, and can therefore be described and pedagogically manipulated linguistically coherently, e.g. by analysis of article, preposition, or verb morphological errors. Error Analysis also offers the conceptual justification for the organization of the accuracy evaluation in this study around linguistically named categories of errors, rather than seeing all errors as a mass, and it provides the basis for the expectation that these categories will differ systematically in terms of their formal regularity and hence their suitability to automated error detection, as found in the literature reviewed above.

The second framework is the validity argument-based approach to the validation of automated writing evaluation, which was most clearly developed in the work of Ranalli et al. (2017) based on previous validity theory, in which the accurate linguistic feedback provided by an AWE tool is one of a number of components of a validity argument for employing this tool for the purpose of formative assessment; this must also include evidence of interpretation and action of the feedback by the learner and evidence of positive effects on the learner's production of the text. This notion of accuracy is a key part of the overall notion of validity in this study, but the present study is not being used as evidence to answer questions of pedagogical utility beyond accuracy. Together, Error Analysis provides the categorical structure for the evaluation, and the argument-based AWE validity framework provides the interpretive frame in which the resulting accuracy findings should be understood as part of the evidence needed to assess the value of these tools in the classroom.

4. Data Analysis and Discussion

Table 1 shows precision (P), recall (R) and F0.5 results for the three AI writing feedback tools and the seven error categories in this study, where three hundred sentences derived from the corpus had been human-annotated, sampled in a stratified manner.

Table 1

Precision, Recall, and F0.5 Scores by Tool and Error Category

Error Category	Tool A P	Tool A R	Tool A F0.5	Tool B P	Tool B R	Tool B F0.5	Tool C P	Tool C R	Tool C F0.5
Spelling	.95	.92	.94	.93	.89	.92	.91	.94	.92
Verb Tense/Agreement	.84	.78	.83	.80	.74	.79	.82	.85	.83
Punctuation	.88	.81	.86	.85	.79	.84	.79	.83	.80
Preposition	.71	.63	.69	.68	.60	.66	.74	.70	.73
Article/Determiner	.66	.55	.63	.61	.52	.59	.70	.66	.69
Word Choice	.58	.47	.55	.54	.44	.51	.68	.61	.66
Word Order	.52	.41	.49	.49	.38	.46	.64	.58	.63

Table 1 shows a clear and theoretically interpretable trend for all three tools: For mechanical errors, accuracy is always at the highest level, while for meaning- and context-dependent errors, it is always at the lowest level, thus addressing RQ2. A high level of accuracy in spelling errors was achieved overall, with Tool A achieving a precision of .95 and an F0.5 score of .94, followed closely by Tools B and C; this is not surprising because spelling correction is largely based on comparison with a fixed lexical dictionary and uses only a small amount of contextual or syntactic reasoning. The relatively strong results for punctuation and verb tense or agreement errors also followed a similar pattern, agreeing with the idea that subject-verb agreement and standard punctuation rules are largely rule-based and can be mapped on to syntactic patterns which are fairly local (Leacock et al., 2010).

In contrast, the accuracy rates for article and determiner errors, word choice errors and word order errors were significantly lower in all three tools, ranging as low as .49 for word order errors for Tool B and .55 for word choice errors for Tool A. This is in line with a well-known trend in the computational linguistics

literature that, errors in articles and prepositions are particularly hard to automatically correct because the correct answer may be based on information at a higher level in the discourse, such as whether the NP refers to previously mentioned or shared knowledge (Leacock et al., 2010). A similar problem lies with word choice errors: the decision to use a certain lexical item in a text often involves a semantic and pragmatic assessment of the author's intended meaning, which a purely statistical or shallow-syntactic model is less likely to make, a point that was made in the justification for separating them in the ERRANT framework from the other categories of grammar errors (Bryant et al., 2017).

A second interesting cross-tool effect is that the large language model feedback system Tool C performed better than the two more typical grammar-checking tools on context-dependent error types, such as preposition, article, word choice and word order errors, and more similarly to, or slightly worse than, Tools A and B on mechanical error types, such as spelling and punctuation errors. While this is consistent with the current abilities of large language models, modelling long-range relationships of context and meaning, it does not mean that the models are better equipped to make the kind of discourse-sensitive judgements that article and word-choice correction requires, as simpler, dictionary- or pattern-based approaches can be equally or more reliable in cases where mechanical corrections are narrowly rule-governed. Using the pooled number of true positives, false positives, and false negatives across the 21 tool-by-category cells, a chi-square test of independence confirmed that accuracy outcomes were significantly related to the combination of tool and error category, $\chi 2(12, N = 2{,}100) = 184.6$, $p < .001$, supporting the claim that the differences observed in Table 1 are not likely to be due to chance and that tool choice and error category interact to influence accuracy of AI feedback.

The results of this study directly address questions RQ1 and RQ3. To answer RQ1, the results showed that there is a notable but not systematic accuracy difference between the output of the AI tool and the human gold standard, from very good accuracy (close to expert) for spelling to rather low accuracy for word order, with no single numeric measure of accuracy being appropriate to describe the reliability of an AI writing feedback tool in aggregate; the latter was reflected by the category-level differences in accuracy found across the various aspects of the writing task and is critical to consider but is not captured by a single numeric accuracy measure presented commonly in marketing materials or in some previous evaluations. In response to RQ3, the pattern of results indicates that there is a need for a differentiated pedagogical response rather than a blanket policy of trust or distrust of AI feedback: teachers may want to encourage relatively autonomous use of AI feedback, for mechanical categories like spelling and punctuation, where AI feedback is relatively reliable, whereas they might want to explicitly teach the learners to treat AI suggestions for article, word choice and word order errors as provisional prompts for reflection, as Ranalli found (2018) that L2 learners often lack the metalinguistic knowledge necessary to critically evaluate AI feedback. Ferris and Ellis made the general point that feedback specificity and accuracy work together to determine its pedagogical value (2006, 2009).

These results are also important when placed in the context of the precision favoring logic of the F0.5 metric. As the results indicate, word choice and word order have comparatively low scores, which can also be seen as evidence of false positives, meaning that AI tools made suggestions to change phrasing that were linguistically unnecessary or that led to changes in acceptable, but not native-like, phrasing, a concern raised by Koltovskaia (2020) and O'Neill and Russell (2019) with respect to Grammarly, where they noted that students became overly dependent on the AI's suggestions that proved to be linguistically inappropriate or even wrong. The relative precision-to-recall ratio for most categories in this study indicates that the design of Tool A was geared towards “cautious” flagging of fewer but more accurate errors, while the design of Tool C was more geared towards “borderline” flagging of more accurate errors, which may be a desired approach of some developers and instructors for an instructional purpose depending on what they look to achieve in terms of benchmarked data.

5. Conclusion and Suggestions

5.1. Conclusion

The aim of this study was to examine the accuracy of three commonly used AI writing feedback tools—with the view of using a rigorous and corpus-based methodology adapted from grammatical error correction research—in determining and correcting the error types found in the three human annotated corpora. The study, which used a stratified sample of 300 sentences from the FCE and NUCLE corpora and compared the accuracy of the AI-generated writing feedback against gold standard human writing annotations by precision,

recall and F0.5 scores, shows that AI writing feedback tools achieve high, near-human level accuracy for mechanical and rule-governed error categories (such as spelling, punctuation and verb tense or agreement), but significantly lower accuracy for context-dependent and meaning-based error categories (such as article and determiner usage, word choice, word order and similar categories). The pattern in this category was statistically significant and consistent across all three tools, but was relatively more favored across the more context-dependent categories when compared with the two more traditional grammar-checking tools.

The results of this study validate that the use of AI writing feedback tools in the classroom or in research cannot simply be assumed to be a reliable measure of linguistic correctness regardless of the type of error, and that any use of such tools must be guided by a clear, scientifically valid knowledge of where their outputs are reliable and where they should be verified by humans. The study also illustrates the methodological transferability of evaluation techniques that have been developed in computational linguistics to applied and corpus-based studies of accuracy measures in the field of AI feedback tools, providing a replicable model for future accuracy evaluations beyond the perception-based and outcome-based approach employed so far. Overall, the findings indicate that the accuracy of AI writing feedback systems is genuine but uneven, following a pattern of learning errors that was previously known but rarely examined in the context of writing in L2 writing feedback systems, formally regular errors versus discourse-dependent errors.

5.2. Suggestions

From the results, some suggestions are offered to language teachers, tool makers and future researchers. Language teachers are urged to include explicit teaching of the differential reliability of AI writing feedback in their writing pedagogy, to have learners accept mechanical suggestions (e.g., spelling and punctuation) with reasonable certainty while critically evaluating suggestions to improve their articles, word choice, or word order, ideally through classroom exercises that require students to defend or refute specific suggestions that come from the AI, based on their own growing metalinguistic awareness, as recommended by Ranalli (2018) and Ferris (2006) for the importance of encouraging students to engage with feedback rather than passively accepting it.

A key remaining hurdle to tool reliability, particularly in more advanced LLM-based writing systems, is the lack of model refinement that focuses on categories specific to the task domain of discourse, and the present study findings suggest that these categories continue to present a primary challenge. More model refinement work is needed specifically for these categories, perhaps integrating broader document-level context, explicit discourse and information-structure modeling, and more sophisticated semantic representations of the learner's intent; category-specific benchmarking, in addition to aggregate accuracy claims, should be a standard element of tool evaluation and marketing transparency. However, for an institution that has embraced AI writing feedback tools at scale, it is recommended to test these tools on writing samples that are relevant to the institution's context and have been annotated before large-scale adoption, as there are many other factors beyond learner proficiency, L1 background, and genre that affect accuracy profile that were not fully sampled in this study.

To conclude, the present design of this study could be expanded along multiple lines for future research: larger and more linguistically diverse L2 corpora, comprising subsamples of learner accuracy in L1, would enable L2 accuracy patterns to be compared with a wider variety of L2 learner populations and help test the generalizability of the patterns found in the present study across different L2 learner groups; longitudinal designs that also measure L2 learner uptake and engagement along with corpus-based accuracy measures would help clarify L2 accuracy outcomes and their actual impact in the classroom, building on the case study approach suggested by Koltovskaia (2020); comparative evaluations of a wider range of L2 accuracy tasks, prompting strategies, and tuning approaches using the LLM-based tools could determine whether the present study's findings of a category-specific advantage for Tool C is representative of LLM-based tools in general or specific to the particular tool and prompting design evaluated here. Combined, these directions will further strengthen the evidence base for integrating AI writing feedback tools into L2 writing instruction in a pedagogically and empirically informed way.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. Journal of Technology, Learning, and Assessment, 4(3), 1–30.

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. Computer Assisted Language Learning, 34(5–6), 853–876.

Bitchener, J., & Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing. Routledge.

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 793–805). Association for Computational Linguistics.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. AI Magazine, 25(3), 27–36.

Chen, C.-F. E., & Cheng, W.-Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. Language Learning & Technology, 12(2), 94–112.

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

Dahlmeier, D., & Ng, H. T. (2012). Better evaluation for grammatical error correction. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 568–572). Association for Computational Linguistics.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 22–31). Association for Computational Linguistics.

Dale, R., & Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In Proceedings of the 13th European Workshop on Natural Language Generation (pp. 242–249). Association for Computational Linguistics.

Dikli, S. (2006). An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment, 5(1), 1–35.

Ellis, R. (2009). A typology of written corrective feedback types. ELT Journal, 63(2), 97–107.

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. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 3–33). John Benjamins.

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

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. Journal of Technology, Learning, and Assessment, 8(6), 1–43.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44, 100450.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. Morgan & Claypool Publishers.

Napoles, C., Sakaguchi, K., & Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 2, pp. 229–234). Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (pp. 1–14). Association for Computational Linguistics.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference (pp. 572–581). Lancaster University.

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. Journal of Computer Assisted Learning, 38(2), 599–620.

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. Australasian Journal of Educational Technology, 35(1), 42–56.

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47(5), 238–243.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? Computer Assisted Language Learning, 31(7), 653–674.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. Educational Psychology, 37(1), 8–25.

Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157–180.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (pp. 180–189). Association for Computational Linguistics.