Annotating Errors in English Learners' Written Language Production: Advancing Automated Written Feedback Systems

Steven Coyne[1,2(✉)], Diana Galvan-Sosa[3], Ryan Spring[1], Camélia Guerraoui[1,2,4], Michael Zock[5], Keisuke Sakaguchi[1,2], and Kentaro Inui[1,2,6]

[1] Tohoku University, Sendai, Japan
coyne.steven.charles.q2@dc.tohoku.ac.jp
[2] RIKEN, Wako, Japan
[3] ALTA Institute, Computer Laboratory, University of Cambridge, Cambridge, UK
[4] INSA Lyon, Villeurbanne, France
[5] CNRS, LIS, Aix-Marseille University, Marseille, France
[6] MBZUAI, Abu Dhabi, United Arab Emirates

Abstract. Recent advances in natural language processing (NLP) have contributed to the development of automated writing evaluation (AWE) systems that can correct grammatical errors. However, while these systems are effective at improving text, they are not optimally designed for language learning. They favor direct revisions, often with a click-to-fix functionality that can be applied without considering the reason for the correction. Meanwhile, depending on the error type, learners may benefit most from simple explanations and strategically indirect hints, especially on generalizable grammatical rules. To support the generation of such feedback, we introduce an annotation framework that models each error's error type and generalizability. For error type classification, we introduce a typology focused on inferring learners' knowledge gaps by connecting their errors to specific grammatical patterns. We collect a dataset of annotated learner errors and corresponding human-written feedback comments, each labeled as a direct correction or hint. With this data, we evaluate keyword-guided, keyword-free, and template-guided methods of generating feedback using large language models (LLMs). Human teachers examined each system's outputs, assessing them on grounds including relevance, factuality, and comprehensibility. We report on the development of the dataset and the performance of the systems investigated.

Keywords: Computer-Assisted Language Learning · Written Corrective Feedback · Writing Assistance · Educational Technology

1 Introduction

Millions of people worldwide study English as a second language (L2), driven in part by its role in professional and academic domains. Research in education
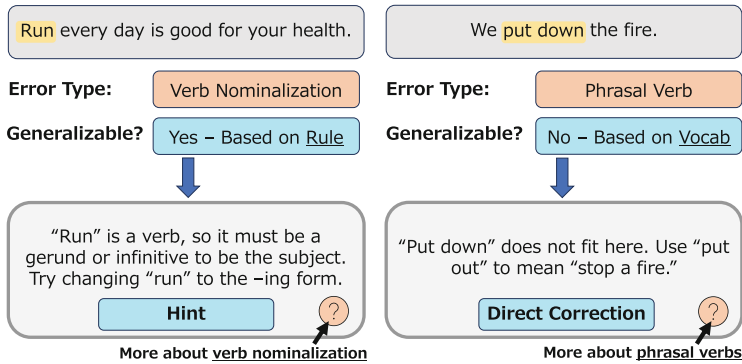
Fig. 1. Example of feedback utilizing our framework, with feedback strategy differentiated based on error type and generalizability.

and applied linguistics has shown that learning is enhanced by targeted, specific, and timely feedback [11], with meta-analysis finding a beneficial impact of written corrective feedback (WCF) on L2 learners' language development [2,14]. However, writing WCF is time-consuming for educators, limiting its availability.

Advances in natural language processing (NLP) offer promising solutions to this challenge, potentially enabling automated systems that can provide learners with consistent and accessible WCF. However, the development of effective automated feedback systems requires careful consideration of pedagogical principles exemplified by human teachers' practices, which we examine in Sect. 2. Based on this insight, we seek to annotate data to model 1) the teacher's interpretation of what knowledge gap an error represents, and 2) the practice of adapting feedback strategy to the details of the error.

Our research makes four contributions to the field of automated WCF: First, we develop an annotation framework that explicitly models both error characteristics and feedback strategies. As detailed in Sect. 3.2, we focus on the **error type**, its **generalizability** ("treatability"), and the **directness** of a feedback comment. Second, we introduce an error typology that better aligns with educational perspectives than existing frameworks, enabling more targeted feedback generation. Third, we gather a dataset of learner errors and feedback comments annotated with our framework, described in Sect. 3. Fourth, we assess the effectiveness of our approach through experiments with large language model (LLM)-based feedback generation using keyword-guided systems, a template-guided system, and a keyword-free system (detailed in Sect. 4.1), followed by manual evaluation of the outputs by practicing English teachers (Sect. 4.2).

We report our results in Sect. 4.3, finding that all systems perform well in this task, with keyword systems and the keyword-free system performing similarly. While the template system performed relatively well when appropriate templates were available, template coverage gaps resulted in mixed or incoherent feedback comments. Furthermore, when no appropriate template option was available, the template system struggled to refrain from giving feedback, select-
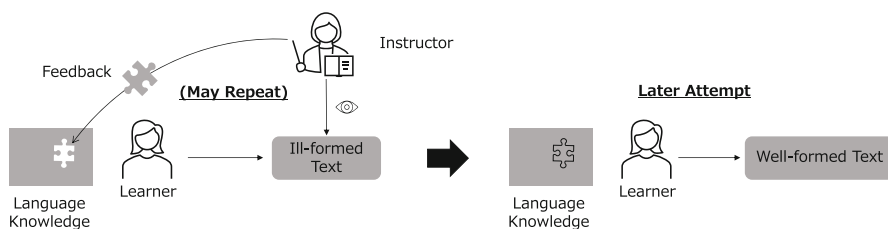
Fig. 2. The cycle of feedback in a learning setting.

ing suboptimal templates instead. However, the template system had the highest level of agreement on directness with human feedback writers.

To support future research in automated feedback, we release our guidelines and resources, available at https://github.com/coynestevencharles/annotating-errors-wcf.

2 Principles of Written Corrective Feedback

Applied linguists have studied corrective feedback for decades. Works such as Bitchener et al. [1], Ferris [9], Sheen [23], and Ellis [6] provide overviews of the topic, describing a number of dimensions of feedback and the factors contributing to a given feedback comment. In this section, we examine WCF as an act of communication and identify factors that teachers use to determine the content and delivery strategy of their comments. Specifically, we seek to identify "key" factors that can be annotated in learner writing corpora in a practical way.

The Feedback Cycle. Feedback in a language education setting is part of a cycle of communication between a learner and an instructor or peer. At first, the learner has partial knowledge of a concept and attempts to apply it. The result may be flawed, such as a sentence with an error. Based on the error, an instructor infers a knowledge gap and intervenes with feedback to help the learner improve. The learner processes the feedback and may make a second attempt. They may be successful on the new attempt, or the cycle may continue as long as their output is monitored. Figure 2 illustrates this learning cycle.

Alignment with this feedback cycle is an important aspect of this work, as it emphasizes the development of the learner's skills over potentially multiple attempts. The assumption of a feedback cycle is one motivation to limit the explicitness of feedback and prefer hints that invoke reflection and self-correction.

Strategies for Giving Feedback. Teachers providing feedback make various decisions about content and delivery, including what errors to target, how explicit to be, and what form the feedback will take [1]. In this work, we focus on metalinguistic feedback [6], which provides information about the causes of errors

and how to fix them. It can vary considerably in complexity, directness of an edit suggestion (if included), and the extent of additional information known as "elaboration" [16, 24]. In the context of this work, the most important distinction to make is the difference between direct corrections and hints.

Teacher decisions about feedback are influenced by many factors, which we divide into three general groups. *Task and setting factors* include the educational context of the writing task (e.g., ESL vs. EFL) and assignment type (e.g., daily writing vs. academic publication). *Learner factors* include age, language level, first language, and past patterns of errors. Finally, there are *error factors*, which depend on the textual product itself. We focus on error factors in this work, as it is more practical to gather text than metadata about real-world context.

One key error factor is the **error type** of the learner's error. As feedback should be targeted and specific to the error in question [11], it is natural for feedback to differ between, e.g., a tense error and a spelling error.

An additional key error factor is the **generalizability** of the error, which has also been referred to as *treatability*. Ferris [8] describes treatability as whether "there are rules to consult" for an error, or whether the error is "non-idiomatic" or "idiosyncratic." Ferris, Hyland, and Hyland [10] reported that teachers gave direct feedback to *untreatable* errors 65.3% of the time, and more indirect feedback 33.6% of the time, whereas for *treatable* errors the feedback was 36.7% direct and 58.7% indirect. This pattern reflects research suggesting that direct corrections are more suitable when the error is based on a non-generalizable, potentially lexical issue, whereas indirect feedback or elaborated hints are more effective for generalizable errors based on broader grammatical rules [2, 18].

3   Annotating Data for Feedback Generation

In this section, we introduce our framework for annotating English learner writing data to facilitate educational feedback comment generation tasks. We first review previous error typologies and annotation schemes, highlighting the need for a new approach. Then, we detail our annotation scheme, which captures both error characteristics and feedback strategies.

3.1   Related Work

When designing our annotation framework and error typology, we considered existing error tag sets in NLP, corpus linguistics, and applied linguistics.

There are several typologies of errors used in the field of grammatical error correction (GEC). These include the system used in the NUCLE dataset [5], the typology used in the Cambridge Learner Corpus [21], and ERRANT [3], currently the most widely applied typology of errors in GEC research.

However, these typologies primarily model surface-level errors for the purpose of correction, without capturing the underlying language knowledge gaps necessary for effective feedback. A tag like "Missing Preposition" indicates the
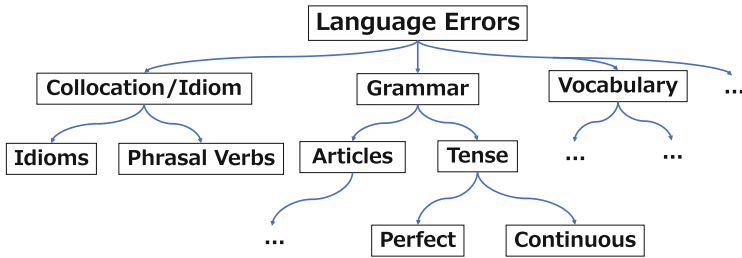
Fig. 3. Sample of our hierarchical learner error typology. Here, "Idioms", "Phrasal Verbs," "Perfect" and "Continuous" serve as terminal tags.

surface error, but does not specify whether it involves a phrasal verb, a prepositional phrase, or another grammatical pattern. This distinction is crucial for providing feedback addressing the specific concept the learner is struggling with.

Recent research in error explanation [13,25] focuses on generating technical descriptions of errors, but these explanations are not written with language learners in mind, and not all function as metalinguistic feedback comments.

Concerning error generalizability and feedback directness, we are not aware of any dataset featuring such annotations for language learners' errors. Furthermore, we are not aware of any well-accepted, comprehensive list of errors and whether they are *generalizable/treatable*. Presenting a dataset with explicit generalizability and feedback directness labels is thus a contribution of this work.

3.2   Annotation Scheme

Error Annotations

Error Type. After examining previous error typologies (see Sect. 3.1), we were motivated to define a new typology of learner errors that better reflects the *topic* of WCF likely to be given in response to the error. That is, the typology targets the perceived language knowledge gap underlying a given error. A sample of the typology can be seen in Fig. 3, and the full typology can be found in the Appendix. Errors are divided into six collections, such as "Vocabulary" and "Grammar," each containing sets of tags and sub-tags. The hierarchical nature of the typology allows three things: (a) to group similar errors; (b) to analyze the output (e.g., classification success rates for each collection or downstream analysis of learner revision by error type); and (c) to toggle collections or individual tags to provide focused feedback. In total, there are 81 terminal tags.

When naming tags, we prioritized alignment with concepts familiar to teachers of English as an additional language, such as terms common in textbook indices. In fact, as these tags represent relevant educational concepts, it is possible to map them to such resources. Thus, in our proposed feedback, shown in Fig. 1, we include resource links, as our framework indirectly supports them.

Error Generalizability. Annotators judge whether an error represents a relatively predictable rule of the language, or is more idiosyncratic, perhaps specific to one lexical item, following the principles outlined in Sect. 2.

Feedback Annotations. In addition to describing the learner's error, annotators provide WCF in response to it. This process includes defining the comment location, writing the feedback comment, and assessing its directness.

Comment Highlight. As with previous works [13,20,25], we annotate text spans to identify error locations, or more accurately, the scope of a feedback comment. A comment highlight must contain all words to be edited, but may extend further to provide appropriate context. For example, in "We put *down* the fire," the minimal highlight would be "down," but our guidelines suggest highlighting "put down" to complete the lexical unit of the phrasal verb. Since the feedback is likely to discuss the two highlighted words as a set, our comment highlight includes both, adapting to the error in question.

Explanation and Edit Suggestion. The feedback is divided into two parts: the explanation "identifies **what is wrong** and **why**," while the edit suggestion "tells the learner **what to do** to fix the issue." This separation enables targeted handling of either field. Otherwise, the default assumption is that the WCF consists of the explanation followed immediately by the edit suggestion.

Directness. Annotators judge whether their edit suggestion provides a direct correction or a hint. We define a direct correction as one that provides the exact text of the suggested edit (e.g., "Change 'eat' to 'ate'") or specifies the exact word(s) and location of an insertion or deletion, while a hint provides a metalinguistic clue (e.g., "Change 'eat' to the past tense").

3.3   Preliminary Dataset Collection

To confirm the robustness of our framework and the quality of our annotation guidelines, we conducted a pilot annotation study. Two annotators with English teaching experience annotated a subset of data in parallel in three batches. Both of them were co-authors of this work, with one of them being the first author.

The base data consisted of examples taken from the training set of a corpus of learner writing and corrections, EXPECT [7]. The instances were randomly selected as a representative sample across learner levels A1-C2 present in the dataset. Annotation was performed in three batches of 114, 114, and 228 instances, for a total of 456 annotated instances. The annotators discussed disagreements and updated the guidelines between the batches.

Inter-Annotator Agreement. We measure agreement for categorical annotations such as error type, generalizability, and directness with exact match rate

Table 1. Inter-annotator agreement observed during the pilot annotation study.

Annotation	Agreement Metric	Batch 1	Batch 2	Batch 3
Error Tag	Exact Match	63.16%	69.30%	76.32%
Error Tag	Krippendorff's α	0.601	0.677	0.794
Comment Highlight	Exact Match	18.42%	51.75%	54.25%
Comment Highlight	Pairwise Token F1	0.375	0.699	0.778
Generalizability	Exact Match	70.18%	74.56%	80.26%
Directness	Exact Match	62.28%	70.18%	80.26%
Rejections	Krippendorff's α	0.366	0.541	0.645

and Krippendorff's Alpha [15]. In the case of error type, we compare the terminal tag, the most fine-grained tag used for a given example.

For the comment highlights, we report exact matches and partial matches. Partial matches are evaluated with pairwise token-level F1 scores.

For metrics other than exact match rate, instances rejected by both annotators are not included in the calculations, including the denominators of averages. Instances rejected by one annotator but annotated by the other are considered mismatches (i.e., null vs. not null), and the instance is assigned a score of 0.

Table 1 shows agreement across batches. Scores on all metrics improved between each batch, suggesting that the discussions and guideline updates were productive. The agreement achieved in the pilot study suggests that our annotation scheme is well-defined and can be effectively applied by human annotators.

4   Experiments

To empirically validate the usefulness of this framework for feedback generation, we explore methods to incorporate our annotations into automated WCF pipelines. The details of each approach can be seen in Sect. 4.1. Human raters evaluated the feedback from each, providing perspectives on its pedagogical validity. The human rating experiment is discussed in Sect. 4.2.

4.1   Feedback Generation

We compare three approaches, all of which incorporate an LLM for text generation: keyword-guided generation using error tags from our typology and existing typologies, template-based generation using fillable templates grouped by tag, and a baseline of keyword-free feedback generation. All systems used the same underlying LLM, `gpt-4o-2024-11-20` [22], with temperature 0 and a response format of `"json_object"`. As input, all systems received the learner's original sentence with the error and correction marked with asterisks and the comment highlight marked with angle brackets. Systems that consider error type tags received the input instance's tag as well. As output, systems return separate `feedback_explanation` and `feedback_suggestion` fields.

```
### Example 1

Example 1 Input:
source: The responsibility of <*the* educational institutions> is to make sure that he/she won't be in
danger in the swimming pool instead of dissuading him/her from getting close to the water.
corrected: The responsibility of *[NONE]* educational institutions is to make sure that he/she won't be
in danger in the swimming pool instead of dissuading him/her from getting close to the water.
error_tag: {One of: "Missing/Unnecessary Article" (Ours), or "Article" (EXPECT), or "U:DET (Determiner is
not needed)" (ERRANT)}

Example 1 Output:
feedback_explanation: The article "the" is not necessary because you are talking about all educational
institutions in general.
feedback_suggestion: Remove the article "the."

### Example 2

Example 2 Input:
source: That is why I totally agree with Richardson's modality dealing with this important issue which is
present in students', schools' and <parents *[NONE]* lives> nowadays.
corrected: That is why I totally agree with Richardson's modality dealing with this important issue which
is present in students', schools' and parents *'* lives nowadays.
error_tag: {One of: "Possessive" (Ours), or "Possessive" (EXPECT), or "M:NOUN:POSS (Noun possessive is
missing)" (ERRANT)}

Example 2 Output:
feedback_explanation: When something belongs to someone, it is necessary to use a possessive.
feedback_suggestion: Change "parents" to a possessive form to show whose lives we are talking about.
```

Fig. 4. The two static few-shot examples as they appear in our prompt. Additional examples follow the same format. `error_tag` is omitted from the keyword-free setting.

For this experiment, we isolate the feedback writing task and set all prerequisite "upstream" information, such as successful error detection, classification, and highlighting, to the ground truth "oracle" data from the human annotations. In a real-world setting, such information must first be obtained from the learner's text. However, by keeping this information stable across all the feedback sources we compare, we can better focus on the characteristics of the feedback content itself, without attributing any differences between settings to a prior stage.

The 228 instances comprising the first two annotation batches are designated as a "Training" set. The other half, the 228 instances from batch 3, are designated as the "Test" set. To avoid potentially flawed data, we exclude all instances rejected by either annotator, leaving the Training set with 189 instances and the Test set with 197. Then, for each instance, we randomly selected one of the two annotations, leaving 50% representation of each human feedback writer.

Keyword-Guided Systems. We test a paradigm of keyword-guided feedback generation, in which the model is prompted to consider an error tag when writing feedback. We define three such systems by providing tags from three different typologies: our typology outlined in Sect. 3.2, the tags from the original EXPECT dataset (available from our base data), and ERRANT tags obtained by comparing the source and corrected sentences for each instance.

The system is provided with up to four few-shot examples of inputs and corresponding feedback comments drawn from the Training set. Two of the examples are static: one with an article error and one with a possessive error. These examples were chosen because all three typologies had clear tags for the errors in question. They can be seen in Fig. 4. Up to two other examples could be provided as well. These were randomly selected from the instances that share the
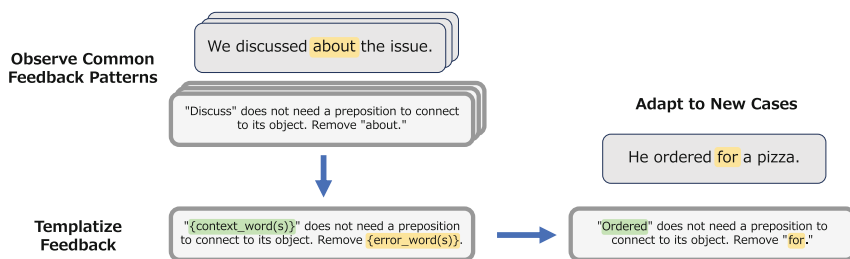
Fig. 5. Visualization of the template construction process. Templates are written by hand based on patterns in our data. Here, `{error_word(s)}` represent what must be removed or edited, and `{context_word(s)}` are used to justify explanations.

error type of the input instance, using the tags from the typology being evaluated. When there are no annotated instances that share a tag with the input, the two static examples remain as a general demonstration of the task.

Some additional adaptations were made for the ERRANT setting. In a few cases, ERRANT assigned multiple tags to an instance in the data. This occurred in 11 of the 386 non-rejected instances used in the experiment, a rate of 2.84%. In these cases, few-shot examples were selected from the set of all instances that share an ERRANT tag with the input, favoring one example each from two different tags if possible. Additionally, since ERRANT tags are expressed in codes such as `M:NOUN:POSS`, we append a short natural language description after each tag used in the prompt. For these, we use the ERRANT tag descriptions from Li and Lan [19], and add our own for the `OTHER` tags, which they omit.

Keyword-Free System. The keyword-free system uses a prompt that is nearly identical to the keyword-guided systems, minus all mention of error types. It uses the same two static examples as the keyword-guided system, plus two examples drawn randomly from the Training set. Since error tags were not considered in the few-shot selection process, this system always received four examples.

Template-Guided System. Works such as Lai and Chang [17] and Coyne [4] describe using templates to control feedback generation. A template is selected, and any slots are filled by lookup mechanisms or neural models. Potentially, this can mitigate issues described in older works in feedback generation, such as Hanawa et al. [12], who report generative outputs that mix multiple training examples or fail to adapt to the specific words in the input sentence. To explore such an approach, we include a template-guided system that uses our error types.

We create feedback templates by manually grouping existing feedback comments by error type tag, observing common patterns, and writing a template based on each "archetype" found, as depicted in Fig. 5. Each archetype represents a set of instances under the umbrella of a given error tag that are similar enough that a template could substitute for any of them. An error tag is assumed to have several such archetypes and, thus, several potential templates.

Table 2. Feedback Rating Criteria. "Out of scope" content includes e.g., assumptions about the learner or AI chat-like phrases such as "Okay, here's my feedback:"

Rating Criterion	Annotation Type
Comment is relevant to the error	Binary Radio
Comment is factually correct	Binary Radio
Comment explains what is wrong and why	Binary Radio
Comment explains what to do to fix the error	Binary Radio
Comment is comprehensible to a CEFR B1-B2 academic learner	Binary Radio
Comment contains "out of scope" content	Binary Radio
Whether the comment is direct, a hint, or NA	Radio (3 options)
Overall quality of the comment	Likert (1–5)
Rater Comments	Text Box

We start by converting the examples in our annotation guidelines into templates. We then examine the Training set, adding any previously unseen archetypes. This resulted 149 templates, 120 from the guidelines and 29 from the data. We do not repeat this for the Test set, in order to keep it "unseen." This process achieved 92.63% coverage of the Training set (remaining errors were judged to be rare and specific, so no template was composed), and 76.76% of the Test set.

For the purpose of the template system experiment, each instance was assigned a ground-truth correct template. At inference time, the system is shown all templates associated with the input instance's error tag and asked to select the most appropriate option. Concretely, the system outputs a `template_id`, and then writes the feedback fields by filling the template. Instances with no template coverage were assigned a `"None"` template, and the system was expected to select this option and refrain from giving feedback in these cases.

4.2 Human Evaluation

To assess the feedback for pedagogical validity, we arranged a task in which annotators rated feedback comments on several criteria, as seen in Table 2.

We recruited four raters, who each had at least 7 years of experience teaching English as an additional language, including writing instruction in high school and university settings. The raters were not co-authors of this paper, but belong to the same university as the first author. They were not provided with the source of each feedback comment or otherwise directed to favor "our" systems. The raters were paid an average of $19.31 USD per hour of work performed.

The raters were provided with a document detailing the task and each rating field, which is available in the Appendix. They attended a paid training session in which the first author demonstrated the task and interface, answered questions, and monitored ratings on practice instances. Annotation was then performed in three batches with quality reviews based on a sample of 10% of the batch plus

the set of a) all instances with a rater comment, b) all instances where human-written feedback was rated as low-quality, and c) all instances where the rater disagreed with the reference directness label for human or template feedback.

4.3   Results

Feedback Generation Results. The template-guided system demonstrated mixed performance in template selection and generation tasks. While it correctly identified the ground truth template in 76.65% of cases, yielding an F1-score of 0.76, it struggled to handle cases where no template was appropriate. When the correct action was to refrain from providing feedback (selecting the ''None'' template), the system often incorrectly attempted to fill an available template, resulting in low recall (0.37). While the system did correctly identify some cases requiring no feedback (all ''None'' outputs were correct), this performance suggests limited robustness in identifying coverage gaps and withholding feedback.

Additionally, we observed template filling errors (e.g., including template syntax such as curly braces in the final text) in 4.57% of outputs. While these outputs can be filtered out programmatically before reaching learners, they represent an additional failure mode that must be taken into consideration.

Table 3. Average ratings for each feedback source. Most values are relatively close.

	Relevant	Factual	What & Why	What to Do	Comp.	Scope ↓	Overall
Human	**1.000**	0.972	0.987	**1.000**	0.952	0.008	4.449
Keyword: Ours	**1.000**	0.970	0.992	**1.000**	0.970	0.008	4.487
Keyword: ERRANT	0.997	0.967	0.992	**1.000**	**0.982**	**0.003**	4.475
Keyword: EXPECT	0.997	**0.975**	0.990	**1.000**	0.975	0.005	**4.500**
Keyword-free	0.995	0.970	**0.997**	**1.000**	**0.982**	0.005	4.495
Templates	0.977	0.921	0.944	0.994	0.980	0.023	4.184

Human Evaluation Results. The results of the human rating task can be seen in Table 3. All systems were rated fairly well in this task, with mean quality ratings between 4 and 5 points. No output from any system was reported as toxic or inappropriate. Keyword-guided systems and the keyword-free system performed similarly, and the specific typology used was not a significant factor. Interestingly, they were rated comparably to the human feedback writers.

Figure 6 shows the breakdown of quality ratings by source. The template system has the highest proportion of feedback comments rated 1 or 2. Examining such cases, the most common pattern is that there were no completely valid templates for the instance, but the system attempted to fill one instead of selecting ''None'' and refraining from writing feedback. The average quality rating is 4.40 when the template is chosen correctly, 3.51 when chosen incorrectly, and 2.88 when a template is chosen and filled when ''None'' is the correct choice.
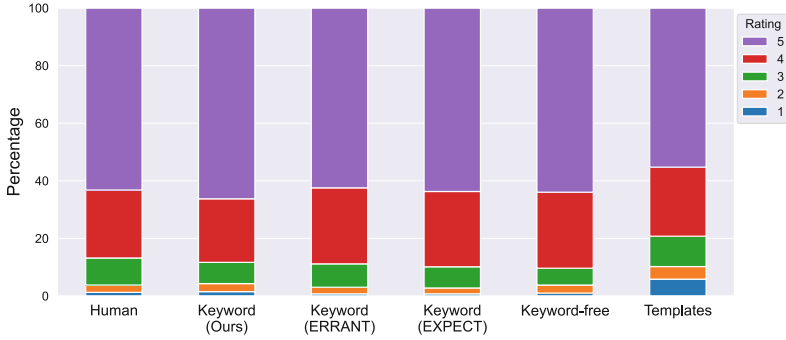
Fig. 6. Distribution of overall quality ratings for different feedback sources.

Despite some language in the prompt to encourage hint feedback, and the presence of in-context examples providing hints, the keyword-guided and keyword-free systems almost always provided direct corrections. The human feedback writers provided hints for 40.86% of instances, but these systems ranged from 0 to 3%. The template-guided system was closer to human practices in this regard, giving hints in 39.77% of instances, for an F1 score of 0.78 for hint feedback.

4.4 Discussion

All models received high ratings, suggesting that LLMs, when properly constrained, can generate valid feedback in many cases. However, even the best-performing system did not achieve 100% factuality. It is imperative to exercise caution generating "pedagogical" text and promote students' AI literacy skills.

For most settings, the feedback was overwhelmingly direct. This did not preclude it from being rated well, but does suggest an area where the AI and human feedback writers differed significantly on qualitative grounds.

The keyword-guided and keyword-free results suggest that the typology used was not a significant factor in feedback quality in this experiment. Further investigation is needed to distinguish whether this is due to a similar informativeness level of all typologies, or a bias or keyword insensitivity in the base model.

The quality issues resulting from template coverage gaps present a challenge for such systems. Nevertheless, our template system was effective for controlling directness, and thus shows some promise for contexts where this is a priority.

5 Conclusion

We define a framework to annotate data with key factors of errors and feedback and explore their impact on LLM-based automated WCF. We found that human teachers rated all systems fairly well, error tags had little effect on feedback quality, and that a template-guided approach reflected human directness decisions

but faced challenges handling template coverage gaps. Limitations include a lack of student ratings or learning outcomes and the use of oracle information for error location and classification. In future work, we will investigate student responses to feedback, methods to control directness without sacrificing generation quality, and performance in fully automated settings requiring error detection.

Appendix. Available at github.com/coynestevencharles/annotating-errors-wcf.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bitchener, J.: Written Corrective Feedback, pp. 207–225. Cambridge Handbooks in Language and Linguistics, Cambridge University Press (2021)
2. Brown, D., Liu, Q., Norouzian, R.: Effectiveness of written corrective feedback in developing l2 accuracy: a Bayesian meta-analysis. Lang. Teach. Res., 13621688221147374 (2023). https://doi.org/10.1177/13621688221147374
3. Bryant, C., Felice, M., Briscoe, T.: Automatic annotation and evaluation of error types for grammatical error correction. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 793–805. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1074
4. Coyne, S.: Template-guided grammatical error feedback comment generation. In: Bassignana, E., Lindemann, M., Petit, A. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 94–104. Association for Computational Linguistics, Dubrovnik, Croatia (2023). https://doi.org/10.18653/v1/2023.eacl-srw.10
5. Dahlmeier, D., Ng, H.T., Wu, S.M.: Building a large annotated corpus of learner English: the NUS corpus of learner English. In: Tetreault, J., Burstein, J., Leacock, C. (eds.) Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 22–31. Association for Computational Linguistics, Atlanta, Georgia (2013). https://aclanthology.org/W13-1703
6. Ellis, R.: A typology of written corrective feedback types. ELT J. **63**(2), 97–107 (2008). https://doi.org/10.1093/elt/ccn023
7. Fei, Y., Cui, L., Yang, S., Lam, W., Lan, Z., Shi, S.: Enhancing grammatical error correction systems with explanations. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7489–7501. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.acl-long.413
8. Ferris, D.: The case for grammar correction in L2 writing classes: a response to truscott (1996). J. Second. Lang. Writ. **8**(1), 1–11 (1999). https://doi.org/10.1016/S1060-3743(99)80110-6

9. Ferris, D.: Treatment of Error in Second Language Student Writing, Second Edition. University of Michigan Press/ELT, Ann Arbor, MI (2011). https://doi.org/10.3998/mpub.2173290
10. Ferris, D., Hyland, K., Hyland, F.: Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction, pp. 81–104. Cambridge Applied Linguistics, Cambridge University Press (2006). https://doi.org/10.1017/CBO9781139524742.007
11. Goodwin, B., Miller, K.: Good feedback is targeted, specific, timely. Educ. Leadership J. Dept. Supervision Curriculum Dev. N.E.A **70**, 82–83 (2012)
12. Hanawa, K., Nagata, R., Inui, K.: Exploring methods for generating feedback comments for writing learning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.T. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9719–9730. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.766
13. Kaneko, M., Okazaki, N.: Controlled generation with prompt insertion for natural language explanations in grammatical error correction. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 3955–3961. ELRA and ICCL, Torino, Italia (2024). https://aclanthology.org/2024.lrec-main.350
14. Kang, E.Y., Han, Z.: The efficacy of written corrective feedback in improving L2 written accuracy: a meta-analysis. Mod. Lang. J. **99**, 1–18 (2015). https://doi.org/10.1111/modl.12189
15. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. SAGE Publications, Inc., Thousand Oaks, California, fourth edition edn. (2019). https://doi.org/10.4135/9781071878781
16. Kulhavy, R.W., Stock, W.A.: Feedback in written instruction: the place of response certitude. Educ. Psychol. Rev. **1**(4), 279–308 (1989). https://doi.org/10.1007/BF01320096
17. Lai, Y.H., Chang, J.: TellMeWhy: learning to explain corrective feedback for second language learners. In: Padó, S., Huang, R. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pp. 235–240. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-3040
18. Lee, I.: Research into practice: written corrective feedback. Lang. Teach. **46**(1), 108–119 (2013). https://doi.org/10.1017/S0261444812000390
19. Li, X., Lan, Y.: Large language models are good annotators for type-aware data augmentation in grammatical error correction. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics, pp. 199–213. Association for Computational Linguistics, Abu Dhabi, UAE (2025). https://aclanthology.org/2025.coling-main.14/
20. Nagata, R.: Toward a task of feedback comment generation for writing learning. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3206–3215. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1316

21. Nicholls, D.: The Cambridge learner corpus: error coding and analysis for lexicography and ELT. In: Proceedings of the Corpus Linguistics 2003 Conference, pp. 572–581 (2003). https://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf
22. OpenAI: GPT-4o system card (2024). https://doi.org/10.48550/arXiv.2410.21276
23. Sheen, Y.: Corrective Feedback, Individual Differences and Second Language Learning, Educational Linguistics, vol. 13. Springer Netherlands, Dordrecht (2011). https://doi.org/10.1007/978-94-007-0548-7
24. Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008). https://doi.org/10.3102/0034654307313795
25. Song, Y., Krishna, K., Bhatt, R., Gimpel, K., Iyyer, M.: GEE! grammar error explanation with large language models. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024, pp. 754–781. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.findings-naacl.49