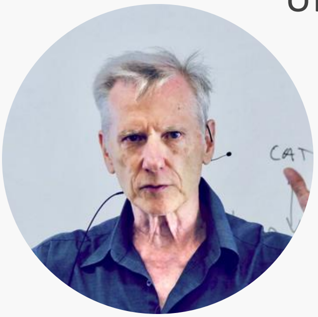
Annotating Errors in English Learners' Written Language Production: Advancing Automated Written Feedback Systems

Steven Coyne
Tohoku University, RIKEN

Diana Galvan-Sosa
ALTA Institute, Computer Laboratory, University of Cambridge

Ryan Spring
Tohoku University

Camélia Guerraoui
INSA Lyon, Tohoku University, RIKEN



Michael Zock
CNRS, LIS, Aix-Marseille University

Keisuke Sakaguchi
Tohoku University, RIKEN

Kentaro Inui
MBZUAI, Tohoku University, RIKEN

TOHOKU UNIVERSITY

RIKEN

UNIVERSITY OF CAMBRIDGE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

amU Aix Marseille Université

جامعة محمد بن زايد للذكاء الاصطناعي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Contents

Setting and Motivation

- Setting: English language learning.
- Learners write English text.
- Teachers write **written corrective feedback (WCF)**.

Run every day is good for your health.

“Run” is a verb, so you can’t use it as the subject.
Change it to a noun by using the –ing form.

Challenges:

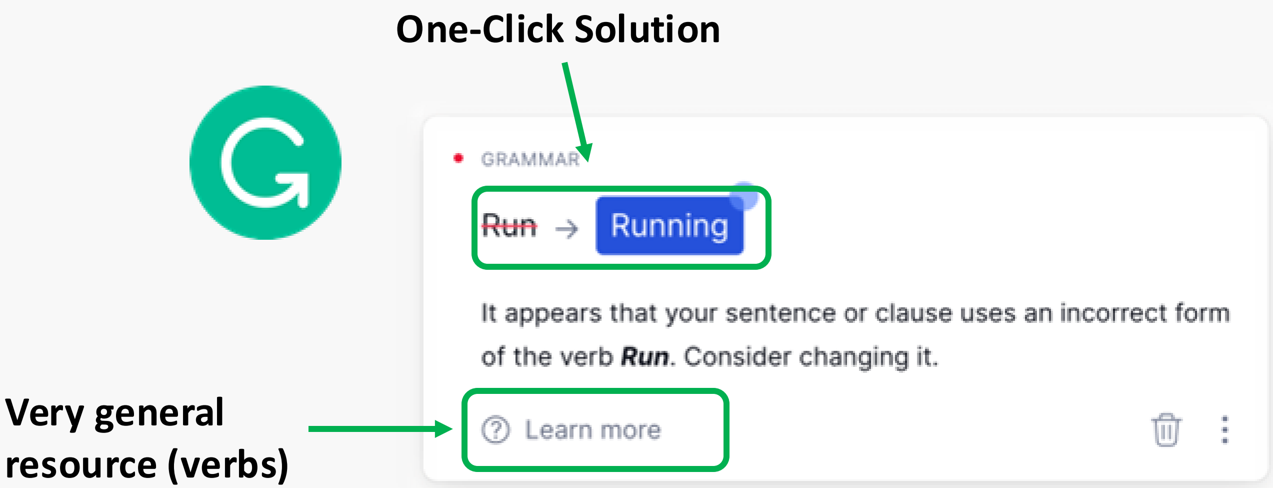
- Labor-intensive.
- Access to instructors is not equal.

Can we automate this?

Existing Systems: Writing Assistance Software

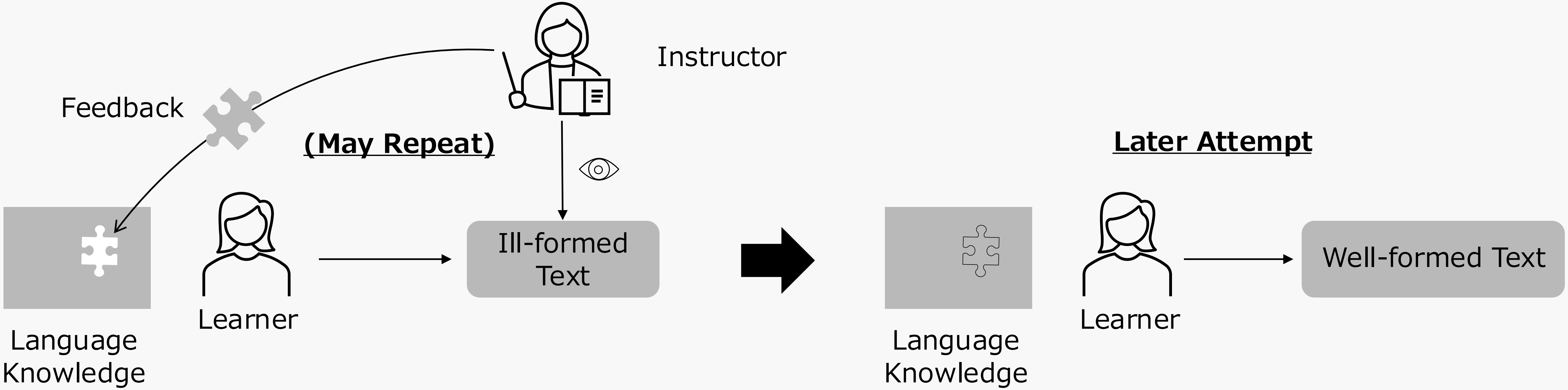
- Feedback from tools like Grammarly focuses on **revision**, not **learning**.
- All feedback includes “click-to-fix” direct corrections.
- Meanwhile, teachers use a variety of strategies based on context, not just direct corrections.

Ex. Run every day is good for your health.



The Feedback Cycle

- Effective learning involves a cycle: Attempt → Feedback → Reflection → New Attempt.
- Teachers **infer a knowledge gap** and choose whether and how to intervene.
- Instead of giving the answer, they may provide a **hint** to encourage **reflection** and **self-correction**.
  - This is where the writing assistants are misaligned.



Modeling the Teacher's Choices

- **How do we build automated WCF systems that can align with teacher practices?**

- **Our Approach:**
  - **Explicitly annotate data with the factors that influence teachers.**
  - **Use this information when generating feedback.**

- We selected two key factors to focus on in this study:
  - **Error Type** (e.g., conditionals vs. spelling)
  - **Error Generalizability** (Is the error based on a rule?)
    - See “Treatability” (Ferris 1999)

- For feedback comments, we focus on aligning the use of **hints** vs. **direct corrections**.

Feedback Generation with Our Approach

Run every day is good for your health.

Feedback Generation with Our Approach

Run every day is good for your health.

Error Type: Verb Nominalization

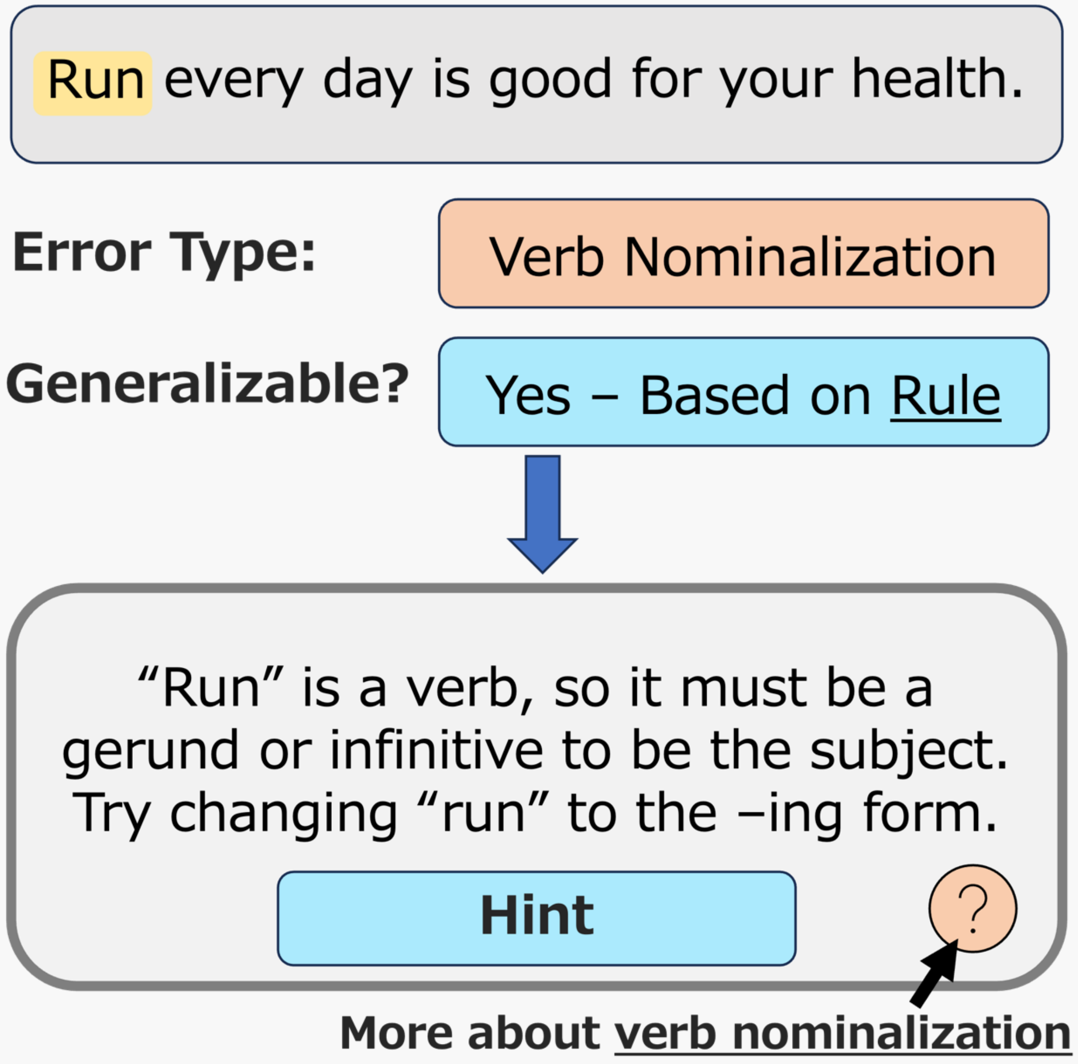
Feedback Generation with Our Approach

Run every day is good for your health.

Error Type: Verb Nominalization

Generalizable? Yes – Based on <u>Rule</u>

Feedback Generation with Our Approach



Feedback Generation with Our Approach

Run every day is good for your health.

We put down the fire.

Error Type: Verb Nominalization

Generalizable? Yes – Based on Rule

"Run" is a verb, so it must be a gerund or infinitive to be the subject. Try changing "run" to the –ing form.

Hint

?

More about verb nominalization

Feedback Generation with Our Approach

Run every day is good for your health.

Error Type: Verb Nominalization

Generalizable? Yes – Based on Rule

“Run” is a verb, so it must be a gerund or infinitive to be the subject. Try changing “run” to the –ing form.

Hint

?

More about verb nominalization

We put down the fire.

Error Type: Phrasal Verb

Generalizable? No – Based on Vocab

Feedback Generation with Our Approach

Run every day is good for your health.

Error Type: Verb Nominalization

Generalizable? Yes – Based on Rule

"Run" is a verb, so it must be a gerund or infinitive to be the subject. Try changing "run" to the –ing form.

Hint

?

More about verb nominalization

We put down the fire.

Error Type: Phrasal Verb

Generalizable? No – Based on Vocab

"Put down" does not fit here. Use "put out" to mean "stop a fire."

Direct Correction

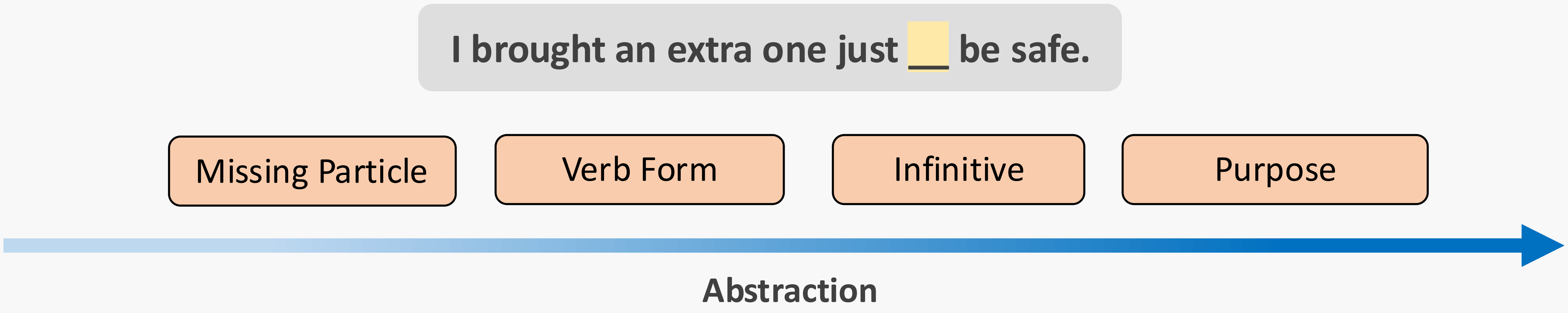
?

More about phrasal verbs

Contents

Annotation Challenges

- **Generalizability:** Somewhat inconsistent lists in the literature; No known accessible dataset
- **Error Type:** Granularity and scope issues:



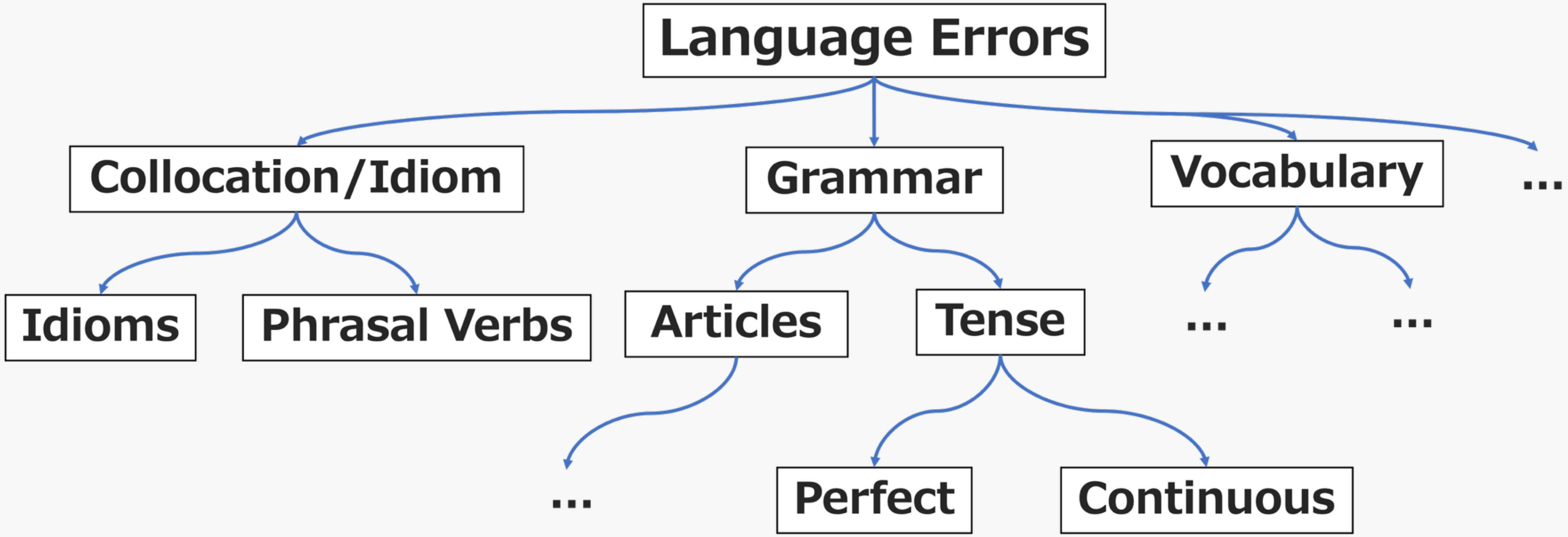
- **Our goal:** Target the underlying learning gap for the most effective feedback.
- Labels should be **useful as keywords** and sufficiently informative.

Existing Typologies

- Established Typologies like ERRANT are great for Grammatical Error Correction (GEC)
- Focuses on edit operations and parts of speech (e.g., "Missing Preposition").
- However, this doesn't specify the **underlying grammatical pattern** the learner struggled with.
- We need a typology designed for error-to-feedback, rather than just error-to-correction.

A New Typology

- We created a new, hierarchical error typology for this task.
- Targets the perceived language knowledge gap behind an error.
- Tag names align with terms familiar to teachers and textbooks - can serve as hooks to link to resources.



Annotation Process

- Two annotators with 5+ years of English teaching experience each annotate 456 instances.
- Base corpus: EXPECT (Fei et al., 2023), based on W&I (Yannakoudakis et al., 2018).

- Example of an annotation:

source: <If my mom *was* here>, she would know what to do.
corrected: If my mom *were* here, she would know what to do.
error_tag: `Conditional`
error_is_generalizable: True
feedback_explanation: In this conditional clause, you can't use "was" with "would."
feedback_suggestion: Check which type of conditional you want to use, and change the tense of the verb.
feedback_is_direct: False

Annotation Process: Agreement

- Annotated the instances in three batches, refining guidelines between each batch.
- Agreement scores consistently improved for all annotation types.
- Suggests the framework is well-defined and can be applied consistently
- Dataset and full guidelines are available online in the appendix.

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

Contents

Experiment: Can an LLM Generate Good Feedback?

- Goal: Use our annotated data to guide an LLM (GPT-4o) in generating feedback.

- Simplified Setup: We provide the model with "oracle" information:
  - The original sentence & its correction.
  - The highlighted error location.
  - The ground-truth error type.

- This isolates the final feedback generation step when comparing systems.

- Half the data is ”train” (usable for few-shot examples), and half is “test” (can include unseen error types)

Systems/Pipelines Used

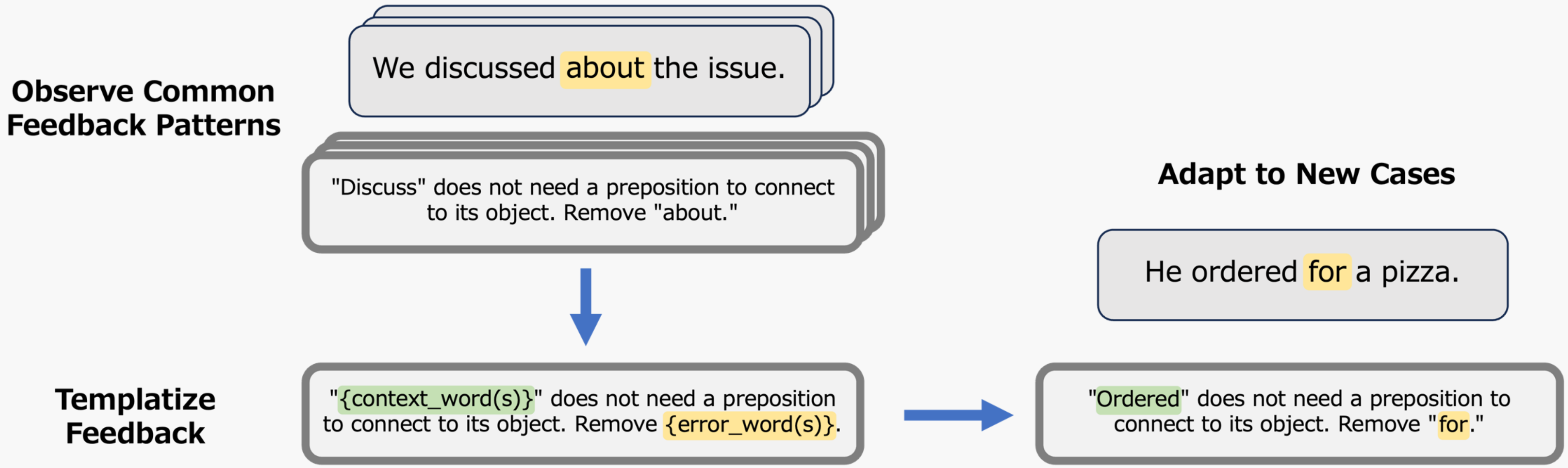
- **Three Keyword-Guided Systems**
  - Prompt includes an error tag.
  - Tags used: Ours, ERRANT, or EXPECT.
- **Keyword-Free System**
  - Prompt has no error tag; a baseline.
- **Template-Guided System**
  - Uses our error tags to select and fill a pre-written template.
- All systems use a few-shot approach with 2-4 examples.

Learning English gives the ability in live abroad.
Learning English gives the ability to live abroad.

Replace Particle	Infinitive	Preposition
ERRANT Tag	**Our Tag**	**EXPECT Tag**

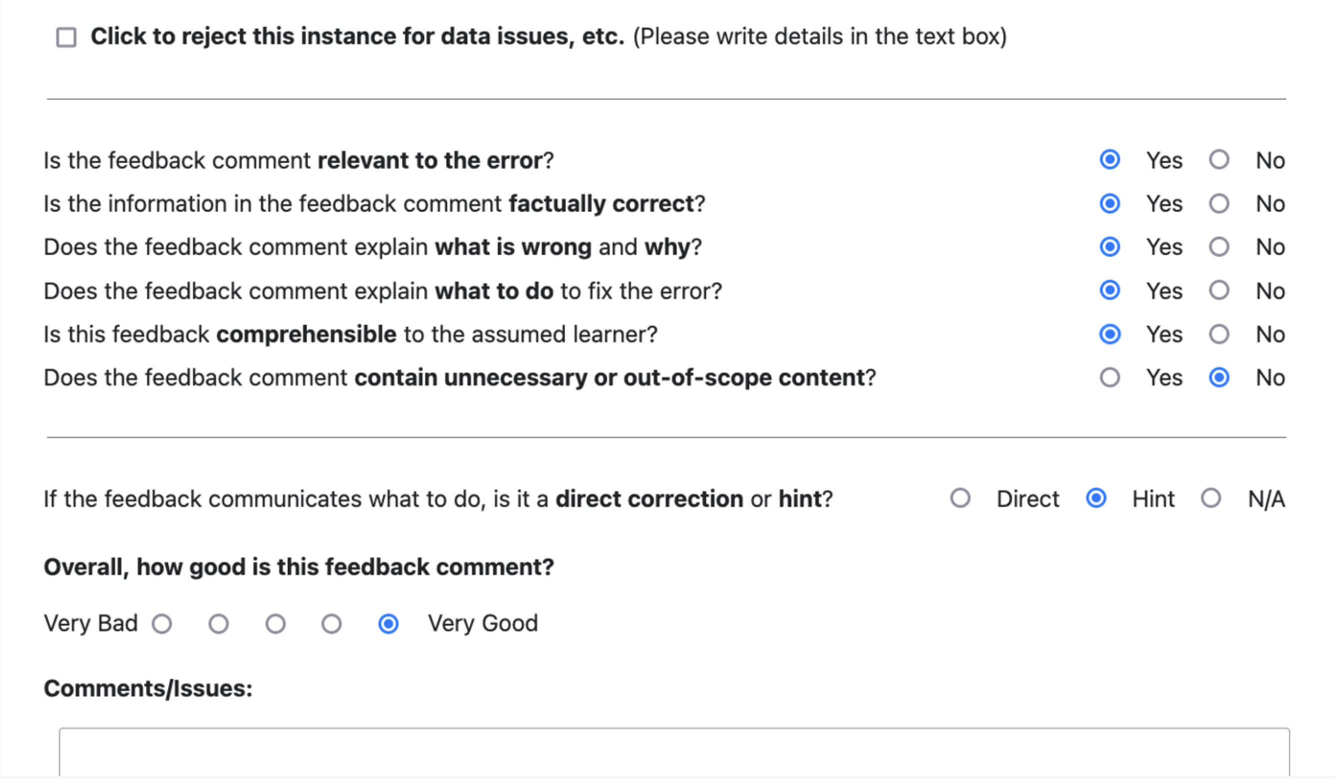
The Template-Guided System

- Step 1: Manually group feedback comments from our training data by error tag.
- Step 2: Identify common patterns ("archetypes") and write a fillable template for each.
- Step 3: At inference time, the LLM selects the best template for a given error and fills in the blanks
  - **If no template is appropriate, it should select "None"**



Human Evaluation

- Raters: Four experienced English teachers (≥7 years experience). Two per instance (2312 ratings).
- Rated feedback from all systems (plus the original human-written feedback) in a blind setting.
- 1-5 Likert scale for quality, plus factuality, relevance, comprehensibility, and directness judgements.



Contents

Results: Feedback Quality

- All systems performed well, with mean scores between 4.18 and 4.50 (out of 5).
- Keyword-guided and keyword-free systems were rated comparably to human-written feedback.
- No toxic or inappropriate outputs were generated.

	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

Results: Does the Typology Matter?

- No significant difference in quality ratings between the three keyword typologies
- Hypothesis: The base LLM may be powerful enough to infer the core issue from the text itself, making it less sensitive to the specific keyword provided.

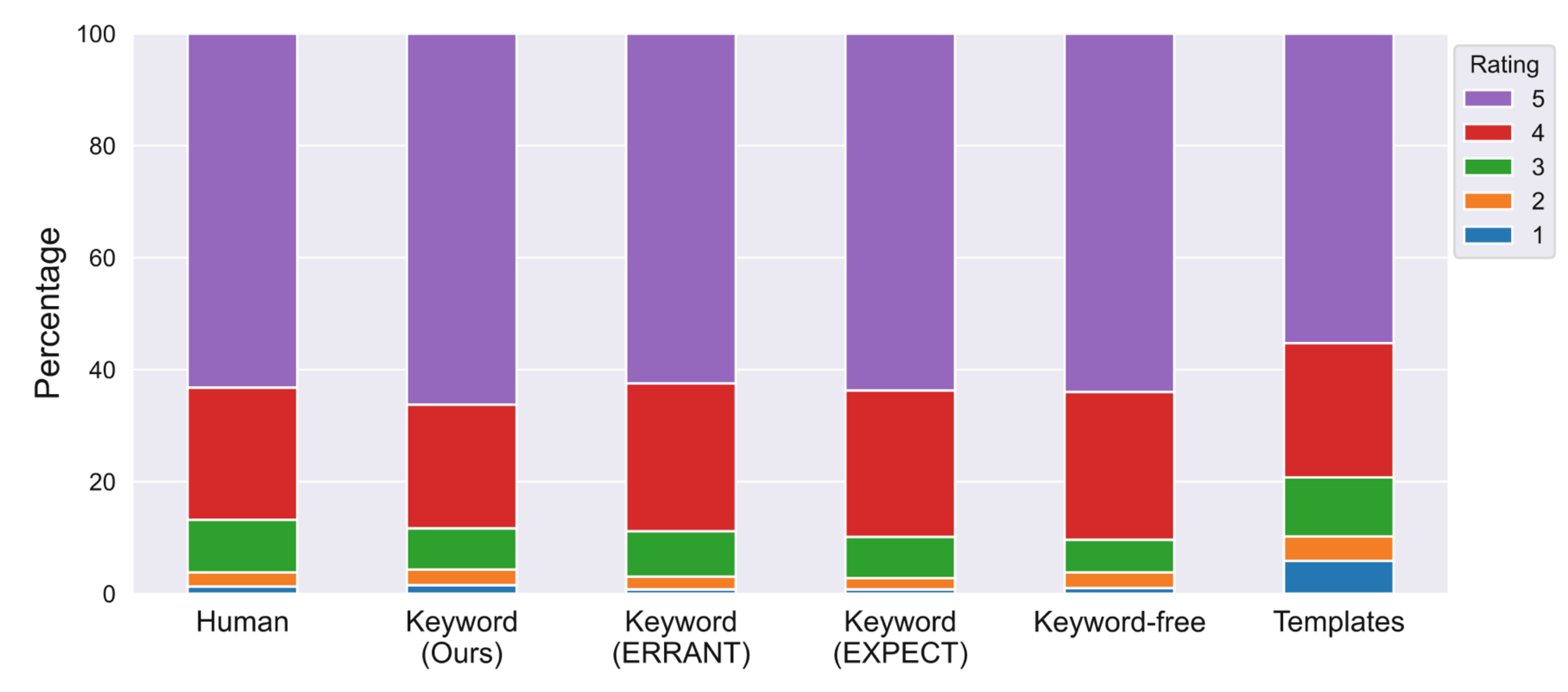
	Relevant	Factual	What & Why	What to Do	Comp.	Scope ↓	Overall
Human	**1.000**	0.972	0.987	**1.000**	0.952	0.008	4.449
Keyword: Ours	**1.000**	0.970	0.992	**1.000**	0.970	0.008	4.487
Keyword: ERRANT	0.997	0.967	0.992	**1.000**	**0.982**	**0.003**	4.475
Keyword: EXPECT	0.997	**0.975**	0.990	**1.000**	0.975	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

Results: Directness Alignment

- Humans: Provided hints in 40.9% of cases, mostly for generalizable errors.
- Keyword/Keyword-Free AI: Almost always gave direct corrections (0-3% hints).
- Result: The models did not replicate human hint-giving behavior, despite prompting, showing a strong bias towards direct corrections for all errors.

Results: Template System Performance

- The template system more closely matched human behavior, providing hints in 39.8% of cases.
- It also had the highest proportion of low-quality ratings (1s and 2s)
- This was mostly due to a failure to select "None" when no template was appropriate



Contents

Discussion

- The impact of error tags on quality ratings was seemingly minimal
  - A good typology is still useful for e.g., grouping errors for analysis or for resource recommendations.
- GPT-4o had a strong "directness bias" not easily overcome by simple prompting.
  - Direct feedback could be rated highly by the teachers even if written for a generalizable error.
- Templates offer more control over style and directness but can be brittle, especially around coverage gaps. They also require manual labor to create.
- LLMs are capable of generating pedagogically sound WCF, but there remains much work to do to fully align them with teacher behaviors.

Limitations

- Did not explore adapting the feedback to the learner's level. This is another major factor.
- The feedback style assumes a very academic learner in general – not appropriate for all learning contexts.
- The experiment used "oracle" error information, skipping challenges like isolating errors from raw text.
- Human evaluation experiments were performed with teachers, but not students.
- The creation of templates requires expert human labor, which is a scalability challenge.

Future Work

- Explore methods to control directness without relying on templates
- Explore methods to adapt to learner level
- Implement and evaluate a fully automated pipeline (error detection → classification → feedback).
- Analyze student interactions from a real-world deployment (e.g., feedback views, revision success).

Future Work

- Explore methods to control directness without relying on templates.
- Explore methods to adapt to learner level
- Implement and evaluate a fully automated pipeline (error detection → classification → feedback).
- Analyze student interactions from a real-world deployment (e.g., feedback views, revision success).

Deployment Underway at Tohoku University with ~2000 B1-B2 student users

Conclusions

- We introduced a framework for annotating learner errors with a focus on pedagogical feedback
- We introduced a new error typology focusing on the error-to-feedback context
- We created and released a dataset with annotations for error type, generalizability, and feedback directness
- We found that LLMs can generate feedback that teachers rate highly
- Templates were the most reliable way to control for directness, but they could be brittle

Thank You for Listening!

- We welcome any questions you have!
- Contact: coyne.steven.charles.q2@dc.tohoku.ac.jp
- Resources available at: https://github.com/coynestevencharles/annotating-errors-wcf

Paper Link **Github Link**