Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System

Mariia Soliar[1], Leona Colling[1,2], Stephen Bodnar[3], Detmar Meurers[1,2,3]

[1]Leibniz-Institut für Wissensmedien (IWM) Tübingen, Germany
{m.soliar, l.colling, d.meurers}@iwm-tuebingen.de
[2]Department of Linguistics, University of Tübingen, Germany
[3]Tübingen Center for Digital Education (TüCeDE), University of Tübingen, Germany
stephen.bodnar@uni-tuebingen.de

Abstract

Intelligent Tutoring Systems (ITS) can record learner interactions in fine-grained detail at scale. This opens the door to data-driven methods for investigating system performance and identifying points for improvement. In this paper, we draw on authentic log data from an English language ITS (N_{logs} = 5646, $N_{students}$ = 368) to investigate the performance of its feedback algorithm. In step 1 of our analysis, we profiled feedback accuracy by exploring how well the system provided error-specific feedback to malformed student answers in gap-filling grammar exercises using an expert-created set of feedback generation rules. We then identified frequently occurring student errors that triggered incorrect or unspecific feedback and refined the rule set used to detect and respond to these errors with correct specific feedback. In step 2, we validated the rule modifications on an unseen dataset. Comparing the performance of the initial and updated rule sets, we find significant improvement that generalizes to unseen data. Our study thus illustrates how an empirical evaluation of authentic data can complement feedback creators' expertise by informing rule refinement decisions that yield significant and generalizable performance improvements to feedback in ITS systems.

1 Introduction

Among other trending approaches to computer-assisted language learning (CALL, Huang et al., 2023; Zerkouk et al., 2025), Intelligent Tutoring Systems (ITS) have demonstrated a notable capacity to improve learning outcomes (Heift, 2010; Meurers et al., 2019), enhance learner engagement, motivation, and foster positive social outcomes (Ukenova and Bekmanova, 2023). Both educational practice and Skill Acquisition Theory (DeKeyser, 2007) motivate the need for deliberate practice with feedback, which can be facilitated by performing error detection and providing automatic corrective feedback (Amaral et al., 2011; Ruiz et al., 2023).

Feedback plays a crucial role in language acquisition, and therefore its correctness and relevance is of high importance (Loewen, 2013). When it comes to feedback in ITSs, Cavalcanti et al. (2021), in their review of automatic feedback generation systems in online learning environments, concluded that automatic feedback can be as effective, or sometimes even more effective, than manual instructor feedback.

Another beneficial property of ITSs is the ability to record large amounts of user-system interactions, enabling educational data mining to draw conclusions about both learner performance and system behavior (Romero and Ventura, 2013, 2020) and, therefore, to evaluate the performance of feedback generation (Koedinger et al., 2013).

In recent years, the focus of using AI in education has shifted significantly towards Large Language Models (LLMs) and the integration of generative AI (Dong et al., 2024; Degraeuwe and Goethals, 2024; Baidoo-Anu and Ansah, 2023; Dai et al., 2023). Research shows that LLMs can generate useful formative feedback and the accuracy of the latter is promising, but false positives are still a concern, especially in educational settings, as they might undermine learners' confidence and lead to misconceptions (Jaganov et al., 2025). Moreover, for closed exercise types and semi-open exercises (e.g., gap filling with grammatical forms) where the set of correct answers is known, a rule-based or hybrid approach to generating feedback can follow the "don't guess if you know" advice for achieving transparency and robustness.

Therefore, as Stamper et al. (2024) suggest, one should not disregard previous knowledge and years of experience using expert-generated and traditional data-driven approaches that allow high

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 190–207

levels of control over feedback provision, a system feature that is particularly crucial for beginner learners who tend to rely heavily on system feedback (Fu et al., 2022; Yang and Li, 2024).

In this paper, we analyze feedback generation performance of the FeedBook (Rudzewitz et al., 2017), a language ITS that primarily focuses on semi-open gap-filling grammar exercises (e.g., "[...] he now? He is at home now". Target answer: "Where is") and provides dynamic scaffolding feedback. We aim to showcase the value of authentic interaction logs for evaluating and refining a rule-based feedback system, such as the one in the FeedBook by addressing the following research questions (RQs):

1. How can systematic analysis of authentic logs be applied to evaluate and improve rule-based feedback generation?

2. How well does the approach generalize to unseen data?

The results of our experiment show that log analysis can be used effectively to identify rules to refine or add, and that even a small number of changes can yield significant performance improvements. We also employ statistical analysis to demonstrate the generalizability of the approach to unseen data.

2 Related Work

ITSs play an increasingly significant role in modern education, and therefore, the importance of evaluating these systems to ensure their technical reliability, usability, and overall effectiveness increases proportionally (Mousavinasab et al., 2021; Marouf et al., 2024; Latif et al., 2026).

Siemer and Angelides (1998) were among the first to specify an evaluation framework for ITSs and suggested a two-part evaluation: external and internal. External evaluation is concerned with the system's influence on the student, e.g., on learning achievement and learning affect. As for internal evaluation, the goal is to analyze the system from within, to evaluate its functionality and efficiency.

Most ITS evaluations (e.g., Bór, 2023; Wu et al., 2022; Erümit et al., 2019; Karaci et al., 2018) focus on external system effectiveness by analyzing usability, learner performance, or motivation rather than evaluating the intelligent algorithms employed. When it comes to feedback evaluation and enhancement in ITSs, Banihashem et al. (2022) confirm that learning achievement and pedagogical gains are the primary goals of such studies. For example, Van der Kleij et al. (2015) analyzed the effect of more elaborate feedback on learning outcomes compared to simpler feedback and the provision of correct answers. Lavolette et al. (2015) analyzed the accuracy of error detection and the respective feedback but the study focused on the effect of immediate feedback versus delayed feedback and not on fine-tuning or enhancement of the feedback itself. Inn-Chull (2019) and Fazilatfar et al. (2024) explored the effect of computer-assisted feedback on learner uptake and analyzed learner data to tailor feedback strategies (e.g., example-based or metalinguistic) based on their pedagogical effectiveness. Horbach et al. (2022) also examined the effect of automated informative tutorial feedback on learner writing performance.

Today, research often focuses on LLM-generated feedback, mainly featuring short answer and free writing assessment (Stahl et al., 2024; Jaganov et al., 2025; Emirtekin, 2025). The approaches to providing grammar feedback in these studies are partially inspired by Grammatical Error Correction (GEC) systems, and a significant amount of recent work has evaluated them (e.g., Volodina et al., 2023; Kobayashi et al., 2024). However, in their comprehensive survey Bryant et al. (2023) emphasize that most GEC systems focus on correcting errors without explaining them (e.g., Omelianchuk et al., 2024), which limits their educational value. Although some resources for generating explanatory feedback are emerging (e.g., Fei et al., 2023; Ye et al., 2025), more work is needed to develop robust, explainable GEC and to evaluate the generated feedback (Nagata, 2019; Nagata et al., 2020, 2021).

When it comes to data, Leacock et al. (2014) highlighted the importance of authentic learner texts for training GEC and feedback tools, enhancing their effectiveness in CALL and automatic writing evaluation contexts. This benefit has recently been emphasized, for example, by Song et al. (2024) in their work on Grammar Error Explanation. They propose a two-step pipeline that uses fine-tuned and prompted LLMs trained on real-world learner corpora in order to first edit erroneous input and subsequently prompt GPT-4 for explanations of the edits.

Greer and Mark (2016) also highlighted the benefits of analyzing the fine-grained data col-

lected by an ITS for its evaluation. For instance, Heift (2013) used learner interaction logs to improve the *Error Priority Queue*, i.e., to determine the relevance of feedback messages to the learning goal whenever several errors were present in student answer.

Interestingly, our research on semi-open gap-filling grammar exercises revealed that there are very few works on feedback generation for such exercises or evaluation thereof apart from the work by Rudzewitz et al. (2018). Although free production is the ultimate goal of language learning, constrained constructed responses - such as gap-fills - are effective in building explicit language knowledge and accuracy (Ellis, 2006). This learning effect, as mentioned above, can be enhanced with scaffolding feedback. However, most research involving gap-filling is concerned with the generation of the questions and the gaps themselves (e.g., Kiros Bitew et al., 2023; Hill and Simha, 2016; Knoop and Wilske, 2013), or distractors for those gaps (e.g., Yoshimi et al., 2023; Panda et al., 2022; Yeung et al., 2019), rather than feedback. These studies are also rarely associated with an ITS.

To address this gap, we adopt the data-mining approach and use authentic logs of learner interactions to perform an internal evaluation of the feedback mechanism in the FeedBook and demonstrate the benefits of the approach for gap-filling exercises.

3 Feedback Mechanism in FeedBook

The FeedBook is a web-based ITS designed for 7th grade English as a second language (ESL, L2) learners in German secondary schools. It is publicly available for use in ESL classrooms and is under ongoing development within research projects.[1] Its main function is to provide learners with language exercises and to give immediate, individualized scaffolding feedback (Rudzewitz et al., 2017). Feedback becomes available in two stages: offline *feedback generation* and real-time *feedback provision* (see Rudzewitz et al. (2018) for a detailed description of the algorithm).

FeedBook's **feedback generation** mechanism utilizes various Natural Language Processing (NLP) tools, such as Apache Unstructured Information Management application (UIMA) (Ferrucci et al., 2009), ClearNLP (Choi and Palmer, 2012), Morpha (Minnen et al., 2001), and Sfst (Schmid, 2005) to analyze and annotate exercises, including the encoded target answer (TA; i.e., the expected correct answer), in order to predict possible learner errors and provide appropriate and helpful feedback options. TAs can be full but single sentences, phrases or single words. If a TA comprises a partial sentence, the remaining parts (i.e., to the left and right of the gap) are provided as *context* to the NLP pipeline. To account for the phenomenon of multiple admissibility (Katinskaia et al., 2019), some exercises have one or several alternative TAs.

The system utilizes a multi-layer rule-based approach (see Appendix A) that takes TAs and transforms them into ill-formed or well-formed but otherwise incorrect anticipated learner answers (*hypotheses*). The transformations are made through a variety of second language acquisition (SLA) theory-based, pedagogically driven rules, e.g.,

> overregulation of irregular verbs in simple past: *did* > *doed*, *went* > *goed*;
> incorrect tense or person agreement of auxiliary verbs: *do* > *did*, *was* > *were*.

Each hypothesis carries the transformed surface string information, the respective error diagnosis, and the context and matches it with the most specific feedback message for later display to learners.

In terms of **feedback provision**, once a learner enters an erroneous answer, the system compares it with the generated hypotheses, identifies the closest match and displays its associated feedback message as a popup. The messages are teacher-crafted metalinguistic error explanations tailored to our target population, e.g., *"You've forgotten an auxiliary verb ('Hilfsverb', e.g. 'do'/'did')."*. The system provides one feedback message at a time, adjusting it when a learner attempts to correct their answer. The messages cover a wide range of grammar errors, spelling and capitalization errors. If the system cannot diagnose the error, it falls back to a *default feedback* message: *"This is not what I'm expecting. Please, try again."*.

4 Evaluation Approach

In this paper, we propose an approach to systematically use authentic data to enhance a rule-based feedback system for semi-open gap-filling grammar exercises and evaluate the generalizability of

[1]The system and information about the related projects are available at `https://feedbook.website/`.

improvements. The approach uses two datasets: a training dataset to identify refinement points in a data-driven way, and an unseen testing dataset to assess generalizability.

As shown in Figure 1, we suggest first establishing a baseline by processing authentic data through a pipeline that mimics the feedback system and then annotating the output to evaluate its correctness and granularity. The analysis focuses on three aspects: 1) identifying and correcting faulty rules to improve precision [2], 2) extending existing rules to increase recall, and 3) introducing previously missing rules to further improve recall.

After implementing the identified changes and adjusting the system accordingly, the data is processed again to establish improvements in precision and recall. Once a significant improvement has been confirmed, we suggest testing the updated system on unseen data to avoid overfitting.

We demonstrate the approach using the FeedBook, its feedback mechanism, and data from its use in real classrooms.

4.1 Data

The FeedBook has been used in German schools for four large-scale field trials, each focusing on a different aspect of the system. For our analysis, we drew on user data from two consecutive trials (Parrisius et al., 2022; Berens et al., 2024). [3]

For a concise analysis, we focused on a single learning target, 'Questions', selected due to the complexity of interrogative syntactic structures, their known difficulty for L2 learners, and their presence at all stages of learning (Mackey, 1999). We included only data units from semi-open gap-filling exercises, where typed student answers (SAs) provided valuable insight into actual learner needs by revealing weaknesses in students' productive grammatical skills (see Appendix E for examples of the exercises). Each log data unit contains SA, TA, and *context*:

SA (userInput): *is*
TA (targetAnswer): *Was*
context: *he late? Yes, he was late.*

We removed units with correct SAs, fully duplicate units, and units where capitalization was the only issue with the SA as such cases were considered correct during the second data collection.

The two data collections served as the *training* and the unseen *testing* dataset, respectively. For a more precise analysis, we divided each dataset into three subsets, based on the TA length: Word (e.g., *Where, do, is*), Phrase (e.g., *did you do*), and Sentence (e.g., *What does Jenny want to do?*).

In the second data collection, the number of users increased, leading to a much larger number of data units. To simulate one iteration of 10-fold validation per subset, we extracted a stratified random sample for the testing data (see Appendix B). Table 1 summarizes the final dataset sizes. [4]

	Training			**Testing**		
	Units	TAs	Ex.	Units	TAs	Ex.
Wrd	720	48	15	80	26	8
Phr	2083	60	14	236	43	14
Sent	2273	48	8	254	43	8
Total	5076	156	37	570	112	30

Table 1: The sizes of training and testing datasets. Wrd: word, Phr: phrase, Sent: sentence, TAs: target answers, Ex.: exercises.

The training dataset contains 5076 data units from 267 students and 37 exercises. In the Sentence subset, seven TAs have alternative correct answers (e.g., *Where are you from?* for *Where do you come from?*). They are not counted as separate TAs but accounted for in the feedback algorithm and our evaluation environment (see Section 4.2). The testing dataset contains 570 data units from 101 students and 30 exercises, with nine and 17 alternative TAs in the Phrase and the Sentence subsets, respectively.

The initial hypothesis generation rule set for questions was inspired by established SLA theories on developmental sequences that postulate that learners acquire grammatical structures in a constrained and ordered manner (Petersen, 2010; Pienemann et al., 1988). These theories provided the conceptual foundation for identifying question-related linguistic phenomena that are both developmentally important and frequently problematic for L2 learners. Building on this foundation, together with an expert teacher, we structurally selected, extended, and prioritized the

[2] We define precision as provision of incorrect vs. correct specific feedback and recall as specific vs. default feedback.

[3] Ethical and ministerial approval to collect data for research purposes was granted.

[4] Data is available for download at https://osf.io/3hrud/overview?view_only=ede0c00a7e36489884bd8cf788bd16eb.

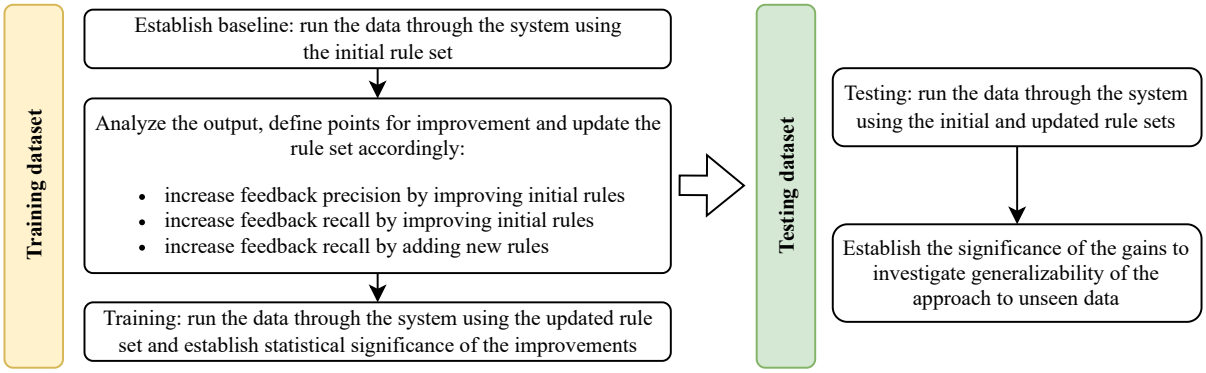


Figure 1: Proposed approach for employing authentic logs to evaluate and improve feedback in an ITS system.

rules. As a result, the initial rule set for questions contained eight rules implementing transformations such as missing or incorrect auxiliary verbs or question words, missing question marks, incorrect tense or person agreement of the auxiliary verb "do", incorrect form of the main verb when an auxiliary is present (see Appendix C). Some rules rely on syntactic dependency, others on the token position in a sentence.

4.2 Methods

As a first step, we created an environment that simulates FeedBook's feedback generation and provision including the SA pre-processing steps (e.g., string normalization). Unlike the actual system, our simulation takes one TA-SA pair at a time as input, generates hypotheses for the given TA, and then matches the SA with a feedback message. If the TA has alternatives, both the system and our simulation generate hypotheses for all given TA variations.

We ran four simulations to obtain the following performance data for analysis (S - seen, training data, U - unseen, testing data):

- **baseline**$_S$ (training data with initial rules)
- **update**$_S$ (training data with updated rules)
- **baseline**$_U$ (testing data with initial rules)
- **update**$_U$ (testing data with updated rules)

The output included SA, TA, and a respective fine-grained question-specific (QS) feedback label (e.g., "missing question word"). For this initial analysis, we combined all QS feedback into one score. Based on the baseline$_S$ results, we identified rules to refine or add and implemented the changes. Next, we performed a comparative analysis of baseline$_S$ and update$_S$ expecting an increased proportion of QS feedback and reduced default feedback. Then we compared the baseline$_U$ with the update$_U$, expecting the same trend as evidence for generalizability to unseen data.

Two expert annotators manually labeled the fine-grained feedback[5] to evaluate its correctness and report on precision and recall based on the following annotation schema:

- "optimal" – correct error diagnosis (true positives, see Figure 2 a)); or default feedback provided for nonsensical input (true negatives, see Figure 2 b));
- "incorrect" – incorrect specific diagnosis of an error (false positives, see Figure 2 c));
- "unspecific" – default feedback where a specific one could (with a new rule) or should (with an existing rule) have been given (false negatives, see Figure 2 d)).

It is important to point out that "unspecific", i.e. default feedback should not be considered incorrect but only a less optimal variant of feedback that the system can provide to the students. The diagnostic process for the feedback provision is optimal if the system correctly identifies the specific error contained in the SA. If the system only detects the presence of an error, the process is still considered successful, as immediate correctness feedback is one of the mechanisms by which ITSs achieve learning gains comparable to human tutors (VanLehn, 2011).

To assess the significance of the impact that rule improvements had on system's performance,

[5]In case of conflict, the feedback was labeled by a third independent annotator and the majority vote was used for the final scores.

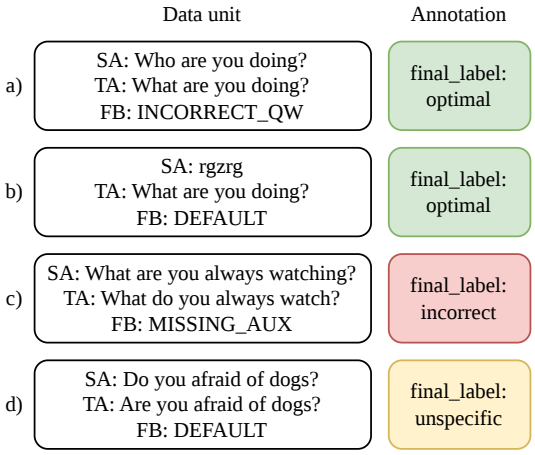


Figure 2: Examples of data units and their respective annotations. SA: student answer, TA: target answer, FB: feedback, QW: question word, AUX: auxiliary verb.

we use McNemar's test with continuity correction (Agresti, 2012).

5 Baseline Evaluation and Rule Set Enhancement

Table 2 shows the raw numbers of the feedback from the baseline$_S$ evaluation as well as the annotated correctness for all three subsets. The inter-annotator agreement, measured using Cohen's Kappa, is $\kappa = 0.97$, indicates almost perfect agreement (Cohen, 1960; see Appendix D). The corresponding precision and recall scores are summarized in Table 3.

The baseline$_S$ results revealed that the overall success rate of the feedback provision is very high - 94.37%. However, whilst only a small proportion of errors were diagnosed incorrectly (5.63%), a significant proportion did not receive optimal feedback, i.e. a specific diagnosis (59.83%), indicating a substantial need to improve recall. To address this, we closely analyzed both SAs and TAs in each subset. Based on the distribution of linguistic constructs in TAs (see Table 4), we investigated SAs containing errors related to the targeted constructs and observed the following: in all subsets, errors such as incorrect question word, missing or incorrect auxiliary, and incorrect agreement of auxiliary "do" are the most common and are rarely diagnosed correctly (only 10.92% to 27.12% per error type). We summarize the numbers in Table 5.

Further investigation showed that errors such as incorrect agreement of the main verb "be" or its replacement by auxiliary verbs, or missing main verb were not diagnosed due to the lack of respective rules. Moreover, the Phrase and Sentence subsets contained 54 SAs where only word order was incorrect (e.g., *you do have* instead of *do you have*). It was particularly insightful to observe that such malformations as switching the tense forms of auxiliary and main verbs in simple past questions (e.g., *Do you saw?* instead of *Did you see?*) or the conjugation of "be" with the help of "do" (e.g., *does be* instead of *is*) were more common than our experts anticipated. In addition, students often omitted or added tokens to their answers.

The annotations of baseline$_S$ also showed that most cases labeled as "incorrect" received spelling feedback where it should have been default feedback for nonsense SAs (66.81%) or QS feedback (6.64%), i.e., incorrect question word or incorrect auxiliary verb. Although we addressed the latter issue by enhancing the rule set, improving the spell-checking algorithm (not a part of the hypothesis generation) was not in focus for this paper.

Based on these observations, we refined four initial rules (i.e., missing and incorrect auxiliary verb, incorrect agreement thereof, incorrect question word), and added four new rules introducing transformations such as missing main verb, incorrect tense and person agreement of the verb "be", verb "be" being replaced by verbs "do" or "have", switching the tenses of auxiliary "do" and main verb in past simple tense. An additional check to account for word-order permutations was implemented.

It is worth noting that in the Phrase and Sentence subsets, 3 and 18 SAs, respectively, were grammatically and semantically correct but did not match any of the provided TAs. The system diagnosed these cases as spelling or default, and thus, they contributed to the counts for "incorrect" and "unspecific" categories. This observation suggests the refinements should include not only the feedback generation rules but also the alternative correct answers or fuzzy-matching. Addressing this was outside the scope of this paper, but is a promising direction for future work and a perfect phenomenon to explore the benefits of LLM-based or hybrid approaches.

Feedback	QS		Spelling		DF		Total		
	✓	✗	✓	✗	✓	◆	✓	✗	✓+◆
Initial rules									
Word	66	5	132	85	6	426	204	90	630
Phrase	202	42	571	61	33	1174	806	103	1980
Sentence	126	17	587	76	30	1437	743	93	2180
Total	394	64	1290	222	69	3037 (59.83%)	1753 (**34.54%**)	286 (5.63%)	4790 (94.37%)
Updated rules									
Word	179	0	132	85	6	318	317	85	635
Phrase	448	1	571	61	33	969	1052	62	2021
Sentence	266	23	587	75	30	1292	883	98	2175
Total	893	24	1290	221	69	2579 (50.81%)	2252 (**44.37%**)	245 (4.83%)	4831 (95.17%)

Table 2: Results for initial and updated rules on **training data**: Raw numbers of provided feedback annotated as optimal (✓), incorrect (✗) and unspecific (◆) per subset (word, phrase, and sentence). QS: question-specific, DF: default feedback.

	Precision	Recall
baseline$_S$	0.86	0.37
update$_S$	0.90	0.47
baseline$_U$	0.81	0.35
update$_U$	0.92	0.55

Table 3: Precision and recall scores.

Ling.constr. in TA	Word	Phr	Sent
QW only	**189**	-	-
QW + be	25*	**161**	**190**
QW + do/does/did	-	**493**	**1368**
do/does/did	**139**	**1323**	-
be	**296**	-	-
Y/N + do/does/did	-	-	**607**
Y/N + be	-	-	108
modals	-	106	-
main verb	71	-	-
Total:	720	2083	2273

Table 4: Numbers of data units, i.e. erroneous SAs, clustered by linguistic constructs focused in the TAs. Top-3 constructs per input type are highlighted in bold. QW: question word, QM: question mark. *25 data units in Word subset were two-word phrases.

Updated Rule Implementation

We employed the syntactic dependency approach in all new rules and in four initial rules that received an update (see Appendix A). In addition, we extended the *Incorrect Auxiliary Rule* by configuring it to generate hypotheses with incorrect

Error	Word	Phr	Sent
Incorr. QW	43 / **115**	39 / **161**	14/**78**
Missing QW	1 / 3	13 / 60	5/23
Incorr. agr. "do"	16 / **42**	67 / **333**	31/**175**
Incorr. auxiliary	2 / **46**	25 / **161**	4/77
Missing auxiliary	0 / 1	43 / **267**	8/**159**

Table 5: Number of selected student errors in baseline$_S$; diagnosed correctly/total errors made. Top-3 errors per input type are highlighted in bold. QW: question word, agr.: agreement.

grammatical number (the initial rule only transformed the lemma and the tense) and for the *Question Word Replacement Rule* we extended the list of question words based on log data observations.

5.1 Results Update$_S$

Table 2 shows the raw numbers of the feedback provided in the update$_S$ as well as the annotated correctness for all three subsets. The inter-annotator agreement is $\kappa = 0.98$ (almost perfect agreement). The system with updated rules provided less default feedback (-9.02 percentage points (p.p.)) and more QS feedback (+9.83 p.p.) compared to the baseline$_S$, suggesting an overall positive effect of the refinements.

Interestingly, the number of "incorrect" cases decreased for the Word and Phrase subsets but increased by six for the Sentence subset. 12 "incorrect" cases in Sentence and one in Phrase received a new "missing main verb" diagnosis in

contrast to the default (or spelling in one case) in the baseline$_S$. Ten "incorrect" cases in Sentence were diagnosed as "incorrect word order". These results suggest that the implemented changes are largely beneficial, but might still introduce errors.

Initial rule set	Updated rule set		Total
	✓	✗ + ◆	
✓	1752	23	1775
✗ + ◆	500	2801	3301
Total	2252	2824	5076

Table 6: Contingency table for training data.

Overall, we observed an increase in precision and an even greater increase in recall (see Table 3). The results of McNemar's test (see Table 6) showed that the improvement of the feedback performance with the updated rule set is statistically significant: McNemar's χ^2 = 433.22, df = 1, p-value $< .0001$, Cohen's $g = 0.46$ (large effect size).

6 Results of Generalizability Evaluation

Turning now to the testing dataset (U), table 7 shows the raw numbers of the feedback provided in the baseline$_U$ and in the update$_U$ as well as the annotated correctness for all three subsets. The inter-annotator agreement for the two runs are κ = 0.96 and κ = 0.97, respectively (almost perfect agreement). The updated rules consistently produced less default feedback (-16.85 p.p.) and more QS feedback (+14.46 p.p.) compared to baseline$_U$. Although some QS cases became incorrect with the update, the overall "incorrect" feedback decreased (-3.15 p.p.). The corresponding precision and recall scores are summarized in Table 3. These results suggest a positive trend and confirm the observations on the training data.

The results of McNemar's test (see Table 8) show statistical significance of the improvements on unseen data, indicating good generalizability of the approach: McNemar's χ^2 = 101.68, df = 1, p-value $< .0001$, Cohen's $g = 0.47$ (large effect size).

7 Discussion

The results show that our data-driven approach substantially improves feedback performance, in particular the specificity of provided feedback. Comparing the baseline and updated datasets, we observe more instances of specific feedback, fewer instances of default feedback, and higher recall and precision scores. These findings show that even limited rule enhancements motivated by user interaction logs yield notable gains. Thus, regarding **RQ 1**, we find that the authentic logs provided valuable insights into actual learner needs, which, in combination with a thorough analysis of the expected target answers, enabled targeted refinement of the rule set. The evaluation results on the testing data show that the improvements to the rules are stable and, even with unseen data, increase precision and recall of the feedback performance. This suggests a positive answer to **RQ 2**: our data-driven enhancement to rule-based feedback is likely to generalize well and perform similarly on unseen data.

Although the increase in precision and recall was higher on the unseen data, an unexpected outcome, further examination of the unseen dataset identified two causes. One is a higher percentage of incorrect question-specific feedback in baseline$_U$ (3.68%) which allowed for more room for improvement relative to baseline$_S$ (1.26%). A second reason is the simpler linguistic structures in the testing data due to changes to FeedBook's exercise generation and exercise selection between the first and second data collection, which potentially allowed for better NLP parsing and therefore more accurate diagnoses by the feedback system. Based on these observations, we expect the updated system to perform similarly on unseen data if the aforementioned factors are controlled for or the testing data is sampled from the same dataset as the training data.

Furthermore, the results of our student answer investigations revealed a great variety of errors and misconceptions that are challenging to handle by a system or anticipate in general. We also observed a wider range of errors in full sentence answers, concluding that while our rule-based approach is precise and easily adjustable for shorter gap-fills, an approach using LLMs could be beneficial for longer input.

8 Conclusion and Outlook

In this paper, we evaluated and improved a rule-based feedback mechanism in an ITS using authentic learner interaction logs. Our analysis demonstrated that such real-world data provides valuable insights into actual learner difficulties and can meaningfully guide refinements of rule-based feedback. The targeted rule up-

	Initial rules								
Feedback	**QS**		**Spelling**		**DF**		**Total**		
	✓	✗	✓	✗	✓	◆	✓	✗	✓+◆
Word	18	20	3	3	3	33	24	23	57
Phrase	32	1	33	19	7	144	72	20	216
Sentence	38	0	50	0	0	166	88	0	254
Total	88	21	86	22	10	343	184	43	527
						(60.18%)	(**32.28%**)	(7.54%)	(92.46%)
	Updated rules								
Word	40	0	3	3	3	31	46	3	77
Phrase	70	1	33	19	7	106	110	20	216
Sentence	92	2	50	0	0	110	142	2	252
Total	202	3	86	22	10	247	298	25	545
						(43.33%)	(**52.28%**)	(4.39%)	(95.61%)

Table 7: Results for initial and updated rules on **testing data**: Raw numbers of provided feedback annotated as optimal (✓), incorrect (✗) and unspecific (◆) per subset (word, phrase, and sentence).

	Updated rule set		Total
Initial rule set	✓	✗ + ◆	
✓	183	4	187
✗ + ◆	115	268	383
Total	298	272	570

Table 8: Contingency table for the testing data.

dates led to a reduction in default and incorrect feedback, while increasing both precision and recall. These improvements generalized well to unseen data, confirming that authentic logs help strengthen rule-based systems while retaining pedagogical control over generated feedback.

Although we focused on improving an existing rule set, the proposed methodological flow has wider potential. Beyond rule-based feedback, future work could apply the same data-driven evaluation approach to LLM-based feedback generation or explore hybrid models that leverage both expert knowledge and data-driven flexibility. Training LLMs on learner data to produce new rules, or applying the approach to other exercise types, linguistic structures, or learner populations, represents promising avenues for future work.

Our work also lays the foundation for a direct comparison between the refined rule-based approach and an LLM-based one. While we argue that rule-based feedback may be particularly advantageous for gap-filling grammar exercises, especially due to its transparency and reliability, this assumption has yet to be empirically tested. Comparing both approaches within the same evaluation framework will allow a better understanding of their respective strengths and the contexts in which each approach is most suitable.

Finally, although our analysis focused on internal evaluation, authentic data also offers valuable insights for external evaluation by indicating how many learners would benefit from specific refinements. Overall, our findings demonstrate that leveraging real learner data enables systematic, generalizable improvements to feedback mechanisms and supports the development of future feedback technologies in CALL.

9 Limitations

Data. Due to differences in data collection, the data from two periods varied significantly, which might influence the results but was not controlled for in this experiment. The most relevant differences are as follows:

- In terms of feedback, in the first data collection metalinguistic information in the feedback popups was replaced with the default message for some students while others received full access to the metalinguistic information during a field trial. There were no such treatment variations in this regard during the second data collection.
- All the exercises for the first data collection were manually crafted and implemented in the system. Later, however, only the exercise specifications such as vocabulary, grammar

context and others, were designed manually, while the exercises themselves were generated automatically based on these specifications (Heck et al., 2022). As we discuss in Section 7, this may result in more exercises with similar syntactic structures in the testing data and therefore, cause better feedback performance.

- In the first data collection, students had access to all the exercises at once and could select exercises from any difficulty level. In the second data collection, however, an algorithm assigned the most appropriate difficulty level to each student according to their performance (Colling et al., 2023). Such macro-adaptive selection could potentially lead to more students completing easier exercises and, therefore, to more data with simpler linguistic structures in the testing dataset.
- In real-life application, some of the student answer processing (e.g., removal of extra white space before punctuation, replacing characters of one encoding by the ones in another) currently happens in the front-end but did not take place when the training data was collected. Thus, some student answers that the system treated as correct in the second data collection were still included in the first data collection as malformed and contributed to the numbers with incorrect spelling or default feedback in our baseline.
- In contrast to the first data collection, the system from the second data collection does not consider capitalization a potential source of error in order to facilitate tablet and smartphone usage and to avoid over-punishing students on automatic device behavior, i.e. capitalizing initial characters in an input field.

Feedback mechanism. Most of the feedback rules rely on dependency parsing and annotations provided by NLP tools, such as UIMA. However, these tools are sometimes incorrect in their annotations, which can lead to unexpected results, e.g., default feedback for *do – does – did* confusion because of incorrect lemma value in the annotation provided by the UIMA Java SDK 2.9.0 version. We acknowledge that the latest version might yield better results and we encourage the use of a newer NLP pipeline in general.

Rules selection. In this paper, we selected rules to improve based on their performance in the baseline$_S$, prioritizing rules that performed poorly with certain linguistic constructs related exclusively to interrogative sentences (e.g., question words or auxiliary verbs in the simple past and simple present tense). However, we acknowledge that another approach to rule selection, for instance, to triage learner errors according to how frequently they go undiagnosed, might have shown even greater gains, especially in recall.

Moreover, many student answers were not ill-formed, but well-formed questions that are incorrect due to the context, exercise instruction, and nature of the exercise itself. We acknowledge that the data units with such student answers had a certain impact on precision and, particularly, recall scores (most of these cases received default feedback). However, the current analysis did not focus on feedback related to, for example, tense confusion or erroneous use of pronouns.

Increasing precision. Most of the incorrect feedback provided in the training baseline was spelling feedback. It was erroneously given to student answers with non-question grammatical issues, answers that contained context (text around the target answer but not part of it), nonsense answers, and only in a few cases to answers with question-specific errors. Increasing the feedback precision by accounting for all the observed issues was out of the scope of this paper, but is a promising venue for further enhancement of the feedback algorithm.

Acknowledgments

This work is based on data from the Interact4School and AI2Teach projects that were financially supported by the German Ministry of Education and Science (BMBF DLR, funding code: 01JD1905A) and by the AIM (Akademie für Innovative Bildung und Management Heilbronn-Franken gemeinnützige GmbH), respectively.

We thank Walid El Hefny for his help with data extraction as well as Daniela Verratti Suoto and Sonja Gelzenleuchter for their thorough annotation efforts. We also thank Matthew Pattemore for his support in revising and polishing the paper and Kate Belcher for insightful discussions and contributions to shaping the final version. Last but not least, we thank the reviewers for their thoughtful comments and constructive feedback.

References

Alan Agresti. 2012. *Categorical data analysis*, volume 792. John Wiley & Sons.

Luiz Amaral, Detmar Meurers, and Ramon Ziai. 2011. Analyzing learner language: towards a flexible natural language processing architecture for intelligent language tutors. *Computer Assisted Language Learning*, 24(1):1–16.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Seyyed Kazem Banihashem, Omid Noroozi, Stan van Ginkel, Leah P Macfadyen, and Harm JA Biemans. 2022. A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, page 100489.

Florian Berens, Katharina Wendebourg, Mareike Kholin, Leona Colling, Julia Schmidt-Peterson, Manuel Hopp, Tanja Heck, Stephen Bodnar, Walid El Hefny, Florian Nuxoll, Christoph Deeg, Katja Krey, Josef Schrader, Hannes Schröter, Benjamin Nagengast, Detmar Meurers, and Ulrich Trautwein. 2024. AI2Teach: Effectiveness of a teacher training on technology-supported teaching and learning. Registry ID: 20365.1v1. Pre-registration of the study design.

Dorina Bór. 2023. Improving the user experience of an intelligent tutoring system for proving mathematical statements: Master thesis.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.

Jinho D Choi and Martha Palmer. 2012. Fast and robust part-of-speech tagging using dynamic model selection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 363–367.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Leona Colling, Tanja Heck, and Detmar Meurers. 2023. Reconciling Adaptivity and Task Orientation in the Student Dashboard of an Intelligent Language Tutoring System. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 288–299.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Jasper Degraeuwe and Patrick Goethals. 2024. Leading by example: The use of generative artificial intelligence to create pedagogically suitable example sentences. In *Swedish Language Technology Conference and NLP4CALL*, pages 33–48.

Robert DeKeyser. 2007. Skill acquisition theory. In Bill VanPatten and Jessica Williams, editors, *Theories in Second Language Acquisition: An introduction*, volume 97113. Routledge.

Bingyu Dong, Jie Bai, Tao Xu, and Yun Zhou. 2024. Large language models in education: A systematic review. In *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 131–134. IEEE.

Rod Ellis. 2006. Current issues in the teaching of grammar: An SLA perspective. *TESOL quarterly*, 40(1):83–107.

Emrah Emirtekin. 2025. Large Language Model-Powered Automated Assessment: A Systematic Review. *Applied Sciences*, 15(10):5683.

Ali Kürşat Erümit, İsmail Çetin, Mehmet Kokoç, Temel Kösa, Vasif Nabiyev, and Emine Selin Aygün. 2019. Designing a usability assessment process for adaptive intelligent tutoring systems: A case study. *Turkish Online Journal of Qualitative Inquiry*, 10(1):141–179.

Ali Mohammad Fazilatfar, Mohadeseh Sedghi, and Mohammad Javad Ali Beigi. 2024. An attempt to improve grammatical structure retention: Impact of computer-mediated feedback on iranian EFL learners. *Journal of Studies in Language Learning and Teaching*, 1(2):289–306.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. *arXiv preprint arXiv:2305.15676*.

David Ferrucci, Adam Lally, Karin Verspoor, and Eric Nyberg. 2009. Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard.

Qing-Ke Fu, Di Zou, Haoran Xie, and Gary Cheng. 2022. A review of AWE feedback: types, learning outcomes, and implications. *Computer Assisted Language Learning*, 37(1-2):179–221.

Jim Greer and Mary Mark. 2016. Evaluation methods for intelligent tutoring systems revisited. *International Journal of Artificial Intelligence in Education*, 26:387–392.

Tanja Heck, Detmar Meurers, and Florian Nuxoll. 2022. Automatic exercise generation to support macro-adaptivity in intelligent language tutoring systems. *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, page 162.

Trude Heift. 2010. Prompting in CALL: A longitudinal study of learner uptake. *Modern Language Journal*, 94(2):198–216.

Trude Heift. 2013. Multiple learner errors and meaningful feedback. *CALICO Journal*, 20(3):533–548.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, pages 23–30.

Andrea Horbach, Ronja Laarmann-Quante, Lucas Liebenow, Thorben Jansen, Stefan Keller, Jennifer Meyer, Torsten Zesch, and Johanna Fleckenstein. 2022. Bringing automatic scoring into the classroom–measuring the impact of automated analytic feedback on student writing performance. In *Swedish Language Technology Conference and NLP4CALL*, pages 72–83.

Xinyi Huang, Di Zou, Gary Cheng, Xieling Chen, and Haoran Xie. 2023. Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1):112–131.

Choi Inn-Chull. 2019. Exploring the potential of a computerized corrective feedback system based on a process-oriented qualitative error analysis. *J Eng Teach Movie Media*, 20(1):89–117.

Timur Jaganov, John Blake, Julián Villegas, and Nicholas Carr. 2025. Large Language Model-Driven Dynamic Assessment of Grammatical Accuracy in English Language Learner Writing. *arXiv preprint arXiv:2505.00931*.

Abdulkadir Karaci, Zeynep Piri, Halil İbrahim Akyüz, and Göksal Bilgiici. 2018. Student perceptions of an intelligent tutoring system: A technology acceptance model perspective. *Online Submission*, 182(22):31–36.

Anisia Katinskaia, Sardana Ivanova, and Roman Yangarber. 2019. Multiple admissibility in language learning: Judging grammaticality using unlabeled data. In *Workshop on Balto-Slavic Natural Language Processing*, pages 12–22.

Semere Kiros Bitew, Johannes Deleu, A Seza Doğruöz, Chris Develder, and Thomas Demeester. 2023. Learning from Partially Annotated Data: Example-aware Creation of Gap-filling Exercises for Language Learning. *arXiv e-prints*, pages arXiv–2306.

Fabienne M Van der Kleij, Remco CW Feskens, and Theo JHM Eggen. 2015. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, 85(4):475–511.

Susanne Knoop and Sabrina Wilske. 2013. WordGap - Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. In *Proceedings of Second Workshop NLP Computer-Assisted Language Learning*, pages 39–47.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77. Association for Computational Linguistics.

Kenneth R. Koedinger, Emma Brunskill, Ryan S.J.d. Baker, Elizabeth A. McLaughlin, and John Stamper. 2013. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41.

Ehsan Latif, Vincent Liu, and Xiaoming Zhai. 2026. A systematic review of intelligent and robot tutoring systems: evolution, pedagogical design, and AI-driven classification. *Smart Learning Environments*, 13(1):1.

Elizabeth Lavolette, Charlene Polio, and Jimin Kahng. 2015. The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, 19(2).

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated grammatical error detection for language learners*. Morgan & Claypool Publishers.

Shawn Loewen. 2013. The role of feedback. In *The Routledge handbook of second language acquisition*, chapter 3, pages 24–40. Routledge.

Alison Mackey. 1999. INPUT, INTERACTION, AND SECOND LANGUAGE DEVELOPMENT: An Empirical Study of Question Formation in ESL. *Studies in Second Language Acquisition*, 21(4):557–587.

Ahmad Marouf, Rami Al-Dahdooh, Mahmoud Jamal Abu Ghali, Ali Osama Mahdi, Basem S Abunasser, and Samy S Abu-Naser. 2024. Enhancing education with artificial intelligence: The role of intelligent tutoring systems. *International Journal of Engineering and Information Systems (IJEAIS)*, pages 10–16.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–233.

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared task on feedback comment generation for language learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33. Association for Computational Linguistics.

Subhadarshi Panda, Frank Palma Gomez, Michael Flor, and Alla Rozovskaya. 2022. Automatic generation of distractors for fill-in-the-blank exercises with round-trip neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 391–401.

Cora Parrisius, Ines Pieronczyk, Carolyn Blume, Katharina Wendebourg, Diana Pili-Moss, Mirjam Assmann, Sabine Beilharz, Stephen Bodnar, Leona Colling, Heiko Holz, et al. 2022. Using an intelligent tutoring system within a task-based learning approach in English as a foreign language classes to foster motivation and learning outcome (Interact4School): Pre-registration of the study design.

Kenneth A Petersen. 2010. *Implicit corrective feedback in computer-guided interaction: Does mode matter?* Georgetown University.

Manfred Pienemann, Malcolm Johnston, and Geoff Brindley. 1988. Constructing an acquisition-based procedure for second language assessment. *Studies in second language acquisition*, 10(2):217–243.

Cristobal Romero and Sebastian Ventura. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, 3(1):12–27.

Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, and Detmar Meurers. 2017. Developing a web-based workbook for English supporting the interaction of students and teachers. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 36–46.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating feedback for English foreign language exercises. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 127–136.

Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2023. Supporting individualized practice through intelligent CALL. In *Practice and automatization in second language research*, pages 119–143. Routledge.

Helmut Schmid. 2005. A programming language for finite state transducers. In *Finite-State Methods and Natural Language Processing (FSMNLP)*, volume 4002 of *Lecture Notes in Computer Science*, pages 308–309. Springer.

Julika Siemer and Marios C Angelides. 1998. A comprehensive method for the evaluation of complete intelligent tutoring systems. *Decision support systems*, 22(1):85–102.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! Grammar Error Explanation with Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing LLM-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.

Aru Ukenova and Gulmira Bekmanova. 2023. A review of intelligent interactive learning methods. *Frontiers in Computer Science*, 5.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16.

Chih Hung Wu, Hao-Chiang Koong Lin, Tao-Hua Wang, Tzu-Hsuan Huang, and Yueh-Min Huang. 2022. Affective mobile language tutoring system for supporting language learning. *Frontiers in Psychology*, 13:833327.

Lu Yang and Rui Li. 2024. ChatGPT for L2 learning: Current status and implications. *System*, 124:103351.

Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025. Corrections Meet Explanations: A Unified Framework for Explainable Grammatical Error Correction. *arXiv preprint arXiv:2502.15261*.

Chak Yan Yeung, John SY Lee, and Benjamin K Tsou. 2019. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the 17th annual workshop of the Australasian language technology association*, pages 159–164.

Nana Yoshimi, Tomoyuki Kajiwara, Satoru Uchida, Yuki Arase, and Takashi Ninomiya. 2023. Distractor generation for fill-in-the-blank exercises by question type. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 276–281.

M Zerkouk, M Mihoubi, and B Chikhaoui. 2025. A comprehensive review of AI-based intelligent tutoring systems: Applications and challenges. *arXiv preprint arXiv:2507.18882*.

A Rule Change Example

Figures 3 and 4 provide snippets of simplified pseudocode to illustrate a change from the token-position approach to the syntactic dependency one in a rule. The example shows *Question Replacement* rule that searches for a question word in a target answer (TA) and, if it identifies one, replaces the question word with a different one from a provided list of question words.

Figure 5 shows the different components and the application flow of a feedback rule in the FeedBook.

```
[...]
questionWords = "Whose", "Which", "Where",
"Whom", "What", "When",
"Why", "Who", "How"
[...]
isRuleApplicable(Configuration of TA):
    get a stringTA from Configuration
    if stringTA is null
        return false
    split stringTA by white space
    get the firstToken from stringTA
    if firstToken in questionWords
        if lastToken is "?"
            return true
        else return false
    else return false

applyRule(annotatedTA, indexTA):
    get the firstToken from annotatedTA
    replace firstToken with questionWord
    set correct value of firstToken
    set incorrect value of firstToken
    return extended annotatedTA copy
[...]
```

Figure 3: *Question Word Replacement* rule with token-position approach (simplified pseudo-code)

B Sampling Strategy for Unseen Data

As there was more data collected in the second period we did not take the entire dataset for evaluation but instead selected representative samples. We simulated one iteration of a 10-fold validation for each subset to create the testing (unseen) data.

Sampling for each subset contained two steps: 1) Random sample of exercises; 2) Stratified random sample of attempt data units from sampled exercises.

For step 1, to match the same number of exercises present in the training dataset, and thus have a similar distribution of different exercise contexts, we randomly selected exercises from the

```
[...]
isRuleApplicable(Configuration of TA):
    for every dependency in TA span:
       if dependencyType is "advmod" and
        questionWords contains dependent
            return true
    return false

applyRule(annotatedTA, indexTA):
    for every dependency in TA span:
       if dependencyType is "advmod" and
        questionWords contains dependent
            replace dependent
          set correct value of dependent
         set incorrect value of dependent
    return extended annotatedTA copy
[...]
```

Figure 4: *Question Word Replacement* rule with syntactic dependency approach (simplified pseudo-code)

second period data to be included in the testing dataset. For Word, the training dataset contained 15 exercises. However, due to the adaptive exercise selection feature introduced between data collections (see Section 9), there were fewer instances with one-word target answers. Thus, all eight available exercises were included for the Word subset.

In step 2, a stratified sampling approach on the remaining data was conducted with the different exercises as strata. During the sampling, one constraint was implemented: Once an attempt from a student was included in the sample, the entire sequence of attempts for this student and this particular exercise was included to capture entire interaction sequences as present in the training dataset. For Phrase and Sentence, this resulted in minimally more data units included in the sample than calculated according to the 90:10 ratio.

C Feedback Messages

Table 9 provides the set of teacher-crafted feedback messages for question-specific learner errors including such general messages as default, spelling, and capitalization.

Table 10 provides the set of feedback messages for the rules added according to the learner needs identified during profiling of training data.

D Inter Annotator Agreement

Table 11 shows the inter-annotator agreement results for the two independent raters.

E Gap Filling Exercises in FeedBook

Figure 6 shows examples of semi-open gap filling grammar exercises of input types "word", "phrase", and "sentence" for learning target "Questions" in FeedBook.

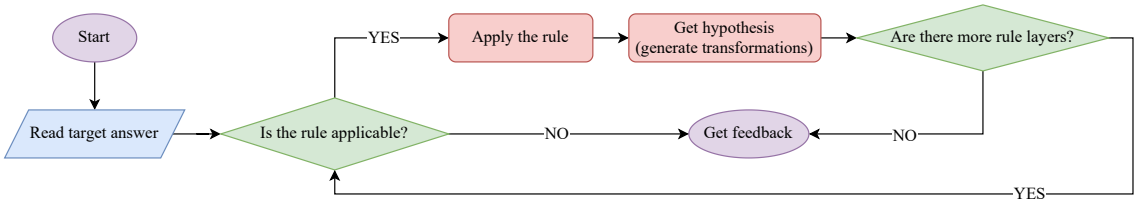


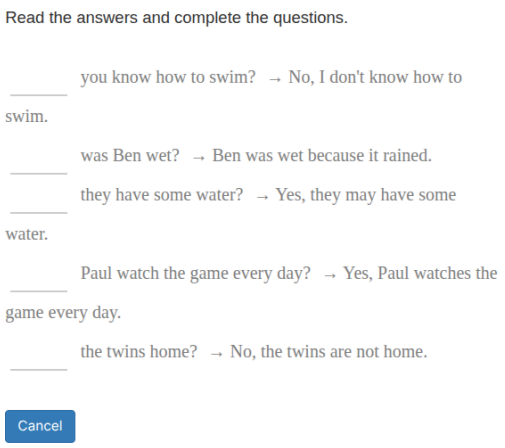
Figure 5: Feedback generation approach in the FeedBook system. Each rule consists of three methods: 1) Check whether the rule is applicable to the TA, i.e. relevant linguistic information is present in the TA to perform a specific transformation (e.g., *Question Word Replacement* is not applicable for "Did you sleep well?" but applicable for "Where did you see him?"); 2) Apply transformations to the TA in order to generate ill-formed and well-formed incorrect variations, and 3) get the error diagnosis for future feedback mapping.

Error	**Message**
Incorrect question word	Use another question word.
Missing question mark	You've forgotten to use the '?'.
Missing question word	You've forgotten to use a question word.
Missing auxiliary word	You've forgotten an auxiliary verb ('Hilfsverb', e.g. 'do'/'did').
Incorrect auxiliary word	That's not the correct auxiliary verb ('Hilfsverb').
Auxiliary and main verb are both in past tense	In questions with 'did' and a main verb (e.g., "Did you go?"), the main verb needs to be in the infinitive.
Incorrect agreement of auxiliary verb 'do'	That's not the correct form of an auxiliary verb ('Hilfsverb').
Auxiliary and main verb are both in third person	In questions with 'does'/'do' and a main verb (e.g., "Where does he live?"), the main verb needs to be in the infinitive.
Spelling error	The following words in your answer seem to contain spelling errors: [the word from the student answer with a spelling mistake].
Capitalization error	There is a problem with capitalization in your answer.
Default feedback	This is not what I am expecting - please try again.

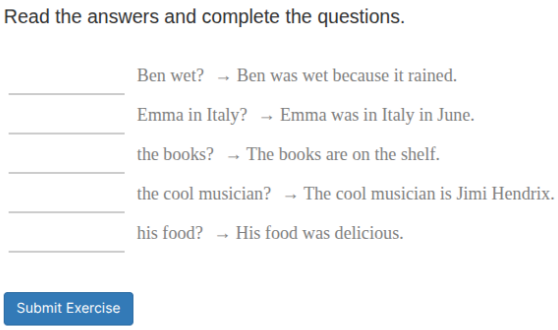
Table 9: Teacher-crafted feedback messages for the initial rules under analysis.

Error	**Message**
Main verb in past tense, auxiliary verb in infinitive	There seems to be a mistake in inflection. Look at this example: "What did you eat for breakfast?" and try to use the same structure in your question. Pay attention to the form of the auxiliary verb ('Hilfsverb', e.g. 'did') and the main verb.
Missing main verb	You forgot the main verb.
Main verb 'be' replaced by an auxiliary verb	You do not need an auxiliary verb ('Hilfsverb') here, because the main verb is a form of the verb 'to be'. Remember to use the correct form of the verb 'to be' here.
Incorrect agreement of the verb 'be' as an auxiliary or main verb	You should use a different form of the verb 'to be'.
Incorrect word order	The words in your answer are in the wrong order.

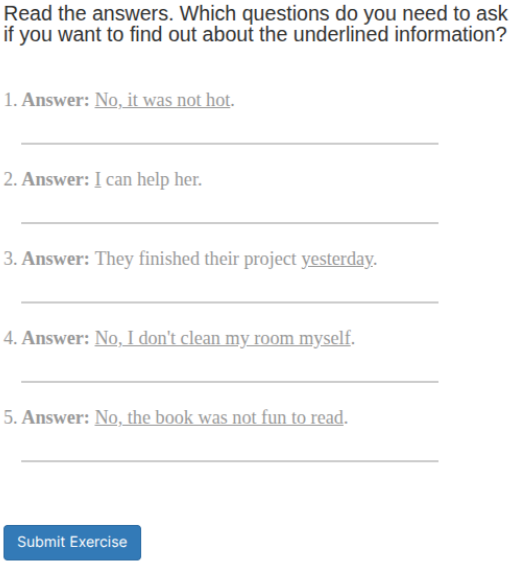
Table 10: Feedback messages for the added rules and a message for the additional word order check.



(a) Gap filling exercise with single word input.



(b) Gap filling exercise with phrase input.



(c) Gap filling exercise with full sentence input.

Figure 6: Examples of semi-open gap filling grammar exercises for learning target “Questions” in FeedBook.

baseline$_S$ (Total: 5076)	correct-specific	correct-non-specific	incorrect	unspecific
correct-specific	1676	0	12	0
correct-non-specific	0	69	0	0
incorrect	67	0	215	0
unspecific	0	1	0	3036
Inter-annotator agreement, Cohen's Kappa = **0.97** (Cohen, 1960)				
update$_S$ (Total: 5076)	correct-specific	correct-non-specific	incorrect	unspecific
correct-specific	2157	0	14	0
correct-non-specific	0	56	0	0
incorrect	22	0	235	0
unspecific	0	14	0	2578
Inter-annotator agreement, Cohen's Kappa = **0.98**				
baseline$_U$ (Total: 570)	correct-specific	correct-non-specific	incorrect	unspecific
correct-specific	169	0	7	0
correct-non-specific	0	7	0	3
incorrect	0	0	38	2
unspecific	0	1	0	340
Inter-annotator agreement, Cohen's Kappa = **0.96**				
update$_U$ (Total: 570)	correct-specific	correct-non-specific	incorrect	unspecific
correct-specific	286	0	8	0
correct-non-specific	0	8	0	2
incorrect	1	0	18	0
unspecific	0	0	0	248
Inter-annotator agreement, Cohen's Kappa = **0.97**				

Table 11: Inter Annotator Agreement.