ChatBack: Investigating Strategies of Providing Synchronous Grammatical Error Feedback in a GUI-based Language Learning Social Chatbot

Kai-Hui Liang[1], Sam Davidson[2], Xun Yuan[1], Shehan Panditharatne[1], Chun-Yen Chen[3],
Ryan Shea[1], Derek Pham[1], Yinghua Tan[3], Erik Voss[1], Luke Fryer[4], Zhou Yu[1,3]

[1]Columbia University, [2]University of California, Davis, [3]Articulate.AI,
[4]The University of Hong Kong,
{kaihui.liang, xy2569, zy2461}@columbia.edu, *ssdavidson@ucdavis.edu*

Abstract

The increasing use of AI chatbots as conversation partners for second-language learners highlights the importance of providing effective feedback. To ensure a successful learning experience, it is essential for researchers and practitioners to understand the optimal timing, methods of delivery, and types of feedback that are most beneficial to learners. Synchronous grammar corrective feedback (CF) has been shown to be more effective than asynchronous methods in online writing tasks. Additionally, self-correction by language learners has proven more beneficial than teacher-provided correction, particularly for spoken language skills and non-novice learners. However, existing language-learning AI chatbots often lack synchronous CF and self-correction capabilities. To address this, we propose a synchronous conversational corrective feedback (CCF) method, which allows self-correction and provides metalinguistic explanations (ME). Our experiments examine the effects of different feedback presentation methods and self-correction on users' learning experiences and intention to use the system.Our study suggests that in chatbot-driven language-learning tools, corrective feedback is more effectively delivered through means other than the social chatbot, such as a GUI interface. Furthermore, we found that guided self-correction offers a superior learning experience compared to providing explicit corrections, particularly for learners with high learning motivation or lower linguistic ability.

1 Introduction

The growing prevalence of AI chatbots as conversational partners for second-language learners emphasizes the vital role of delivering effective feedback to enhance the overall learning experience. As researchers and practitioners work to optimize computer-based conversational language learning, it is essential to determine the optimal timing, methods of delivery, and feedback types that contribute to the most successful outcomes. Prior research has shown that synchronous corrective feedback (CF) for grammatical errors is more effective than asynchronous methods in online writing tasks (Shintani and Aubrey, 2016). However, the best form of synchronous CF in AI chatbot systems has yet to be determined. Furthermore, self-correction by language learners has proven to be more beneficial than teacher-provided correction (Brown, 2009), especially for spoken language skills and for learners with more than limited L2 proficiency. Despite this evidence, numerous current language-learning AI chatbots lack diverse synchronous CF and self-correction features. And while past research has shown that learners' proficiency levels significantly influence their preferences (Orts and Salazar, 2016; Yang, 2016; Wiboolyasarin et al., 2022), the optimization of feedback strategies to adapt to users with varying proficiencies and motivations in language-learning chatbots remains unexplored. To address this limitation, we propose a AI chatbot for language learning with synchronous conversational corrective feedback (CCF), and investigate the effect of the feedback form and self-correction with metalinguistic explanations (ME). Specifically, we explore the following two research questions:

RQ1: How do the forms of CF delivery, specifically, feedback from the conversational partner (i.e., the chatbot) and a separate role (i.e., a GUI), impact the learning experience, including conversational enjoyment, negative emotions, self-efficacy, perceived usefulness, and intention to use the system? We hypothesize that: **H1**: Learners prefer receiving feedback from a separate role rather than from the conversation partner.

RQ2: How does the process of self-correction (compared to explicit feedback without self-correction) impact the learning experiences, including conversational enjoyment, negative emotions, self-efficacy, perceived usefulness, and intention to

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 83–99
July 13, 2023 ©2023 Association for Computational Linguistics

use the system? Specifically, what are the effects on people with different linguistic ability and learning purposes? We hypothesize that: **H2.1**: Learners with lower linguistic ability prefer receiving guided self-correction compared to those with higher proficiency. And **H2.2**: Learners with serious learning purposes prefer receiving guided self-correction relative to those who report other learning motivation.

2 Related Work

2.1 Chatbots as Conversational Partners for L2 Learners

A major challenge for second language instructors and students is finding adequate opportunities for students to practice conversational skills. A possible solution is the use of AI-driven chatbots to fill this gap. For example, Fryer and Carpenter (2006) discuss how chatbots can be used to increase opportunities for students to practice their second language. Fryer and Carpenter (2006) also point out that students who are reticent to speak with human interlocutors are often able to talk more freely with a computer. Similarly, Huang et al. (2022) states that chatbots "encourage students' social presence by affective, open, and coherent communication." This interaction is driven by recent advances in generative AI and chatbot design that have improved the dialogue flow of chatbots as well as their adaptability to individual user attributes (Li et al., 2022). In the present work we combine scripted dialogue with generative AI to create a chatbot which is able to effectively interact with users.

2.2 Automatic Corrective Feedback for L2 learners

Providing CF to students is an extremely time-consuming prospect for instructors (Shintani, 2016), and the automation of feedback can free up instructor time to focus on rhetorical and conversational skills (Li et al., 2015). Particularly, automated CF (ACF) can provide the type of real-time feedback to students that is impossible for instructors to provide, allowing students to immediately take advantage of the proposed suggestions and gain more confidence in their independent expressive abilities (Barrot, 2021). Heift and Hegelheimer (2017) further explains that ACF enables "learner self-study and practice of the target language by identifying and explaining error sources" and allows for self-revision.

In the present work, we test two alternate types of CF: explicit and implicit feedback, in the context of an educational chatbot for language learning. Previous work had shown that providing metalinguistic explanations without explicit corrections, which we term guided self-correction, tends to result in better student engagement and immediate gains in target-form usage (Sauro, 2021) and may improve long-term learning outcomes in writing tasks (Gao and Ma, 2019; Barrot, 2021). (Penning de Vries et al., 2020) investigates the use of ACF in a spoken language system, and finds speaking practice with ACF benefits users' learning goals. However, these feedback methods have not previously been tested in the context of language learning chatbots, a gap that the present paper seeks to address.

An additional key aspect of the present work is our testing alternate strategies for presenting feedback to language learners. Specifically, we test whether students prefer receiving CF directly from the chatbot as part of the conversational flow, or from another source such as the GUI window. While previous work has looked at student reactions to the timing of CF (Deeva et al., 2021), student control over feedback (Deeva et al., 2021), and level of explicitness (Sarré et al., 2021; Sauro, 2021), few studies investigate the effect of method of feedback presentation on engagement and learning experience. As such, this study is the first to investigate the impact of strategies for providing feedback on learning experiences and self-efficacy in the setting of a language learning chatbot.

2.3 Grammatical Error Correction & Classification models

Much recent progress has been made in the task of Grammatical Error Correction (GEC). To date, this work has largely focused on student essays (Ng et al., 2014; Bryant et al., 2019). For example, Omelianchuk et al. (2020)'s GECToR reframes the GEC task as a sequence labeling task rather than a sequence transformation task. Other promising models are proposed by Stahlberg and Kumar (2021) and Rothe et al. (2021), who achieve strong results on the JFLEG (Napoles et al., 2017) and CoNLL-2014 (Ng et al., 2014) datasets, respectively. Furthering this work, Qorib et al. (2022) achieves state-of-the-art results on several datasets by combining successful GEC models, such as Omelianchuk et al. (2020) and Rothe et al. (2021)

using a simple logistic regression algorithm. More recently, Fang et al. (2023), Wu et al. (2023), and Coyne and Sakaguchi (2023) have investigated the application of pretrained large language models, such as GPT-3, to GEC benchmark tasks. We emphasize that the above-referenced works primarily target correcting written student essay data. We, on the other hand, seek to apply GEC to the dialogue domain, and thus previously proposed GEC models may not work as effectively as demonstrated in prior art.

The present work also relies on error classification models to ensure that the correct type of feedback is presented to users. ERRANT (Bryant et al., 2017) is a rule-based algorithm to discriminate error categories by their part-of-speech (POS) tags. As an improvement to ERRANT, SERRANT (Choshen et al., 2021) improves the type accuracy by utilizing SErCL (Choshen et al., 2020) rules when ERRANT is not informative. SErCL defines errors by combining the Universal Dependencies (Nivre et al., 2016) tags of the target item before and after correction.

3 Study Method

3.1 Recruitment and participants

For this study, we recruited native Mandarin speakers as participants. To find users genuinely interested in conversing with a chatbot and improving their English grammar, we used social media for recruitment, rather than relying on school classes or Amazon Mechanical Turk. Our demographic recruitment criteria included being a native L1 Mandarin speaker aged 18 years or older. We also sought participants having an interest in discussing travel (the topic of the study) in English via text message while receiving grammatical error feedback. Participation in the study was entirely voluntary and unpaid.

175 participants completed the conversation and post-survey, with the following socio-demographic profile. The average age of respondants was 32 years, with the large majority having post-secondary education. Participants have studied English for an average of 15.7 years. Most participants reported self-improvement or having fun as their motivation for engaging with our system. Of those users who participated, 120 users produced one or more targeted errors while using the system. A full breakdown of sociodemographic details can be found in Appendix B.

Conversation
Post-survey
Asynchronous CF

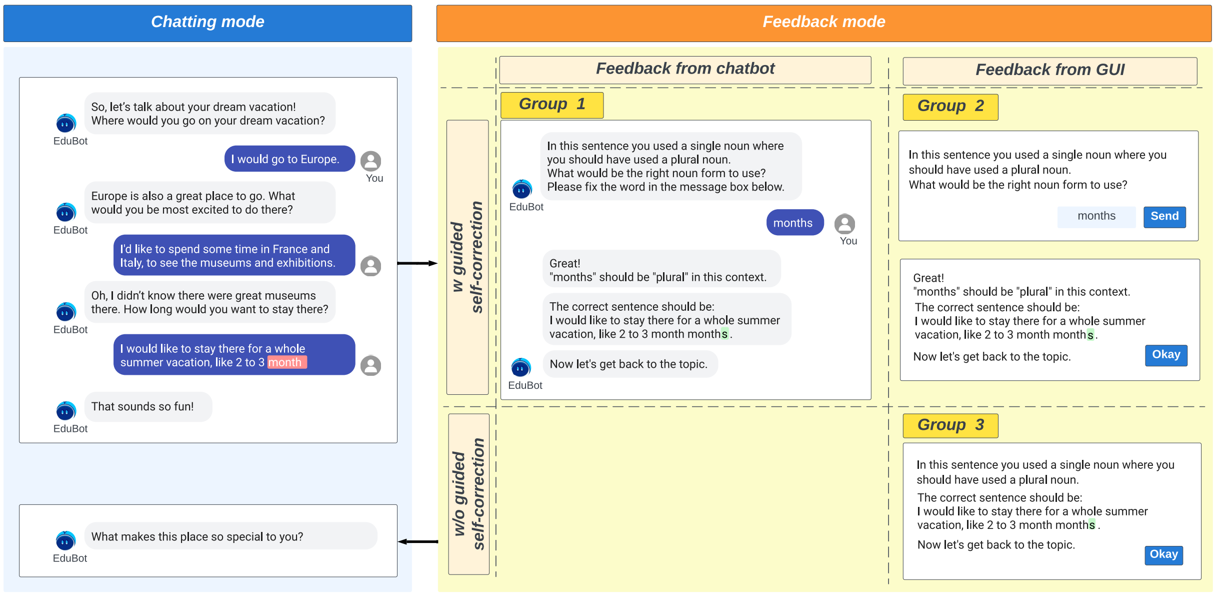
Figure 1: User study procedure

3.2 Procedure

Figure 1 depicts the user study procedure. Participants were randomly allocated to one of three experimental groups, each implementing a unique grammatical error feedback strategy. The study initiated with a travel-themed conversation with the chatbot. If participants made grammatical errors, as detected by our GEC model, the system offered feedback in accordance with their group's strategy. To ensure that grammar errors could be identified, users were required to type at least three words per turn and encouraged to use complete sentences. They also needed to complete a minimum of 12 dialogue turns, corresponding to the length of the scripted responses. After the conversation, users completed a post-survey collecting their socio-demographic information, English learning background, motivations, and subjective experiences with the system. To incentivize survey completion, participants who finished the survey received asynchronous grammar feedback, including a conversation summary and grammar error corrections for their responses. Both the system UI and post-survey were in Mandarin.

3.3 Conversation and grammar error feedback

As shown in Figure 2, the conversation alternates between chatting and feedback modes for all experimental groups. It starts with a chatting mode discussing travel with users. Whenever a user makes a grammatical error from the targeted error types (as defined in Section 3.3.1 below), the system first acknowledges their response and then switches to feedback mode. In Group 1, users receive feedback directly from the chatbot (i.e., the interlocutor) via guided self-correction. In Groups 2 and 3, however, users receive feedback via a pop-up window on the system GUI (i.e., separate from the interlocutor) to distinguish it from the conversation. While Group 2 receives guided self-correction, group 3 only receives explicit error correction without an opportunity to self-correct. (See 3.3.2 for more details.) Once the feedback is completed, the system switches back to chatting mode and resumes the ongoing conversation. In case of a non-targeted



So, let's talk about your dream vacation! Where would you go on your dream vacation?

I would go to Europe.

Europe is also a great place to go. What would you be most excited to do there?

I'd like to spend some time in France and Italy, to see the museums and exhibitions.

Oh, I didn't know there were great museums there. How long would you want to stay there?

I would like to stay there for a whole summer vacation, like 2 to 3 month

That sounds so fun!

What makes this place so special to you?

In this sentence you used a single noun where you should have used a plural noun.
What would be the right noun form to use?
Please fix the word in the message box below.

months

Great!
"months" should be "plural" in this context.

The correct sentence should be:
I would like to stay there for a whole summer vacation, like 2 to 3 month months.

Now let's get back to the topic.

In this sentence you used a single noun where you should have used a plural noun.
What would be the right noun form to use?

months

Great!
"months" should be "plural" in this context.
The correct sentence should be:
I would like to stay there for a whole summer vacation, like 2 to 3 month months.
Now let's get back to the topic.

In this sentence you used a single noun where you should have used a plural noun.
The correct sentence should be:
I would like to stay there for a whole summer vacation, like 2 to 3 month months.
Now let's get back to the topic.

Figure 2: Conversation and feedback flow

error (i.e., an error detected by the GEC model but not explicitly handled by our feedback generator), the system simply highlights the error in the GUI and displays the corrected form at the appropriate location in the user's previous utterance, without disrupting the chatting mode.

3.3.1 Targeted error types

Our current feedback generation method generates feedback for five common types of grammatical errors frequently made by English learners. The error types are defined according to the SERRANT framework (Choshen et al., 2021). The error types we target are as follows:

- `VERB:SVA`: Subject-verb agreement errors.
- `VERB:TENSE`: Incorrect verb tense usage.
- `VERB:FORM`: Verb form errors. For example, using an infinitive verb when a conjugated form is needed.
- `NOUN:NUM`: Noun number errors. For example, a user saying "I like cat" instead of "I like cats".
- `DET`: Misuse or omission of a determiner, such as "the" or "a".

We target these errors because they are among the most common errors identified in the ErAConD dataset, indicating a high prevalence of these error types in L2 English learner conversations. We also consulted with professional second language educators who agreed that these error types are among the most frequently seen in their students' speech. Finally, to avoid overwhelming students with feedback and disrupting the conversation too frequently, we chose this relatively small set of errors to target for the purposes of this study; we plan to add additional error types in future work.

3.3.2 Grammar error feedback strategies

When the user makes a targeted error, we generate CF that includes metalinguistic explanations, hints, and corrected forms. We use the term "metalinguistic" to reference a student's capacity to "reflect on and manipulate the structural features of language" (Nagy and Anderson, 1995). In the context of the present work, we define "metalinguistic explanation" as feedback which contains explicit information about the student's language use, such as pointing out that the student used an incorrect verb tense. Depending on the experimental group, the feedback presented to the user can consist of one or more of the following types:

1. Error identification: This specifies the portion of the user's utterance that contains the error without providing the correct form.
2. Implicit metalinguistic clues: This includes a metalinguistic suggestion about the type of error made, followed by prompts that encourage the user to self-correct, with additional guidance. There are two levels of this type of feedback: Level 1 provides a simple metalinguistic suggestion for the user's first attempt, while level 2 provides a more detailed metalinguistic explanation for the second attempt.

3. Explicit correction: This provides an explicit statement of the corrected form.

We present these suggestions in different ways depending on the experimental setting. The first type of feedback, which we refer to as *guided self-correction*, begins with feedback types 1 and 2, and progresses to type 3 only if the student is unable to self-correct after two attempts. In this approach, the user is first provided the identified error portion (e.g. “In this sentence you made a mistake on the verb ‘are’. ”), along with a metalinguistic suggestion (level 1) and an opportunity to self-correct (e.g. “What verb form should you have used? For example, "sees" and "saw" are different forms of "see".”). If the user is unable to self-correct, they are given a second chance with a more detailed metalinguistic suggestion (level 2) (e.g. “Not quite. Think about subject-verb agreement. How should your verb be changed to agree with the subject "He"? ”) If the user is still unable to self-correct after two attempts, we then present the explicit correction containing the corrected form. (e.g. ‘“Good try, but not quite. It’s tricky, I know. The correct verb form here is "is". Remember to make your verbs agree with their subjects.”) This guided self-correction feedback approach is presented to experimental groups 1 and 2, as shown in Figure 2. The second type of feedback, which we refer to as *explicit feedback*, consists only of providing type 1 and type 3 feedback (see group 3 in Figure 2).

3.4 Measurement

3.4.1 Linguistic ability

Linguistic ability includes various aspects. In this study, we focus on learners’ lexical competence in their produced utterances. We measure lexical diversity using the VocD method (McKee et al., 2000) [1] and assess lexical sophistication with the English Vocabulary Profile (EVP), aligning vocabulary usage with CEFR levels. Both metrics are evaluated with the online tool Text Inspector (Bax, 2012), with the medium of text designated as "writing." While the Text Inspector tool also provides language proficiency levels based on the CEFR framework, we do not rely on this information in our study. The tool’s original design primarily targets writing tasks and may not be as suitable for evaluating language proficiency in textual conversation. For a comprehensive evaluation of the results, please refer to Appendix D.

3.4.2 Post-conversation surveys

Upon the completion of each conversation, we gathered self-reported ratings from users on five distinct constructs related to users’ attitudes toward the system: negative emotion toward the feedback (frustration and annoyance), self-efficacy (confidence in grammar usage and expressive ability), perceived usefulness of the grammatical CF and suggestions, enjoyment using the system, and future intention to use the system. To ensure the reliability and validity of these constructs, we utilized a set of two measurement items, each rated on a 5-point Likert scale, for each construct. These measurement items were adapted from previous research studies (See Table 9) and subsequently modified to better suit the context of language learning chatbots. Figure 5 shows the survey results for each item. Hypotheses related to each construct and detailed descriptions of the constructs are shown in Appendix F.

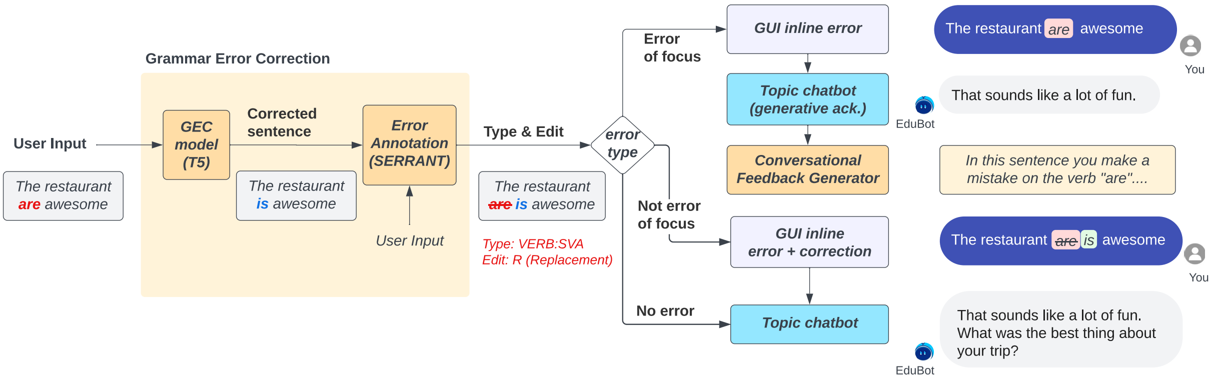
4 System

4.1 Overview

Figure 3 presents the system pipeline in chatting mode. At each turn, user input is first processed by the grammar error correction (GEC) module. If any targeted errors are identified, the system switches to feedback mode. The system first highlights the portion of the user’s utterance that contains errors with red backgrounds. Then, the topic chatbot acknowledges the user’s response using its generation model. Subsequently, the conversational feedback generator provides grammatical feedback to the users. The feedback content and form of delivery will vary depending on the group’s feedback strategies. For non-targeted error types, the topic chatbot will continue the conversation while the system will highlight the user’s error and display the corrected form on the GUI at the user’s previous response. If there are no grammar errors in the user’s input, the topic chatbot continues the conversation without highlighting or interruption.

The process in feedback mode, where targeted types are being addressed, proceeds as follows: For the group without guided self-correction (group 3), the system switches back to chatting mode immediately after providing explicit grammatical feedback at the same turn. For groups with guided self-correction (groups 1 and 2), the feedback mode continues to the next turn until the correction process concludes. During feedback mode in subsequent turns, the GEC module checks if users are able

[1] https://textinspector.com/help/lexical-diversity/



The restaurant are awesome

That sounds like a lot of fun.

In this sentence you make a mistake on the verb "are"....

The restaurant ~~are~~ is awesome

That sounds like a lot of fun. What was the best thing about your trip?

Figure 3: System pipeline in chatting mode: Grammar error correction & response generation flow

to successfully self-correct their errors. If users self-correct successfully, the feedback generator acknowledges the correction and the system returns to chatting mode where the topic chatbot continues the conversation. If they don't, they are given a second chance where the feedback generator provides a more detailed metalinguistic hint. If they fail to self-correct after two attempts, the feedback generator provides explicit feedback the system switches back to chatting mode. Otherwise, the feedback continues.

4.2 Topic chatbot

The topic chatbot combines scripted dialogue with a generative model to create a topic-oriented chatbot capable of effectively interacting with users. At every dialogue turn, the chatbot first generates a response and subsequently concatenates it with the scripted responses. Scripted dialogue is employed for experimental control purposes, primarily to pose questions designed to elicit more grammatical errors and to ensure consistency in the topics presented to users across different experimental groups. Conversely, the generative model is used to acknowledge user responses in a more natural manner by dynamically responding to user input.

The script encompasses 12 dialogue turns covering travel preferences, past travel experiences, and dream vacations. We employ Blenderbot3 3B as our generative model, which possesses various conversational skills and long-term memory. To reduce latency, Blenderbot's internet access was disabled during experiments. After completing the scripted portion of the conversation, if users decide to continue the conversation, the chatbot's responses will rely solely on the generative model.

4.3 Grammatical Error Feedback

4.3.1 Grammar error correction

Table 1: Performance of GEC model. TP, FP and FN denote the average number of true positives, false positives and false negatives among 5 runs of cross-validation, respectively.

Model	TP	FP	FN	Prec	Rec	$F_{0.5}$
GECToR	24.6	14.4	174.0	0.63	0.12	0.34
T5 (Ours)	43.8	34.6	154.8	0.56	0.23	0.43

Figure 3 illustrates the grammar error correction process, which consists of two main steps: grammar error correction and error annotation. First, we use a grammar error correction (GEC) model to generate corrected sentences based on user-input sentences. The GEC model is a T5 (Raffel et al., 2020) model trained for grammar correction[2]. We fine-tuned the model on the ErAConD dataset (Yuan et al., 2022), a GEC conversation dataset between L2 English learners (of at least intermediate proficiency level) and an educational chatbot. We selected level 3 errors (as defined in the ErAConD dataset) as our training data since they are most likely to result in misunderstanding. The resulting fine-tuned model achieves an overall $F_{0.5}$ of 0.43 evaluated by 5-fold cross-validation, as shown in Table 1. Detailed results by error type are shown in Appendix Table 10. While our reported $F_{0.5}$ is substantially lower than SOTA GEC models designed for written text, there is no established baseline for dialog GEC. Note that the precision of 0.56 doesn't mean that half of the edits generated are incorrect. In fact, there are many equally valid ways to correct a given grammar error; however, when

[2] https://huggingface.co/deep-learning-analytics/GrammarCorrector

calculating precision using a test dataset, we can only compare system-generated corrections with the one or two human-annotated gold edits. If the machine-generated correction does not match the gold annotation, it will negatively impact evaluation performance, even if the correction is a completely legitimate alternative. As a result, current evaluations tend to underestimate the performance of GEC models. Rozovskaya and Roth (2021) provides an in-depth study of this issue. While the current model is effective for the present study, we are working to improve the GEC model for future iterations of our system.

After error correction by the GEC model, SERRANT compares the user input sentence with the corrected version to extract edits and classify error types. For most categories, there are three possible operations to specify user input errors: Missing (M), Replacement (R), and Unnecessary (U), indicating whether tokens should be inserted, substituted, or removed, respectively. Subsequently, we filter out trivial grammar error types (e.g., punctuation) and reapply the edits to the original sentences.

4.3.2 Grammar error feedback presentation

Grammar errors can be presented in three different forms: 1) GUI inline highlighting on the user's utterance, 2) conversational feedback presented in the form of a chatbot response from the feedback generation module, and 3) conversational feedback presented in a pop-up window from the feedback generation module.

As discussed in Section 3.3.1, our feedback generation module explicitly targets five error types, while other error types detected by our GEC model are referred to as "non-targeted". For targeted errors, the error is first presented in the form of GUI inline highlighting on the user's previous response. Then, after the topic chatbot acknowledges the user's content, conversational feedback is presented in a form that depends on the experiment group. For group 1, the feedback is presented by the chatbot, while for groups 2 and 3, it is presented in a pop-up window. For non-targeted errors, only GUI inline highlighting is shown without any additional feedback.

To generate conversational feedback, we rely on a number of feedback templates that can be modified based on the specifics of the respective error. For example, if SERRANT tags an error as `R:NOUN:NUM`, indicating a replacement operation ('R') resulting from a difference in noun number between the original input and the correction, we populate a template with noun number information to generate feedback such as "In this sentence, you used a single noun when you should have used a plural noun", as shown in Figure 2. We use a similar approach to populate feedback templates for error types such as subject-verb agreement, verb tense, verb forms, and determiners.

5 Results

5.1 Dialog statistics

Table 2 displays the distribution of participants across each experimental group. Among the 175 participants, 154 encountered at least one error, with 120 experiencing at least one targeted error. In this study, our survey analysis focuses on the 120 users who encountered targeted errors, since the primary experimental treatment involved the feedback delivery strategy for these errors.

Table 3 offers statistics for users who had targeted errors in their conversations, with a sample size of 120. On average, users engaged in 15.1 dialog turns (i.e. 15.1 responses from users), each consisting of 10.1 tokens. Each conversation contained 3.4 turns with any error, 1.6 turns with non-targeted errors exclusively, and 1.8 turns with targeted errors. The average number of errors per dialog amounted to 4.3. We also analyzed the most frequently occurred error types among all 175 participants, with the top ten including the five targeted error types as well as preposition, spelling, noun, and verb errors (see Appendix E for comprehensive error type counts).

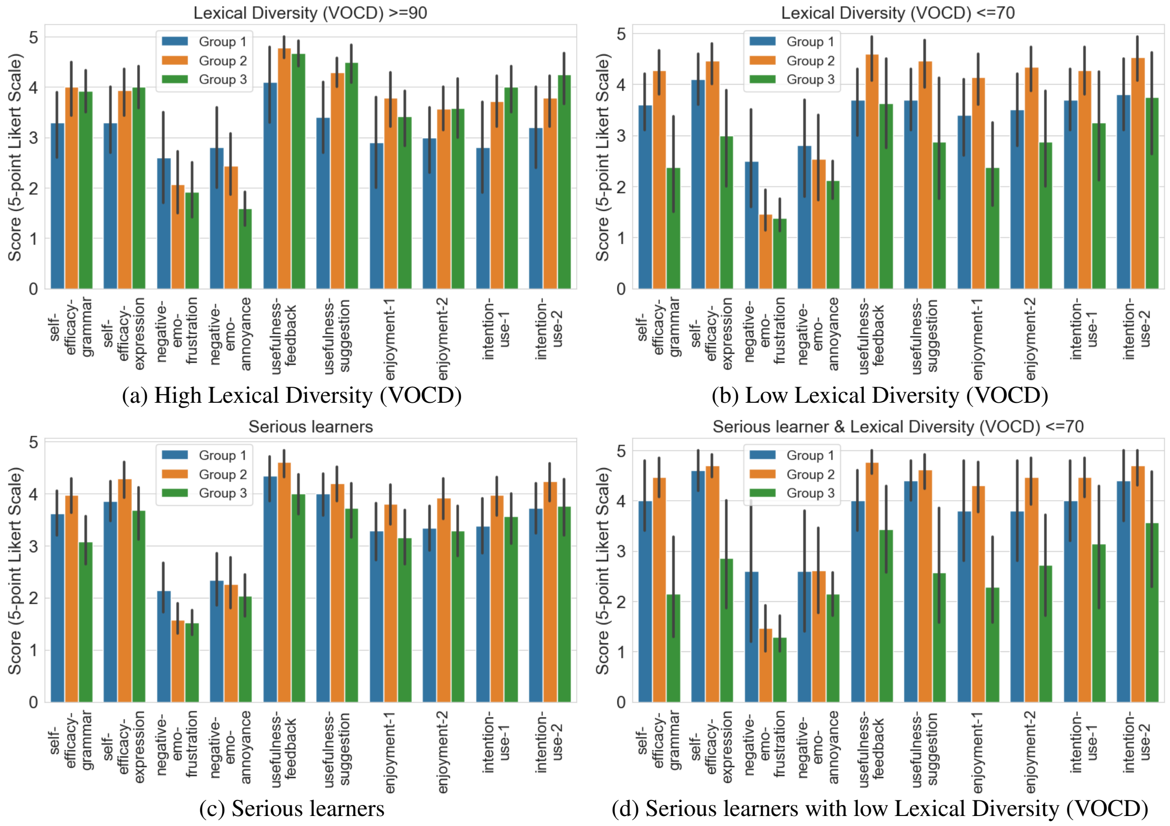
Regarding learners' lexical competence, we assessed their lexical diversity, which had a mean (M) value of 84.8 (SD = 27.0) and a median of 80.25. The range of lexical diversity scores ranged from 37.1 to 200 (see Appendix D for more details).

Table 2: Numbers of participants in each group

Group	All	W/ any err.	W/ targeted err.
Group 1	49	43	33
Group 2	66	60	48
Group 3	60	51	39
Total	**175**	**154**	**120**

5.2 Survey results

Figure 5 Shows the survey results of all dialogs with targeted errors. We performed two-tailed t-tests between groups (Groups 1 and 2 for RQ1,



(a) High Lexical Diversity (VOCD)

(b) Low Lexical Diversity (VOCD)

(c) Serious learners

(d) Serious learners with low Lexical Diversity (VOCD)

Figure 4: Survey results of learners with with different lexical diversities and motivation.

Table 3: Dialog statistics

Item	M ± SD	Mdn.	Range
# of dialog turns	15.1 ± 5.2	13	13-47
# of tokens per turn	10.1 ± 4.4	9	4-29
# of turns w/ any error	3.4 ± 2.2	3	1-16
# of turns w/ non-targeted errors only	1.6 ± 1.7	1	0-10
# of turns w/ targeted error	1.8 ± 1.0	1	1-6
# of errors per dialog	4.3 ± 3.6	3	1-31

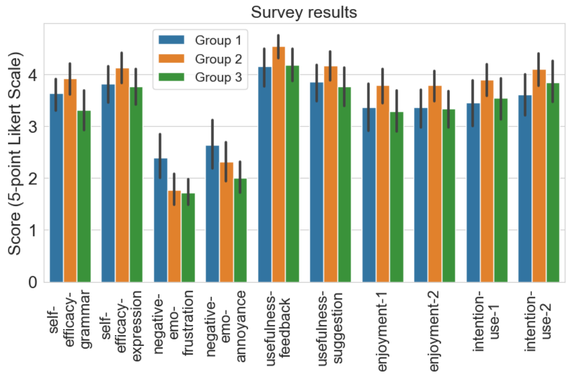


Figure 5: Survey results

and Groups 2 and 3 for RQ2), and use Welch t-test when the sample sizes are unequal, as recommended by Zimmerman (2004).

5.2.1 Effects of the form of feedback delivery

The results presented in Figure 5 demonstrate that users experienced higher frustration levels when interacting with Group 1 than with Group 2 ($t(58.61) = 2.26, p < .05$). Our findings suggest that feedback provided by the dialogue agent leads to greater frustration than feedback delivered from another role, such as the GUI, even when the content and timing of the feedback are identical.

5.2.2 Effects of guided self-direction

Figure 5 shows that users gained more self-efficacy in their grammar skills when interacting with Group 2 compared to Group 3 ($t(77.88) = 2.51, p < .05$). These results suggest that guided self-correction may be beneficial for enhancing users' confidence in their English grammar skills during conversations.

Effects of user's linguistic ability To examine the influence of guided self-correction on users with varying linguistic abilities, we analyzed survey data from participants with higher and lower lexical diversities (VocD >= 90 and VocD <=70, respectively). The threshold values were determined based on the median VocD score (80) with a range of plus or minus 10. Our results indicate that users with higher lexical diversity found guided self-correction (Group 2) more annoying compared to the absence of guided self-correction (Group 3). This could be because users with higher lexical competence might have already understood the corresponding metalinguistic rules, making guided self-correction redundant and less efficient than explicit feedback.

Effects on users' motivation To investigate the effects on users with varying motivations, particularly their level of commitment to improving their English conversation skills, we excluded approximately one-third of users who reported using the system out of curiosity or for fun and defined the remaining users as "serious learners". Our findings (Figure 4c) reveal that serious learners not only experienced significantly higher levels of confidence in their grammar skills with guided self-correction ($t(46.57) = 2.96, p < .01$), but also perceived the feedback to be more useful compared to the absence of guided self-correction ($t(40.54) = 2.47, p < .01$). Moreover, we conducted a further analysis on serious learners with low lexical diversity (VOCD <= 70) (Figure 4d) and found that when receiving guided self-correction, they reported higher enjoyment in conversation ($t(9.14) = 3.46, p < .01$ for enjoyment-1 and $t(8.28) = 2.84, p < .05$ for enjoyment-2), increased self-efficacy in both grammar skills ($t(8.21) = 4.20, p < .01$) and expressing ideas ($t(6.61) = 3.01, p < .05$), and perceived the grammatical corrective feedback ($t(6.78) = 2.70, p < .05$) and suggestions ($t(6.94) = 3.03, p < .05$) to be more useful compared to the absence of guided self-correction.

6 Conclusion

Results from this preliminary study provide evidence that learners may prefer getting corrective feedback from a separate role, instead of from the conversation partner to reduce frustration. In addition, guided self-correction may provide better learning experiences than the absence of self-correction, especially for learners with lower lexical competence or more serious learning motivation. These findings highlight the importance of considering users' individual differences when designing language-learning chatbots, and the need for personalized feedback mechanisms that cater to individual users' need.

7 Limitations

7.1 Assessment of learner's linguistic ability and future research

In this study, the assessment of learners' linguistic ability was limited to analyzing the learners' produced utterances in a single short conversation. Also, it was analyzed with the online tool TextInspector, which was primarily designed for evaluating writing tasks rather than textual conversation. While this provides some insight into their language proficiency, a more comprehensive assessment of learners' language proficiency could offer a deeper understanding of how it influences their preference toward different feedback strategies. Future research should consider incorporating additional measures to evaluate learners' language proficiency comprehensively. This could involve utilizing standardized tests for receptive and productive skills and conducting detailed assessments of vocabulary, grammar, and discourse abilities.

7.2 Effect of participants' language proficiency

In this study, survey data were collected from participants capable of engaging in a conversation about travel with at least 12 turns from each side. Participants without the ability to meet this requirement were automatically excluded and did not complete the post-survey. Previous research (Van Beuningen et al., 2012) indicates that learners with limited proficiency may prefer explicit corrective feedback, as they may face challenges in independently arriving at correct answers. However, it should be noted that due to the inherent study design, some learners with limited proficiency might not have been included in the sample.

7.3 Effect of the GEC model performance

During the experiment, there were no existing GEC (Grammar Error Correction) models specifically designed for conversational grammar errors. As a result, we developed our own GEC model using a small dataset of GEC dialogues. To enhance the performance of the GEC model in future iterations, we are actively working on collecting additional conversational GEC datasets. By incorporating more diverse and extensive data, we aim to improve the accuracy and effectiveness of the GEC model. The enhanced performance of the GEC model is anticipated to have an impact on the effectiveness of different feedback strategies. A more proficient GEC model could potentially yield better user experiences, resulting in higher intentions to use the system. The availability of improved GEC capabilities will enable more precise and tailored feedback, enhancing the overall effectiveness of the system.

7.4 Effect of different feedback strategies

In this study, all feedback strategies used were interruptive, potentially disrupting the conversation flow. However, learners with higher linguistic ability may prefer fewer interruptions, such as preferring no self-correction than self-correction. Additionally, it is important to acknowledge that individual learners may have different preferences and learning styles. To address this, future systems could consider non-intrusive feedback strategies. For example, grammar errors could be highlighted with a background color, and optional metalinguistic explanations could be provided on-demand. This allows learners to access guidance without forcefully interrupting the conversation, catering to their preferences and maintaining a smoother learning experience.

References

Martínez Agudo and Juan de Dios. 2013. An investigation into how efl learners emotionally respond to teachers' oral corrective feedback. *Colombian Applied Linguistics Journal*, 15(2):265–278.

Jessie S Barrot. 2021. Using automated written corrective feedback in the writing classrooms: effects on l2 writing accuracy. *Computer Assisted Language Learning*, pages 1–24.

S Bax. 2012. Text inspector. *Online text analysis tool.*

Alan V Brown. 2009. Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The modern language journal*, 93(1):46–60.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805.

Leshem Choshen, Dmitry Nikolaev, Yevgeni Berzak, and Omri Abend. 2020. Classifying syntactic errors in learner language. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 97–107.

Leshem Choshen, Matanel Oren, Dmitry Nikolaev, and Omri Abend. 2021. SERRANT: a syntactic classifier for english grammatical error types. *CoRR*, abs/2104.02310.

Steven Coyne and Keisuke Sakaguchi. 2023. An analysis of gpt-3's performance in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. 2021. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Luke Fryer and Rollo Carpenter. 2006. Bots as language learning tools. *Language Learning & Technology*, 10(3):8–14.

Jianwu Gao and Shuang Ma. 2019. The effect of two forms of computer-automated metalinguistic corrective feedback.

Trude Heift and Volker Hegelheimer. 2017. Computer-assisted corrective feedback and language learning. *Corrective feedback in second language teaching and learning*, pages 51–65.

Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2022. Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Jinrong Li, Stephanie Link, and Volker Hegelheimer. 2015. Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27:1–18.

Yu Li, Chun-Yen Chen, Dian Yu, Sam Davidson, Ryan Hou, Xun Yuan, Yinghua Tan, Derek Pham, and Zhou Yu. 2022. Using chatbots to teach languages. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 451–455.

Mingxu Liu. 2013. English bar as a venue to boost students' speaking self-efficacy at the tertiary level. *English Language Teaching*, 6(12):27–37.

Gerard McKee, David Malvern, and Brian Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing*, 15(3):323–338.

William E Nagy and Richard C Anderson. 1995. Metalinguistic awareness and literacy acquisition in different languages. technical report no. 618.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR–Grammatical Error Correction: Tag, Not Rewrite. *arXiv preprint arXiv:2005.12592*.

Sara Orts and Patricia Salazar. 2016. Efl students' preferences towards written corrective feedback: An exploratory study on age and level of proficiency. *The Grove-Working Papers on English Studies*, 23.

Bart WF Penning de Vries, Catia Cucchiarini, Helmer Strik, and Roeland Van Hout. 2020. Spoken grammar practice in call: The effect of corrective feedback and education level in adult l2 learning. *Language Teaching Research*, 24(5):714–735.

Muhammad Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sonny Rosenthal and Rabindra A Ratan. 2022. Balancing learning and enjoyment in serious games: Kerbal space program and the communication mediation model. *Computers & Education*, 182:104480.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*.

Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698.

Tracii Ryan and Michael Henderson. 2018. Feeling feedback: students' emotional responses to educator feedback. *Assessment & Evaluation in Higher Education*, 43(6):880–892.

Raafat George Saadé, Weiwei Tan, and Fassil Nebebe. 2008. Impact of motivation on intentions in online learning: Canada vs china. *Issues in Informing Science & Information Technology*, 5.

Cédric Sarré, Muriel Grosbois, and Cédric Brudermann. 2021. Fostering accuracy in l2 writing: Impact of different types of corrective feedback in an experimental blended learning efl course. *Computer Assisted Language Learning*, 34(5-6):707–729.

Shannon Sauro. 2021. Computer-mediated corrective feedback and the development of l2 grammar. *UMBC Education Department Collection*.

Natsuko Shintani. 2016. The effects of computer-mediated synchronous and asynchronous direct corrective feedback on writing: a case study. *Computer Assisted Language Learning*, 29(3):517–538.

Natsuko Shintani and Scott Aubrey. 2016. The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *The Modern Language Journal*, 100(1):296–319.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47.

Ting Sun and Chuang Wang. 2020. College students' writing self-efficacy and writing self-regulated learning strategies in learning english as a foreign language. *System*, 90:102221.

Catherine G Van Beuningen, Nivja H De Jong, and Folkert Kuiken. 2012. Evidence on the effectiveness of comprehensive error correction in second language writing. *Language learning*, 62(1):1–41.

Kanokpan Wiboolyasarin, Ruedee Kamonsawad, Nattawut Jinowat, and Watcharapol Wiboolyasarin. 2022. Efl learners' preference for corrective feedback strategies in relation to their self-perceived levels of proficiency. *English Language Teaching Educational Journal*, 5(1):32–47.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Juan Yang. 2016. Learners' oral corrective feedback preferences in relation to their cultural background, proficiency level and types of error. *System*, 61:75–86.

Xun Yuan, Derek Pham, Sam Davidson, and Zhou Yu. 2022. Eracond: Error annotated conversational dialog dataset for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 76–84.

Donald W Zimmerman. 2004. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181.

A Supplementary Materials

The detailed experiment results related to this paper are available in the following GitHub repository: https://github.com/KaihuiLiang/chatback_gec_feedback

In the following sections, we have selected the most critical aspects of these results for a concise understanding.

B Sociodemographics of participants

Table 4: Sociodemographics

	All users (N=175)			Users with targeted errors (N=120)		
Sociodemographics	**n (%) or M ± SD**	**Mdn.**	**Range**	**n (%) or M ± SD**	**Mdn.**	**Range**
Age (years)	32.0 ± 13.7	26	18-70	32.1 + 13.2	26.5	18-70
Gender						
Women	99 (56.6%)			73 (60.8%)		
Men	66 (37.7%)			38 (31.7%)		
Prefer not to say	10 (5.7%)			9 (7.5%)		
Education						
Graduate	90 (51.4%)			61 (50.8%)		
Undegraduate	73 (41.7%)			52 (43.3%)		
High school	9 (5.1%)			5 (4.2%)		
others	3 (1.7%)			2 (1.7%)		
Motivation						
Self improvement	69 (39.4%)			50 (41.7%)		
For fun	62 (35.4%)			39 (32.5%)		
Pass tests	15 (8.6%)			12 (10.0%)		
others	12 (6.9%)			10 (8.3%)		
Talk to friends/families	5 (2.9%)			3 (2.5%)		
Travel	5 (2.9%)			3 (2.5%)		
Learn cultures	4 (2.3%)			2 (1.7%)		
Job opportunities	3 (1.7%)			1 (0.8%)		
Learning duration	15.7 ± 9.9	14	0-55	16 ± 9.4	15	0-50

C Dialog statistics and grammar error counts

Table 5: Dialog statistics

Dialog stats.	**All users (N=175)**			**Users w/ targeted err. (N=120)**		
Item	**M ± SD**	**Mdn.**	**Range**	**M ± SD**	**Mdn.**	**Range**
# of dialog turns	14.8 ± 4.6	13	13-47	15.1 ± 5.2	13	13-47
# of tokens per turn	9.8 ± 4.4	9	3-31	10.1 ± 4.4	9	4-29
# of turns w/ any error	2.7 ± 2.3	1	0-6	3.4 ± 2.2	3	1-16
# of turns w/ non-targeted errors only	1.5 ± 1.7	1	0-10	1.6 ± 1.7	1	0-10
# of turns w/ targeted error	1.2 ± 1.2	2	0-16	1.8 ± 1.0	1	1-6
# of errors per dialog	3.4 ± 3.5	3	0-31	4.3 ± 3.6	3	1-31

D Participants' lexical competence and language proficiency levels

Table 6: Users' lexical competence. All scores are measured by TextInspector based on users' responses.

	All users (N=175)			Users with targeted errors (N=120)		
Lexical competence	**M ± SD**	**Mdn.**	**Range**	**M ± SD**	**Mdn.**	**Range**
Lexical Diversity						
VocD	81.8 ± 27.8	78.5	0-200	84.8 ± 27.0	80.25	37.1 - 200
MTLD	76.8 ± 27.5	73.6	0-176.4	78.8 ± 25.9	74.7	30.1-176.4
Lexical Sophistication: English Vocabulary Profile (EVP)						
C2 type %	0.3 ± 0.6	0	0-2	0.3 ± 0.6	0	0-2
C1 type %	0.5 ± 0.7	0	0-4	0.5 ± 0.7	0	0-3
B2 type %	2.1 ± 1.9	1.7	0-8	2.0 ± 1.8	1.7	0-6
B1 type %	7.2 ± 3.2	6.8	0-16	7.1 ± 3.3	6.7	0-16
A2 type %	15.4 ± 4.5	15	5-30	15.8 ± 4.6	15.5	7-30
A1 type %	63.3 ± 6.6	63.4	46-80	63.0 ± 6.5	63.2	47-80
C2 token %	0.2 ± 0.4	0	0-2	0.2 ± 0.4	0	0-2
C1 token %	0.3 ± 0.5	0	0-3	0.4 ± 0.5	0	0-2
B2 token %	1.5 ± 1.3	1.2	0-5	1.4 ± 1.2	1.3	0-5
B1 token %	5.2 ± 2.4	5	0-11	5.2 ± 2.5	5	0-11
A2 token %	11.8 ± 3.4	11.4	5-23	12.0 ± 3.4	11.5	6-23
A1 token %	71.9 ± 5.4	72	53-85	71.9 ± 5.2	71.9	53-85

Table 7: Users' language proficiency levels. All scores are measured by TextInspector based on users' responses. The overall CEFR represents the holistic score derived from all available metrics. The "VocD - CEFR level" indicates the CEFR level determined by the VocD score, while the "MTLD - CEFR level" represents the CEFR level determined by the MTLD score.

	Overall CEFR level		VocD - CEFR level		MTLD - CEFR level	
Level	**All users (N=175)**	**Users with targeted err. (N=120)**	**All users (N=175)**	**Users with targeted err. (N=120)**	**All users (N=175)**	**Users with targeted err. (N=120)**
C2	1 (0.6%)	0	0	0	0	0
C1+	0	0	0	0	7 (4.0%)	6 (5.0%)
C1	4 (2.3%)	4 (3.3%)	0	0	5 (2.9%)	4 (3.3%)
B2+	26 (14.9%)	19 (15.8%)	18 (10.3%)	16 (13.3%)	9 (5.1%)	6 (5.0%)
B2	47 (26.9%)	31 (25.8%)	23 (13.1%)	16 (13.3%)	12 (6.9%)	10 (8.3%)
B1+	56 (32.0%)	38 (31.7%)	23 (13.1%)	17 (14.2%)	17 (9.7%)	13 (10.8%)
B1	32 (18.3%)	24 (20.0%)	12 (6.9%)	7 (5.8%)	12 (6.9%)	7 (5.8%)
A2+	7 (4.0%)	3 (2.5%)	0	0	0	0
A2	2 (1.1%)	1 (0.8%)	29 (16.6%)	18 (15.0%)	0	0
A1	0	0	0	0	38 (21.7%)	25 (20.8%)
N/A	0	0	70 (40.0%)	46 (38.3%)	75 (42.9%)	49 (40.8%)

E Grammar error type counts

Table 8: Grammar error type counts in utterances of all participants. Targeted errors are highlighted with a yellow background. "op." denotes operations: R for Replacement, M for Missing, U for Unnecessary. The error types are defined according to the SERRANT framework (Choshen et al., 2021).

Error type (w/ op.)	**Count**	**%**	**Error type (w/o op.)**	**Count**	**%**
R:NOUN:NUM	70	11.7	PREP	71	11.9
R:SPELL	62	10.4	NOUN:NUM	70	11.7
R:VERB:FORM	47	7.9	DET	62	10.4
R:VERB:SVA	38	6.4	SPELL	62	10.4
M:DET	38	6.4	VERB:FORM	60	10
R:PREP:WC	34	5.7	VERB:SVA	38	6.4
M:PREP	20	3.3	NOUN	34	5.7
R:OTHER	20	3.3	VERB:TENSE	22	3.7
R:VERB:TENSE	17	2.8	VERB	21	3.5
R:NOUN:WC	16	2.7	OTHER	20	3.3
U:DET	15	2.5	OTHER:MW	14	2.3
U:PREP	15	2.5	PRON	11	1.8
R:OTHER:MW	14	2.3	AUX:MW	9	1.5
U:NOUN	14	2.3	VERB:MW	8	1.3
M:VERB:FORM	12	2.0	NOUN->VERB	7	1.2
R:DET:WC	9	1.5	VERB:INFL	5	0.8
R:AUX:MW	9	1.5	NOUN:INFL	5	0.8
R:VERB:WC	9	1.5	NOUN->PRON	4	0.7
R:VERB:MW	8	1.3	ADV	4	0.7
R:NOUN->VERB	7	1.2	ADJ	4	0.7

F Survey constructs

Table 9 shows all survey questions and references.

Negative emotions For negative emotions towards feedback, we measured users' negative emotions, specifically their levels of frustration and annoyance when receiving immediate corrections during the conversation. Our hypotheses were that users would experience fewer negative emotions in two scenarios: 1) when receiving corrections from the GUI, which is a separate role from the chatbot; and 2) when not required to correct themselves.

Self-efficacy Regarding self-efficacy, we measured the level of self-efficacy that users gained after the conversation, specifically their confidence in their grammar skills and their ability to express ideas in English conversations. Our hypotheses were that users would experience a greater increase in self-efficacy when: 1) corrections were given through the GUI, which would provide a less frustrating experience; and 2) they were given the opportunity for guided self-correction, allowing them to actively participate in the learning process and gain a better understanding of their mistakes.

Usefulness For usefulness, we measured the level of perceived usefulness of the grammatical CF by users. Our hypothesis was that guided self-correction would be perceived as more useful than without.

Enjoyment Regarding enjoyment, we measured the level of enjoyment that users experienced while conversing with the chatbot. Our hypothesis was that receiving grammatical correction feedback from the GUI would be more enjoyable than from the chatbot, as the interruptive feedback would be given from a separate role rather than the conversation partner. Additionally, we hypothesized that higher proficiency learners would find having a conversation without guided self-correction more enjoyable, as they would require less self-correction and experience fewer interruptions.

Intention to use Lastly, we asked users if they intended to use the system again, using one item that was reverse-coded for a sanity check. Our hypothesis was that users would have a higher intention to use the system if they experienced less negative emotion, gained more self-efficacy, perceived the system as more useful, and enjoyed the conversation more.

Table 9: Survey questions

Construct	Item abbr.	Question	Reference
Self-efficacy	self-efficacy-grammar	I think my grammar skills in English conversations improved after using the system	(Sun and Wang, 2020) (Liu, 2013)
	self-efficacy-expression	I feel more confident expressing my ideas in English conversations after using the system.	
Negative Emotion	negative-emo-frustration	I feel frustrated when the system immediately corrects my grammar mistakes	(Ryan and Henderson, 2018) (Agudo and de Dios, 2013)
	negative-emo-annoyance	I feel annoyed when the system immediately corrects my mistakes	
Usefulness	usefulness-feedback	I think the grammar correction feedback during the chat is useful.	(Agudo and de Dios, 2013)
	usefulness-suggestion	I get useful suggestions about how to improve my grammar in English conversations	
Enjoyment	enjoyment-1	I enjoyed talking with the chatbot.	(Saadé et al., 2008)
	enjoyment-2	Talking with the chatbot was pleasant.	
Intention to use	intention-to-use-1	I would like to use this system again.	(Rosenthal and Ratan, 2022)
	intention-to-use-2	I am not interested in using this system again.	

G GEC model performance

Table 10: Performance of our T5 GEC model by grammar error type following ERRANT's error code.

Type	TP	FP	FN	Prec	Rec	$F_{0.5}$
M:ADJ	0	1	0	0	1	0
M:ADV	1	0	6	1	0.14	0.45
M:CONJ	0	0	5	1	0	0
M:CONTR	0	0	6	1	0	0
M:DET	7	13	37	0.35	0.16	0.28
M:NOUN	0	2	1	0	0	0
M:NOUN:POSS	0	0	3	1	0	0
M:OTHER	1	1	25	0.5	0.04	0.15
M:PART	0	0	1	1	0	0
M:PREP	6	2	16	0.75	0.27	0.56
M:PRON	0	7	22	0	0	0
M:VERB	2	5	14	0.29	0.13	0.23
M:VERB:FORM	5	7	8	0.42	0.38	0.41
M:VERB:TENSE	1	1	4	0.5	0.2	0.38
R:ADJ	1	2	10	0.33	0.09	0.22
R:ADJ:FORM	1	1	4	0.5	0.2	0.38
R:ADV	3	0	9	1	0.25	0.63
R:CONJ	0	0	1	1	0	0
R:DET	9	0	17	1	0.35	0.73
R:MORPH	8	1	29	0.89	0.22	0.55
R:NOUN	2	4	31	0.33	0.06	0.18
R:NOUN:INFL	2	1	3	0.67	0.4	0.59
R:NOUN:NUM	17	16	32	0.52	0.35	0.47
R:NOUN:POSS	0	0	1	1	0	0
R:OTHER	3	15	119	0.17	0.02	0.08
R:PART	0	0	6	1	0	0
R:PREP	26	10	45	0.72	0.37	0.60
R:PRON	0	0	15	1	0	0
R:SPELL	55	26	120	0.68	0.31	0.55
R:VERB	3	3	29	0.5	0.09	0.27
R:VERB:FORM	29	12	24	0.71	0.55	0.67
R:VERB:INFL	1	0	1	1	0.5	0.83
R:VERB:SVA	18	3	6	0.86	0.75	0.83
R:VERB:TENSE	5	3	36	0.63	0.12	0.34
R:WO	0	1	15	0	0	0
U:ADJ	0	0	1	1	0	0
U:ADV	2	2	3	0.5	0.4	0.48
U:DET	2	4	15	0.33	0.12	0.24
U:NOUN	1	2	9	0.33	0.1	0.22
U:OTHER	1	19	8	0.05	0.11	0.06
U:PART	0	0	1	1	0	0
U:PREP	4	5	6	0.44	0.4	0.43
U:PRON	0	0	5	1	0	0
U:SPACE	0	0	15	1	0	0
U:VERB	2	4	7	0.33	0.22	0.30
U:VERB:FORM	0	0	2	1	0	0
U:VERB:TENSE	1	0	1	1	0.5	0.83