RLens: A Computer-aided Visualization System for Supporting Reflection on Language Learning under Distributed Tutorship

Meng Xia
School of Computing, KAIST
Daejeon, Republic of Korea
mengxia@andrew.cmu.edu

Yankun Zhao*
HKUST
Hong Kong SAR
yzhaock@connect.ust.hk

Jihyeong Hong*
School of Computing, KAIST
Daejeon, Republic of Korea
z.hyeong@kaist.ac.kr

Mehmet Hamza Erol*
School of Computing, KAIST
Daejeon, Republic of Korea
mhamzaerol@kaist.ac.kr

Taewook Kim
School of Computing, KAIST
Daejeon, Republic of Korea
taewook@u.northwestern.edu

Juho Kim
School of Computing, KAIST
Daejeon, Republic of Korea
juhokim@kaist.ac.kr

ABSTRACT

With the rise of the gig economy, online language tutoring platforms are becoming increasingly popular. These platforms provide temporary and flexible jobs for native speakers as tutors and allow language learners to have one-on-one speaking practices on demand, on which learners occasionally practice the language with different tutors. With such distributed tutorship, learners can hold flexible schedules and receive diverse feedback. However, learners face challenges in consistently tracking their learning progress because different tutors provide feedback from diverse standards and perspectives, and hardly refer to learners' previous experiences with other tutors. We present RLens, a visualization system for facilitating learners' learning progress reflection by grouping different tutors' feedback, tracking how each feedback type has been addressed across learning sessions, and visualizing the learning progress. We validate our design through a between-subjects study with 40 real-world learners. Results show that learners can successfully analyze their progress and common language issues under distributed tutorship with RLens, while most learners using the baseline interface had difficulty achieving reflection tasks. We further discuss design considerations of computer-aided systems for supporting learning under distributed tutorship.

CCS CONCEPTS

• **Human-centered computing** → **Visual analytics**; **Interactive systems and tools**.

KEYWORDS

distributed tutorship; language learning; learning progress visualization; learning reflection; tutoring system

*The authors contributed equally to this research.

L@S '22, June 1–3, 2022, New York City, NY, USA.

ACM ISBN 978-1-4503-9158-0/22/06…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

ACM Reference Format:
Meng Xia, Yankun Zhao, Jihyeong Hong, Mehmet Hamza Erol, Taewook Kim, and Juho Kim. 2022. RLens: A Computer-aided Visualization System for Supporting Reflection on Language Learning under Distributed Tutorship. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1–3, 2022, New York City, NY, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

With the rise of the gig economy, temporary and flexible jobs are prevalent toward efficient resource allocation [1, 59]. As instances of the gig economy, online language tutoring services (e.g., Cambly [1], Preply [2], and italki [3]) that provide part-time jobs for native speakers to work as tutors and enable language learners to have one-on-one lessons with tutors on demand are becoming increasingly popular [28, 65]. In contrast to the fixed instructors in the conventional language learning classroom, learners can select different tutors every learning session. The learning experience in such kind of online language tutoring services was newly identified as "distributed tutorship", in which learners distribute their learning time with different tutors, implying learning discontinuously in time with different tutors [63]. For example, in Ringle[4], a popular online English tutoring platform, 40% of 15,959 learners change to new tutors every session; 44% of learners change to new tutors while reverting to previous tutors sometimes; and only 16% of learners change to new tutors and then fix on one tutor [63].

Distributed tutorship brings learners convenience in scheduling tutoring sessions and benefits in receiving diverse feedback. However, higher distributedness is suggestively correlated with lower learning gains and poses challenges for learners to reflect on their learning progress [63]. Feedback discontinuity [11, 20] is one of the issues in distributed tutorship. In traditional learning, fixed instructors can provide continuous feedback to learners by pointing out their recurring bad habits or suggesting incremental improvements based on observing longitudinal learning practices. In distributed tutorship, it is hard for learners to receive such feedback since tutors have limited access and motivation to check a learner's session history with other tutors. In particular, language

[1] https://www.cambly.com/english?lang=en
[2] https://preply.com/
[3] https://www.italki.com/
[4] https://www.ringleplus.com/en/student/landing/home

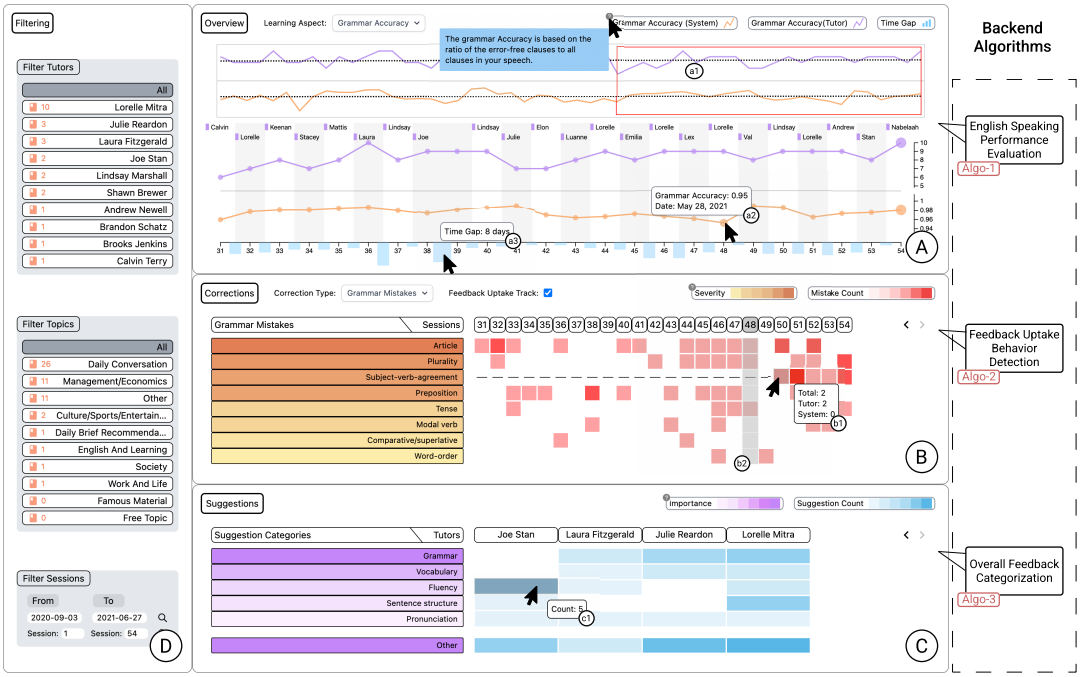

Figure 1: RLens System: Overview (A) shows the overall learning progress with both tutors' scores and system scores; Correction View (B) ranks common language issues and track learners' feedback uptake behaviors; Suggestion View (C) groups suggestions from different tutors; Transcript View maps tutor feedback to transcript (shown in Figure 4), and Filter Panel (D) for filtering by tutors, topics, and date. Algo1 - Algo4 are backend algorithms that drive data visualizations (Algo4 in Figure 4).

learning is a long process [6], where learners need to reflect on their learning practices over time to correct their common problems (e.g., tense errors, redundant filler words) [43, 51]. However, little research has investigated learners' challenges in reflecting on cumulative language learning practices under distributed tutorship and how computer techniques can assist the reflection process.

Previous studies proposed to improve the feedback quality by asking tutors to check the samples of other tutors' feedback before grading assignments [55–57]. However, this process is tedious and involves privacy issues for a tutor to listen to the learner's previous audio recordings or check other tutors' feedback in online language tutoring platforms. To reduce tutors' workload and avoid privacy concerns, we propose a computer-aided visualization system, RLens, that utilizes natural language processing (NLP) and data visualization techniques to automatically analyze different tutors' feedback and learners' speaking transcripts for assisting learners' reflection under distributed tutorship.

First, by interviewing 16 English learners who experienced distributed tutorship, we identified four major challenges (i.e., grading inconsistency, feedback discontinuity, unorganized feedback, lacking context for feedback understanding) that learners face in reflecting on their learning progress. We then implemented RLens to address these challenges. Specifically, to mitigate the challenge of different tutors having different grading standards, RLens calculates learners' speaking performance (Algo1) based on transcripts throughout the sessions and shows the computed scores in Overview (Figure 1A). To solve the feedback discontinuity, Correction View (Figure 1B) helps learners identify common language issues by ranking the language issues pointed out by different tutors based on their frequency and recency. It further detects learners' feedback uptake behaviors (i.e., learners' corrective actions to the feedback [35]) across sessions (Algo2) and demonstrates them using a heat map. In particular, we propose an algorithm to extract atomic corrections (e.g., suggested words) from tutors' feedback and track feedback uptake behaviors in each learning session using masked language modeling [7]. Suggestion View (Figure 1C) groups different tutors' suggestions using natural language inference techniques [36] (Algo3) and uses a heat map to show where to focus on. Transcript View (Figure 4) helps the learner to understand the context of the feedback by mapping tutors' feedback to the transcripts based on the sentence similarity (Algo4). A Filter Panel (Figure 1D) is integrated into RLens to filter tutoring sessions.

We evaluated RLens in a between-subjects study with 40 real-world learners by asking them to reflect on their actual learning data. Results show that learners can successfully analyze their progress and common language issues under distributed tutorship with RLens, while most learners using the baseline interface had difficulty achieving reflection tasks. Our contributions are:

- A computer-aided visualization system facilitating learners' reflection on the learning process under distributed tutorship.
- A user study showing the effectiveness of reflecting learning progress with RLens, and a set of design considerations for computer-aided learning systems under distributed tutorship.

2 RELATED WORK

This section reviews previous work on online language tutoring platforms, feedback quality control with multiple tutors, and computer-assisted systems for reflection in language learning.

2.1 Online Language Tutoring Platforms

Gig economy has gained popularity by providing temporary and flexible jobs for efficient resource allocation and it brings new practices and opportunities in learning and teaching [63]. Online language tutoring is an emerging type of language learning in gig economy [27]. This mechanism provides jobs to native speakers to work as tutors and allows language learners to have one-on-one lessons from native speakers with low time and distance barriers [28]. Most of previous studies have explored the basic characteristics of different stakeholders in online language tutoring platforms as opposed to the education mode. Some focused on the tutors' perspectives on the online tutoring setting [54, 67]. For example, a study of English tutors from three countries (Poland, Portugal, and Turkey) showed that tutors perceived online tutoring as a source of income, helping, and professional development [54]. Other studies investigated learners' demographics, goals, expectations [28], and motivations [66]. For example, the analysis of 121 application forms from a private tutoring platform in Russia revealed that the majority of learners are adults, and their motivations for having online tutoring services are work-related and examination-related [28].

Recent work investigated how the learning outcome is influenced by the new learning mode, distributed tutorship (i.e., learners occasionally practice the language with different tutors) [63]. It demonstrated that distributed tutorship is highly active and suggestively correlates with lower learning gains. However, most studies mentioned above are analytical in nature. We attempt to bridge the gap between these analyses and real-world learners through a system to address learning challenges under distributed tutorship.

2.2 Feedback Quality Control with Multiple Tutors

While reflection under distributed tutorship is not well studied, issues in feedback quality control when learning from multiple tutors have been explored before, such as grading inconsistency [57] and feedback discontinuity [19, 63]. Grading inconsistency refers to inconsistent marking standards and feedback quality amongst different tutors. Researchers introduced SPARK, a software tool for tutors to give feedback by comparing average marks and other tutors' feedback [55, 56], or an advanced version, SPARK+, to additionally support discussions among tutors to address grading inconsistency [57]. Similar issues and methods are mentioned in the peer grading in MOOCs [34].

Feedback discontinuity means that the feedback given by different tutors across learning sessions lacks coherence and does not focus on the same learning goal, as previously reported in a medical education program [19]. The authors reported that only 16% of the written feedback given by geographically distributed supervisors mentioned students' clinical performance over time continuously. The authors suggested giving more detailed feedback and having communication among tutors before giving feedback. Another work also mentioned that learners need to receive continuous feedback in the acquisition of expert performance [11].

Table 1: The session background of the 16 participants in our needfinding interviews.

Participants	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
#of sessions	77	67	33	54	17	81	137	105	134	399	35	23	76	24	126	59
# of tutors	37	53	11	33	13	41	72	57	71	70	26	19	48	18	70	54

However, previous methods are not applicable in online language tutoring to solve the grading inconsistency and feedback discontinuity issues. They are tedious and not scalable [57], while raising privacy issues when a tutor listens to the learner's previous audio recordings or check other tutors' feedback. In addition, many tutors are part-time workers and can devote only limited time to online tutoring [54]. Instead of introducing more workload to tutors and avoiding privacy issues, we propose using NLP techniques to automatically organize tutors' feedback and trace learners' learning progress based on tutors' feedback.

2.3 Computer-Assisted Systems for Reflection in Language Learning

Computer-assisted tools have been developed to facilitate learning language skills: writing, speaking, listening, and reading. A computer-supported collaborative prewriting tool was developed for enhancing young L2 learners' writing performance [30]. In speaking, recent work proposed using exaggerated audio-visual corrective feedback to help learners with pronunciation [3]. In listening, a computer-assisted shadowing trainer was developed for self-regulated foreign language listening practice [47]. Finally, in reading, a computer-supported ubiquitous learning environment was designed for vocabulary learning [42]. Some of these tools utilize NLP techniques to analyze and evaluate learners' language learning skills. They also utilize different types of visualizations to provide visual feedback to learners for reflection. However, most tools are designed for English learning at the level of a word, a sentence, or a single session instead of supporting reflection of learning progress over multiple sessions.

Visualization could effectively present learning data to promote self-reflection [17, 61]. The language learning process is a sequence of learning events in each tutoring session and previous studies visualize event sequences by placing events along a horizontal time axis, such as Lifelines [44], CloudLines [29] and TimqueSlice [68]. Inspired by these techniques, we propose a set of designs including a timeline-based heat map to show the learning progress.

3 FORMATIVE STUDY

To understand learners' practices and challenges when reflecting on their learning progress under distributed tutorship, we conducted semi-structured interviews with 16 learners on ENGLearn (a pseudonym for the anonymous review process), a popular online English tutoring platform that provides 1:1 speaking and writing sessions with native English speakers.

Participants As shown in Table 1, we selected people who have experienced distributed tutorship (i.e., more the one tutor) on the platform. In total, 16 (four males, 12 females; ages between the 20s and 50s) learners whose first language is not English participated.

Their educational background ranged from having college credits to having a master's degree. There were 12 participants who reported their self-evaluated speaking skills, with eight intermediate and four advanced. Participants received USD 20 each as compensation for participating in a 60-minute interview.

Interview Questions and Analysis Procedure The interview is semi-structured and questions include but are not limited to: (1) How often do you check your feedback given by the tutor? (2) What do you review and how long does it take? (3) How do you evaluate yourself when feedback is given by different tutors? (4) How do you calibrate your progress when you have taken multiple lessons with different tutors? Have you encountered any difficulties? (5) If we design an interface to help you review your learning history across sessions, what functions would you want to have? Two of the authors transcribed and analyzed the interview results using content analysis [37]. Key findings are summarized as follows.

Challenges of Reflection under Distributed Tutorship *C1: Grading inconsistency.* Based on the tutor scores, learners have difficulty knowing whether their performance has improved since they think scores provided by different tutors are of different standards. We found that although ENGLearn provides grading criteria, this issues still exists. All 16 participants mentioned this situation, *"...the scores depend on the tutors too much, so they can't be a clear evaluation metric. " (P13).*

C2: Feedback discontinuity. Learners are uncertain whether they applied what they had learned and unaware of their common errors, given that detailed corrections provided by different tutors are not tracked and mentioned in subsequent learning sessions. Fifteen out of the 16 participants mentioned that they preferred having sessions with fixed tutors for receiving continuous feedback. For example, *"with the tutor I have seen, I can have a conversation that follows the previous conversation, and they know more about my common mistakes and habits, so they know more about in which aspect I have improved, so I wish to meet same tutors again." (P5)*

C3: Unorganized feedback. Learners cannot easily organize the feedback to know which aspect to focus because feedback provided by different tutors spans different perspectives. Some participants (5 out 16) mentioned how feedback includes diverse perspectives, and another two said that tutors' feedback is too much to organize. *"However, the feedback is written by the tutor, so it's very subjective. Some tutors select one or two things they think are important and write it in the feedback; some tutors divide the feedback into five metrics such as grammar and vocabulary." (P6)*

C4: Lacking context for feedback understanding. Since the tutor feedback is given using a separate doc, learners cannot easily interpret detailed feedback without context information. Some participants (6 out of 16) mentioned that they need to refer to the script and audio to understand the tutor's feedback. However, it is hard to find the exact position in the transcript .

Design Requirements Based on the challenges and existing literature on continuous feedback [11, 63], we extracted learners' needs and derived design requirements for a computer-aided system for learning progress reflection.

R1: Provide a data-driven assessment on learning performance along with tutors' scores over time. To address the grading inconsistency (C1), we propose to provide a computed score of learners' performance based on their learning data (e.g., audio-to-text transcription) in addition to tutors' scores. This notion was mentioned by P9, P10, P12, P13.

R2: Identify common language issues and track feedback uptake behavior. All participants desired to track their common issues and improvements. To address feedback discontinuity (C2), since it is difficult to have a fixed tutor in the online tutoring system [63], we propose to apply NLP techniques to track and rank their language issues, rank them to find common ones, and detect learners' feedback uptake behaviors. For example, the system can track how the learner apply the suggested vocabulary after the tutor's correction.

R3: Organize tutor feedback automatically into different categories. To address unorganized feedback (C3), we propose to group feedback into different categories automatically to highlight the focus area and common suggestions based on their frequency.

R4: Map the tutor feedback to transcripts. To facilitate learners to understand the correction within its usage context (C4), for each correction pointed by the tutor, we propose highlights and edits of the correction in the transcript.

R5: Provide intuitive visualizations to present the learning progress (C1-C4). Visualization could effectively present learning data to promote self-reflection [17, 61]. In addition, five participants wanted a visual representation of their learning progress.

4 SYSTEM DESIGN

To meet these requirements, we designed an interactive visualization-based system powered by a set of data-driven algorithms. We first introduce the algorithms and then the visualizations.

4.1 Algorithms for Evaluating Language Learning Progress

We propose four algorithms (Algo1-Algo4) to satisfy the four requirements (R1-R4) and drive the visual interface (Figure 1, Figure 4 based on existing research and discussions with three experienced tutors (with an average of three years of experience) from ENGLearn. These algorithms take tutors' written feedback and learners' speech transcripts as input sources. For each tutoring session on ENGLearn, learners receive an audio recording, audio-to-text transcription, scores on English speaking performance given by the tutor, and the tutor's written feedback. The tutor's written feedback usually contains the overall feedback and in-depth corrections.

Algo1: English Speaking Performance Evaluation To evaluate the English speaking performance computationally (R1), we adopt the metrics proposed in previous English education research [9, 10, 21, 49]: complexity, accuracy, and fluency. Since there is no fixed and optimized measurement for each metric, we select commonly used measurements from the literature. In particular, we calculate *Vocabulary Complexity* based on the measure of textual lexical diversity (MTLD) [40], which calculates the average length of sequential words a speaker can produce that keep the type-token ratio (TTR) higher than x. x is set to 0.72 by referring to previous work [14, 41]. TTR is the ratio of the number of different words (i.e., types) to the total number of words (i.e., tokens) [4]. *Grammar Accuracy* is calculated as the ratio of error-free C-Units to the total number of C-Units, where C-Unit is defined as the minimal communication unit (e.g., "Yes.") [15, 22]. In terms of *Fluency*, we calculate the Mean

Length of Run [23], the average number of syllables per utterance without any pause, where the threshold for pause identification is set to 250ms in accordance with previous cases [26, 45].

Algo2: Feedback Uptake Behavior Detection Feedback uptake behavior refers to learners' corrective actions according to the feedback [35]. We focus on the feedback uptake behavior for corrective feedback (e.g., two apple -> two apples) because uptake behavior for high-level feedback (e.g., watch more English movies) is difficult to track through speech data without additional resources. Tutors' written corrective feedback on ENGLearn contains the original sentence spoken by the learner, the corrected sentence by the tutor, and the correction type, namely grammar, vocabulary, and fluency. To detect feedback uptake behavior (R2), we propose the following pipeline: (1) detecting which language issue is pointed out by the tutor (e.g., a subject-verb disagreement in grammar); (2) detecting whether learners still have this issue or corrected it in their subsequent speaking sessions. We introduce how we detect feedback uptake behaviors for each correction type as follows.

Grammar. For each grammar error pointed out by the tutor, we consider grammar errors of the same type in learners' subsequent sessions in detecting the feedback uptake behavior. To this end, we first detect which type of grammar error is pointed out by the tutor by comparing the original sentence and the corrected sentence using an open-source grammar checker (https://github.com/languagetool-org/languagetool). Second, we count the number of that type of error in subsequent lessons using the grammar checker. In particular, we fine-tuned the grammar checker to suit the characteristics of spoken English and tolerate auto speech recognition errors by ignoring the grammar errors caused by capitalization, punctuation, homophones, repeats, pauses, false starts, corrections, interjections, and stutters [12].

Vocabulary. For vocabulary corrections, we consider two types of feedback uptake behaviors. The first type is forgetting to apply the suggested expression (vocabulary or phrase), in which the learner used the original expression when the suggested one can be used. The second type is applying the suggested expression correctly. For example, for a pair of original and suggested expression: "request" and "require", the first type of uptake behavior will be detected if the learner said "The job requests at least two years of related experience." And the second type of uptake behavior is detected if there is a sentence like "This document requires your signature." in the speaking transcripts.

To achieve this goal, we first extract the pair of original and suggested expressions given the original and suggested sentences. The detailed steps are as follows. (1) Find out the word differences between two sentences using ERRANT [2, 13] as the fundamental algorithm. For example, suppose that the original sentence is "She always tries to think positively." and the suggested sentence is "She is always so optimistic." Then the difference found is changing "tries to think positively" to "is so optimistic". (2) Enumerate two lists of possible expressions from different parts in the original and suggested sentence respectively, where each expression must contain at least one of a noun, verb, adjective, or adverb. For example, we have list 1 from the original sentence: {"tries", "think", "positively", "tries to", "to think", "think positively", "tries to think", "to think positively", "tries to think positively"}, and list 2 from the suggested sentence: {"optimistic", "so optimistic", "is so optimistic"}. (3) Pick one expression from each list to form a pair and check the expressions in which pair has the most similar meaning in context using a sentence transformer (MPNet-base-v2 model [46, 50]). After matching, "think positively" and "optimistic" is the pair that turns out to be the most similar, hence is extracted.

Second, we detect the feedback uptake behavior (i.e., forgetting to apply the suggested expression or applying the suggested expression correctly) in the subsequent lessons. Given that the use of vocabulary is highly dependent on the contextual semantics, we utilize the ALBERT-xxlarge-v2 model [31, 58] pretrained using the masked language modeling (MLM) objective [7], which allows the model to fuse both of the left and right context of a masked word. The model thus can take contextual semantics into consideration. Specifically, for a pair of original and suggested words, once we find the original word (or its variants/derivatives) in transcripts, we first mask the word, then use the model to predict the masked token in the sentence. If the suggested word (or its variants/derivatives) is one of the predicted words, we identify this occurrence of the original word as a recurring error. We apply a similar idea for detecting correct applications of a suggested word. We mask the suggested word found and see if the original word is in the predicted word candidates list. Our uptake behavior detection algorithm is evaluated on a random sample (10%, 126 sentences) from the corpus of written feedback in 20 learners' data. This evaluation set is labeled by three experienced tutors (three years of experience on average). Our detection algorithm shows the precision of 91.49% and the recall of 92.47% for detecting vocabulary uptake behaviors.

Fluency. Since there is hardly consensus on fluency evaluation [48], we chose the frequency of filler words to show the fluency feedback uptake behavior, which is pointed out as one of the most common issues by the three experienced tutors. We adopt five common filler words from previous work: "uh", "um", "like", "you know", and "I mean" [32] and detect them in learners' transcripts. An occurrence of "uh", "um", or "like" is counted as a filler word when its part of speech is an interjection. "You know" or "I mean" is counted as a filler word when "know" or "mean" only has one left child in the dependency tree.

Algo3: Overall Feedback Categorization We categorize each sentence of the overall feedback from different tutors to help learners easily know where to focus based on the frequency on different categories (R3). To determine the categories to classify, we first ran topic modeling using BERTopic [18] on around 30,000 feedback sentences, and we identified 49 topics as our candidate categories. Then, with the help of the tutors, we manually selected seven categories: "grammar", "vocabulary", "pronunciation", "fluency", "sentence structure", "compliment", and "greeting", considering how useful and actionable the feedback is as the main criteria. We further combine "compliment" and "greeting" to "other" category. Finally, we use an ensemble of a pretrained RoBERTa-large-MNLI model [33, 58], a pretrained XLM-RoBERTa-large-XNLI model [5, 58], and a pretrained XLM-RoBERTa-large-XNLI-ANLI model [5, 58] to perform sentence classification for each feedback sentence. The classification accuracy of our ensemble model on a labeled sampled (575 sentences, 10% of all the data used in the user study in section 5) tutors' feedback corpus is 74.09%, where the corpus is labeled by three authors.

Algo4: Feedback-Transcript Mapping To match a sentence pointed by the tutor to a sentence in the transcript (R4), we utilize the similarity detection function from an NLP library spaCy (https://spacy.io) .

4.2 Visualization System

Leveraging the algorithmic pipeline introduced in Section 4.1, we designed visualizations of RLens (Figure 1) to present the learning progress (R5). It contains four views and one filter panel.

Overview Overview (Figure 1A) is designed to help learners understand their overall learning progress by presenting both tutors' scores and computed scores calculated by our algorithms (Sec. 4.1). We show both scores because tutors' scores are meaningful to some degree but may be not consistent, and computed scores can be used for a reference as they are evaluated with same rules. Since the two scores use different schemes and require different y-axes, we juxtapose two line charts for comparison instead of superimposing them [16]. The purple line represents the tutors' scores, and the yellow line represents the computed scores. Two line charts share the x-axis to show the session number. Learners can track the tutor change to evaluate the tutors or scores associated with the classes of some tutors, or the distributiveness of their tutorship by observing the change of the white and grey background of the line chart. We also use a window (Figure 1A_a1) with regression lines to provide learners a quick overall trend of their scores in all sessions. The bar chart (Figure 1A_a3) between two sessions shows the time gap (days).

Correction View Correction View (Figure 1B) ranks the common language issues and demonstrates the feedback uptake behavior for learners to prioritize what aspects of their language learning should be improved further. As introduced in Algo2 in Section 4.1, feedback uptake behaviors are analyzed from three aspects: grammar errors, vocabulary advice, and fluency suggestions. For each aspect, we use an independent tab to display the corresponding information. For example, as shown in the **Grammar Mistakes** tab in Figure 1B, grammar errors are grouped into different categories and ranked based on severity. The **severity** is calculated based on the prediction of the error count in the next session using the regression line to simulate the trend of the error count by considering time and frequency (i.e., the more recent and more frequent error type is ranked higher). To help learners to perceive the overall trend of the grammars errors across sessions, we visualize the frequency of errors using a heat map, a widely used, effective, and simple visual technique in showing the frequency of learning activities [25, 38, 39, 60]. Tiles with a higher frequency are encoded with a darker shade of red. We use a short line "-" in the tile to indicate that no tutor has pointed out the error yet. In addition, if a learner wants to check the issues pointed out by the tutor without the feedback uptake detected by the system, they can disable the feedback uptake toggle beside the drop-down menu.

The **Vocabulary Advice** tab, shown in Figure 2A, lists the vocabulary advice based on their severity (estimation of the times the user forgot to apply the suggested expressions in the next session). The format of vocabulary advice is "original word - suggested word". Different from the grammar errors, we use a two-color heat map for each vocabulary advice to show how learners address that feedback in one learning session. The shade of green of the top-half

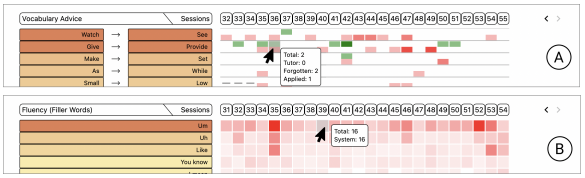

Figure 2: Corrections View: Vocabulary Advice Tab (A) and Fluency (Filler Words) Tab (B)

Figure 3: Suggestions View (when clicked)

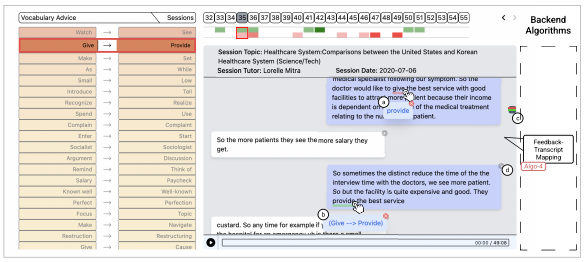

Figure 4: Transcript View with Algo4 the driven algorithm

of a tile indicates how frequently that learner applied the suggested expression correctly; the shade of red of the bottom-half of a cell indicates that how frequently that learner used the original expression in the learning context when the suggested expression can be applied instead of the original one. For example, in Figure 2A, the learner applied vocabulary advice "provide" correctly once and used the original words without applying the suggested word twice in session 36. In the **Fluency (Filler words)** tab shown in Figure 2B, the system lists the common filler words (i.e., uh, um, I mean, like, you know) and ranks them based on their severity. For each filler word, we use a heat map to show its occurrence in each session, similar to the Grammar Mistake tab.

Suggestion View Suggestion View (Figure 1C) groups all the tutors' suggestions into six categories as introduced in Section 4.1. To help learners prioritize the suggestions, the system ranks all the categories based on the number of sentences mentioned by tutors in each category. Moreover, we use a heat map for each category to show each tutor's contribution (i.e., number of sentences) to that category. The darker the blue, the more sentences are mentioned by a tutor for that category. When the learner clicks a tile, the sentences belonging to the category and tutor are shown in a pop-up (Figure 3).

Transcript View Transcript View is designed to understand the learning context by mapping the correction and the feedback uptake behavior to the transcripts. For example, as shown in Figure 4, when the learner clicks vocabulary advice in Correction View (e.g., Give -> Provide) and further clicks session 35, Transcript View pops up to show the feedback uptake locations in the session's

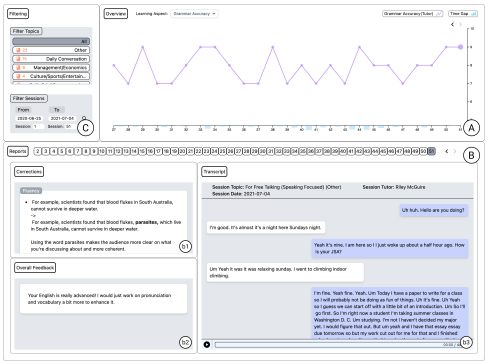

Figure 5: Baseline System:Overview (A), Report View (B), and Filter Panel (C)

transcript. In particular, for an immediate localization of the errors, it uses a red and green bar (Figure 4_c) over the scrollbar to show where the feedback uptake is in the transcript. It also highlights places in the transcript where a learner forgot to apply the suggested vocabulary using red underlines (Figure 4_a) and applied the suggested vocabulary correctly using green underlines (Figure 4_b). Learners can check the corrected expression by clicking the underlined error(Figure 4_a). It works similarly to Grammar Mistakes and Fluency tabs. Meanwhile, learners can click the play button (Figure 4_d) to listen to the audio of each turn in the conversation to recall the learning context.

Filter Panel RLens provides a Filter panel (Figure 1D) to provide learners with a focused view of their learning progress from different perspectives (e.g., tutors, topics, sessions, dates, etc.).

5 EVALUATION

We evaluated the usefulness and efficacy of RLens in assisting learning progress reflection under distributed tutorship.

5.1 Study Design

We conducted a between-subjects study on a Baseline system and RLens. Since there is no support for distributed tutorship in previous systems, we constructed a Baseline system (Figure 5) to simulate the learners' dashboard currently provided by the ENGLearn platform and other language tutoring platforms [27, 63]: the tutor scores (Figure 5A), the tutor feedback (Figure 5B_b1, b2), and the speaking transcript (Figure 5B_b3). We did not use ENGLearn directly because it contains additional information (e.g., tutors' pictures, advertisements) other than the experimental variables that might confound our results. We applied the same UI elements (e.g., layout, fonts) to both systems and tried to minimize visual and usability differences between them. We acknowledge that Baseline is an overly simple system, and therefore our goal is not to see if RLens beats the Baseline but rather to understand and analyze how people use RLens in-depth, in comparison with the Baseline.

Tasks We ask people to look at their progress data and try to make sense of it for reflection. In particular, we derived seven tasks from the challenges found in our needfinding stage and previous research on language learning reflection [64]. **T1:** Please describe your overall learning progress.**T2:** Please identify your common language issues in the English learning process. **T3:** Please describe whether you have corrected your common language issues in the learning process. **T4:** Please describe the common aspects in tutors' overall feedback. **T5:** Please describe how you check the transcript using the system for learning. **T6:** Please describe the reasons for ups and downs in scores showing in Overview. **T7:** Please describe how you will use this system in learning reflection if it is deployed. To simulate real reflection scenarios, we required participants to reflect on their own learning data by loading their session data from ENGLearn to the system upon participants' consent.

Measures We adopted a four-layer taxonomy to evaluate our system based on Weibelzahl's work [53], where the authors proposed the evaluation pipeline of interactive systems. We systematically evaluate the effectiveness, informativeness, usability, and intuitiveness of RLens in assisting learners with reflection under distributed tutorship. Moreover, since people's trust and perceived accuracy is an important metric in an AI-infused system [8], we also evaluate learners' trust in information provided in RLens. The questionnaire can be seen in Table 2.

Participants We recruited learners from ENGLearn by posting an advertisement on the platform's website. We eliminated learners with fewer than 25 sessions to guarantee sufficient experience on distributed tutorship. Furthermore, we paired learners based on the number of sessions they had and their session/tutor ratio (i.e., # of sessions/ # of tutors) to guarantee that participants in both Baseline and RLens groups have a similar learning experience and distributed tutorship experience. Finally, we had 40 (12 males, 28 females) participants, with 20 in each group (B1-B20 in Baseline and A1-A20 in RLens). All participants' first language was not English. Baseline group (7 male, 13 female) had a mean age of 33.5 (min 25, max 53), a mean number of sessions of 67.05 (min 31, max 145), a mean session-tutor ratio of 1.78 (min 1.05, max 3.22), and the distribution self-reported English speaking proficiency is low (3), good (6) and intermediate (11). RLens group (5 male, 15 female) had a mean age of 35.25 (min 27, max 52), a mean number of sessions of 65.85 (min 27, max 185), a mean session-tutor ratio of 1.91 (min 1.23, max 4.56), and the distribution self-reported English speaking proficiency is low (2), good (10) and intermediate (8). The recruitment and user study procedures were approved by the Institutional Review Board (IRB) at our university, and each participant received approximately USD 38 as compensation for participating in a 90-minute study session.

Procedures The user study was conducted remotely through Zoom. It contained five steps and lasted around 90 minutes: First, we introduced the background of the study, and participants read and signed the consent form. We then demoed and introduced the interface. Participants were asked to explore the system and complete eight learning reflection tasks using the think-aloud strategy for about 40 minutes. Upon task completion, participants completed a questionnaire with 7-point Likert questions derived from existing literature [62]. Lastly, we asked debriefing questions about their opinions on the most or least helpful features and suggestions for the system.

Hypotheses Based on previous learning dashboard evaluation [52, 62], we present the following hypotheses:

H1: RLens is more effective in helping learning progress reflection under distributed tutorship than Baseline. Specifically, RLens is more helpful for learners to clearly understand the learning progress

Table 2: A questionnaire was designed to cover five aspects: effectiveness (Q1-Q7), informativeness (Q8-Q9), usability (Q10-Q11), visualization & interaction (Q12-17), and trust (Q18-Q21). All are 7-point Likert scale questions. Q14-Q21 are only applicable to RLens.

Questions
Q1: The scores in the system help me to have a clear understanding of whether I have improved under distributed tutorship.
Q2: The system helps me in being aware of my common language issues under distributed tutorship.
Q3: The system helps me to know whether I have corrected my common errors under distributed tutorship.
Q4: The system helps me to organize different tutors' suggestions for future guidance.
Q5: The system helps me to know the learning context of tutors' feedback under distributed tutorship.
Q6: The system helps me to analyze and reflect on my learning progress under distributed tutorship.
Q7: I would like to recommend this system to others if they learn from different tutors.
Q8: The information needed is easy to access to reflect on my learning progress
Q9: The information is sufficient to reflect my learning progress under distributed tutorship.
Q10: It is easy to learn the system.
Q11: It is easy to use the system.
Q12: Overall, the visualization designs in the system are intuitive.
Q13: Overall, the interactions in the system are intuitive.
Q14: The visualization in Overview is intuitive.
Q15: The visualization in the Correction View is intuitive.
Q16: The visualization in the Suggestion View is intuitive.
Q17: The visualization in Script View is intuitive.
Q18: I trust the computed score provided by the system.
Q19: I trust the feedback uptake behavior detected by the system.
Q20: I trust the grouping of suggestions by the system.
Q21: I trust the mappings for the corrections to the script.

(*H1a*), be aware of common errors (*H1b*) and correction behaviors (*H1c*), organize tutors' suggestions (*H1d*), understand learning context (*H1e*), and analyze learning progress (*H1f*). Therefore, learners are more willing to recommend RLens to others (*H1g*).

H2: The information for learning progress reflection in RLens is more accessible (*H2a*) and sufficient (*H2b*) than Baseline for learning progress reflection under distributed tutorship.

5.2 Results and Analysis

Two authors analyzed users' interactions in the tasks, verbal reasons for the ratings, and post-study interviews. We performed the Mann-Whitney U (rank) test on the questionnaire items (Q1-Q9) to test statistically significant differences between the conditions. Overall, participants reported RLens to be significantly more effective and informative to perform the learning progress reflection tasks under distributed tutorship than Baseline. Besides, it received positive ratings towards usability as well as visual & interaction design, though lower than Baseline while no significance is shown. In addition, people generally trusted the algorithm outputs with highly positive ratings.

H1. Effectiveness Overall, the participants thought RLens is significantly more effective in helping with learning progress reflection under distributed tutorship than Baseline.

Clear understanding of learning progress. Participants reported that scores (including system scores and tutor scores) presented in RLens ($Mean = 5.95, SD = 1.504$) are more helpful for them to clearly understand their learning progress than only tutors' scores in Baseline ($Mean = 4.4, SD = 1.536$). Significance has been found in the Mann-Whitney U test ($U = 87.0, p < 0.001$, **H1a supported**). Based on participants' verbal feedback in T1 (i.e., describe the overall learning progress), 15 out of the 20 participants in Baseline and 16 out of the 20 participants in RLens thought having only the tutor scores as a source does not feel like to be a representative for their actual learning progress. In RLens, most participants preferred to have both tutors' scores and computed scores. 14 out of the 20 learners thought that computed scores in RLens provided a more objective evaluation since they fluctuated less in some conditions during the tutor change. In addition, some participants used the two scores in a complementary manner to understand their learning progress. *"I will trust tutor's score for fluency because I think it should be graded by a person, not a system. For grammar, I will trust the system score because it is objectively being right or wrong. (A18)"*

Awareness of common language issues. Participants reported they were significantly more aware of common language issues using RLens ($Mean = 6.45, SD = 0.945$) than Baseline ($Mean = 3.65, SD = 1.531$). The Mann-Whitney U test further reveals the significance ($U = 25.0, p < 0.001$, **H1b supported**). Notably, participants' answer to T2 (i.e., identifying common language issues) was extremely different between Baseline and RLens. Participants in Baseline could hardly give the answer by reporting that they are not sure about their common errors or answered the questions based on their memory in an uncertain manner. Many of them did not check the Baseline system since they reported that it is time-consuming to find common errors by clicking each session. However, in RLens, all participants answered the question by checking different tabs in the correction view first and then confirmed their answers with their memory. Their answers pointed the specific errors, e.g., tense errors in grammar. Some participants also realized the common errors that they overlooked before, e.g., *"I thought I use 'like' or 'I mean' the most, but there are lots of 'uh' and 'you know,' which I did not recognize before. (A14)"*

Correction of common language issues. RLens ($Mean = 5.8, SD = 1.542$) was more helpful for participants to know whether they have corrected their common errors than Baseline ($Mean = 3.3, SD = 1.261$), with Mann-Whitney U test ($U = 39.5, p < 0.001$, **H1c supported**). According to T3 (i.e., describe whether common language issues have been corrected), in Baseline, 18 out of 20 participants failed to answer how they have proceeded with their common errors, and they were not sure whether they had applied tutors' corrections in their learning progress. Only two participants checked Correction View one by one and answered the question based on whether a specific error was mentioned by the tutor again. However, in RLens, all the participants used the heat map to answer their progress and whether they corrected their common errors. For example, A1 noted that *"Filler words become lighter in the recent session, and I can see improvements in fluency. "*. RLens received

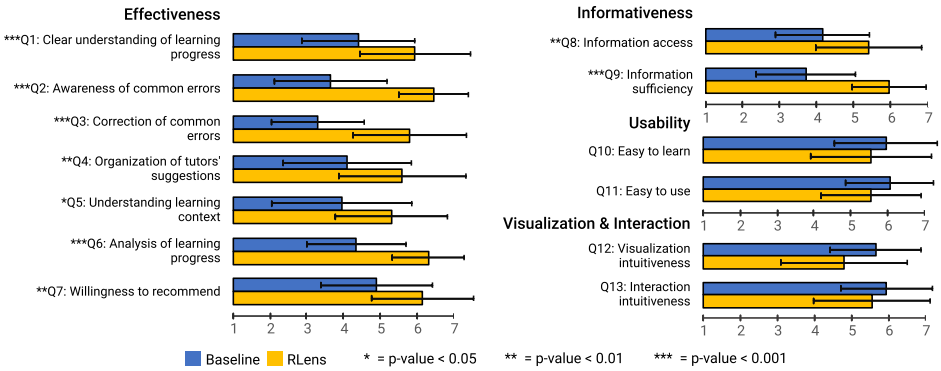

Figure 6: Means and standard errors of Baseline and RLens on effectiveness, informativeness, usability, and visualization & interactions on a 7-point Likert scale ($* : p < .05, ** : p < .01, * * * : p < .001$).

positive feedback regarding the heat map that shows progress, as A8 said, *"For me, it is really useful and efficient. (the color)."*.

Organization of tutors' suggestions. Participants reported that RLens ($Mean = 5.6, SD = 1.729$) is more helpful for them to organize tutors' feedback for future guidance on English learning than Baseline (($Mean = 4.1, SD = 1.744$), with Mann-Whitney U test showing significance ($U = 102.5, p < 0.01$, **H1d supported**). For T4 (describe common aspects in tutors' feedback), many participants (15 out of 20) in Baseline had difficulties answering the common suggestion given by different tutors. In RLens, 19 out of 20 participants answered the question with Suggestion View. One participant still felt the workload is heavy to check all sentences from one category: *"It is nice to have overall feedback gathered all together, but it is hard to read it one by one. (A11)"*. In addition, some participants mentioned there might be conflicts in tutors' suggestions, which they wished to spot in RLens.

Understanding learning context. The mapping of tutor feedback to the transcript in RLens ($Mean = 5.3, SD = 1.525$) helps learners better understand their learning context than Baseline ($Mean = 3.95, SD = 1.905$). The Mann-Whitney U test shows the significance ($U = 117.5, p = 0.012 < 0.05$, **H1e supported**). Participants in Baseline found T5 (i.e., how to use the transcript) is hard. B6 said that *"If I am looking at an article issue, I need to find it in the script. Then there is the cognitive effort required. "* In RLens, participants all tried the mapping function by clicking the red cells and checking feedback context in the transcript view. Two participants (A12, A20) mentioned that Transcript View could be further simplified to show only sentences with errors.

Analysis of learning progress. Participants found RLens ($Mean = 6.3, SD = 0.979$) significantly better in supporting their analysis of the learning progress under distributed tutorship than Baseline ($Mean = 4.35, SD = 1.348$). The Mann-Whitney U test confirmed the significance ($U = 43.5, p < 0.001$, **H1f supported**). In T6 (i.e., describe the reasons for learning improvements and decreases), three learners in Baseline gave up analyzing their learning progress and reported they did not have enough information. For other learners in Baseline, the major pattern for analysis was that they referred to their memory to explain why they had higher scores or lower scores in some classes. Participants in RLens exhibited a variety of strategies to analyze their learning progress. For example, four participants used Correction View to analyze their learning progress. Seven participants used both Overview and Correction view to reason about their progress. They checked Overview for the overall trend and then referred to Correction View to find the common errors. Also, A18 utilized all views to achieve the analysis. This participant first checked the lowest scores in Overview and searched the Correction View for all the errors in the corresponding session, and checked the suggestions.

Overall, participants were more willing to recommend RLens ($Mean = 6.15, SD = 1.387$) to other learners for learning under distributed tutorship than Baseline ($Mean = 4.9, SD = 1.518$). Significance is found in the Mann-Whitney U test ($U = 94.5, p = 0.002 < 0.01$, **H1g supported**).

H2. Informativeness Compared with Baseline, RLens received significantly higher ratings in informativeness, including information access and sufficiency. *Information accessibility.* Participants found it was easier to access the information needed for learning progress reflection in RLens ($Mean = 5.4, SD = 1.429$) than Baseline ($Mean = 4.15, SD = 1.268$). Significance has been found in the Mann-Whitney U test ($U = 90.0, p = 0.001 < 0.01$, **H2a supported**). *Information sufficiency.* The information provided by RLens ($Mean = 5.95, SD = 0.999$) in learning reflection under distributed tutorship was perceived as more sufficient than Baseline ($Mean = 3.7, SD = 1.342$). We also observe a significant difference with the Mann-Whitney U test ($U = 43.5, p < 0.001$, **H2b supported**). According to participants' reasons provided in Q8 and Q9, we found that whether having the access to tutors' information to be one of the primary reasons for the rating difference.

Usability Overall, Participants thought RLens was easy to learn ($Mean = 5.55, SD = 1.638$) and easy to use ($Mean = 5.55, SD = 1.356$). Baseline received a relatively higher score than RLens in ease of learning ($Mean = 5.95, SD = 1.395$) and ease of use ($Mean = 6.05, SD = 1.191$) while no significance was found.

Visualization & Interaction RLens received $Mean = 4.8, SD = 1.704$ for overall visualization intuitiveness, and Baseline received $Mean = 5.65, SD = 1.226$. Further analysis on the intuitiveness scores for each view, as shown in Figure 7 (Overview: $Mean = 5.95, SD = 1.638$; Correction View: $Mean = 6, SD = 1.487$; Suggestion View: $Mean = 5.1, SD = 1.804$; Transcript View: $Mean = 5.8, SD = 1.576$) shows that the scores for individual views are

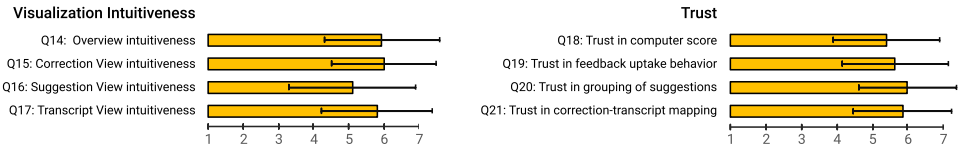

Figure 7: Means and standard errors of RLens on visualization and trust on a 7-point Likert scale

relatively high, and participants in RLens reported that each view is easy to understand. However, when individual views are combined as the whole system, the large amount of information presented might have reduced the overall intuitiveness. A12 said that the whole system is complex for his age (53). Participants thought that the interactions in both RLens ($Mean = 5.55, SD = 1.572$) and Baseline ($Mean = 5.95, SD = 1.234$) were intuitive.

Trust: Overall, participants reported that they trust the information calculated by the algorithms. As shown in Figure 7, their ratings for trust in computed score in the Overview is $Mean = 5.4, SD = 1.501$; feedback uptake behavior in Correction View is $Mean = 5.65, SD = 1.496$; groups of tutors' feedback in Suggestion View is $Mean = 6, SD = 1.376$; and mappings for the correction to the transcript is $Mean = 5.85, SD = 1.387$. Most participants held a positive attitude towards the computer-generated information because they did not spot errors during the exploration process. Some spotted system errors in the user study (e.g., "focus -> topic" was extracted from "His answers got out of the focus." -> "His answers was off the topic."), and they rated the view where they spotted the error with a low score. However, they said this feeling did not affect their trust in other views/information in the post interviews.

6 DISCUSSION

This section discusses design considerations we learned from the study and limitations that we can address in future work.

6.1 Design Considerations

DC1: Organize information from the dimension of tutor. Our user study showed that learners want to access learning data with particular tutors when learning with multiple tutors. For example, some participants compared the grading standard of different tutors to understand their actual learning performance because they were worried some tutors are excessively generous or harsh. In addition, many participants also evaluated their learning progress with particular tutors to decide which tutor they would like to select for future learning sessions.

A computer-aided learning system for distributed tutorship should provide information organized from the dimension of the tutors. When learning under distributed tutorship, the diversity of tutors becomes an important factor for learners to evaluate their learning progress and make future decisions on learning.

DC2: Utilize computer as a reference tutor. Distributed tutorship can provide flexible learning schedules and diversified feedback and language styles. However, tutors in distributed tutorship face challenges in maintaining the same grading standards, detecting detailed language issues, and giving continuous feedback by tracking learners' progress. In the user study, some participants only trusted computed scores, while most participants checked and compared both human tutors' scores and computed scores. Moreover, most participants used RLens as a reference tutor to find their common language issues. Our proposed method received favorable responses by combining high-quality feedback from different human tutors and the consistent tracking ability of the system to provide a continuous learning experience.

We suggest that future learning tools designed for distributed tutorship utilize the computer as a reference tutor to provide data-driven assessment and give continuous feedback, complementing the role of human tutors.

DC3: Provide access to all information but surface actionable information. Participants in the user study exhibited various strategies to utilize different views to analyze their learning progress, and they appreciated the access to all the learning information from different levels. They particularly liked the severe language issues highlighted in the Correction View. As pointed out by previous research, learning dashboards should provide learners with actionable suggestions [24, 52]. In particular, when learning with multiple tutors, different tutors point out diverse issues and give various suggestions, and it is challenging for learners to figure out the priority and distill actionable insights. A few participants mentioned that the excessive amount of information presented in the system made it difficult to prioritize, and the current grouping algorithm of tutors' feedback in Suggestion View could be further refined to solve the conflicts of different tutors' feedback and make the suggestions more actionable.

Future learning systems for distributed tutorship should surface actionable information by distilling common feedback, resolving conflicts through algorithms, and highlighting them in the interface.

6.2 Generalization

Our pipeline and designs can be generalized to other platforms that might have distributed tutorship dynamics and feedback culture—for example, online skill practice (e.g., writing) using P2P skill-sharing communities (e.g., Clascity [5]), freelance markets, (e.g., Upwork [6]). These communities/platforms are developed to help individuals freely share their skills and receive feedback from one another; thus, users on these platforms also potentially experience distributed tutorship. The algorithms and visualization designs of RLens can be generalized to assist users in keeping track of learning progress and organizing diverse feedback in broader domains.

6.3 Limitations

Our work has several limitations that have to be considered. First, the accuracy of the feedback uptake behavior algorithms and the feedback categorization algorithm can be further improved, and

[5] https://clascity.com/
[6] https://www.upwork.com/

they are only tested on a small sample we labeled due to the lack of a ground truth dataset. Second, the multi-view dashboard of the system might impose a steep learning curve and information overload. However, as being the first system in the space to address distributed tutorship, the current interface is not meant to be a complete solution on its own but rather a prototype built to investigate how reflection on progress can be supported in various data-driven ways. Different parts of our work could be selectively applied, simplified, etc. to different platforms and learning contexts since each view/algorithm can be used in a modular manner. Third, we have not introduced pronunciation metrics in RLens because the target user group has a relatively advanced speaking level and fewer pronunciation issues, while this metric is also important to be added. In addition, more language elements like grammar complexity and dialogue dynamics can be further considered.

7 CONCLUSION AND FUTURE WORK

In this work, we proposed RLens, a computer-aided visualization system that allows learners to analyze and reflect their language learning progress under distributed tutorship. It utilizes NLP and information visualization techniques to empower learners to understand their learning progress and recognize common errors. We conducted a between-subjects study with 40 real-world learners. Results show that learners can successfully analyze their progress and common language issues under distributed tutorship with RLens. We further discuss design considerations of computer-aided learning systems with multiple tutors. In the future, we plan to deploy RLens in real-world online language tutoring platforms to test its usability and algorithmic accuracy in the long term.

REFERENCES

[1] Ali Alkhatib, Justin Cranshaw, and Andrés Monroy-Hernandez. 2018. Laying the Groundwork for a Worker-Centric Peer Economy. *arXiv preprint arXiv:1807.08189* (2018).

[2] 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)*. Association for Computational Linguistics, Vancouver, Canada, 793–805. https://doi.org/10.18653/v1/P17-1074

[3] Yaohua Bu, Tianyi Ma, Weijun Li, Hang Zhou, Jia Jia, Shengqi Chen, Kaiyuan Xu, Dachuan Shi, Haozhe Wu, Zhihan Yang, et al. 2021. PTeacher: a Computer-Aided Personalized Pronunciation Training System with Exaggerated Audio-Visual Corrective Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[4] A.M. Colman. 2015. *A Dictionary of Psychology*. Oxford University Press.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116* (2019).

[6] Jill G De Villiers, Jill De Villiers, Peter A De Villiers, and Peter A DeVilliers. 1978. *Language acquisition*. Harvard University Press.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186.

[8] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[9] Rod Ellis et al. 2003. *Task-based language learning and teaching*. Oxford University Press.

[10] R. Ellis and Gary P. Barkhuizen. 2005. *Analysing Learner Language*. Oxford University Press.

[11] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological review* 100, 3 (1993), 363.

[12] Hawa Fadhila. 2013. *Errors In Speaking English Made By Students Ofenglish Department Of Muhammadiyah University Of Surakarta*. Ph.D. Dissertation. Universitas Muhammadiyah Surakarta.

[13] Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 825–835.

[14] Gerasimos Fergadiotis, Heather Harris Wright, and Samuel B Green. 2015. Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research* 58, 3 (2015), 840–852.

[15] Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied linguistics* 21, 3 (2000), 354–375.

[16] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309.

[17] Sten Govaerts, Katrien Verbert, Erik Duval, and Abelardo Pardo. 2012. The student activity meter for awareness and self-reflection. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 869–884.

[18] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[19] Pam Harvey, Natalie Radomski, and Dennis O'Connor. 2013. Written feedback and continuity of learning in a geographically distributed medical education program. *Medical teacher* 35, 12 (2013), 1009–1013.

[20] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[21] Alex Housen and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics* 30, 4 (2009), 461–473.

[22] Diana L Hughes, LaRae McGillivray, Mark Schmidek, et al. 1997. *Guide to narrative language: Procedures for assessment*. Thinking Publications Eau Claire, WI.

[23] Noriko Iwashita, Annie Brown, Tim McNamara, and Sally O'Hagan. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied linguistics* 29, 1 (2008), 24–49.

[24] Jelena Jovanović, Shane Dawson, Srećko Joksimović, and George Siemens. 2020. Supporting actionable intelligence: reframing the analysis of observed study strategies. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 161–170.

[25] Igor Jugo, Bozidar Kovacic, and Vanja Slavuj. 2015. Integrating a Web-based ITS with DM tools for Providing Learning Path Optimization and Visual Analytics. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015, Madrid, Spain, June 26-29, 2015*. 574–575. http://www.educationaldatamining.org/EDM2015/proceedings/poster574-575.pdf

[26] Jimin Kahng. 2014. Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning* 64, 4 (2014), 809–854.

[27] Olga Kozar. 2015. Discursive practices of private online tutoring websites in Russia. *Discourse: Studies in the cultural politics of education* 36, 3 (2015), 354–368.

[28] Olga Kozar and Naomi Sweller. 2014. An exploratory study of demographics, goals and expectations of private online language learners in Russia. *System* 45 (2014), 39–51.

[29] Milos Krstajic, Enrico Bertini, and Daniel Keim. 2011. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2432–2439.

[30] Yu-Ju Lan, Yao-Ting Sung, Chia-Chun Cheng, and Kuo-En Chang. 2015. Computer-supported cooperative prewriting for enhancing young EFL learners' writing performance. *Language Learning & Technology* 19, 2 (2015), 134–155.

[31] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*.

[32] Charlyn M. Laserna, Yi-Tai Seih, and James W. Pennebaker. 2014. Um . . . Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology* 33, 3 (2014), 328–338.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[34] Heng Luo, Anthony Robinson, and Jae-Young Park. 2014. Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning Journal* 18, 2 (2014).

[35] Roy Lyster and Leila Ranta. 1997. Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in second language acquisition* 19, 1 (1997), 37–66.

[36] Bill MacCartney. 2009. *Natural language inference*. Stanford University.

[37] Philipp Mayring. 2015. *Qualitative Content Analysis: Theoretical Background and Procedures*. Springer Netherlands, Dordrecht, 365–380. https://doi.org/10.1007/

978-94-017-9181-6_13

[38] Riccardo Mazza and Vania Dimitrova. 2004. Visualising student tracking data to support instructors in web-based distance education. In *Proceedings of the 13th International Conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*. 154–161. https://doi.org/10.1145/1013367.1013393

[39] Riccardo Mazza and Vania Dimitrova. 2007. CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses. *International Journal of Human-Computer Studies* 65, 2 (2007), 125–139. https://doi.org/10.1016/j.ijhcs.2006.08.008

[40] Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. Dissertation. The University of Memphis.

[41] Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42, 2 (2010), 381–392.

[42] Hiroaki Ogata, Chengiu Yin, Moushir M El-Bishouty, and Yoneo Yano. 2010. Computer supported ubiquitous learning environment for vocabulary learning. *International Journal of Learning Technology* 5, 1 (2010), 5–24.

[43] Selami Ok. 2014. Reflections of ELT Students on Their Progress in Language and Vocabulary Use in Portfolio Process. *English Language Teaching* 7, 2 (2014), 53–62.

[44] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. LifeLines: visualizing personal histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 221–227.

[45] Yvonne Préfontaine, Judit Kormos, and Daniel Ezra Johnson. 2016. How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing* 33, 1 (2016), 53–73.

[46] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[47] Mohi Reza and Dongwook Yoon. 2021. Designing CAST: a computer-assisted shadowing trainer for self-regulated foreign language listening practice. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[48] Marian J Rossiter. 2009. Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review* 65, 3 (2009), 395–412.

[49] Peter Skehan et al. 1998. *A cognitive approach to language learning*. Oxford University Press.

[50] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *NIPS*.

[51] William L Tarvin and ALI YAHYA AL-ARISHI. 1991. Rethinking Communicative Language-Teaching: Reflection and the EFL Classroom. *Tesol Quarterly* 25, 1 (1991), 9–27.

[52] Katrien Verbert, Xavier Ochoa, Robin De Croon, Raphael A Dourado, and Tinne De Laet. 2020. Learning analytics dashboards: the past, the present and the future. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 35–40.

[53] Stephan Weibelzahl. 2001. Evaluation of adaptive systems. In *International Conference on User Modeling*. Springer, 292–294.

[54] DOROTA WERBIŃSKA, ELIF BOZYIĞIT, LUÎS GUERRA, Małgorzata Ekiert, and SERHAN KÖSE. 2019. English language teachers' conceptualizations of one-to-one private tutoring: An international phenomenographic study. *Glottodidactica. An International Journal of Applied Linguistics* 46, 2 (2019), 175–196.

[55] Keith Willey and AP Gardner. 2010. Improving the standard and consistency of multi-tutor grading in large classes. In *ATN Assessment Conference*. Institute for Interactive Media and Learning, University of Technology Sydney.

[56] Keith Willey and Anne Gardner. 2010. Perceived differences in tutor grading in large classes: Fact or fiction?. In *2010 IEEE Frontiers in Education Conference (FIE)*. IEEE, S2G–1.

[57] Keith Willey, Anne Gardner, et al. 2011. Getting tutors on the same page. In *Australasian Association for Engineering Education Conference 2011: Developing engineers for social justice: Community involvement, ethics & sustainability 5-7 December 2011, Fremantle, Western Australia*. Engineers Australia, 454.

[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 38–45.

[59] Chris F Wright, Nick Wailes, Greg J Bamber, and Russell D Lansbury. 2017. Beyond national systems, towards a 'gig economy'? A research agenda for international and comparative employment relations. *Employee Responsibilities and Rights Journal* 29, 4 (2017), 247–257.

[60] Jinyue Xia and David C. Wilson. 2018. Instructor Perspectives on Comparative Heatmap Visualizations of Student Engagement with Lecture Video: Comparative Heatmap Visualizations of Student Video Engagement. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education, SIGCSE 2018, Baltimore, MD, USA, February 21-24, 2018*. 251–256. https://doi.org/10.1145/3159450.3159487

[61] Meng Xia, Yuya Asano, Joseph Jay Williams, Huamin Qu, and Xiaojuan Ma. 2020. Using Information Visualization to Promote Students' Reflection on" Gaming the System" in Online Learning. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 37–49.

[62] Meng Xia, Mingfei Sun, Huan Wei, Qing Chen, Yong Wang, Lei Shi, Huamin Qu, and Xiaojuan Ma. 2019. Peerlens: Peer-inspired interactive learning path planning in online question pool. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[63] Meng Xia, Yankun Zhao, Mehmet Hamza Erol, Jihyeong Hong, and Juho Kim. 2021. Understanding Distributed Tutorship in Online Language Tutoring. *arXiv preprint arXiv:2112.03500* (2021).

[64] Hisako Yamashita and Satoko Kato. 2012. The Wheel of Language Learning: A tool to facilitate learner awareness, reflection and action. *Advising in language learning: Dialogue, tools and context* (2012), 164–169.

[65] Hui-Chin Yeh and Wei-Yun Lai. 2019. Speaking progress and meaning negotiation processes in synchronous online tutoring. *System* 81 (2019), 179–191.

[66] Kevin Wai Ho Yung and Ming Ming Chiu. 2020. Secondary school students' enjoyment of English private tutoring: An L2 motivational self perspective. *Language Teaching Research* (2020), 1362168820962139.

[67] Kevin Wai Ho Yung and Rui Yuan. 2020. 'The most popular star-tutor of English': Discursive construction of tutor identities in shadow education. *Discourse: Studies in the Cultural Politics of Education* 41, 1 (2020), 153–168.

[68] Jian Zhao, Christopher Collins, Fanny Chevalier, and Ravin Balakrishnan. 2013. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2080–2089.