Developing a Pedagogically Oriented Interactive Reading Tool with Teachers in the Loop

Mihwa Lee, Björn Rudzewitz and **Xiaobin Chen**
Hector Research Institute of Education Sciences and Psychology,
LEAD Graduate School and Research Network
University of Tübingen, Germany
{mihwa.lee, bjoern.rudzewitz, xiaobin.chen}@uni-tuebingen.de

Abstract

Reading is an essential life skill and crucial for students' academic success. Particularly, the need for students to read in English as a second language (L2) has grown due to its global significance. However, L2 readers often have limited opportunities for meaningful, interactive reading practice with immediate support. This paper introduces *ARES*, a pedagogically oriented, web-based intelligent computer-assisted language learning (ICALL) system designed to enhance the L2 reading experience, developed through an action research approach involving practitioners. ARES offers a range of interactive features for students, including not only the autonomous identification of vocabulary and more than 650 language means, but also making them interactively explorable in the text, providing detailed explanations and practical examples in contexts. To support effective teaching, ARES employs a Large Language Model (LLM) for generating tailored reading comprehension questions and answer evaluations, with teachers in the loop, achieving human and Artificial Intelligence (AI) collaboration. We present the development and application of the system from both technical and pedagogical perspectives to advance L2 learning research and refine educational tools.

1 Introduction

In today's increasingly globalized world, the growing necessity for students to read in L2 English underscores the importance of proficient L2 reading skills (Vettori et al., 2023). Learning to read in L2 is complex, as learners must grasp literacy in an unfamiliar language (Verhoeven, 2011). Thus, it is important to support L2 learners' reading development, especially in school contexts where L2 learning most often takes place. However, school teachers often face challenges in providing interactive and meaningful learning experiences for a large number of students due to limited time and highly heterogeneous students with different proficiency levels, native languages, and learning preferences in the same class.

Digital environments, such as ICALL systems, offer unique opportunities for new ways of learning and teaching (Amaral and Meurers, 2011). These systems have been shown to enhance learning engagement (Liu et al., 2016) and achieve better language acquisition (Oberg and Daniels, 2013) through features such as automatic feedback (Ai, 2017), intelligent tutoring (Choi, 2016), and personalized support (Heilman et al., 2010). Despite these advancements, a lot of previous systems are falling short on integrating the AI technologies (e.g., LLMs) or on addressing the practical day-to-day needs of L2 teachers.

In order to address these gaps and enhance real-life usage of ICALL systems in classrooms, we designed and developed an ICALL system that systematically and automatically provides various interactive support for L2 reading, targeting young learners of English as a foreign/second language (EFL/ESL). The development of the system is grounded in theories of text comprehension in second language acquisition (SLA), leveraging the affordances of current language technologies. The general goal of the system is to provide school teachers with a tool to easily create reading activities with interactive and individualized support for their students. Currently, the system provides a web-based platform that features (1) provision of annotations and glossing of vocabulary and language means, (2) automatic generation and evaluation of reading comprehension questions, and (3) easy management of student classes, assignments and their submissions, as well as feedback on assignments. In this article, we introduce the design rationale and development of the system with its different elements in detail. We conclude with an outlook on the design of a study that investigates perceptions of the system in German secondary schools. By exploring these dimensions, we aim to showcase how the Natural Language Processing (NLP) and AI technologies can be used to support L2 reading learning and teaching.

2 Background

2.1 Vocabulary and Grammar Knowledge in L2 Reading Comprehension

Reading is a complex cognitive activity that requires the integration of information from the text and the reader's background knowledge. Successful reading comprehension (RC) depends on skilled processing of the visually presented text (Verhoeven, 2011). It requires a wide range of linguistic as well as non-linguistic skills including word recognition, linguistic knowledge, discourse-level meaning making, reading strategies, inferring, and comprehension monitoring (Grabe, 2014). Current theories on RC typically involve conceptual representations with several interdependent layers. There is typically a local-level representation based on text-based information (i.e., vocabulary, grammar) and a high-level representation where the content of the text becomes integrated into the reader's larger conceptual structure (i.e., integrating the textual information across sentences) (Jung, 2009; Kintsch, 1988; Kintsch and van Dijk, 1978). During the construction of semantic structures at these various levels, a reader's vocabulary and grammatical knowledge influences the entire reading process (Jung, 2009). In particular, the parsing mechanism, driven by this vocabulary and grammar knowledge, operates on text segments assembled locally. Consequently, if readers generate inaccurate or incomplete representations of these local text segments, their overall comprehension of the text can be significantly impaired (Jung, 2009; Koda, 2007). Lexical-syntactic knowledge is critical in the construction of the local-level representation, where text-based propositions are built to eventually support the high-level representation (Choi and Zhang, 2021; Kintsch, 1988). Knowledge of vocabulary and grammar thus helps with the construction of text-based information and eventually facilitates in-depth comprehension.

Following Alderson's (1984) discussion of whether L2 reading is a reading problem or a language problem, SLA researchers have been interested in the importance of vocabulary and grammar knowledge in an effort to understand the process of L2 RC. A plethora of empirical studies have been conducted to gain a better understanding of how vocabulary and grammar knowledge affect L2 RC, whose results generally support the primacy of both L2 vocabulary and grammar knowledge in L2 RC (Choi and Zhang, 2021). For instance, in a longitudinal study examining the relation of oral language proficiency and decoding skills to L2 RC among Dutch-speaking young EFL learners, Droop and Verhoeven (2003) found that both vocabulary and morphosyntactic knowledge had an equally strong correlation to L2 RC, especially at the initial stage when the learners had relatively low L2 proficiency. Recent meta-analyses in L2 RC (Chen and Mei, 2024; Choi and Zhang, 2021) also demonstrate that L2 vocabulary and grammar knowledge are the two strongest predictors of L2 RC. Hence, it is important to accommodate both types of knowledge in the design of teaching of L2 reading. However, vocabulary and grammar knowledge varies a lot among individuals, requiring support for their development be highly personalized. From an instructional perspective, however, due to the time constraint and students' heterogeneity, it is almost impossible for teachers to pinpoint vocabulary and grammatical knowledge that each learner does not understand while they are reading.

2.2 Computer-based Development of L2 Reading Comprehension

Technological applications in L2 reading range from basic digital texts such as e-readers with limited interactivity to online dictionaries to collaborative annotation. Reviews of L2 RC literature (Saeidi and Yusef, 2012; Sawaki, 2001) have shown that specially designed software, ICALL systems, online lessons, animated texts, use of multimedia contexts, interactive multi-modal materials, online dictionaries, e-books and hypertext/hypermedia environments have been used to enhance L2 RC. Here, we describe two features that are highly relevant to our system.

Online Dictionaries Primarily used for looking up unknown words in reading, writing, and vocabulary learning activities, online dictionaries often in the form of electronic glossing have been considered highly feasible, individual learning materials (Çolak and Balaman, 2022) as they "provide controlled opportunities for linguistic input for the learner and interaction with the computer" (Chapelle, 2003, p. 25). One of the prominent examples of electronic glossing is Amazon's Kindle, which provides users with a dictionary function that presents the definitions of words at the bottom of the screen (Lee and Lee, 2015). Another example is *Readlang* (Ridout, 2013), a commercial platform that provides instant translation of words in texts in multiple languages. In fact, it has been shown that online dictionaries such as glossing enhance L2 RC as well as L2 vocabulary acquisition, as found in a meta-analysis of studies on both electronic and textual glosses (Taylor, 2009). Studies also revealed that L2 learners prefer computerized glossing to its paper counterparts (Bowles, 2004). Traditional online dictionaries, however, constrain the selection of an appropriate meaning among all the possible meanings as well as providing a wider range of information such as collations, as they in general list only straight definitions. Previous literature suggests that examples illustrating syntax, collocation, usage and context are more helpful in clarifying meaning than straight definitions (McAlpine and Myles, 2003). Furthermore, to the best of our knowledge, there has been no attempt to integrate a dictionary on language means (i.e., explanation of forms) into language learning applications.

Feedback Feedback is information communicated to learners to modify their thinking or behaviors to close the gap between their actual performance and the target performance (Hattie and Timperley, 2007), thus aiming to improve learning (Shute, 2008), as well as enhance emotions and motivation during learning (Fong et al., 2019). The need for feedback on learner production has been well documented in SLA research (Mackey, 2006). Feedback can be categorized into three types: Knowledge-of-Response (KOR) feedback that only includes verification, Knowledge-of-Correct-Response (KCR) feedback that additionally includes the correct answer, and Elaborated Feedback (EF) that also includes extra-instructional information (Swart et al., 2022) such as explanations (e.g., "In the text, the author does not state that...."), follow up questions (e.g., "Why does the author of the text think...?"), location or hint of the correct information in the text (e.g., "Check the part in the text again where the author mentions...."), or a combination of multiple types of information (Finn et al., 2018). Among them, EF can be used to guide and direct the L2 reader, thereby providing additional support. Bown (2017), borrowing words from Mitchell et al. (2013), attests that "from a sociocultural view of L2 acquisition, this support can be considered as a form of scaffolding: a 'process of supportive dialogue which directs the attention of the learner to key features of the environment, and which prompts them through successive steps of a problem' (Mitchell et al., 2013, p. 25)". In fact, in the field of educational sciences, several meta-analyses (Bangert-Drowns et al., 1991; der Kleij et al., 2015; Wisniewski et al., 2020) have demonstrated positive effects of EF over other simpler types of feedback. Despite the potential of EF in L2 RC, only a few attempts have been made to implement it in ICALL systems (Bown, 2017, 2018; Murphy, 2007, 2010). However, most of these research prototypical systems have not been tested widely in schools practically.

Overall, there have been several attempts to integrate features that support L2 RC (e.g., *Readlang*, Bown, 2018; Murphy, 2010), but most existing systems focus on a single aspect of L2 reading support (e.g., vocabulary) and fall short in offering comprehensive, pedagogically grounded support throughout the entire L2 reading process. This poses challenges for a practical implementation in classroom settings. Moreover, most of these systems were research-oriented and not designed for actual widespread classroom usage, further complicating their adoption and effectiveness. Our work seeks to address this gap between research, foreign language pedagogy, and real-life classroom usage by developing *ARES*, a pedagogically oriented, web-based ICALL system designed to enhance L2 reading experience. In the following section, we present the system architecture and each feature of the system in detail.

3 ARES

ARES (Annotated Reading Enhancement System) is designed as a multi-layer web application that strikes a balance between usability and flexibility. The system implements a responsive design that adapts the display for all devices and platforms. Therefore, it works seamlessly across multiple platforms, requiring only a computer, tablet, or smartphone with a web browser and internet access. Using NLP tools, the system supports students by identifying and providing glossing on vocabulary and language means with examples, which they can consult as needed while reading the assigned texts. For teachers, ARES automates the process of generating questions for assignments and providing individual feedback to each student response by implementing a pre-trained LLM (ChatGPT 4o[1]), significantly reducing their workload and allowing them to focus more on communicative activities in classrooms.

Involving teachers or stakeholders in education research whose results will be used in schools is considered very important because schools and teachers should not only be treated as consumers of the research results (Farley-Ripple et al., 2018). Successful research that has a practical impact in schools is always the outcome of bi-directional efforts. This bi-directional effort is not a one-off process, but it will involve multiple iterations of interactions between the researchers and the teachers. Consequently, we decided to use a multi-cycle action research paradigm to guide the development and research process. The action research model (Figure 1) is a systematic, collective, collaborative, and self-reflective scientific inquiry aimed at improving educational practices and addressing the practical concerns of teachers (Kemmis and McTaggart, 1988; Rapoport, 1970), where a key characteristic of action research is the involvement of stakeholders, including teachers, students, and researchers.

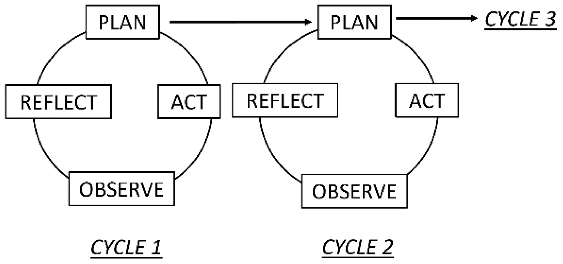


Figure 1: Action Research Model (Kemmis and McTaggart, 1988)

In the following subsections, the system that has been developed in the first phase of the action research paradigm is described in more detail, both from the teacher perspective and the student perspectives.

3.1 System Architecture

Utilizing a software-as-a-service (SaaS) approach, ARES provides the software through the cloud, allowing system developers to update the application with new features and fix bugs without requiring users to download updates from app stores. The system is built on a Java backend deployed in a Jetty server. For the display layer, we use the Bootstrap framework, which provides a highly extensible component-based design for an optimized display. In order to enable Learning Analytics, all user activities such as button clicks, lookups of language means, reading comprehension question attempts, assignment submissions, viewing of specific feedback messages, and any other relevant user actions are logged through xAPI[2], an interoperability specification for recording user interactions, and stored in a Learning Record Store (LRS).

3.2 Home Interface

Based on discussion with the involved stakeholders, the home pages that users first see when they log in offer the most commonly used functionalities as a starting point for efficient usage.

Teacher Home There are three main sections that teachers can select from, described in detail below:

- **Classes** Teachers can create, delete, edit classes and manage students.
- **Assignments** Teachers can manage assignments and check the results of each assignment.
- **Texts** Teachers can browse, upload, and edit texts.

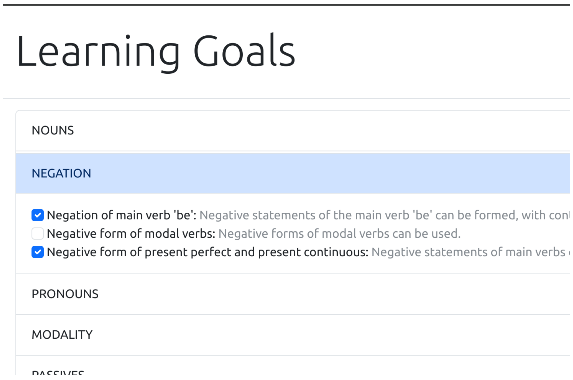
To address the challenge English teachers face in finding texts appropriate for their students' English levels, the system includes a "text bank" with reading materials covering 12 topics (e.g., History, Travel and Nature, Technology). These materials are crafted by experienced ESL/EFL teachers ensuring users always have access to relevant content from a variety of themes, addressing a need by teachers to search for material to prepare their lessons. The initial target audience is classes in German secondary schools (Gymnasium) with proficiency levels roughly equivalent to A2-B1 according to the Common European Framework of

[1] https://chatgpt.com/

[2] https://xapi.com/

Reference for Languages (CEFR) (Council of Europe, 2020). The texts are tailored to match these proficiency levels. Additionally, teachers have the option to upload their own texts, which they can later edit or delete as needed. When creating an assignment, teachers receive automatically generated suggestions for comprehension questions generated from the LLM. With the goal of keeping teachers in the loop, we designed the system so that teachers always hold the ultimate decision-making power, and are supported by the system's suggestions and tools. They can post-edit these suggestions, confirm them, or add their own questions manually, to ensure that teachers' expertise is involved in the process. On the technical side, we conducted an iterative approach to refine the prompt for question generation. The full final version of the prompt implemented in the system is attached in the Appendix.

Teachers can decide which annotations on language means to show students (section 3.3), allowing them to tailor assignments and annotations to specific learning goals and ensure appropriateness for their students' proficiency levels (see Figure 2). The motivation behind this customization is that reading texts often contain a wide range of language means and grammatical structures, and it is often hard for teachers to selectively control students' focus on a certain language mean in reading texts. By enabling teachers to customize annotations of language means based on learning goals, the system ensures that reading materials support the target structures, making the learning process more efficient and tailored to pedagogical needs.


Learning Goals

☑ Negation of main verb 'be': Negative statements of the main verb 'be' can be formed, with con

□ Negative form of modal verbs: Negative forms of modal verbs can be used.

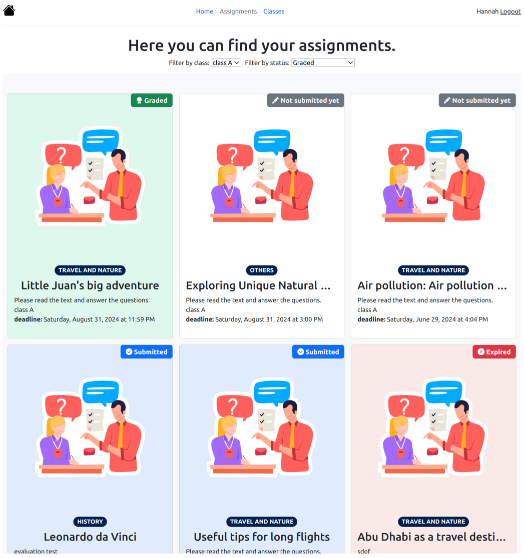
☑ Negative form of present perfect and present continuous: Negative statements of main verbs

Figure 2: Selection of Annotations of Language Means

Student Home The system presents two main options that students need most on their start page:

- **Classes** Students can see classes they are enrolled and join a class using a 4-digit access code provided by the teacher.
- **Assignments** Each assignment card indicates the status of an assignment using different background colors and badges (see Figure 3).

Upon clicking or touching the assignment card, students are forwarded to the reading interface.



Here you can find your assignments.

Filter by class: class A Filter by status: Graded

Little Juan's big adventure
Please read the text and answer the questions.
class A
deadline: Saturday, August 31, 2024 at 11:59 PM

Exploring Unique Natural ...
Please read the text and answer the questions.
class A
deadline: Saturday, August 31, 2024 at 3:00 PM

Air pollution: Air pollution ...
Please read the text and answer the questions.
class A
deadline: Saturday, June 29, 2024 at 4:04 PM

Leonardo da Vinci
evaluation test

Useful tips for long flights
Please read the text and answer the questions.

Abu Dhabi as a travel desti...
sdaf

Figure 3: Student Assignments Page

3.3 Reading Interface

The main features of the interface are an on-demand annotation on language means that is based on the English Grammar Profile (EGP)[3] and an on-demand vocabulary lookup based on the LLM. Given their relevance to the overall goal of the system, the following subsections describe these functions in detail.

3.4 Annotations on Language Means

The annotation function of language means acts as an instant glossing on forms, allowing students to click on any word (or section of a sentence) within a reading text to access its detailed explanation with example sentences and the corresponding CEFR level of the grammatical structure. When a text is uploaded to the system, it is automatically analyzed and indexed by an NLP tool our research group has created to extract form-based language means from the EGP. The EGP is a comprehensive database listing over 650 language means spanning the entire range of CEFR levels. It is based

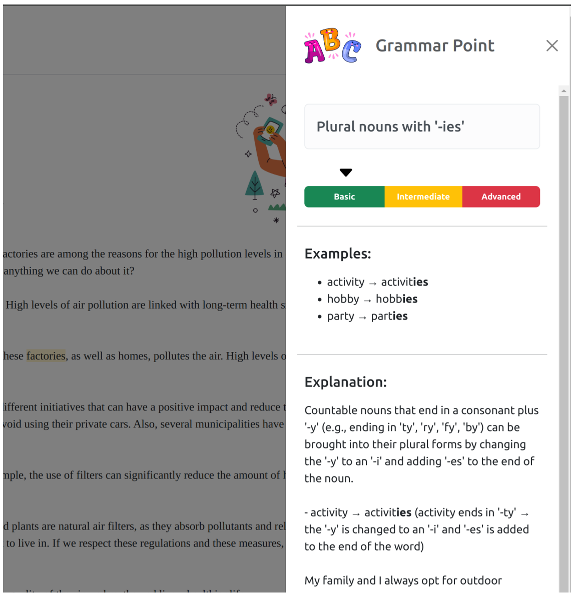
[3] https://www.englishprofile.org/english-grammar-profile/egp-online

on an extensive analysis of the Cambridge Learner Corpus, providing insights into the typical grammar usage at each proficiency level (O'Keeffe and Mark, 2017). For each language mean, we asked experienced teachers to write an explanation and examples in both a student-directed and a more concise teacher-directed way. Along with an indication of the CEFR level, this information is shown to the users, with the respective variant of the explanation selected on the user's role (see Figure 4).

The pipeline for this function is based on the further development of the pipeline introduced in Quimal et al. (2021). It is based on the Unstructured Information Management Architecture (UIMA, Ferrucci and Lally 2004)[4], an open-source Apache framework used in large-scale text processing applications. It includes three main components: an NLP preprocessing module, an annotator built using UIMA's Rule-based Text Annotation (Ruta)[5] framework, and an application to run the pipeline for analyzing texts. The NLP preprocessing module employs tools like Standard CoreNLP (Manning et al., 2014)[6] and DKPro Core (de Castilho et al., 2016)[7] for tasks such as tokenization, part-of-speech tagging, and dependency parsing. The Ruta annotator applies regular expression-based rules after the pre-processing to identify specific language means, tagging them with information like construction type and position in the text, ensuring robust and scalable text processing.

3.5 Vocabulary Lookup Function

The system offers an instant vocabulary glossing for students. It enables students to click on any word within a reading text and immediately access comprehensive vocabulary information about that word. When a student clicks on a word in the reading text, the system identifies and extracts the clicked word as a token and its surrounding sentence as a context. The LLM is then applied to analyze the word both as an isolated token and within the context of the sentence to understand its specific usage and meaning, including the general definition, meaning in the specific context, collocations, related vocabulary, morphosyntactic elements of the word, and additional information (see Figure 5). In order to make sure that the students understand the relevant information of the clicked vocabulary, there is an option for them to see a translation of the explanation.



actories are among the reasons for the high pollution levels in
anything we can do about it?

High levels of air pollution are linked with long-term health s

hese factories, as well as homes, pollutes the air. High levels o

ifferent initiatives that can have a positive impact and reduce t
void using their private cars. Also, several municipalities have

nple, the use of filters can significantly reduce the amount of h

d plants are natural air filters, as they absorb pollutants and rel
to live in. If we respect these regulations and these measures,

Grammar Point

Plural nouns with '-ies'

Examples:

- activity → activit**ies**
- hobby → hobb**ies**
- party → part**ies**

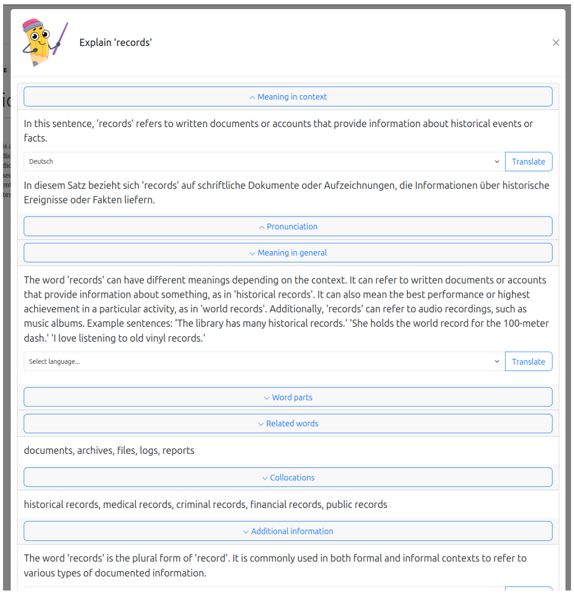
Explanation:

Countable nouns that end in a consonant plus '-y' (e.g., ending in 'ty', 'ry', 'fy', 'by') can be brought into their plural forms by changing the '-y' to an '-i' and adding '-es' to the end of the noun.

- activity → activit**ies** (activity ends in '-ty' → the '-y' is changed to an '-i' and '-es' is added to the end of the word)

My family and I always opt for outdoor

Figure 4: Grammar Lookup



Explain 'records'

In this sentence, 'records' refers to written documents or accounts that provide information about historical events or facts.

Deutsch

In diesem Satz bezieht sich 'records' auf schriftliche Dokumente oder Aufzeichnungen, die Informationen über historische Ereignisse oder Fakten liefern.

The word 'records' can have different meanings depending on the context. It can refer to written documents or accounts that provide information about something, as in 'historical records'. It can also mean the best performance or highest achievement in a particular activity, as in 'world records'. Additionally, 'records' can refer to audio recordings, such as music albums. Example sentences: 'The library has many historical records.' 'She holds the world record for the 100-meter dash.' 'I love listening to old vinyl records.'

documents, archives, files, logs, reports

historical records, medical records, criminal records, financial records, public records

The word 'records' is the plural form of 'record'. It is commonly used in both formal and informal contexts to refer to various types of documented information.

Figure 5: Vocabulary Lookup

3.6 Questions and Rating Functionality

For assignments that accompany RC questions, these questions are displayed below the text. Students have the flexibility to complete the assignment without answering all the questions. At the end of the assignment, when students click on the submit button, the system presents a dialogue box

[4] https://uima.apache.org/

[5] https://uima.apache.org/ruta.html

[6] https://stanfordnlp.github.io/CoreNLP/

[7] https://dkpro.github.io/dkpro-core/

asking students to rate the difficulty and interestingness of the text using a 5-star Likert scale with an option to leave free-form comments about the assignment, which will be provided to the teacher, offering insights into both the overall and individual perceptions of the assigned text to teachers, text authors, and researchers.

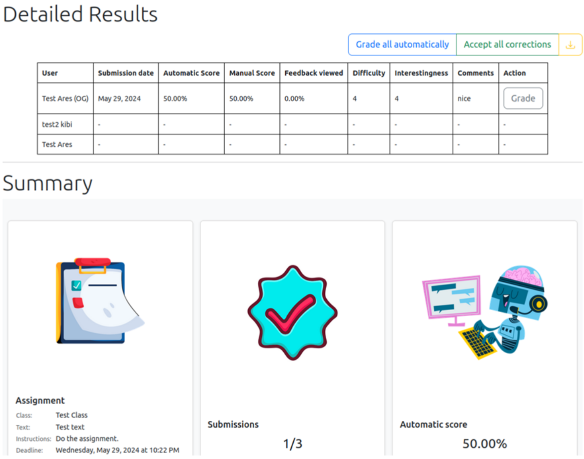
Once the students submit the assignment, the system forwards them to the Assignments page (section 3.2). However, students keep the right to access the reading interface even after submitting the responses in order to give them chances to review the finished assignments and view the teacher's feedback.

4 Evaluation Interface

Teacher Evaluation On the selection of the assignment in the Teacher Home (section 3.2), the system directs them to the Evaluation Interface, which consists of two main sections as shown in Figure 6. The upper section of the page displays information about individual student submissions in a table format, including the time of submission, automatic score (calculated by the system), manual score (assigned by the teacher), percentage of the feedback read by the student, difficulty rating, interestingness rating, and comments (see Figure 6). With the purpose of reducing the teachers' workload, we equip the system with functionalities that automate grading by integrating the LLM. Upon clicking the "Grade all automatically" button above the submission table, all student responses are sent to the LLM in a parallelized way for processing. The LLM evaluates the student responses against a target response for each question while also provided with the reading text as context. As the output of this process, the teacher sees a percentage score of correct responses displayed under the "Automatic score" column. Teachers can then transfer these automatic scores to the "Manual score" column by clicking the "Accept all corrections" button. The full final version of the refined prompt to the LLM is attached in the Appendix.

In order to keep teachers in the loop, we allow teachers to review and modify the automated scores by clicking the "Grade" button within the submission table, which redirects them to the individual submission page. Here, detailed evaluation information (questions, student responses, target answers, automatic scores, and automatic feedback) is displayed, allowing teachers to adjust scores and feedback as needed. If the teacher agrees with the automated grading, they may utilize the "Copy all" button to transfer the automated scores and feedback to the manual grading section. Alternatively, for more granular adjustments, the "Copy" button allows for the selective adoption of scores on an individual question basis. Eventually, what students see is what teachers confirm at the end. This way, although we reduce teachers' burden of grading, we at the same time make sure that teachers are in full control of what students see.

The lower section of the page provides a summative assessment of the assignment, including the number of submissions, average automatic score, average manual score, average interestingness rating, and average difficulty rating. The average automatic and manual scores are updated automatically based on the teacher's grading of individual submissions. The evaluation data of both the class as a whole and individual students can be downloaded as a CSV file for the teacher to bring to class for further review and discussion.



Detailed Results

User	Submission date	Automatic Score	Manual Score	Feedback viewed	Difficulty	Interestingness	Comments	Action
Test Ares (OG)	May 29, 2024	50.00%	50.00%	0.00%	4	4	nice	Grade
test2 kibi	-	-	-	-	-	-	-	-
Test Ares	-	-	-	-	-	-	-	-

Summary

Assignment

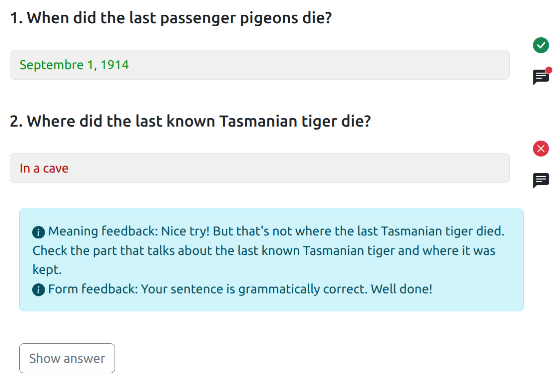
Class:	Test Class
Text:	Test text
Instructions:	Do the assignment.
Deadline:	Wednesday, May 29, 2024 at 10:22 PM

Submissions
1/3

Automatic score
50.00%

Figure 6: Assignment Grading Overview Page

Student Evaluation Students see only the manual evaluations confirmed by the teacher during the grading process. It is important to note that the evaluation display is only accessible to students once the teacher has entered the manual evaluation. Each answer is accompanied by different colors and icons to indicate binary feedback (correct/incorrect) (see Figure 7). Under the binary feedback icon, a chat button icon is available, which students can click to open or close the teacher's feedback for each response. The system tracks which feedback has been viewed by the students and informs teachers about which students have read which feedback, providing insight into student engagement and enabling more targeted support.



1. When did the last passenger pigeons die?

Septembre 1, 1914 ✓

2. Where did the last known Tasmanian tiger die?

In a cave ✗

Meaning feedback: Nice try! But that's not where the last Tasmanian tiger died. Check the part that talks about the last known Tasmanian tiger and where it was kept.

Form feedback: Your sentence is grammatically correct. Well done!

Figure 7: Feedback for Students

5 Conclusion and Outlook

Grounded in theories of text comprehension in SLA, and leveraging the affordances of language and AI technology, we present ARES, a web-based language learning system designed to support L2 RC of young EFL/ESL learners with teachers in the loop. The system provides on-demand help functions, such as glossing of vocabulary and language means, allowing students to interactively engage with texts, as well as EF on RC questions. These features not only aid students in understanding English reading texts but also alleviate teachers' workloads by automating time-consuming tasks such as question generation and evaluation. Furthermore, ARES facilitates direct interaction between students and teachers outside the classroom, enabling flexible assignment and feedback processes.

We acknowledge certain limitations in our system. First, there are challenges regarding the classification accuracy of language means (see Section 3.4). To tackle this challenge, a member of our research team is conducting a study to assess the system's classification accuracy by comparing the results of our automatic classification with labels provided by human annotators. Second, it is important to note that LLMs still lack the same level of understanding and context awareness as humans (Ray, 2023). Although they can perform a variety of tasks within seconds, LLMs struggle due to tendencies toward hallucination (Nye et al., 2023). However, this challenge is precisely why we designed the system to involve teachers in the process, ensuring they confirm outputs before students see them, rather than relying solely on raw LLM-generated results. Although teachers might occasionally miss inaccuracies produced by the LLM, the system still significantly reduces their workload, allowing them to focus more on communicative activities in the classroom. Nevertheless, we are currently working on investigating the feasibility of leveraging the LLM to generate short answer questions and feedback. Using a human-authored evaluation method, we are investigating the linguistic and pedagogical quality of these LLM-generated outputs. For the evaluation criteria of the questions, we will employ a nine hierarchical criteria rubric (e.g., *Understandable, Grammatical, Answerable, Clear*) used in previous studies (Horbach et al., 2020; Moore et al., 2022; Steuer et al., 2021), which has been shown to be comprehensive, easy to interpret, and includes the pedagogical aspects of a question (Moore et al., 2022). For the evaluation criteria of the feedback, we will employ a four criteria rubric (*Readily applicable, Readability, Relational, Specificity*) that is formulated based on previous work on the human-authored evaluation of the quality of machine-generated feedback (Jia et al., 2021; Liang et al., 2024; Pinger et al., 2018; van der Lee et al., 2021).

Since the first version of the system is deployed, a study investigating teachers' and students' perceptions of the system is currently taking place in two intact English classes at secondary schools in southwest Germany with the purpose of evaluating the system's usability and students' interaction with the system. Over a four-week period, students will read two texts weekly as part of their homework assigned by teachers. System perceptions will be assessed through a self-report questionnaire of comprehensive evaluation of technology adapted from Lai et al. (2022). In addition to the survey data, log data will be analyzed to explore the learning behavior within the context of real-world ICALL system use.

ARES is currently available under https://ares.kibi.group.

Acknowledgements

This project is funded by the German Ministry of Education and Science (BMBF) under the funding number 01IS22076.

References

Haiyang Ai. 2017. Providing graduated corrective feedback in an intelligent computerassisted language learning environment. *ReCALL*, 29(3):313–334.

Charles J. Alderson. 1984. Reading in a foreign language: A reading problem or a language problem. In Charles J. Alderson and Alexander H. Urquhart, editors, *Reading in a foreign language*, pages 1–25. Longman, London.

Luiz A. Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23:4–24.

Robert L. Bangert-Drowns, Chen-Lin C. Kulik, James A. Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2):213–238.

Melissa A. Bowles. 2004. L2 glossing: To CALL or not to CALL. *Hispania*, 87(3):541–552.

Andy Bown. 2017. Elaborative feedback to enhance online second language reading comprehension. *English Language Teaching*, 10(12):164–171.

Andy Bown. 2018. Supporting online L2 academic reading comprehension with computer-mediated synchronous discussion and elaborative feedback. *The Reading Matrix*, 18(1):41–63.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, , and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84.

Carol A. Chapelle. 2003. *English Language Learning and Technology: Lectures on Applied Linguistics in the Age of Information and Communication Technology*, volume 7. John Benjamins Publishing, Amsterdam.

Huilin Chen and Huan Mei. 2024. How vocabulary knowledge and grammar knowledge influence L2 reading comprehension: a finer-grained perspective. *European Journal of Psychology of Education*, pages 1–23.

Inn-Chull Choi. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364.

Yunjeong Choi and Dongbo Zhang. 2021. The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching*, 59(1):1–30.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion Volume*. Cambridge University Press, Cambridge.

Mienke Droop and Ludo Verhoeven. 2003. Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38(1):78–103.

Elizabeth Farley-Ripple, Henry May, Allison Karpyn, Katherine Tilley, and Kalyn McDonough. 2018. Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, 47(4):235–245.

David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Bridgid Finn, Ruthann Thomas, and Katherine A. Rawson. 2018. Learning more from feedback elaborating feedback with examples enhances concept learning. *Learning and Instruction*, 54:104–113.

Carlton J. Fong, Erika A. Patall, Ariana C. Vasquez, and Sandra Stautberg. 2019. A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*, 31(1):121–162.

William Grabe. 2014. Key issues in L2 reading development. In *Proceedings of the 4th CELC Symposium for English Language Teachers-Selected Papers*, pages 8–18.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1):73–98.

Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1753–1762.

Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. All-in-one: Multi-task learning BERT models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895*. `https://arxiv.org/pdf/2110.03895`.

Jookyoung Jung. 2009. Second language reading and the role of grammar. *Working Papers in TESOL and Applied Linguistics*, 9(2):29–48.

Stephen Kemmis and Robin McTaggart. 1988. *The Action Research Planner*, 3rd edition. Deakin University Press, Geelong.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182.

Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.

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

Keiko Koda. 2007. Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57:1–44.

Jennifer W. M. Lai, John De Nobile, Matt Bower, and Yvonne Breyer. 2022. Comprehensive evaluation of the use of technology in education – validation with a cohort of global open online learners. *Education and Information Technologies*, 27:9877—9911.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Hansol Lee and Jang Ho Lee. 2015. The effects of electronic glossing types on foreign language vocabulary learning: Different types of format and glossary information. *Asia-Pacific Education Researcher*, 24(4):591–601.

Zhiping Liang, Lele Sha, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2024. Towards the automated generation of readily applicable personalised feedback in education. In *International Conference on Artificial Intelligence in Education*, pages 75–88.

Chen-Chung Liu, Pin-Ching Wang, and Shu-Ju D. Tai. 2016. An analysis of student engagement patterns in language learning facilitated by Web 2.0 technologies. *ReCALL*, 28(2):104–122.

Alison Mackey. 2006. Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3):405–430.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Janice McAlpine and Johanne Myles. 2003. Capturing phraseology in an online dictionary for advanced users of English as a second language: a response to user needs. *System*, 31(1):71–84.

Rosamond Mitchell, Florence Myles, and Emma Marsden. 2013. *Second language learning theories*, 3rd edition. Routledge, London.

Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *Proceedings of the 17th European Conference on Technology Enhanced Learning, EC-TEL 2022*, pages 243–257.

Philip Murphy. 2007. Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology*, 11(3):107–129.

Philip Murphy. 2010. Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL*, 22(2):112–134.

Benjamin D. Nye, Dillon Mee, and Mark G. Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023)*, pages 78–88.

Andrew Oberg and Paul Daniels. 2013. Analysis of the effect a student-centred mobile learning instructional method has on language acquisition. *Computer Assisted Language Learning*, 26(2):177–196.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Petra Pinger, Katrin Rakoczy, Michael Besser, and Eckhard Klieme. 2018. Implementation of formative assessment–effects of quality of programme delivery on students' mathematics achievement and interest. *Assessment in Education: Principles, Policy & Practice*, 25(2):160–182.

Martí Quixal, Björn Rudzewitz, Elizabeth Bear, and Detmar Meurers. 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*, pages 15–27.

Robert N. Rapoport. 1970. Three dilemmas of action research. *Human Relations*, 23(6):499–513.

Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Steve Ridout. 2013. Readlang. *The EUROCALL Review*, 21(2):64–68.

Mahnaz Saeidi and Mahsa Yusef. 2012. The effect of computer-assisted language learning on reading comprehension in an Iranian EFL context. In *EUROCALL Conference Proceedings: Using, Learning, Knowing*, pages 259–263.

Yasuyo Sawaki. 2001. Comparability of conventional and computerized tests of reading in a second language. *Language Learning and Technology*, 5(2):38–59.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Tim Steuer, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. 2021. On the linguistic and pedagogical quality of automatic question generation via neural machine translation. In *Proceedings of the 16th European Conference on Technology Enhanced Learning, EC-TEL 2021*, pages 289–294.

Elise K. Swart, Thijs M.J. Nielen, and Maria T. Sikkema de Jong. 2022. Does feedback targeting text comprehension trigger the use of reading strategies or changes in readers' attitudes? a meta-analysis. *Journal of Research in Reading*, 45(2):171–188.

Alan M. Taylor. 2009. CALL-based versus paper-based glosses: Is there a difference in reading comprehension? *CALICO Journal*, 27(1):147–160.

Ludo Verhoeven. 2011. Second language reading acquisition. In Michael L. Kamil, P. David Pearson, Elizabeth Birr Moje, and Peter Afflerbach, editors, *Handbook of reading research*, volume 4, pages 661–683. Routledge, New York, NY.

Giulia Vettori, Lidia Casado-Ledesma, S. Tesone, and Christian Tarchi. 2023. Key language, cognitive and higher-order skills for L2 reading comprehension of expository texts in English as foreign language students: a systematic review. *Reading and Writing: An Interdisciplinary Journal*.

Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10:3078.

Fulya Çolak and Ufuk Balaman. 2022. The use of online dictionaries in video-mediated L2 interactions for the social accomplishment of virtual exchange tasks. *System*, 106:102772.

Appendix. Prompts for the LLM

Prompt 1. Question Generation

The query template for asking the LLM to provide two types of reading comprehension questions (factual and inferential). The placeholder fields with angle brackets are to be substituted for the actual data in each query.

```
You are an EFL teacher who teaches English to non-
     native school students between 10-18 years old.
      Provide simple one-sentence short-answer
     reading comprehension questions based on the
     given text to these EFL learners. Do not use
     too difficult words. Literal comprehension
     refers to an understanding of the
     straightforward meaning of the text, such as
     facts, vocabulary, dates, times, and locations.
      Questions of literal comprehension can be
     answered directly and explicitly from the text
     with a few words. Inferential questions ask
     students to infer information from the passage
     where the answer is not directly stated in the
     text. The students have to use their background
      knowledge to make a logical assumption about
     ideas in the passage and normally require a
     full sentence to answer, not a few words.
- text: <reading_text>
- number of factual questions: <number>
- number of inferential questions: <number>
Please provide the questions in JSON format as
     follows:
{
  "questions": [
    {
      "type": <factual_or_inferential>,
      "prompt": "<question>",
      "answer": "<correct_answer>"
    },
  ]
};
```

Prompt 2. Feedback Generation

The query template for asking the LLM to provide feedback and hint to a student's response. The placeholder fields with angle brackets are to be substituted for the actual data in each query.

```
For each question, evaluate each EFL student's
     answer as follows using simple language as the
     students are non-native and kids:
1. Determine if the answer is correct or incorrect
     based on the content only.
2. Provide binary feedback for content ("Correct"/"
     Incorrect").
3. Offer short, kind, and friendly feedback on the
     content of the answers.
4. Give a concrete hint on the content explaining
     why the response was correct or incorrect,
     allowing the student to review part of the text
     , without revealing the target answer. When
     correct, do NOT provide hint.
- text: <reading_text>
Provide evaluation in JSON format using the match of
      answer id:
{
  "evaluation": [
    {
      "question": <question>,
      "answer_id": <answer_id>,
      "answer_text": <student's_answer>,
      "solution": <correct_answer>,
      "binary": <binary_feedback>,
      "feedback": <content_feedback>,
      "hint": <content_hint>
    },
  ]
}
```