

ReadLens: In-Place Contextual Assistance for Second-Language Reading

ANONYMOUS AUTHOR(S)

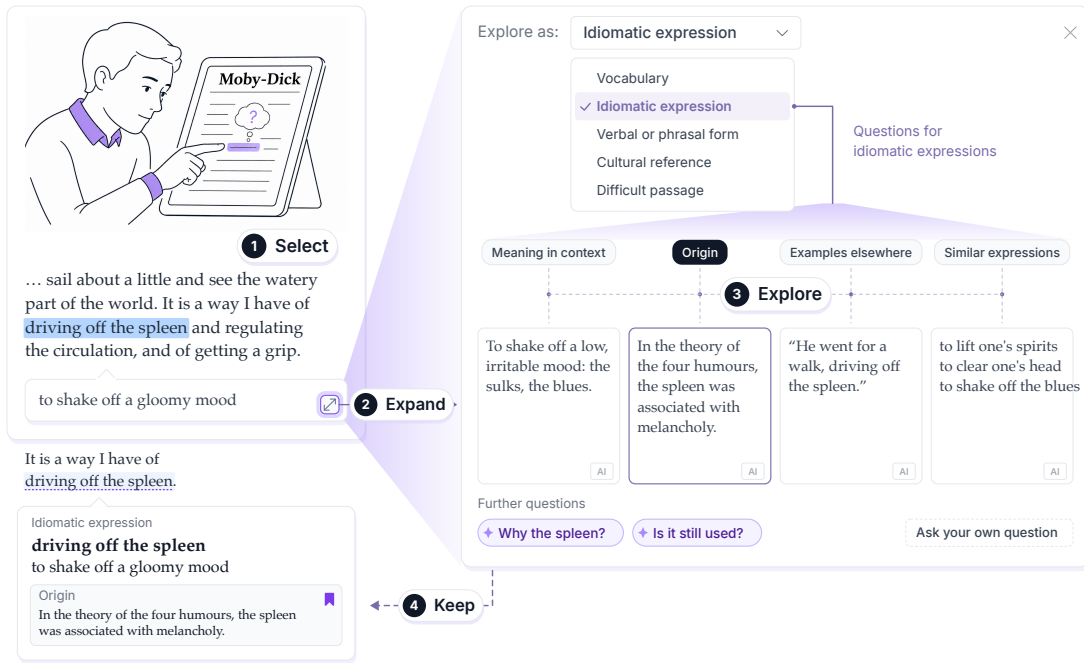


Fig. 1. READLENS provides a brief contextual translation beside a selected span (1). Expanding the response opens a panel with five selection categories (2). The highlighted idiomatic-expression category leads to four questions, each paired here with an illustrative response (3). Readers can save responses in an annotation attached to the source text (4). Interface labels and illustrative responses are shown in English.

Readers of a second language stop to resolve an unfamiliar word or expression. A dictionary answers without the passage. A chat interface uses it but requires leaving the text and composing a question. We examined what readers ask when the questions are already written for them. A survey of 89 readers and interviews with 10 readers and three language experts identified recurring difficulties. READLENS answers a selected word or passage in place and offers questions written for each of these kinds of difficulty. Eighteen readers each read one article with READLENS and one with a dictionary and an LLM chat panel. Readers requested a broader range of explanations with READLENS than with the baseline, including origins, examples, and alternatives. We contribute READLENS and an account of the explanations readers seek when an interface supplies the questions.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Second-language reading, contextual assistance, suggested questions, language models

1 Introduction

While reading an English-language article about a car-free neighbourhood, an Italian reader encounters the sentence “The light rail whisks her to work in 10 minutes.” A word-level lookup of *whisks* may return *frusta*, the Italian word for a kitchen whisk. In this sentence, *whisks* is a verb meaning “takes her quickly.” A conversational assistant can interpret the word from the sentence. Reaching it requires the reader to leave the article, supply the passage, compose a question, and carry the answer back.

People routinely encounter such difficulties while reading in a second language for work, study, and leisure. Their immediate goal is understanding the content, and language learning may occur incidentally as they resolve unfamiliar language. At each difficulty the reader either continues with an uncertain interpretation or stops to clarify it. The cost of that clarification depends on where the assistance appears and on what the reader must do to obtain it. We call assistance *in place* when the reader starts it from a selected span and receives it at that span, without leaving the document.

Digital glosses, brief definitions or translations attached to words, can support vocabulary acquisition and comprehension during second-language reading [18]. Reading.help places generated explanations in place for second-language readers, who select text and choose among explanations of vocabulary, grammar, and passages [7]. In scientific reading, Qlarify offers fixed questions beside a selected passage of an abstract, and its readers used those questions for most of their requests [11]. The designers of these systems wrote the questions for one kind of document, and their readers read in their first language. We examine second-language readers of ordinary articles, with questions drawn from the difficulties that population reports. What such readers request while reading for content, and how that compares with the requests they compose themselves, is the subject of this study.

We surveyed 89 people who studied or read in English and interviewed 10 readers and three language experts. They described uncertainties about vocabulary, expressions, contextual references, and relationships within longer passages, and they differed in how much explanation they wanted before resuming the article. From these accounts we wrote questions that request meanings, examples, comparisons, and explanations of references or passages. We grouped them into five categories of selected language: vocabulary, idiomatic expressions, verbal constructions and phrasal verbs, cultural references, and passages (Table 2). Each category carries the questions written for it.

READLENS provides assistance in place, and the reader controls how far it goes. A context-sensitive Italian translation of the selected English text appears beside the selection. One further action opens the questions written for that selection’s category. Readers can select a suggested question or compose their own. The initial translation stays attached to the selected span, and readers can save follow-up responses in the same annotation.

Our evaluation addresses two research questions:

RQ1. What forms of assistance do readers request with READLENS and a dictionary-plus-LLM baseline during second-language reading?

RQ2. How do readers’ experience and reading outcomes compare between READLENS and the baseline?

We conducted a within-subject study with 18 readers whose first language was Italian. Each participant read two English articles, one with READLENS and one with the baseline, in counterbalanced conditions. The baseline combined a word-level bilingual dictionary with an LLM side panel that automatically received the complete article as context.

During reading, participants submitted 59 requests through READLENS’s prewritten questions and 16 through baseline chat. Eight readers used READLENS to request alternatives, examples, origins, grammar, or relevance, functions absent from the baseline requests. In post-study interviews, readers described unresolved meanings, checks on generated

translations, curiosity, and occasions when the initial translation was sufficient. Readers also related their preferences to independent use of dictionary meanings and to article-level review.

This work contributes:

- a formative account of second-language readers' difficulties, assistance practices, and desired assistance, based on a survey and interviews with readers and language experts;
- READLENS, a reading interface that provides in-place assistance and converts findings about readers' assistance needs into prewritten questions for freely selected words and passages; and
- a descriptive mixed-method account of what readers request through these questions, next to a dictionary-plus-LLM interface, and of how those requests relate to understanding the article as a whole.¹

2 Related Work

Readers can acquire vocabulary while reading to understand a text, a process termed incidental vocabulary learning [31]. In second-language acquisition, Yanagisawa et al. found that glosses, short definitions or translations attached to selected words, increased vocabulary learning compared with reading the same target words without glosses [33]. In materials with glosses, the author selects which words receive explanations, and readers decide whether to consult those explanations. Dictionary access allows readers to choose which words to look up [17]. Consultation also depended on what readers needed to understand. Peters found that words needed to answer comprehension questions received approximately three times as many lookups as other target words and were retained more reliably [27]. Laufer and Hill found that lookup frequency was weakly related to retention, while consulting combinations of lexical information was associated with better retention [20].

The Involvement Load Hypothesis relates vocabulary retention to the need to understand or use a word, the search for its meaning or form, and the evaluation of how a word or meaning fits a context [21]. Yanagisawa and Webb found that learners retained more words when tasks required them to compare meanings or decide how to use a word in context. Needing a word to complete the task also supported learning, while looking up its meaning showed no additional benefit after accounting for these other task requirements [32]. Godfroid et al. found that longer cumulative reading time on unfamiliar words predicted subsequent recognition and recall of their meanings [12]. We therefore distinguish the explanations readers request with READLENS from the vocabulary they subsequently retain.

Digital reading systems make lexical assistance available in texts selected by readers. GLOSSER provides bilingual definitions and corpus examples, and REAP combines documents selected for a learner's reading level with lexical support [15, 24]. SmartReader extends selection to words and passages and uses linguistic analysis to distinguish content words, multiword expressions, grammatical elements, and named entities [3]. Zeeguu supports translation of adjacent words as an expression and retains the source sentence for subsequent vocabulary practice [23]. The effort required to access assistance also relates to consultation. In successive classroom cohorts using a REAP-based interface, students requested translations more frequently when a click opened the translation than when they had to enter or paste text into a separate field [10]. Levine et al. likewise found more lookups with online glosses and dictionaries than with a paper dictionary, without a statistically detectable advantage in comprehension or vocabulary retention [22]. Alevén et al. describe a related challenge in interactive learning environments, where the availability of assistance does not ensure that learners use it effectively [1].

¹Project website: <https://readlens.backtothehuman.com/>.

Draxler et al. observed longer mean reading time with detailed grammar overlays than with translation alone in an e-reader. Compact contextual popups received high usability ratings, and the authors found no statistically significant difference in global reading flow across conditions [9]. Ho et al. found that translations displayed for the currently read passage were perceived as less distracting than continuously visible translations. Color correspondence between source and translated sentences reduced readers' reported difficulty keeping their place [16]. EyeLingo opened word explanations based on gaze and reduced reading time relative to click-based access, although readers described incorrectly triggered explanations as distracting [8]. In scientific reading, ScholarPhi places definitions of terms and symbols beside their point of use and links them to the defining sentence. Readers answered lookup questions faster and consulted less of the document than with a basic PDF reader, with no detected difference in answer accuracy [14]. The effort involved in consulting assistance extends to reading the response and locating the passage it explains.

Readers also use large language models (LLMs) to request interpretations of words and passages from their documents. Information workers interviewed for LingoQ described copying material from work documents into chatbots for lookup, translation, and proofreading [34]. In a reading course for learners of English as a foreign language, an embedded chatbot improved reported reading engagement and use of self-regulated reading strategies relative to the same platform without the chatbot [25]. Subramonyam et al. identify difficulties in anticipating a model's capabilities, expressing an intended outcome, and evaluating whether its response meets that intention [28]. Tankelevitch et al. relate prompt formulation to the work of identifying one's goals and expressing them as instructions [29]. Obtaining assistance through open chat requires readers to identify what requires clarification and formulate that need as a request.

In Reading.help, second-language readers select text and choose among generated explanations of vocabulary, grammar, and passages. The system also recommends potentially difficult passages and provides paragraph summaries. Its design includes concise explanations with optional further detail and aims to support independent interpretation and learning [7]. Its explanations are requested through three fixed options that apply to any selection.

Paper Plain combines definitions and section summaries with curated questions that lead readers to relevant passages in medical papers. Its full interface and question-focused version reduced perceived difficulty and increased perceived understanding, with comprehension meeting the study's non-inferiority criterion relative to a PDF reader [2]. For biology learning, Inquire Biology suggests questions for highlighted textbook passages and offers further questions with its answers [6]. Koć-Januchta et al. subsequently found that the number of questions students composed was associated with retention; use of suggested questions showed no statistically significant association [19]. Qlarify's deployment offered three fixed questions informed by formative work alongside a question generated for the selected text. Readers predominantly selected fixed questions during that deployment. In a subsequent comparison, Qlarify users asked more questions than users of an unstructured question interface [11]. Readers also reported that verbose or inaccurate expansions could interrupt reading.

READLENS builds on contextual explanation for second-language reading [7] and on in-place question-based assistance [2, 11]. Its prewritten questions reflect the linguistic and contextual support needs identified in our formative study of second-language readers. This work concerns readers whose immediate goal is understanding an article and who decide which aspects of a word, expression, or passage to explore before continuing. We examine the assistance they request and its relationship to their experience of reading the article.

3 Formative Study

3.1 Participants and Procedure

We recruited readers through personal contacts, social media, and flyers distributed at language schools and universities in a city in northern Italy. Participants included people studying English and people reading English for other purposes. We administered a survey online through Google Forms in January and February 2026 and received 89 responses. We asked about English proficiency, reading frequency, purposes, devices, difficulties, and support practices through single-choice and multiple-choice items and two open questions. Respondents' mean age was 26.4 years (range 18–62), and 55 were aged 18–24. Table 1 reports their self-described proficiency and language-study status. Forty-nine respondents read in English daily and 16 several times a week.

Table 1. Formative participant characteristics. English proficiency was self-reported using the Common European Framework of Reference for Languages. Studying English follows the survey's current-study category and the reader interviews' combined formal and self-directed study category. Three experts in applied linguistics and English teaching were interviewed separately.

Characteristic	Survey ($n = 89$)	Reader interviews ($n = 10$)
<i>English proficiency</i>		
A1–A2	5	0
B1	10	2
B2	36	3
C1–C2	36	5
Unspecified	2	0
<i>English study status</i>		
Studying English	27	6
Not studying English	60	4
Other	2	0

Ten readers aged 22–35 participated in individual, approximately 40-minute semi-structured interviews conducted in Italian through Microsoft Teams. We began each interview by revisiting the reader's survey responses and asked open questions about seven topics: language-learning background, reading habits and contexts, approaches to different texts and purposes, recurring difficulties, strategies, existing tools, and desired assistance. We varied the order of topics and asked follow-up questions according to participants' responses. In an observed reading activity, each reader also read an English text selected for their self-reported proficiency and verbalized uncertainties encountered during reading. The texts varied across readers, and we used the activity to discuss particular interpretations and support decisions. Interviews were recorded with consent and transcribed.

We conducted three further semi-structured interviews with experts in applied linguistics and English teaching. We asked about students' difficulties, the role of reading in language learning, teaching practices, digital tools, and independent study.

We summarized the survey responses descriptively and used the interview accounts to identify recurring difficulties, support practices, and requirements. In the formative report, we grouped the reader accounts by these topics and considered variation in proficiency and reading purpose. We used the expert accounts as a secondary perspective on the reader accounts when selecting design requirements.

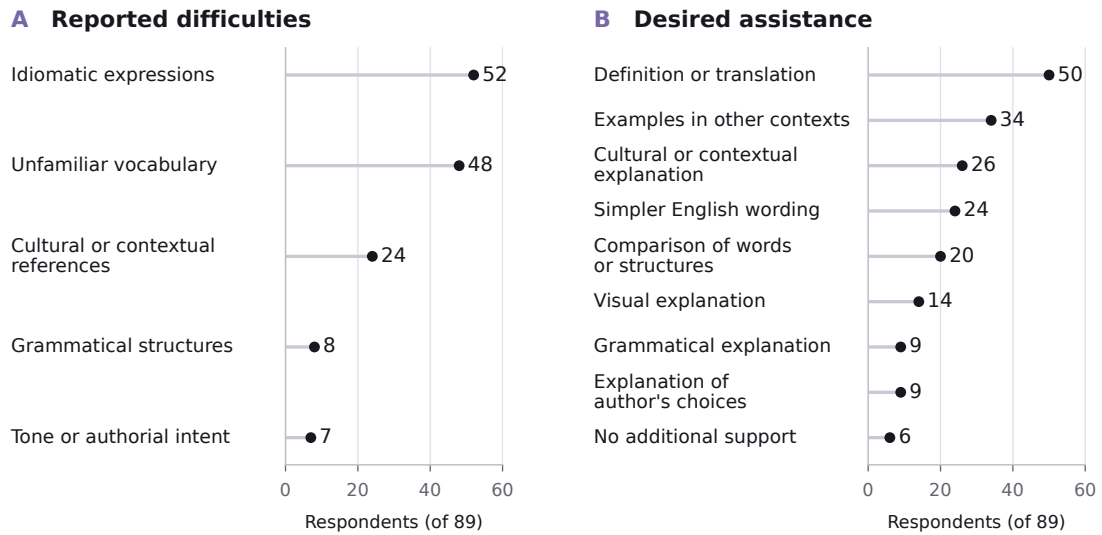


Fig. 2. Formative survey responses for (A) reported difficulties and (B) desired assistance. Both panels use the same respondent-count scale. Each question permitted multiple selections among the 89 respondents. Labels translate the categories documented in the formative report.

3.2 Reading Difficulties and Support Practices

Survey respondents most frequently selected idiomatic expressions and unfamiliar vocabulary as difficulties (Figure 2A). Interviewees described uncertainty about which sense of a familiar word applied in a passage, unfamiliar informal or specialized vocabulary, and expressions whose meaning depended on several words. Interviewees also described references for which a translation left historical, political, or other background unexplained.

Forty-five of the 89 survey respondents reported learning-oriented activities during reading with another primary purpose. Interviewees described using dictionaries, translation tools, and conversational assistants according to the uncertainty they wanted to resolve. Looking up a word could establish a possible meaning, while checking a longer passage could help readers assess their interpretation of a sentence. One interviewee valued the integrated lookup available on an e-reader because they could consult a meaning without leaving the text. Interviewees also described moving between reading and external support or postponing a lookup until later.

Alongside definitions and translations, survey respondents requested examples and contextual explanations (Figure 2B). In the interviews, readers described wanting examples that clarified usage, explanations of why a construction appeared in a sentence, or discussion of a concept encountered in the article. Some interviewees also wished to retain the explanation alongside the passage for later reference.

Interviewees also questioned whether they had understood the relationships between sentences and the meaning of a passage as a whole. One B1 reader described using translation to check an interpretation, and two of the three B2 interviewees discussed difficulty following connections across the text.

3.3 Design Requirements

Interviewees described meanings that depended on several words and valued assistance available in place. We therefore designed assistance that accepts words and longer spans and provides an initial contextual translation beside the selection. Readers can similarly select words or text segments to obtain assistance in SmartReader [3].

Survey respondents requested examples and comparisons, and interviewees wanted explanations of how words and constructions were used. We made further explanation optional through prewritten questions and requests composed by readers. We organized the questions into five selection categories for implementation: vocabulary, idiomatic expression, verbal construction or phrasal verb, cultural reference, and passage (Table 2). We combined distinctions in the language selected with the forms of explanation readers requested to organize the available assistance. A brief response followed by optional detail follows the staged presentation adopted in Reading.help [7], and prewritten questions beside the selection follow the approach of Qlarify [11].

Survey respondents requested cultural or contextual explanations and simpler English wording, and interviewees described difficulty following connections across a text. We included questions about background and relevance for cultural references and questions requesting summaries, detailed explanations, and paraphrases for longer passages.

Interviewees also wished to retain explanations alongside the passage for later reference. We attached the initial translation to the selected text and enabled readers to save further responses in the same annotation. Figure 3 shows these four relationships between reader needs and implemented support.

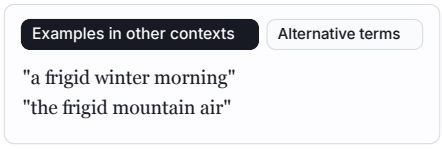
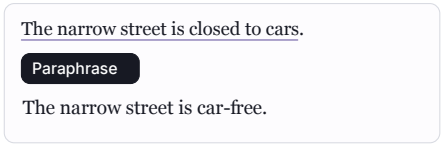
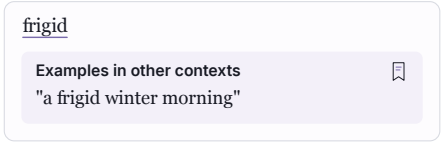
Reader accounts	Implemented support	Illustrative interaction
 Interpret the selected word or expression	Contextual translation → Select a word or span to receive a brief response beside the text.	
 Understand use and differences	Optional questions → Request examples or compare alternative expressions.	
 Understand the passage	Background and passage explanations → Ask about a reference or request an explanation of a longer span.	
 Return to an explanation	Saved annotations → Keep an answer with its selection and reopen it while reading.	

Fig. 3. We used reader accounts to design the support implemented in READLENS. Each row connects a paraphrased need to a design choice and an illustrative interaction. The examples show a contextual translation, a request for examples, a passage paraphrase, and a saved annotation. Labels are shown in English. Text excerpts and responses were constructed for this illustration.

4 ReadLens

READLENS is a web application for reading English articles with contextual language assistance. Readers can select words, expressions, or passages and receive a brief Italian translation beside the selection. An expandable panel offers further questions, and annotations retain translations and saved answers at their source. We designed these interactions according to the requirements described in Section 3.

4.1 Interaction Design

Selecting text produces an Italian translation beside the selection (Figure 4, view 1). The response contains the translation of the selected span, with surrounding text supplied to the model to support selection of the intended sense. Readers can select a single word, an expression, or a passage up to 1,000 characters. Clicking outside the response closes it. The *Approfondisci* control, translated as “Explore further,” opens a panel within the article for additional requests.

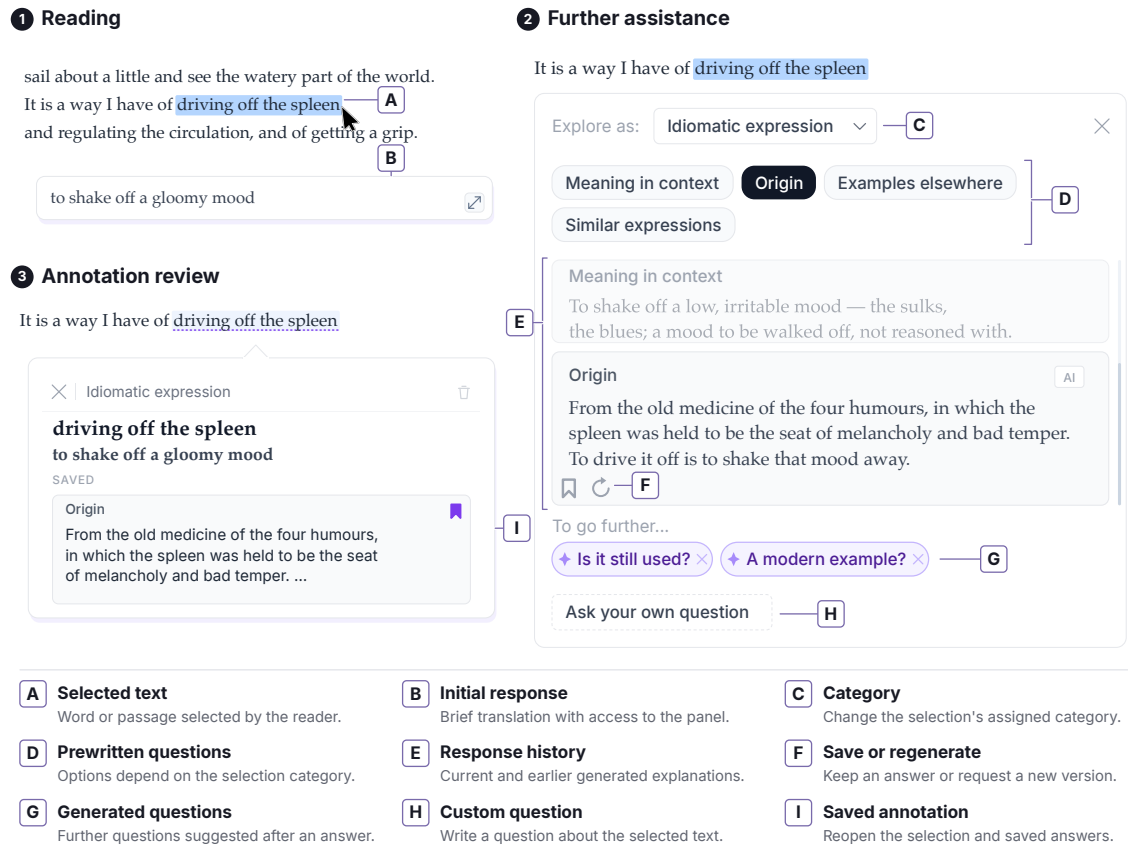


Fig. 4. Interface components of READLENS in the reading view (1), assistance panel (2), and annotation review (3). Letters A–I refer to the component key beneath the views. The key distinguishes prewritten questions, generated follow-up questions, and custom-question entry. Interface labels and illustrative responses are shown in English.

When the panel opens, READLENS assigns the selection to one of five categories. Short selections are classified as vocabulary, an idiomatic expression, a verbal construction or phrasal verb, or a cultural reference. Selections of eight or more words are assigned the passage category. Each category provides three or four prewritten questions (Table 2). For short selections, readers can change the category through a menu and access the associated questions.

Selecting a prewritten question appends a response beneath the question buttons. After each answer, the model proposes two or three further questions about the selection. These generated questions appear below the answer, followed by a text field for reader-written requests. Successive requests use the conversation associated with that selection, so readers can ask about an earlier explanation. Selecting an already answered prewritten question scrolls to its stored response. Explanations are requested in Italian, while quoted language, examples, and English paraphrases retain the source language. Readers can request a regenerated answer and provide feedback on the preceding response.

The initial translation is stored in an annotation linked to the selection’s position in the article. The selected text stays underlined, and hovering over the annotation displays its contents. Readers can add individual answers from the panel to the annotation and remove saved answers. Reopening the annotation restores its earlier requests and responses.

4.2 Implementation

READLENS uses a React and TypeScript client, a Node.js and Express server, and PostgreSQL storage. The client records the selection’s character offsets and includes up to 60 characters on each side as context. A language model generates the initial translation, classifies short selections, and answers panel requests.² Structured response schemas specify the translation field, category, answer text, and generated questions.

Each prewritten question has a category-specific instruction that specifies the requested explanation. The first request includes the selected text and its context; subsequent requests include the conversation history. Passage summaries use only the selected span, while passage explanations and paraphrases identify the selection within its surrounding text. The database stores annotations, requests, generated responses, and timestamped interaction events under participant identifiers.

²The implementation uses the API model identifier gpt-4o-mini. Temperature is 0.2 for translation, 0.1 for classification, and 0.5 for panel responses. Translation and panel calls have output limits of 600 and 1,000 tokens, respectively.

Table 2. Requests available in the panel for each selection category. Labels are translated from Italian.

Category	Available requests
Vocabulary	Meaning in context; examples in other contexts; alternative terms
Idiomatic expression	Meaning in context; origin; examples in other contexts; similar expressions
Verbal construction or phrasal verb	Meaning in context; grammatical explanation; examples in other contexts; alternatives
Cultural reference	Relevance to the text; background; further detail
Passage	Summary; detailed explanation; paraphrase

5 Evaluation

5.1 Participants and Design

Eighteen participants completed two reading blocks, one with READLENS and one with a dictionary-plus-LLM baseline, followed by a delayed vocabulary assessment. We recruited through personal contacts, social media, and flyers distributed at language schools and universities in a city in northern Italy. Recruitment targeted adults whose first language was Italian, who reported B2–C1 English proficiency and regularly read English on a computer. Before the reading session, participants completed the 3,000-, 4,000-, and 5,000-word-frequency bands of the updated Vocabulary Levels Test [30]. All 18 participants met the screening threshold of at least 29 correct answers out of 30 in the 3,000-word band.

Table 3 lists the participants with their condition and article orders, vocabulary-test scores, and questions submitted. Each participant read a different article in each condition. Four assignment groups crossed condition order with article order and contained five, five, four, and four participants. The application recorded consent before the reading session. Its consent information described voluntary participation, withdrawal and data-deletion rights, audio recording of the interview, and storage under participant identifiers.

Table 3. Evaluation participants. Interview labels P01–P18 were matched to interaction logs through the session-completion timestamps recorded by the application, which coincide with the start times of the recorded interviews (P16 by elimination). VLT reports correct answers out of 30 in the 3,000-, 4,000-, and 5,000-word bands. Questions are those submitted through READLENS’s panel and through baseline chat during the reading blocks.

	Condition order	Article order	VLT 3k/4k/5k	READLENS questions	Chat queries
P01	READLENS, baseline	Arctic, car-free	30/30/30	2	0
P02	baseline, READLENS	Arctic, car-free	29/29/28	1	2
P03	READLENS, baseline	car-free, Arctic	29/23/28	3	0
P04	baseline, READLENS	car-free, Arctic	30/27/26	1	3
P05	READLENS, baseline	Arctic, car-free	30/30/30	1	0
P06	baseline, READLENS	Arctic, car-free	29/30/26	3	0
P07	READLENS, baseline	car-free, Arctic	30/30/30	2	0
P08	baseline, READLENS	car-free, Arctic	29/30/29	7	0
P09	READLENS, baseline	Arctic, car-free	29/29/29	0	0
P10	baseline, READLENS	Arctic, car-free	30/30/30	3	0
P11	READLENS, baseline	car-free, Arctic	30/25/26	0	3
P12	baseline, READLENS	car-free, Arctic	30/30/27	0	0
P13	READLENS, baseline	Arctic, car-free	29/29/27	0	0
P14	baseline, READLENS	Arctic, car-free	29/30/30	1	0
P15	READLENS, baseline	car-free, Arctic	30/30/30	4	3
P16	baseline, READLENS	car-free, Arctic	29/29/30	0	0
P17	READLENS, baseline	Arctic, car-free	29/25/28	14	1
P18	baseline, READLENS	Arctic, car-free	30/29/29	17	4

5.2 Conditions and Reading Materials

The baseline displayed the article with a bilingual dictionary available by double-clicking a word and a fixed LLM chat panel beside the text. Dictionary entries included Italian translations, definitions, and examples. The entries were generated offline and served from a local cache. The first chat request automatically included the complete article, and

subsequent requests retained the conversation history.³ The system instruction requested natural Italian, contextual interpretation of idioms and phrasal verbs, and answers of at most two or three sentences.

The two conditions combined different selection ranges, response locations, question-entry mechanisms, context windows, and annotation facilities. The comparison therefore concerns the complete interfaces. Participants read two English-language travel articles, “Photographing one of the most remote places on Earth” and “It’s like being in Greece’: The US neighbourhood where cars are banned.” The study excerpts contained 802 and 861 whitespace-delimited words. Each article had 12 researcher-selected vocabulary targets, including individual words and expressions. The targets defined the vocabulary assessments, while participants could request assistance for any part of either article.

5.3 Procedure

After consent, participants completed a vocabulary pretest covering all 24 vocabulary targets. Each reading block began with practice using the assigned interface on a separate 136-word passage. Participants were instructed to read for understanding and to use assistance as needed. Participants then read the assigned article and completed workload and usability questionnaires, six comprehension questions, and immediate vocabulary assessments for that article’s 12 targets. Participants repeated this sequence with the other interface and article in the second block. A semi-structured interview in Italian followed the two blocks, during which participants discussed further requests, unanswered needs, interruptions, and differences between the reading experiences.

The follow-up repeated the vocabulary assessments for all 24 targets through a self-administered online session. The application scheduled access for calendar days seven through ten after the reading session. Follow-up instructions asked participants to answer from memory without dictionaries, translators, or other support. The consent information described this activity as an online follow-up and disclosed vocabulary testing in the main session. The closing debrief explained that the follow-up measured memory for previously encountered words.

5.4 Measures and Interaction Logs

Interaction logs recorded text selections, dictionary lookups, panel openings, submitted questions, category changes, and annotation actions. Events included the participant, condition, article, study step, and timestamp. We counted questions submitted through READLENS and baseline chat separately from initial selections and dictionary lookups. Within READLENS, we matched selection offsets and annotation identifiers chronologically to distinguish opening the panel after a new selection from reopening an annotation. The question count records requests that generated an answer. Selecting an already answered button returns to the stored response and produces no new question event.

The adapted Vocabulary Knowledge Scale (VKS) used four response levels, from having no recollection of a word to reporting knowledge of its meaning [26]. Participants choosing level three or four supplied a definition or translation. Scores combined the reported level with the accuracy of that definition on a three-point scale. No recollection scored one, and recognition without a reported meaning scored two. A reported meaning scored three when its definition was judged correct and two otherwise. We calculated immediate and delayed changes relative to each target’s pretest score. A four-option Word Recognition Test (WRT) presented each target within its original sentence and asked participants to identify its meaning. WRT scores were the proportion of correct responses among the 12 targets associated with a condition.

³Baseline chat uses the API model identifier gpt-4o-mini, temperature 0.2, and a 500-token output limit.

Six multiple-choice questions per article assessed comprehension, comprising three literal and three inferential questions. Scores were the proportion answered correctly. Workload was measured with the six unweighted NASA Task Load Index subscales, each ranging from 0 to 100, with higher scores indicating greater workload [13]. Usability was measured with the ten-item System Usability Scale (SUS), using Italian wording [4, 5]. The standard SUS scoring transformation produces a score from 0 to 100, with higher scores indicating greater perceived usability.

5.5 Analysis

For RQ1, we summarized the questions submitted and the number of readers using each form of assistance. For RQ2, we summarized vocabulary, comprehension, usability, and workload scores by participant and condition, alongside interview accounts of the reading experience. Reading analyses retained events from the two assigned article-reading blocks for the 18 study participants. Counts included participants who used no assistance. We report counts, proportions, condition means, and mean within-participant differences.

An AI-assisted coding pass examined all 75 submitted questions using their wording and the selected text or article reference. Each question received a function code describing the assistance requested, such as a meaning, an alternative expression, or a summary. A separate scope code identified whether the request concerned a word or expression, a sentence or passage, a topic in the article, or the whole article. The coding record retained a rationale and plausible alternatives for ambiguous requests. An author reviewed the assignments against the question wording and the selected text.

A reader's repertoire was the number of distinct functions requested in a condition. We compared repertoires across all 18 readers and examined the distribution of question counts, each question user's first request, and counts by article and condition order. Two sensitivity groupings applied the same category mergers to both conditions. The first combined meanings, translations, and explanations of unfamiliar referents. The second also combined summaries, paraphrases, passage explanations, and article-content questions. These groupings tested how the descriptive contrast depended on distinctions between closely related requests.

For the 46 READLENS annotations with submitted questions, we reconstructed the selected text, initial translation, subsequent questions, and generated answers. An AI-assisted review recorded what information each answer added or changed and any apparent mismatch with the surrounding passage. Classification events, category changes, and panel openings and closures established which prewritten questions were available. The review treated request sequences as observations of accessed information, and readers' reasons required explicit interview evidence.

An AI-assisted review examined all 18 interview transcripts using a participant-level matrix. The matrix recorded reasons for further assistance, sufficiency of the initial translation, interface preferences, article-level review, and interruptions. Extracts retained the interviewer question and distinguished reported session use from hypothetical preferences and uncertain recollections. We use these accounts to interpret behavioral patterns at the study level. Interview labels were linked to the interaction logs through the session-completion timestamps, which coincide with the recorded start times of the interviews (Table 3). English translations were checked against the Italian transcripts; repetitions and fillers were omitted for readability, and brackets mark editorial clarifications. The coding assignments and translations are subject to author review.

One author rated all 914 definitions supplied at levels three and four as correct or incorrect with respect to the target word. A second rater independently rated a random 20% sample of 182 definitions, giving substantial agreement ($\kappa = 0.69$).

For delayed assessments, we calculated mean time per item from the submission timestamps of consecutive study steps. We flagged a block when this time exceeded 1.5 times the participant’s initial value and the block contained a recorded window-focus or visibility loss. The reference value was the pretest time per item for VKS and the time per item across both immediate assessments for WRT. The delayed VKS sensitivity analysis excluded each flagged reader’s observations from both conditions. The flags describe timing and focus changes and leave uncertainty about whether participants consulted external assistance.

6 Results

6.1 The Questions Readers Requested (RQ1)

Thirteen readers submitted 59 questions with READLENS, and six submitted 16 through baseline chat. Under this coding, meaning or translation was the most frequent function in both conditions (Figure 5). Baseline chat also received requests for an explanation of a referent, summaries, and information about article content.

Combining meanings and referent explanations in both conditions, and grouping summaries, paraphrases, passage explanations, and content questions together, retained 19 READLENS requests for functions absent from baseline chat. These requests came from eight readers and concerned alternatives, examples, origins, grammar, and relevance.

Ten of the 18 readers requested more distinct functions with READLENS (Figure 6). Two readers submitted 31 of its 59 questions, indicating concentration in the aggregate counts. READLENS questions were more numerous for both articles and in both condition positions.

6.2 From Initial Translation to Further Questions (RQ1)

All READLENS questions used initial prewritten buttons. Four readers began with a request for alternatives, examples, or relevance. Of the 46 annotations with questions, 34 received a single question. Successive requests included contextual meaning followed by alternatives and alternatives followed by contextual meaning.

For “traffic-clogged asphalt that disgorge solo commuters into soul-crushingly monotonous suburbs,” R11 first requested a paraphrase, which returned a simpler English formulation, then an Italian explanation relating the passage to car-centred urban design, and finally an Italian summary. The answers also incorporated information from the surrounding paragraph.

Initial translations of “cornhole” retained the English game name. R07 requested relevance to the text, receiving an explanation of the social atmosphere described in the scene, then requested context, receiving a description of throwing beanbags at a board. In an independent interview account, P06 recalled receiving a translation while still wondering what the game was.

For “containers on skates,” R01 requested contextual meaning and then alternative terms. The first answer identified the dog sleds described in the passage. The second offered “sleds,” “trolleys,” “carts,” and “platforms on wheels,” with notes about their use. The wheeled alternatives changed the physical image of transport across ice.

P02 associated READLENS with obtaining assistance “without the stress of having to sit there and write the right question, make the right request.” P06 valued having topics available to explore, and P17 described investigating cultural background.

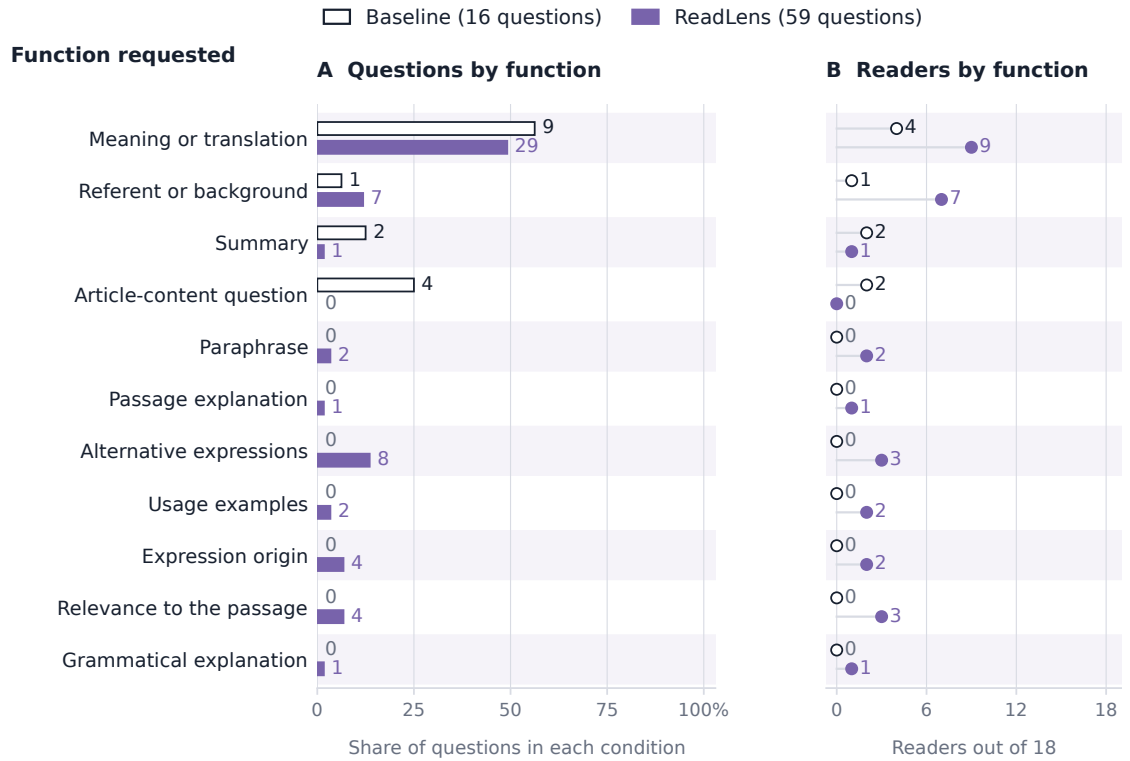


Fig. 5. Functions of submitted questions. **A.** Bars show shares within each condition; labels give question counts. **B.** Points show the number of readers requesting each function, out of 18. A reader can appear in several rows. READLENS questions used initial prewritten buttons; baseline questions were typed. Function assignments follow the coding described in Section 5.

“With the other tool [ReadLens], if I wanted to look into something I was curious about, I could find it more easily because the tool gave me information about it, for example, if I wanted to explore the cultural context.”

(P17)

P17 also reported fatigue during the second reading block, which used the baseline, and perceived that article as more difficult.

P12 compared the interpretations of “whisk” encountered in the two reading blocks.

“I clicked ‘whisk’ and it gave me ‘frusta’ [a kitchen whisk]. With the second [ReadLens], it gave me the correct meaning in context, ‘take,’ ‘transport.’”

(P12)

P05 and P08 described requesting further assistance when an initial READLENS translation seemed inconsistent with the sentence. P08’s account concerned “if bony” in a passage about cod. The corresponding logged sequence began with “se magro” (“if thin”), followed by a request for contextual meaning and an explanation referring to fish bones. P08 recalled, “It did not make sense in that context, so I explored further and it told me ‘full of fish bones.’”

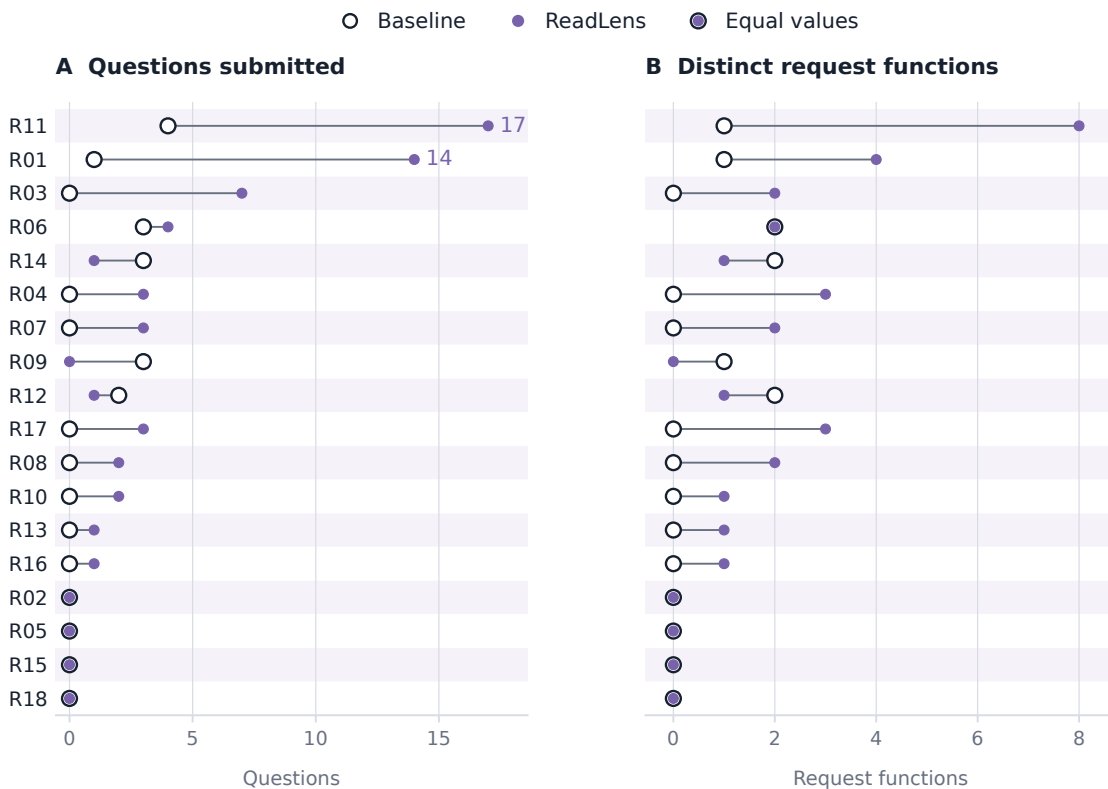


Fig. 6. Paired observations for all 18 readers, including nonuse. **A.** Submitted questions. **B.** Distinct requested functions under the detailed coding. Lines connect the same reader across conditions; concentric markers indicate equal values. Rows are ordered by total questions across both conditions, with ties ordered by log alias. Log aliases R01–R18 are separate from the interview labels P01–P18.

6.3 Sufficiency, Reading Purposes, and Outcomes (RQ2)

P09, P11, and P16 described the initial translation as sufficient for their reading. P16 explained, “I used only the translation and did not use the other features because, at that moment, I did not need them,” while preferring to have further assistance available for more difficult texts. P15 described staying with the translation when it fitted the sentence and requesting synonyms to connect an expression with a familiar word.

P05 and P14 preferred the baseline dictionary because they could usually contextualize a supplied meaning themselves. P05 explained, “I can then fit it back into the context easily.” Both also described using READLENS to investigate a meaning in context. P03 associated the baseline with quick reading and READLENS with more detailed reading.

P04 reported asking baseline chat to summarize the final two technical paragraphs. P15 used baseline chat during and after reading and described requesting the article’s main points.

“[I asked for] the main points of the text, because that is something I habitually do, even when reading in Italian, to help me remember.”

(P15)

P15 preferred baseline chat for orientation and review. P01 associated READLENS with everyday reading and baseline chat with studying, while expressing uncertainty about whether READLENS could accept a whole paragraph.

P18 reported occasional interruptions while exploring explanations, saying, “Sometimes, because I spent a while exploring the explanations, but then I could resume.” P18 tentatively associated these interruptions with the greater number of expressions in the READLENS article and preferred the baseline’s immediacy. P17 reported focusing on words and overlooking the article’s overall meaning, while explaining uncertainty about whether comprehension questions would follow. P04 described stopping mid-sentence for a translation as an interruption and using baseline chat at the end of reading.

Mean usability ratings were higher and workload ratings lower with READLENS, while mean comprehension and delayed word-recognition accuracy were higher with the baseline (Table 4). VKS scores increased from the pretest to both later assessments in each condition, by 0.62 and 0.57 points immediately and by 0.73 and 0.69 points at follow-up. Excluding the three readers flagged for timing and window-focus or visibility changes during the delayed VKS left a mean between-condition difference in delayed gain of -0.01 points.

Table 4. Mean vocabulary, comprehension, and questionnaire scores across 18 readers. Differences are READLENS minus baseline; accuracy differences are in percentage points. Higher SUS indicates greater perceived usability; lower NASA-TLX indicates lower workload.

Measure	READLENS	Baseline	Difference
VKS immediate gain (1–3)	0.62	0.57	0.05
VKS delayed gain (1–3)	0.73	0.69	0.04
Word recognition, immediate (%)	80.6	79.2	1.4
Word recognition, delayed (%)	80.1	84.3	−4.2
Comprehension (%)	88.0	91.7	−3.7
SUS (0–100)	94.3	90.8	3.5
NASA-TLX (0–100)	13.1	16.2	−3.1

7 Discussion

7.1 Suggested Questions and the Assistance Readers Request

For RQ1, readers used READLENS to request alternatives, examples, origins, grammar, or relevance, functions absent from the observed baseline chat even when closely related functions were combined. Correspondence between a menu category and a submitted question also reflects how we organized the interface and does not independently validate the categories as a classification of readers’ needs.

P02 described the effort of writing the right request, and P06 valued having topics available to explore. We distinguish choosing what information could help from expressing that choice as an instruction. Tankelevitch et al. discuss both awareness of task goals and their formulation as instructions in their account of the metacognitive demands of generative AI [29]. With READLENS, a reader could recognize an example or a comparison as relevant and select the corresponding question. Initial requests for alternatives, examples, or relevance leave open whether readers had already intended to seek that information or considered it after seeing the options.

All READLENS questions used the initial prewritten buttons. In scientific reading, Fok et al. also reported predominant use of fixed questions during Qlarify’s deployment [11]. In READLENS, the initial buttons appeared above the responses,

whereas generated questions became available only after an answer and the free-question field appeared below the answer area, so the three controls differed in availability and placement. The logs record the requests readers submitted rather than the controls they considered. A comparison of question formats would need to provide comparable opportunities to encounter each option before attributing differences in use to the wording or origin of the questions.

The aggregate repertoire partly reflects extended use by two readers. The available questions also depended on the selection category, so request frequencies reflected both the selected passages and the explanations offered. A comparison of alternative menus with the same selection and response mechanisms could examine whether readers obtain the explanations they seek and whether offered questions are relevant to their reading purpose.

7.2 Deciding Whether an Interpretation Is Sufficient

In the “cornhole” sequences, the initial translation retained the game’s name, while further answers described the activity or its role in the scene. P06 independently described receiving a translation while still wondering what the game was. A translation can provide an equivalent expression while leaving the reader without the background needed to interpret the passage. P17 also described investigating cultural background out of curiosity. Such an interest can coexist with uncertainty about the text, but a request for background alone does not establish either purpose.

P08 requested further assistance because the initial translation of “if bony” did not fit the passage about cod. The subsequent explanation referred to fish bones, correcting the preceding interpretation of thinness. P08 could identify the inconsistency by comparing that answer with the sentence and used another question to investigate it. Tankelevitch et al. distinguish prompt formulation from evaluating generated outputs and note that evaluation is still necessary when users do not write prompts themselves [29]. In P08’s exchange, a prewritten question supplied the wording needed to request an explanation, while P08 judged whether the answer fitted the passage. A request for further detail can therefore be part of checking the assistance itself, so its frequency reflects verification as well as interest in additional content.

For “containers on skates,” the contextual answer identified the dog sleds described in the article, while the subsequent alternatives included wheeled vehicles. These alternatives differed in their fit to the original scene. Assessing an initial translation is therefore one of the occasions for checking generated assistance; later explanations also require consideration of their relation to the source text. Designers could distinguish alternatives that preserve the meaning of the selected passage from related expressions appropriate elsewhere. A response that revises an earlier interpretation could also identify the changed meaning and the words in the passage that support it.

For RQ2, P09, P11, and P16 described the initial translation as adequate for their immediate reading. P16 preferred to have further assistance available for more difficult material despite not using it during the session. P05 and P14 preferred dictionary meanings they could contextualize independently and also described using contextual explanations for particular expressions. These preferences concerned how much assistance was useful for a passage and how much interpretation readers were willing to perform themselves. We therefore treat sufficiency as a judgment made in relation to the current text and reading purpose. A brief response and optional explanation allow readers to act on that judgment, including by resuming reading. Continuing without another request can accompany adequate assistance, although the absence of a request alone does not establish understanding.

P15 described requesting synonyms to connect an expression with a familiar word. We also observed alternatives followed by contextual meaning and a passage paraphrase followed by an explanation and summary. In second-language acquisition research, Yanagisawa and Webb found a stronger contribution to vocabulary learning from evaluation requirements, including comparing meanings in context, than from searching for a meaning [32]. P15’s account

illustrates a possible use of additional information for such comparison. Relating further questions to learning would require examining how readers use the explanations and what they subsequently retain about the expressions concerned.

7.3 Coordinating Local Clarification and Article-Level Review

P04 used baseline chat to summarize the final two technical paragraphs. P15 used baseline chat during and after reading, including a request for the article's main points, and described requesting main points as a familiar practice even when reading in Italian. These requests concerned reviewing and organizing content across the text, in addition to interpreting unfamiliar language. P03 likewise associated the baseline with quick reading and READLENS with more detailed reading. Support for second-language reading can address these purposes at different moments, with local clarification during reading and broader review when readers want to consolidate the content.

P01 was unsure whether a whole paragraph could be selected in READLENS. P01's account therefore bears on the visibility of passage selection as much as on interest in it. In Reading.help, Chung et al. combine explanations of selected text with paragraph summaries [7]. READLENS also provides summaries for selected passages, while baseline chat receives the complete article. We propose making both the selection range and the supplied context visible when a request is composed. A reader could then distinguish an explanation of the selected expression from a summary or review requiring a larger portion of the article.

P04 described a mid-sentence interruption, and P18 reported spending time exploring explanations before resuming the article. Assistance was in place, and consulting it still required attention to another text. In an e-reader evaluation, Draxler et al. observed longer reading time with detailed grammar overlays than with translation alone and found no statistically significant difference in global reading flow [9]. Time spent on an explanation and the experience of reading continuity therefore require separate consideration. In our interviews, P18 could resume reading and tentatively linked interruptions to the article's expressions. P17 described attending to words while overlooking the article's overall meaning, with uncertainty about the comprehension task also affecting this account.

READLENS already retains translations and saved explanations at their source passages. Designers could use these records to support both immediate return to the interrupted sentence and later review of explanations in article order. A return control could restore the source sentence after an extended answer, while a review view could retain enough surrounding text to relate saved explanations to the article. These proposals would require evaluation with readers who choose when to clarify and when to review. Assessment could examine whether readers recover the sentence they were interpreting and can relate the explanation to the surrounding content.

8 Limitations

We conducted the evaluation with 18 participants reading two English-language travel articles during one session, with a subsequent vocabulary follow-up. Recruitment targeted Italian-first-language readers reporting B2–C1 proficiency. The observed requests reflect these readers, the vocabulary of the selected articles, and the requirements of a researcher-organized session. We developed the question set from the formative survey and interviews within their recruitment scope. Readers with other first languages, lower English proficiency, or specialist reading goals may require different questions and explanations.

The interface conditions varied jointly in selection range, response location, suggested questions, annotation persistence, model instructions, context supplied to the model, and generation parameters. Baseline chat received the complete article, while READLENS supplied the selected text with surrounding context. The observed differences therefore concern the two implementations as used. The observed differences therefore concern the two interfaces

as implemented, and separating the contribution of question suggestions from in-place presentation would require a component comparison. The request repertoire also depends on how the offered options and freely composed requests are grouped and on the number of questions submitted in each condition.

Participants encountered the vocabulary targets in a pretest and knew that vocabulary assessments formed part of the session. P04 explicitly described paying more attention to words during the second reading after experiencing the first assessment. P08 reported checking words remembered from the pretest. Reading therefore took place within an assessment-aware setting. Vocabulary outcomes concern the 24 selected targets and repeated assessments; comprehension outcomes concern six questions per article. We used recorded timing and focus-loss events for sensitivity checks on the self-administered follow-up, with uncertainty about participants' activity outside the application.

The request coding, review of 46 questioned annotations, and participant-level review of all 18 interviews were AI-assisted. Their records retain source excerpts, coding rationales, ambiguous assignments, and qualifications from the interview context. An author checked the resulting assignments against these records. The link between interview labels and interaction logs rests on matching timestamps, and one participant was assigned by elimination. The vocabulary-definition ratings come from one rater, with a second rater on a 20% sample.

Through close reading of the responses, we identified information added or changed relative to the initial translation and the supplied passage. In the “if bony” sequence, a misleading initial response was followed by a contextual clarification. One author rated a random 20% sample of generated responses on a three-level scale. Of 40 sampled initial translations, 33 were rated correct and appropriate to the passage, two partially correct, and five incorrect, the errors concerning single adjectives such as *stark* rendered as *drastico*. All 11 sampled panel answers and all three sampled baseline answers were rated correct. These ratings come from one rater, and the logs record the responses supplied rather than readers' acceptance of them.

9 Conclusions

READLENS converts findings about second-language readers' assistance needs into prewritten questions for selected words and passages. In an 18-participant comparison with a dictionary and LLM side panel, readers used these questions to request forms of assistance, alternatives, examples, origins, grammar, and relevance, that no baseline request contained. In interviews, readers described unresolved meanings, checks on generated answers, curiosity, and sufficient initial assistance as reasons for continuing or stopping. We contribute READLENS and a descriptive account connecting available questions, the assistance readers request, and their judgments of what is needed to understand a passage or review an article.

References

- [1] Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. 2003. Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research* 73, 3 (2003), 277–320. doi:10.3102/00346543073003277
- [2] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30, 5, Article 74 (2023), 38 pages. doi:10.1145/3589955
- [3] Mahmoud Azab, Ahmed Salama, Kemal Oflazer, Hideki Shima, Jun Araki, and Teruko Mitamura. 2013. An NLP-based Reading Tool for Aiding Non-native English Readers. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, Hissar, Bulgaria, 41–48. <https://aclanthology.org/R13-1006/>
- [4] Simone Borsci, Stefano Federici, and Marco Lauriola. 2009. On the Dimensionality of the System Usability Scale: A Test of Alternative Measurement Models. *Cognitive Processing* 10, 3 (2009), 193–197. doi:10.1007/s10339-009-0268-9
- [5] John Brooke. 1996. SUS: A “Quick and Dirty” Usability Scale. In *Usability Evaluation in Industry*, Patrick W. Jordan, Bruce Thomas, Bernard A. Weerdmeester, and Ian L. McClelland (Eds.). Taylor & Francis, London, UK, 189–194. doi:10.1201/9781498710411-35

- [6] Vinay K. Chaudhri, Britte Haugan Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. 2013. Inquire Biology: A Textbook that Answers Questions. *AI Magazine* 34, 3 (2013), 55–72. doi:10.1609/aimag.v34i3.2486
- [7] Sunghyo Chung, Hyeon Jeon, Sungbok Shin, and Md Naimul Hoque. 2025. Reading.help: Supporting EFL Readers with Proactive and On-Demand Explanation of English Grammar and Semantics. arXiv preprint arXiv:2505.14031. doi:10.48550/arXiv.2505.14031
- [8] Jiexin Ding, Bowen Zhao, Yuntao Wang, Xinyun Liu, Rui Hao, Ishan Chatterjee, and Yuanchun Shi. 2025. Unknown Word Detection for English as a Second Language (ESL) Learners Using Gaze and Pre-trained Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. doi:10.1145/3706598.3714181
- [9] Fiona Draxler, Viktoriia Rakytianska, and Albrecht Schmidt. 2022. Maintaining Reading Flow in E-Readers with Interactive Grammar Augmentations for Language Learning. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*. ACM, New York, NY, USA, 53–61. doi:10.1145/3568444.3568459
- [10] Maxine Eskenazi, Yibin Lin, and Oscar Saz. 2013. Tools for Non-native Readers: The Case for Translation and Simplification. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. Association for Computational Linguistics, Atlanta, Georgia, 20–28. <https://aclanthology.org/W13-1503/>
- [11] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, Article 145, 21 pages. doi:10.1145/3654777.3676397
- [12] Aline Godfroid, Jieun Ahn, Ina Choi, Laura Ballard, Yaqiong Cui, Suzanne Johnston, Shinhye Lee, Abdhi Sarkar, and Hyung-Jo Yoon. 2018. Incidental Vocabulary Learning in a Natural Reading Context: An Eye-tracking Study. *Bilingualism: Language and Cognition* 21, 3 (2018), 563–584. doi:10.1017/S1366728917000219
- [13] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Amsterdam, The Netherlands, 139–183. doi:10.1016/S0166-4115(08)62386-9
- [14] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 413, 18 pages. doi:10.1145/3411764.3445648
- [15] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2006. Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension. In *Interspeech 2006*. ISCA, Pittsburgh, PA, USA, 829–832. doi:10.21437/Interspeech.2006-282
- [16] Tien-Yu Ho, Hao-Chuan Wang, and Shong-Hong Lai. 2018. Non-native Language Reading Support with Display of Machine Translation Based on Eye-Tracking and Sentence-Level Mapping. In *Proceedings of the Sixth International Symposium of Chinese CHI*. ACM, New York, NY, USA, 57–63. doi:10.1145/3202667.3202675
- [17] Jan H. Hulstijn, Merel Hollander, and Tine Greidanus. 1996. Incidental Vocabulary Learning by Advanced Foreign Language Students: The Influence of Marginal Glosses, Dictionary Use, and Reoccurrence of Unknown Words. *The Modern Language Journal* 80, 3 (1996), 327–339. doi:10.1111/j.1540-4781.1996.tb01614.x
- [18] Sima Khezrlou, Rod Ellis, and Karim Sadeghi. 2017. Effects of Computer-assisted Glosses on EFL Learners' Vocabulary Acquisition and Reading Comprehension in Three Learning Conditions. *System* 65 (2017), 104–116. doi:10.1016/j.system.2017.01.009
- [19] Marta M. Koć-Januchta, Konrad J. Schönborn, Lena A. E. Tibell, Vinay K. Chaudhri, and H. Craig Heller. 2020. Engaging With Biology by Asking Questions: Investigating Students' Interaction and Learning With an Artificial Intelligence-Enriched Textbook. *Journal of Educational Computing Research* 58, 6 (2020), 1190–1224. doi:10.1177/0735633120921581
- [20] Batia Laufer and Monica Hill. 2000. What Lexical Information Do L2 Learners Select in a CALL Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3, 2 (2000), 58–76. doi:10.64152/10125/25073
- [21] Batia Laufer and Jan Hulstijn. 2001. Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-Induced Involvement. *Applied Linguistics* 22, 1 (2001), 1–26. doi:10.1093/applin/22.1.1
- [22] Adina Levine, Yael Bejarano, Patricia L. Carrell, and Lily Vered. 2004. Comparing Dictionary Definitions and Glosses in Hypertext Foreign Language Reading: Facilitating Foreign Language Reading Comprehension and Vocabulary Acquisition. *The CATESOL Journal* 16, 1 (2004), 59–68. doi:10.5070/B5.36367
- [23] Mircea F. Lungu, Luc van der Brand, Dan Chirtoaca, and Martin Avagyan. 2018. As We May Study: Towards the Web as a Personalized Language Textbook. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3173574.3173912
- [24] John Nerbonne, Lauri Karttunen, Elena Paskaleva, Gabor Proszeky, and Tiit Roosmaa. 1997. Reading More into Foreign Languages. In *Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Washington, DC, USA, 135–138. doi:10.3115/974557.974577
- [25] Mengru Pan, Chun Lai, and Kai Guo. 2025. Effects of GenAI-empowered Interactive Support on University EFL Students' Self-regulated Strategy Use and Engagement in Reading. *The Internet and Higher Education* 65 (2025), 100991. doi:10.1016/j.iheduc.2024.100991
- [26] T. Sima Paribakht and Marjorie Wesche. 1997. Vocabulary Enhancement Activities and Reading for Meaning in Second Language Vocabulary Acquisition. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, James Coady and Thomas Huckin (Eds.). Cambridge University Press, Cambridge, UK, 174–200. doi:10.1017/CBO9781139524643.013

- [27] Elke Peters. 2007. Manipulating L2 Learners' Online Dictionary Use and Its Effect on L2 Word Retention. *Language Learning & Technology* 11, 2 (2007), 36–58. doi:10.64152/10125/44103
- [28] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–19. doi:10.1145/3613904.3642754
- [29] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–24. doi:10.1145/3613904.3642902
- [30] Stuart Webb, Yosuke Sasao, and Oliver Ballance. 2017. The Updated Vocabulary Levels Test: Developing and Validating Two New Forms of the VLT. *ITL – International Journal of Applied Linguistics* 168, 1 (2017), 33–69. doi:10.1075/itl.168.1.02web
- [31] Stuart Webb, Takumi Uchiyara, and Akifumi Yanagisawa. 2023. How Effective Is Second Language Incidental Vocabulary Learning? A Meta-analysis. *Language Teaching* 56, 2 (2023), 161–180. doi:10.1017/S0261444822000507
- [32] Akifumi Yanagisawa and Stuart Webb. 2021. To What Extent Does the Involvement Load Hypothesis Predict Incidental L2 Vocabulary Learning? A Meta-Analysis. *Language Learning* 71, 2 (2021), 487–536. doi:10.1111/lang.12444
- [33] Akifumi Yanagisawa, Stuart Webb, and Takumi Uchiyara. 2020. How Do Different Forms of Glossing Contribute to L2 Vocabulary Learning from Reading? A Meta-regression Analysis. *Studies in Second Language Acquisition* 42, 2 (2020), 411–438. doi:10.1017/S0272263119000688
- [34] Yeonsun Yang, Sang Won Lee, Jean Y. Song, Sangdoo Yun, and Young-Ho Kim. 2026. LingoQ: Bridging the Gap between EFL Learning and Work through AI-Generated Work-Related Quizzes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–23. doi:10.1145/3772318.3791342