



Design Opportunities for Explainable AI Paraphrasing Tools: A User Study with Non-native English Speakers

Yewon Kim
KAIST

School of Electrical Engineering
Daejeon, Republic of Korea
yewon.e.kim@kaist.ac.kr

Thanh-Long V. Le
KAIST

Kim Jaechul Graduate School of AI
Seoul, Republic of Korea
thanhlng0780@kaist.ac.kr

Donghwi Kim*

Samsung Electronics
Seoul, Republic of Korea
dh.tony.kim@samsung.com

Mina Lee[†]

University of Chicago
Department of Computer Science
Chicago, Illinois, USA
mnlee@uchicago.edu

Sung-Ju Lee[†]

KAIST
School of Electrical Engineering
Daejeon, Republic of Korea
profsj@kaist.ac.kr

Abstract

We investigate how non-native English speakers (NNESs) interact with diverse information aids to assess and select AI-generated paraphrases. We develop PARASCOPE, an AI paraphrasing assistant that integrates diverse information aids, such as back-translation, explanations, and usage examples, and logs user interaction data. Our in-lab study with 22 NNESs reveals that user preferences for information aids vary by language proficiency, with workflows progressing from global to more detailed information. While back-translation was the most frequently used aid, it was not a decisive factor in suggestion acceptance; users combined multiple information aids to make informed decisions. Our findings demonstrate the potential of explainable AI paraphrasing tools to enhance NNESs' confidence, autonomy, and writing efficiency, while also emphasizing the importance of thoughtful design to prevent information overload. Based on these findings, we offer design implications for explainable AI paraphrasing tools that support NNESs in making informed decisions when using AI writing systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Writing Assistants, Paraphrasing Tools, Non-native English Speakers

ACM Reference Format:

Yewon Kim, Thanh-Long V. Le, Donghwi Kim, Mina Lee, and Sung-Ju Lee. 2025. Design Opportunities for Explainable AI Paraphrasing Tools: A User Study with Non-native English Speakers. In *Designing Interactive Systems Conference (DIS '25)*, July 05–09, 2025, Funchal, Portugal. ACM, New York, NY, USA, 23 pages. <https://doi.org/10.1145/3715336.3735740>

*Work done while at KAIST.

[†]Equal senior role.



This work is licensed under a Creative Commons Attribution 4.0 International License. *DIS '25, Funchal, Portugal*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1485-6/25/07
<https://doi.org/10.1145/3715336.3735740>

1 Introduction

We have witnessed an enormous shift in the capability of AI writing tools from spell checkers [44, 68] to content generators producing fluent, human-like text [50, 58, 92]. These tools open new avenues for non-native English speakers (NNESs), who face unique linguistic and cultural barriers when writing in English [5, 6, 24, 34, 74, 108]. For instance, NNES-authored emails may unintentionally include awkward or contextually inappropriate expressions, reducing their likelihood of receiving replies [74]. Such expressions may also negatively influence recipients' perceptions of the NNES sender's intelligence and trustworthiness, even when the recipient is aware of the sender's foreign background [108].

Amid these growing challenges faced by NNESs, recent advances in large language models (LLMs) [86, 88, 105] have significantly expanded the scope of AI writing support. Equipped with capabilities to generate fluent, human-like text [120, 121], compared to traditional rule-based feedback tools [89, 119], LLMs now assist with a broad range of writing tasks, including idea brainstorming [38, 59, 92], drafting [31, 71, 99], and paraphrasing [54, 55]. Among these myriad writing tasks that AI can support, the paraphrasing task is shown to improve NNESs' writing by suggesting more fluent paraphrases than their original sentences [23].

While an AI paraphrasing tool could instantly offer paraphrased suggestions, the ultimate responsibility of assessing and selecting the most context-appropriate falls on users. This introduces a significant challenge for NNESs, as they may lack the linguistic proficiency and cultural awareness to evaluate these suggestions accurately [60]. Prior research has identified several types of information aids, such as suggestion quality scores and example sentences, that can assist in evaluating AI-generated suggestions [54, 75, 81, 89]. It has recommended developing writing tools that integrate such aids to provide comprehensive support [55]. However, key questions remain open: *How can information aids be effectively integrated into writing assistants to collectively explain AI-generated paraphrase suggestions?* Answering this question necessitates a deeper understanding of how NNESs interact with and respond to information aids during paraphrasing tasks—an area that remains underexplored.

To address this gap, we investigate how NNEs interact with diverse information aids (interchangeably referred to as “information aids” and “features” throughout the paper) to assess and select paraphrase suggestions. Concretely, we developed PARASCOPE, a research prototype that integrates AI paraphrasing with five types of information aids within a single writing interface. These aids—AI Score, AI Translation, AI Explanation, Example Sentence, and Frequency—represent support information commonly used by NNEs in real-world writing contexts. PARASCOPE logs user interactions with both paraphrase suggestions and the accompanying aids (Section 3). Through a lab-based evaluation of PARASCOPE with 22 NNEs writing academic emails (Section 4), we examined participants’ feature usage patterns, perceptions, and written outcomes (Section 5). We conclude with design implications for explainable AI paraphrasing tools tailored to NNEs (Section 6).

Our findings reveal that while participants generally preferred features used to gauge the overall quality of the suggestions, e.g., AI Score, AI Translation, and AI Explanation, over features offering detailed explanations of specific words or expressions like Example Sentence and Frequency, their preferences varied by proficiency level. NNEs with lower proficiency used significantly more information aids than those with higher proficiency and demonstrated a particular preference for AI Score, a numeric representation of suggestion quality. Although AI Translation—back-translating English suggestions into users’ first languages—was the most frequently used feature, it was not a decisive factor in suggestion acceptance; users relied on a combination of information aids to make informed decisions. Post-interview findings further revealed that integrating information aids positively influenced perceived efficiency, confidence, and autonomy in the decision-making process, although participants noted the potential for information overload if features were integrated without consideration.

Through this study, we discuss design implications for creating explainable AI paraphrasing tools that support NNEs in making informed decisions when collaborating with AI systems. We also explore opportunities for explainable writing assistants in broader contexts, offering directions for future research to enhance writing support across diverse user needs.

We summarize our contributions as follows:

- PARASCOPE, a research artifact prototyping an AI paraphrasing support system that integrates a paraphrasing tool with five information aids to investigate and compare common strategies NNEs use for paraphrasing.
- Empirical findings from an in-lab user study of PARASCOPE with 22 NNEs in an academic email writing context, combining qualitative and quantitative data on how participants engaged with information aids to assess AI suggestions and how these aids shaped their user experience.
- Design implications for explainable AI paraphrasing tools tailored to the needs of NNEs.

2 Related Work

2.1 NNEs’ Challenges in Writing

Writing is regarded as one of the most challenging and complex tasks among NNEs [7]. NNEs’ written English displays lower

linguistic accuracy and complexity compared with native speakers [87, 89, 90]. Their texts tend to include more spelling and grammatical errors and exhibit less lexical diversity [95]. NNEs also face challenges in selecting contextually appropriate words or expressions, primarily due to the gaps in their cultural background [107]. These issues become particularly salient in high-stakes written communication scenarios, such as email and academic writing [10, 35, 40, 51, 108]. In emails, linguistic errors can lead recipients to NNE senders as less intelligent, diligent, conscientious, and cognitively trustworthy [108], and NNEs’ overuse of casual language in academic writings can leave them at a competitive disadvantage [40]. Furthermore, NNEs’ writing is often misclassified as AI-generated by automated detectors due to its lower lexical complexity, raising concerns around fairness [72].

2.2 AI Writing Support Tools for NNEs

While AI writing support tools span a broad range of tasks ranging from idea brainstorming [28, 38, 39, 59, 92, 96, 100, 110] to revision [1, 48, 109], those designed for NNEs often focus on the revision process of writing. These tools aim to aid improvement in writing quality by providing assessment and feedback to the NNEs’ writing, primarily through grammar and spelling checkers [21, 37, 75, 89, 94] and feedback generation tools [14, 46, 52, 61, 79]. Another way to assist NNEs’ revision process is paraphrasing [33, 54–56, 78, 116], which suggests alternative ways of expressing the same content, potentially enhancing the quality of original texts written by NNEs [54].

Despite the usefulness of revision tools, NNEs’ lack of linguistic proficiency hinders them from accurately assessing and selecting tools’ suggestions [55, 60, 64]. As a solution, several revision tools provide information aids (e.g., dictionary results, grammar rules) along with suggestions to help users better understand and reason about the suggestions [21, 54, 55, 75, 89]. For instance, lexical and grammatical suggestions are provided along with informational aids such as dictionary samples and grammar patterns sourced from an academic corpus [21]. On the other hand, LangSmith, a paraphrasing tool for NNEs [54, 55], provides a “typicality score” of paraphrased suggestions, representing the suggestion qualities. Nevertheless, we do not clearly understand which and how supports should be provided [55] to best support NNEs using AI paraphrasing tools. We aim to bridge this gap, offering insights into effective support strategies for NNEs engaged in paraphrasing activities.

2.3 Empirical Studies of Writing with AI

Previous research has explored user behaviors and perceptions of AI writing tools. A major thread of this research has focused on predictive text systems, which aid users in composing text by suggesting the next phrases or sentences they might use [8, 15, 31, 36, 71]. Recent work [36] explored how users write email replies with sentence vs. message-level suggestions. Another study [15] analyzed differences between native and non-native English speakers concerning their perception of and writing behaviors influenced by writing tools’ next phrase suggestions. In parallel, a line of research studied user behaviors by providing AI-based analytics to user written text, such as event sequence visualizations and paragraph summaries [25, 30, 98]. For instance, COALA [25] compared native

and non-native English speakers' collaborative writing behaviors when they are equipped with visual analytics. On the other hand, a few studies have investigated user behaviors regarding paraphrasing tools [55, 67]. Closest to our study, researchers explored behaviors of NNEs with Langsmith [54] and found that NNEs use external resources such as translators and web search results when paraphrasing with AI [55]. However, they focused on the impact of translator use on NNEs' engagement with the tool's outputs. Our study takes a broader and more detailed approach by identifying a comprehensive set of information aids that NNEs need when paraphrasing with AI and comprehensively analyzing user behaviors given these aids. Moreover, we provide insights on designing paraphrasing tools with information aids to support NNEs effectively.

3 PARASCOPE Design and Implementation

3.1 PARASCOPE Design Principles

We designed PARASCOPE as a research prototype to investigate how NNEs interact with diverse information aids when assessing and selecting AI-generated paraphrases. To achieve this, we established high-level design principles of PARASCOPE drawing on prior work [26, 36, 47, 53, 55, 57, 71, 117]. Our design aimed to:

- **DP1 Include core strategies NNEs employ as information aids in a realistic writing context:** Prior studies demonstrate that realistic task contexts elicit natural user behaviors [36, 47, 53, 55]. We aim to develop an AI-assisted paraphrasing tool that integrates multiple types of information aids commonly used by NNEs, and embed it in a writing task representative of real-world scenarios faced by NNE users.
- **DP2 Support open-ended, user-driven exploration:** Prior studies show that minimally constrained interaction reveal emergent usage strategies and inspire design directions [26, 47, 57, 117]. We focus on designing a system that allows participants to freely engage with different information aids during the paraphrasing process in an open-ended manner.
- **DP3 Capture user-system interaction data:** Prior HCI work demonstrates that recording user interactions with a system provides insights into user behavior with interactive systems and reveals patterns that inform future system design [26, 53, 71]. We log detailed user interactions and collect qualitative feedback to understand how participants engage with the system.

3.2 Overall Design of PARASCOPE

PARASCOPE is an AI writing support system that integrates paraphrasing functionality with diverse information aids in a unified interface, enabling NNEs to make informed decisions when assessing and selecting AI paraphrased suggestions. The system consists of two main components: the paraphrasing pane (Figure 1(A)) and the information aids pane (Figure 1(B)). In the paraphrasing pane, users can navigate and select one of four paraphrased suggestions for their input text. In the information aids pane, users can flexibly explore five types of information aids: AI Score, AI Translation, AI Explanation, Example Sentence, and Frequency to evaluate the system's suggestions.

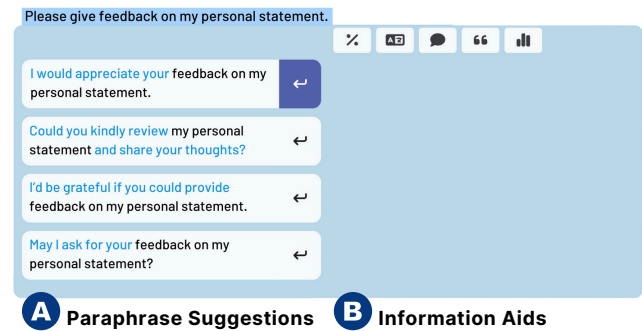


Figure 1: Overview of the PARASCOPE interface. The interface features two panes: (A) a pane displaying four paraphrased suggestions for the user's input text and (B) a pane providing information aids. The right pane (B) remains empty until the user interacts with an information aids button.

3.2.1 Paraphrasing in the email writing task as a use case (DP1). To contextualize PARASCOPE in a meaningful real-world setting, we selected the task of paraphrasing in email writing. Writing emails in English is a common yet challenging activity for NNEs, as it requires clear, professional communication that adheres to linguistic and cultural norms they may be unfamiliar with [108]. Moreover, paraphrasing—an essential activity for refining tone, improving the wording, and enhancing clarity—is widely practiced by NNEs in both practical and educational contexts, often with the aid of AI writing support systems [55, 60]. The practical relevance of paraphrasing and email writing makes this task a suitable context for studying user interactions with AI-generated suggestions and information aids, providing insights to inform the design of writing support systems for NNE users.

3.2.2 UI design and interaction paradigm of PARASCOPE (DP2). We aim to understand how NNEs engage with information aids when assessing and selecting AI-generated suggestions. To achieve this, we focused on two key aspects in designing PARASCOPE:

Usability consistency. Since our primary focus is on studying the use of information aids, we ensured usability consistency [70]—alignment of user experience with other systems that the user is familiar with—in designing the surrounding elements, i.e., the overall UI and interaction methods, of the system. This approach reduces the cognitive overhead associated with adapting to unfamiliar systems, allowing participants to focus on exploring and utilizing the information aids. Specifically, we drew on conventions established by existing writing assistants [44, 54, 71, 112, 113], which provide real-time support directly within text editors. In PARASCOPE, we implemented a pop-up interface activated via a keyboard shortcut. When a user selects a range of text¹ and presses `ctrl + j`, a pop-up box appears directly below the cursor (Figure 1). The system supports navigation of suggestions and information aids using both a mouse (point and click) and a keyboard (arrow keys to navigate UI components, Enter to accept a suggestion or open an information aid). Additionally, we provided multiple paraphrased

¹The system automatically adjusts the user's selection to encompass complete words if the initial selection ends mid-word.

Table 1: List of logged events. Each event is represented as a tuple containing the event name, timestamp, and snapshot of the editor. Text events have associated metadata containing information on inserted or deleted text. Cursor events have associated metadata containing information on start and end indices of cursor selection.

Event name	Trigger	Description
Category: Session		
session-start	System initializes the editor	Start writing session
session-end	User clicks submit button	Finish writing session
Category: Text		
text-insert	User/system inputs (any key)	Insert text
text-delete	User presses delete key	Delete text
Category: Cursor		
cursor-forward	User/system inputs {↓, →} key	Move cursor forward
cursor-backward	User presses {↑, ←} key	Move cursor backward
cursor-select	User presses shift + {↓, →, ↑, ←}	Select range of text
Category: Suggestion		
suggestion-get	User presses cmd/ctrl + j	Request new suggestions
suggestion-open	System fetches suggestions	Show interface w/ suggestions
suggestion-reopen	User presses shift + tab	Reopen interface w/ prev. suggestions
<i>While interface is displayed</i>		
suggestion-select	User presses enter + suggestion is focused User clicks a suggestion	Select suggestion
suggestion-close	User/system inputs esc or (any key) User clicks outside of interface	Hide interface
Category: Information Aid		
<i>While interface is displayed</i>		
info-get	User presses enter + information aid button is focused User clicks the information aid button	Request information aid
info-open	System fetches queried information aid	Show requested information aid
info-expl-select	User presses enter + suggestion is focused + AI Explanation is opened User clicks suggestion + AI Explanation is opened	View explanation associated with suggestion
info-exam-query	User presses enter + Example Sentence is opened + example search bar is focused User clicks the example search button + Example Sentence is opened	Query example sentences
info-freq-query	User presses enter + Frequency is opened + frequency search bar is focused User clicks the frequency search button + Frequency is opened	Query frequencies

suggestions, aligning with established practices in AI paraphrasing tools [49, 54, 55, 112]. We chose the number of suggestions as four, based on the prior work [15] which demonstrated NNESS benefit most from receiving $3 < N < 6$ parallel suggestions.

Flexible, open-ended exploration of information aids. To promote open-ended exploration of information aids, PARASCOPE displays the features—AI Score, AI Translation, AI Explanation, Example Sentence, and Frequency—as horizontally aligned buttons. The UI does not select any feature independently, and the right pane remains empty until a user clicks any button, ensuring all features are equally accessible and users flexibly explore suggestions based on their needs. We provide further details on the design and functionality of each information aid in § 3.3.

3.2.3 *Logging user-system interactions* (DP3). To understand how users interact with information aids to assess and select paraphrased suggestions, we designed a logging mechanism to capture user-system interaction data. Inspired by the prior work [71], we record user-system interactions as events, represented as a tuple containing the event name, timestamp, and snapshot of the current editor and PARASCOPE. An event can be inserting or deleting text within the editor, moving a cursor forward or backward, getting and navigating suggestions and information aids from the system, or accepting or dismissing suggestions. A complete list of all logged events is shown in Table 1. These logs provide granular details

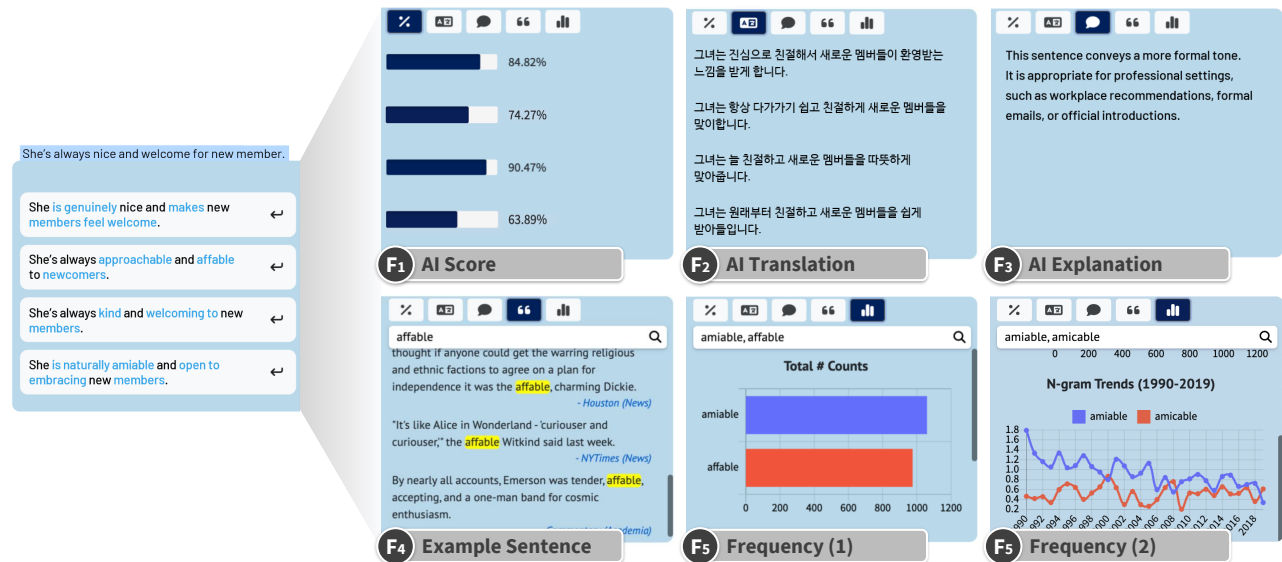


Figure 2: Illustrations of the information aids: AI Score (F_1), AI Translation (F_2), AI Explanation (F_3), Example Sentence (F_4), and Frequency (F_5) for the four paraphrased suggestions of the original text (“She’s always nice and welcome for new member.”). F_1 (AI Score): The scores representing paraphrased suggestions’ quality are displayed next to each suggestion as both bar graphs and numerical values. F_2 (AI Translation): Translated versions of suggestions in users’ first language (in this figure, Korean) are displayed next to each suggestion. F_3 (AI Explanation): Textual explanation about tones and appropriate contexts for each suggestion is displayed. F_4 (Example Sentence): After users search a term using the search box, example sentences containing the term are displayed, allowing users to browse the sentences. F_5 (Frequency): After users search a term(s) using the search box, a bar graph visualizing the frequency of the term(s) is displayed. When users scroll down the prototype, they see a line graph depicting the number of occurrences of the term(s) over the years.

about user-system interactions, enabling the analysis of quantitative usage patterns, such as the preferences for specific information aids and the sequence of information aid usage.

3.3 Design of Information Aids in PARASCOPE

To design information aids that could be useful in evaluating AI-generated suggestions, we adopted a two-fold approach informed by prior research [47]. First, we identified information aids commonly used by NNEs by analyzing functionalities of existing writing support systems from both commercial tools [41–44, 68, 76, 82, 85, 93, 104, 106, 112, 113] and academic studies [16, 49, 52, 54, 61, 75, 81, 89, 119]. Second, we conducted a 1-hour online formative study with 15 NNEs who reported regularly writing English emails and using AI writing assistants [43, 44, 93, 112, 113] to gather information aids a user should be able to use when paraphrasing with AI. The participants’ first language included Korean, Chinese, French, Filipino, and Hindi. All studies were conducted in a semi-structured manner online via Zoom. During the formative study, participants were asked to write a professional email in their naturalistic settings using AI writing assistants or online searches to aid their writing process. After the writing session, the researchers asked participants about their writing experiences, focusing on the tools and functionalities they use to overcome the difficulty of assessing and selecting AI-generated suggestions, as well as their suggestions for improving the writing process with AI tools. Participants were

compensated with approximately 24 USD (35,000 KRW) for participants. The researchers transcribed all audio and conducted open coding and thematic analysis.

Together, we discussed and synthesized our findings into the following list of information aid features that a paraphrasing tool for NNEs should support, until all researchers reached a consensus. Specifically, we identified from the formative study that the major needs of NNEs in assessing and selecting AI-generated suggestions are two-fold: (i) gauging the overall quality of suggestions, and (ii) validating the real-world usage of specific words or expressions that appeared in the suggestions, both stemming from concerns about whether AI-generated texts they select might sound awkward. We analyzed and categorized tools or desired features mentioned by the formative study participants to address each need, and mapped each feature to the system features identified from the survey and literature review. This way, we operationalized user needs into concrete design features, ensuring relevance and applicability to the real world. Accordingly, for each feature, we label it as **GLOBAL** or **LOCAL**, where **GLOBAL** refers to features used to gauge the overall quality of the suggestion, and **LOCAL** refers to features examining the usage of specific words or expressions in suggestions.

3.3.1 AI Score (GLOBAL). AI-based numerical assessments of suggestion quality provide users with metrics for evaluating AI-generated

text, e.g., scores indicating suggestion qualities [44, 54, 113]. Formative study participants expressed similar needs, as noted by two participants. Notably, P8 stopped using Wordtune due to a lack of quality indicators, stating, “There are so many suggestions, but I have no clue which one is the winner,” and suggested incorporating a “ranking system based on model confidence scores.”

Motivated by these observations, we designed AI Score, which represents the quality of AI-generated paraphrases (F_1 in Figure 2). We follow the convention of the existing tools [54, 113] to visualize the computed scores with horizontal bar graphs and numerical values in a percentage alongside AI suggestions. To calculate the scores of paraphrased sentences, we aimed to quantify the following criteria: whether the suggestion is semantically similar to the original sentence, whether it presents diverse alternative expressions to the original text while ensuring syntactic accuracy, as defined by prior works in linguistics [11, 22]. For this purpose, we leveraged ParaScore [97], a language model-based paraphrase evaluation metric. ParaScore outputs higher scores for paraphrases that (i) maintain a similar meaning to the original text and (ii) consist of diverse words and syntactic structures compared with the input, which aligns with our predefined goals.

3.3.2 AI Translation (GLOBAL). Back-translation is widely recognized in the literature as a critical strategy for NNEs to verify their comprehension of English sentences [20, 65]. Participants in our survey reflected this practice, reporting various translation tools [43, 68, 82, 106]. Previous studies [49, 75, 81] have similarly incorporated back-translation in writing tools to enhance comprehension of English suggestions for NNEs. Our formative study revealed similar patterns, with five participants highlighting back-translation as a key step for verifying AI-generated suggestions. For instance, P2 mentioned: “I use Naver Papago—I copy and paste all AI suggestions into it, read texts in my first language, and confirm whether the suggestions sound natural in English or not.”

Based on these observations, we designed AI Translation, a feature that translates paraphrased suggestions into a user’s first language² (F_2 in Figure 2). Drawing on prior work [49, 75], we displayed L1 translations alongside AI-generated suggestions to enable quick and efficient comparison between options. For implementation, we utilized the Naver Papago API [82], a translator frequently mentioned by participants in the survey.

3.3.3 AI Explanation (GLOBAL). We identified the use of natural language explanations in writing support systems, either as a method for interpreting target English text [76] or providing writing feedback [16, 52, 61]. For example, Ludwig [76] provides explanations by clarifying whether a word or phrase is suitable for written English and detailing its usage contexts (e.g., “It is typically used to express a warm greeting to someone who has joined a group, company, team, etc.”). Such explanations are also widely applied in other domains to enhance the interpretability of intelligent interactive agents [3, 17, 18, 77, 101]. Similarly, during our formative interviews, four participants sought natural language explanations on differences between paraphrases and original text, focusing on tone and appropriate usage contexts. P9 expressed a

need for clarifying the tone of suggestions: “I wish the tool could tell me, ‘this option is more polite than the other.’” P5 also mentioned identifying suitable usage contexts: “It would be great if the tool informs me, ‘this option is better for academic writing.’”

Building on the findings, we designed AI Explanation with two goals: generate natural language explanations that (i) compare original and paraphrased text, and (ii) inform users about the tone and appropriate contexts of each suggestion (F_3 in Figure 2). When users click the AI Explanation button, a natural language explanation for the first paraphrased suggestion appears in the right plan. Users can explore explanations for individual suggestions by selecting one at a time, ensuring that only one explanation is displayed at a time to reduce the cognitive load from presenting lengthy texts in a confined space [2]. We implemented AI Explanation using OpenAI’s GPT-3.5 [84] with a few-shot prompting method [13] (see Appendix B.2 for the prompt). GPT-3.5 was selected as it was the best available model at the time of the study and demonstrated strong capabilities in generating human-like responses [27] and linguistic comprehension in few-shot settings [114].

3.3.4 Example Sentence (LOCAL). Participants from the survey often mentioned tools that provide example sentences for the search query [76, 104]. Related works also feature example sentences [75, 89] along with suggestions. For instance, AwkChecker provides example sentences of each suggestion to help users examine the context surrounding phrases to make an informed decision [89]. Similarly, eight participants from the formative study referred to human-authored example sentences from credible sources like news, academic articles, or famous magazines to verify whether words or expressions are used in the real world. P11 mentioned checking example sentences from a native English corpus to verify AI suggestions: “I often refer to the Longman dictionary because it provides accurate information on how words are used in specific contexts. Its examples are taken from the native English corpus, so you can be confident that the words are used in the way that native English speakers use them.”

We design Example Sentence to retrieve example sentences from credible sources (newspapers, academic journals, and magazines) containing a term specified by a user (F_4 in Figure 2). Initially, the search box appears in the right pane of the pop-up box, allowing the user to input a word or expression. Upon submitting it by clicking the search button (magnifier icon) or pressing the enter key, the tool retrieves sentences containing the input term (highlighted in yellow), randomly selects up to 10 example sentences, and shows them along with the reference and genre of the sentences (e.g., New York Times (News)). For the data source, we use the Corpus of Contemporary American English (COCA) [32]. COCA is an extensive collection of over one billion words extracted from approximately 500,000 texts of diverse genres from 1990 to 2019. To ensure we retrieve only the sentences from credible sources, we selected only the sentences from three genres: academic journals, magazines, and newspapers.

3.3.5 Frequency (LOCAL). NNEs often verify whether a phrase is commonly used by analyzing the number of results returned by search engines [45, 115]. Similarly, several writing tools incorporate real-world word usage frequencies into their suggestions [42, 49, 89].

²As all participants in our user study (§4.1) were native Korean speakers, translations were provided in Korean.

For example, TransAhead provides grammatical patterns linked to user input and their frequencies in a reference corpus [49]. Three participants in our formative study also reported using statistical evidence to assess term usage. P11 described using Google search results to evaluate the prevalence of an expression: “*One way to find out if native speakers commonly use an expression is to search it on Google and see how many results you get.*” Similarly, P8 used Google Ngram Viewer [42] to check if the word or expression was trendy.

Inspired by tools that display total frequencies [49, 89] and visualize usage trends [42], we designed Frequency to provide two types of information for user queries: total frequency counts and usage trends over time. After a user queries a term(s) using the search box, the right pane of the system displays (i) a bar graph visualizing the total number of times each term appears in the source data and (ii) a line graph depicting its frequency trend over the years (1990-2019) (F_5 in Figure 2). For implementation, we used the COCA dataset, where we precomputed the number of occurrences of n-grams ($1 \leq n \leq 4$) from the dataset, recording the total count and count per year. This data was stored in our database and retrieved when users submitted queries.

3.4 PARASCOPE Implementation

We implemented PARASCOPE using JavaScript, HTML, and CSS, building upon the CoAuthor interface [71]. We used a Flask server for the backend to preprocess requests from the front-end and forward these requests to the necessary destinations. All log data was stored on a local server and pseudonymized using identifiers recognizable only by the authors.

4 User Study

We conducted a user study with 22 NNEs using PARASCOPE to gain insights into how information aids can be effectively integrated into AI paraphrasing tools. Concretely, we aimed to address the following research questions:

- RQ1:** How do NNEs interact with information aids to assess and select AI paraphrase suggestions in PARASCOPE?
RQ2: What impact does PARASCOPE have on NNEs’ user experience and writing performance?
RQ3: What kinds of interactions and interface features do NNEs wish to have in PARASCOPE?

4.1 Participants

We recruited 22 NNEs through advertisement posts in university online communities. We targeted university students as they are among the most representative NNEs who often write in English, especially emails in academic settings. We had two inclusion criteria: (i) one’s self-assessed English proficiency level is below or equal to B2 (upper intermediate) according to CEFR [83] measurement,³ and (ii) one is familiar with at least one AI paraphrasing tool. This selection was made to explore the behaviors and needs of users with limited English proficiency and to prevent the novelty effect often associated with first-time use of paraphrasing tools. Additionally,

³CEFR levels are defined as basic (A1: Beginner; A2: Elementary), independent (B1: Intermediate; B2: Upper intermediate), and proficient (C1: Advanced; C2: Proficient).

Participant ID	English Proficiency	Academic Status
P1	A2	Master’s student
P2	A2	Undergraduate student
P3	A2	Undergraduate student
P4	A2	Undergraduate student
P5	A2	Undergraduate student
P6	B1	Undergraduate student
P7	B1	Master’s student
P8	B1	Ph.D. student
P9	B1	Ph.D. student
P10	B1	Master’s student
P11	B1	Master’s student
P12	B1	Master’s student
P13	B1	Undergraduate student
P14	B1	Ph.D. student
P15	B1	Undergraduate student
P16	B2	Master’s student
P17	B2	Master’s student
P18	B2	Ph.D. student
P19	B2	Undergraduate student
P20	B2	Undergraduate student
P21	B2	Master’s student
P22	B2	Undergraduate student

Table 2: Detailed background information of the interview participants in the main study (Section 4). The first language of all participants was Korean. For English proficiency, we use self-reported CEFR levels: A1 (beginner) < A2 (elementary) < B1 (intermediate) < B2 (upper intermediate) < C1 (advanced) < C2 (proficient).

we asked applicants to submit their email excerpts optionally, which were utilized to create email scenarios used in the user study.

After applying these criteria, we randomly selected 22 participants (Age=19-34, Mean=24.8, Std=3.9; Female=12, Male=10).⁴ All participants’ first language was Korean. Ten were undergraduates, eight were master’s, and four were PhD students. Five had an English level of A2 (elementary), ten had B1 (intermediate), and seven had B2 (upper intermediate). The detailed demographic data of the participants are shown in Table 2. The study lasted a maximum of 117 minutes (M=105.55, SD=10.9), and participants were compensated with approximately 27 USD (40,000 KRW).

4.2 Study Procedure

The entire process of the user study is shown in Figure 3. Every participant took part in the study online via Zoom. The study began with a brief introduction to the research and the signing of informed consent forms. After the introduction, participants completed a 25-minute-long tutorial task. During the tutorial, one of the authors introduced the tool and each information aid’s functionality, and then participants completed a series of subtasks selecting the best-paraphrased suggestion in the context of email writing, given a pre-written sentence and four paraphrased suggestions with PARASCOPE.⁵ To ensure participants familiarized themselves with each information aid, only one of the five information aids

⁴Among the 25 initially recruited participants, we conducted a pilot study with the first three to address potential technical issues and improve the tool tutorial.

⁵Participants were instructed to access the link to the tool with an incognito window blocking any third-party writing tools. This was to accurately log the information aids participants use and their interactions with the paraphrased suggestions and information aids.

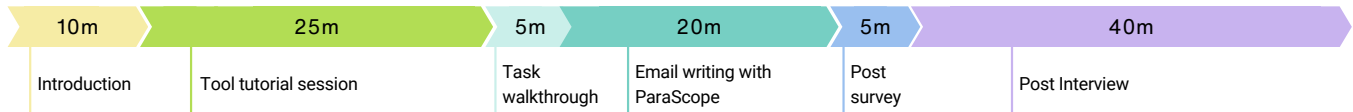


Figure 3: Overview of the user study procedure. After familiarizing themselves with each information aid in the tutorial session, participants wrote an open-ended academic email with PARASCOPE. After the writing, participants engaged in a post-survey and semi-structured interview about their experiences.

was available for each subtask, and they could ask questions about the tools during the tutorial.

After the tutorial, participants proceeded to the email writing task, where they were given an open-ended email writing task with one of the academic scenarios we prepared (Appendix C.1 lists the scenarios). These scenarios were designed to be realistic so that college students could easily relate to them (e.g., inquiring about grade corrections or requesting to audit courses). The scenarios were selected from the email excerpts from the recruitment survey. The scenarios assignment was counterbalanced across participants to minimize bias. The goal of the user study was to investigate how participants used the paraphrasing functionality and the information aids in an open-ended email writing task. As such, we did not require users to engage with every feature; participants were informed that they could freely choose which information aid to use. During the task, the researcher observed and noted their paraphrasing activities and asked about them during the interview.

After the tasks, participants completed a post-survey. The survey included questions about their overall experience using each information aid feature in PARASCOPE and users' overall ratings of the quality and diversity of the paraphrased suggestions. Finally, we conducted semi-structured interviews with all participants on Zoom. During the interview, we asked participants to explain their paraphrasing process using PARASCOPE and share their experiences and suggestions with the paraphrasing functionality and the information aids. Under the consent of the participants, all interviews were recorded and later transcribed for further analysis.

4.3 Analysis

4.3.1 Quantitative Analysis. To investigate participants' usage patterns with PARASCOPE, we conducted a quantitative analysis using the interaction logs collected from the email writing task. We counted the number of paraphrasing events per each writing session, as well as the number of associated information aid usages for each paraphrasing event. We analyzed the acceptance rate of paraphrased suggestions using the Chi-Square test. To identify differences among participants with varying levels of English proficiency, we analyzed the average number of paraphrasing events and information aids used using the Kruskal-Wallis test for overall comparisons and the Mann-Whitney U test for pairwise comparisons. These non-parametric tests were selected as the Shapiro-Wilk test revealed the data was non-parametric. We also investigate the usage trends of each information aid over time using the timestamped information aid usage logs per each paraphrasing event.

Additionally, we evaluated paraphrased sentence pairs (original sentence vs. selected paraphrase) collected in the user study through both automatic human evaluations to investigate whether

using information aids has to do with sentence quality improvement. For the automatic evaluation, we measured grammaticality using the LanguageTool API [69] following a previous work [71] to compute the number of spelling and grammar errors per sentence. The scores were averaged across sentences to compare the quality of sentences before and after paraphrasing. For human evaluation, we focused on two key aspects: fluency and politeness for contextual appropriateness. We presented human evaluators with sentence pairs (original sentence vs. selected paraphrase) and asked them to choose which sentence "sounds more natural" (fluency) and "is more polite" (politeness). We included the question about politeness because the email tasks in the user study involved making requests to a person in authority, and each task description explicitly instructed participants to write the email *politely*. We recruited 24 native English speakers from Prolific [91] and compensated £0.75, which corresponds to £9 per hour. Each evaluator evaluated 28 sentence pairs (14 for fluency and 14 for politeness), and three different evaluators assessed each sentence pair. We analyzed the pairwise judgments using the Wilcoxon signed-rank test. We analyzed the pairwise judgments using the Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test, as confirmed by the Shapiro-Wilk test.

4.3.2 Qualitative Analysis. We transcribed the semi-structured interviews and analyzed them using the constant comparative method [103]. Two authors independently open-coded two interview samples to identify key concepts and patterns. Axial coding was then performed to link these patterns [29], resulting in an initial codebook. The first author coded the remaining interview data while continuously refining the codebook. Upon completing the first coding round, another author coded one interview sample based on the updated codebook for verification, and any issues were resolved through discussion. Throughout the process, all authors regularly discussed emerging themes while triangulating the interview data with quantitative analyses. We include the full codebook in Appendix C.2.

All interviews and qualitative analysis of interview transcriptions were conducted in Korean, and the selected quotes were translated into English. The authors reviewed all translations to ensure accuracy and preserve the integrity of participants' original statements.

5 Results

We present our findings from the user study by explaining observations of NNESS' interaction patterns with information aids (RQ1), potential impacts of integrating information aids on NNESS' user experience and writing performance (RQ2), and participants'

suggestions on the system UI and interaction method (RQ3). We summarize our main findings for three research questions below:

RQ1: Participants favored simple global features, especially among lower-proficiency users. While local features were used less often, they helped resolve difficult choices.

RQ2: Using information aids improved perceived confidence and efficiency, but sometimes caused information overload. Participants also reported potential for increased autonomy in the process and language learning.

RQ3: Participants suggested greater transparency in information aids, interface personalization, editable suggestions, and inclusion of the original sentence for comparison.

While it is not the primary focus of our study, we note that users expressed general satisfaction with the quality (Mean=4.23, Std=0.43) and diversity (Mean=4.36, Std=0.66) of suggestions from our tool (in a 5-point Likert scale, 1=Strongly Disagree, 5=Strongly Agree). This satisfaction with paraphrased suggestions indicates that participants generally did not experience usability problems in their writing task in terms of the paraphrasing outcomes.

5.1 NNESS' Interaction with Information Aids

We explore how participants interacted with information aids during paraphrasing (RQ1). We describe relevant quantitative findings with qualitative insights.

Overall, we collected a total of 258 paraphrasing events from the user study. The average paraphrasing events made per user was 11.73 (Std=5.55, Max=27, Min=4, Median=12.0), with an acceptance rate (Accept = users accept one of the AI paraphrased suggestions, Reject = close without selecting) of 61.63%. Participants used at least one information aid in 193 (=74.81% of all events) paraphrasing events. The average number of information aids used by participants was 1.93 (Std=0.93, Median=2.0).

5.1.1 Motivations for using information aids. During the post-interview, participants noted two primary motivations for using information aids: (i) they were uncertain about the suggestion quality and used the features to *assess* and *rank* the suggestions; (ii) participants with a specific decision in mind referred to the features to *validate* their choice. When leveraging the features, participants considered three aspects of suggestions in their decision-making process: (i) whether suggestions sound overall natural in English, (ii) the appropriateness of the tone, and (iii) real-world applicability of words/expressions. This observation aligns with our formative study findings (§3.3), showing that these aspects are commonly considered in NNESS' writing activities.

5.1.2 Preference for global features. Across all events, we observed significantly higher usage frequencies of **GLOBAL** features (AI Score, AI Translation, AI Explanation) compared to **LOCAL** features (Example Sentence, Frequency) ($\chi^2 = 179.89, p < .001$). Among these, AI Translation was the most frequently used feature (63.18% of all events), followed by AI Score (47.67%) and AI Explanation (40.70%). In contrast, Frequency (8.91%) and Example Sentence (6.98%) were rarely used. To account for potential biases from participants with higher paraphrasing event counts,

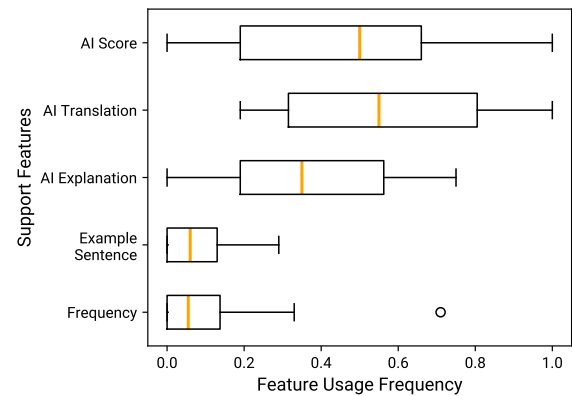


Figure 4: Feature usage frequencies for each information aid. The orange bar indicates the median value. While AI Score (M=0.46, SD=0.33), AI Translation (M=0.56, SD=0.27), and AI Explanation (M=0.37, SD=0.23) were frequently used in a paraphrasing event, usage frequencies of Example Sentence (M=0.08, SD=0.1) and Frequency (M=0.11, SD=0.17) were comparatively low.

we further analyzed feature usage frequencies averaged across individual participants. Figure 4 illustrates these results, confirming a significant difference among feature groups ($\chi^2 = 47.68, p < .001$). Pairwise comparisons using the Mann-Whitney U test showed that all global features were used significantly more frequently than local features ($p < .001$ for all comparisons).

Post-interview analysis revealed that **participants' preferences were largely driven by the simplicity of the information and the ease of interaction.** Notably, 15 out of 22 participants mentioned using information aids to quickly and roughly assess suggestion quality, allowing them to narrow down multiple suggestions to a manageable set for focused evaluation. For example, P3 described using features to *"initially prioritize my preferences"*, while P12 noted a similar approach of *"screening out"* undesirable suggestions to focus on the most viable options.

Eight participants **particularly favored AI Score and AI Translation due to their straightforward presentation**, which is effective in quickly assessing suggestion qualities. P10 said, *"AI Score is presented as a number, and the AI Translation is in my native language, so they're easy to understand and efficient."* Conversely, **local features were less preferred because they were perceived as time-consuming and effortful to use.** For both Example Sentence and Frequency, eight participants reported that deciding on search queries and manually typing them added extra effort. P1 mentioned: *"To search, I have to decide what to look for and type it in—it takes time and effort, so I didn't use it much."* Example Sentence was further criticized for providing indirect and harder-to-comprehend information, as noted by six participants. P7 noted, *"It doesn't give me a clear answer. I have to read, think, and judge the information, which makes it difficult to use."*

Lastly, participants **had mixed opinions on AI Explanation in terms of its simplicity.** Eight participants appreciated its explicit and actionable information, as P2 noted: *"For example sentences, I need to read and interpret the information myself. But AI Explanation directly tells me, 'This is often used here,' or 'This might work better,'"*

which makes it easier because I don't have to think as much." However, four participants found AI Explanation cumbersome due to its length and text-heavy format: "I needed to make quick decisions on suggestions, but since it's presented in long paragraphs, it wasn't easy to skim. If it were in bullet points, it would have been faster and more convenient." (P8).

5.1.3 More reliance on information aids for lower proficiency users. We discovered that, in general, participants with lower English proficiency tend to rely more heavily on information aids when selecting AI paraphrases compared to those with higher English proficiency. Figure 5 shows the number of paraphrasing events per participant, with at least one feature usage. Out of 258 instances of paraphrasing, at least one information aid was used in 193 cases (=74.81%). Notably, the proportion of paraphrasing events with information aids was significantly higher among participants with lower proficiency levels ($\chi^2_{(2)} = 9.60, p = 0.0082$; A2: $M=91.55, SD=9.05$; B1: $M=80.69\%, SD=18.96$; B2: $M=47.57\%, SD=32.51$). We also statistically analyzed the number of information aids used in each paraphrasing event by user proficiency levels. The average number of features used was higher as user proficiency got lower: A2 users, on average, used 2.4 features ($SD=0.61$), B1 used 1.96 features ($SD=0.78$), and B2 used 1.17 features ($SD=0.35$), as shown in Figure 6. Kruskal-Wallis test also revealed that this trend is statistically significant ($\chi^2_{(2)} = 27.79, p < .001$).

From the pairwise statistical analysis of the feature usage frequency data using Mann Whitney U test, we noticed the English proficiency of a participant correlated with AI Score usage frequency: **participants with lower English proficiency levels (A2, B1) were more likely to use AI Score more frequently than those with the highest proficiency levels (B2)** (A2>B2: $U = 2.5, p < .05$; B1>B2: $U = 12.0, p < .05$; A2: $M = 0.72, SD = 0.32$; B1: $M = 0.51, SD = 0.29$; B2: $M = 0.19, SD = 0.21$). Similarly, during the post-interview, we discovered instances where **participants with lower English proficiency put more trust in AI Score**. P1, whose English proficiency is A2 (elementary), mentioned relying on AI Score rather than personal judgments, reasoning that "AI probably has a better understanding of English than I do." In contrast, P22, whose English proficiency is B2 (upper intermediate), was more critical of AI Score and preferred making their own evaluations: "It was hard to understand the criteria the AI used to generate the score. I thought it would be better to rely on my own judgment instead."

5.1.4 Frequent usage but lower acceptance rate for AI Translation. Paraphrasing events with information aids had a higher acceptance rate of 67.36% than those without information aids (44.62%) ($\chi^2 = 14.58, p = 0.0019$). Interestingly, **while AI Translation was the most frequently used feature** (usage frequency of 63.18% in §5.1.2), **it was not associated with higher acceptance rates of suggestions** ($\chi^2 = 0.79, p = 0.375$). In contrast, **AI Score and AI Explanation showed significant correlation with higher acceptance rates of suggestions** ($\chi^2 = 6.34, p = 0.012$ for AI Score, $\chi^2 = 10.01, p = 0.0016$ for AI Explanation).

These results suggest that AI Translation's utility might lie more in comprehending suggestions rather than assisting in comparing suggestions, i.e., deciding which one is the best. While it may help understand the semantic meaning of suggestions as indicated by

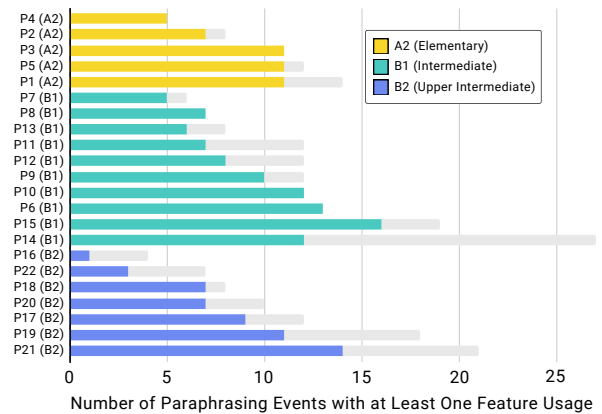


Figure 5: Number of paraphrasing events per user. Out of the total paraphrasing events (gray bars), we show the number of events with at least one support feature usage (colored bars).

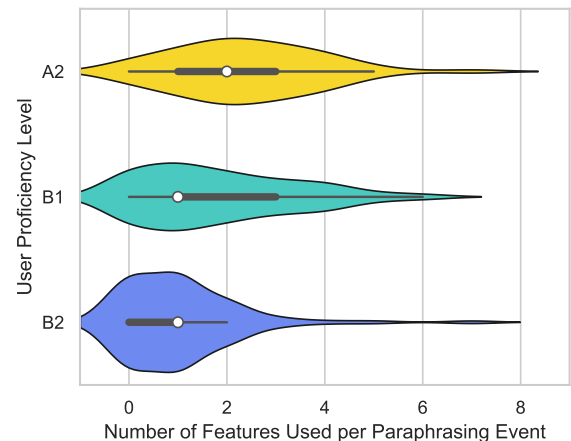


Figure 6: Distribution of the number of features used in one paraphrasing event by user proficiency levels: A2 (elementary), B1 (intermediate), and B2 (upper intermediate). The white dot represents the median and the thick gray bar represents the interquartile range. Values outside the thin line are considered outliers. (A2: $M=2.4, SD=0.61$; B1: $M=1.96, SD=0.78$; B2: $M=1.17, SD=0.35$)

prior works [20, 65], it may not be as critical for determining fluency of suggestions or appropriateness in tone, which were major considerations in NNEs' decision-making process as identified in §5.1.1. Conversely, as AI Score enables direct quantitative comparison among suggestions, it likely supports users in objectively evaluating the relative quality of suggestions. Similarly, AI Explanation, by explicitly explaining tone appropriateness, may guide users in identifying subtle tonal differences and aligning suggestions with the intended communicative goals.

5.1.5 Use of global features in early decision-making, with local features in later stages. While global features played a

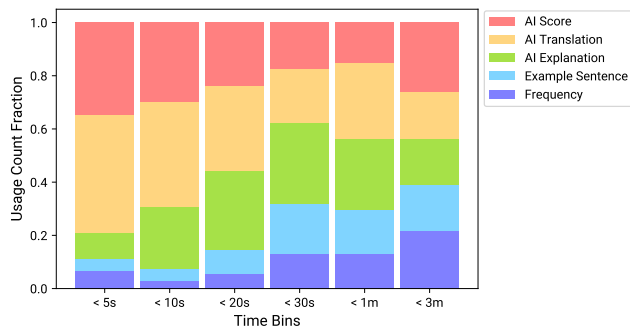


Figure 7: Each support feature’s usage fraction in six phases (time bins) of a paraphrasing event. This figure visualizes which support features participants utilized at what time in a paraphrasing event. Each feature use event is counted within each time bin across all paraphrasing events; for example, the AI Score event is counted to the ‘< 5s’ time bin if the AI Score was clicked two seconds after receiving suggestions. We determined the segmentation of time bins based on the distribution of feature-use events.

dominant role in the decision-making process, we observed an intriguing trend: local features became increasingly utilized in the later stages, as illustrated in Figure 7. The figure provides a temporal breakdown of how different information aids were employed during a paraphrasing event. Notably, as users spent more time, the proportional usage of local features (Example Sentence and Frequency) tended to rise.

Insights from the post-interview revealed that participants **relied on local features to further investigate suggestions or as tie-breakers when initial aids were insufficient**. For example, five participants reported using Example Sentence when the explanation provided by AI Explanation left them uncertain about a suggestion’s appropriateness. P15 explained: “*Even if the AI explanation explicitly states the situations where a suggestion can be used, it doesn’t always cover every possible context. When the explanation seemed ambiguous, I checked the example sentences to confirm the appropriate usage of the expression.*” Similarly, four participants mentioned turning to Frequency as a tie-breaker. P6 described their process: “*After looking at the score and the AI explanation, if I still couldn’t decide on the best option, I used Frequency to choose the expression that was more commonly used.*” In addition, the AI Score feature was also used for tie-breaking, as mentioned by six participants. For instance, P5 noted: “*If I was still unsure which of two sentences to use, I thought, perhaps the one with the higher numerical value would be better.*”

5.1.6 Summary of findings. NNEs used information aids to assess fluency, tone appropriateness, and the real-world applicability of expressions. They preferred global features (AI Score, AI Translation, and AI Explanation) for their simplicity and efficiency, especially among lower-proficiency users. Despite being the most frequently used, AI Translation did not correlate with higher acceptance rates, whereas AI Score and AI Explanation did. Local features

were less frequently used overall but played a role as tie-breakers in the later stages of suggestion evaluation.

5.2 Impact on User Experience and Writing Performance

We describe perceived user experiences of using PARASCOPE in the email writing task among NNEs (RQ2). We report findings from qualitatively analyzing post-interviews, as well as quantitative findings from human evaluation results of user-written sentences.

5.2.1 Informed decision-making with multiple information aids. Having multiple information aids appeared to **benefit participants by enhancing confidence in the decision-making process**, as mentioned by 13 participants. For example, P2 remarked: “*Having access to various types of information that either confirmed or refuted the suitability of a suggestion made me feel more reassured and confident in my choice.*” Eight participants **described their decision-making as a multi-step verification process, using diverse information aids to offer a richer perspective**. For instance, P1 described their approach as starting with the AI Score for an initial evaluation and then refining their decision by cross-referencing other features: “*I used the AI Score to gain initial confidence in the suggestion, but still felt uncertain. So, I checked other features, which gradually strengthened my confidence.*” P22 also mentioned using multiple features leads to more accurate judgments: “*Relying on just one feature might lead to a hasty decision, but using multiple features to make a comprehensive judgment improves the accuracy of my choices.*” Additionally, **seven participants mentioned using one feature to better understand or validate another**. For instance, P11 elaborated on how combining feature clarified the AI Score’s scoring rationale: “*When I only looked at the AI Score, I doubted its reliability and couldn’t understand why it rated a suggestion the highest. But when I checked the AI Translation and Explanation, it made sense, and I could accept the reasoning behind the score.*”

5.2.2 Efficient information-finding process, but possible information overload. Ten out of 22 participants mentioned that having all features in the same interface as the paraphrasing tool **enabled easy and fast access to necessary information without switching between multiple windows or tools**. P21 stated, “*Typically, my writing process involves constantly switching between windows: composing, searching dictionaries, and consulting ChatGPT. This system consolidates these actions within one platform, significantly aiding my workflow.*”

Participants expressed mixed opinions about the cognitive impact of this integration. Twelve participants noted that **having multiple features in one place allowed them to delegate part of the decision-making process** to the system. For example, P15 remarked, “*Rather than deliberating over each suggestion on my own, I find it more convenient to quickly scan the text and let the features handle the finer details of decision-making.*” Similarly, P1 described how their efficiency improved over time: “*As I became more accustomed to using the features, my ability to analyze and make decisions became faster.*” In contrast, five participants found participants found that the **abundance of information aids led**

to stress during decision-making, as they felt pressured to process and use every available resource. For instance, P8 noted: “There was so much to refer to that it actually got in the way of making a decision.” Given all the information aids, P6 felt responsible for checking every suggestion and feature presented in the system: “I feel compelled to use every information aid, even for sentences I could easily move on from. I don’t think this process is efficient, but I am nonetheless convinced that it improves writing.”

5.2.3 Enhanced autonomy in decision-making process. Interestingly, two participants reported that **utilizing information aids increased their sense of autonomy in decision-making**. P1 mentioned, “When there was only AI Score, I relied on it, thinking it would make better choices than me. However, engaging with all five features changed my perspective; I felt more in control of my decisions, as the features seemed to support rather than dictate my choices.” P18 highlighted that this process also augmented their sense of ownership over the final text: “Utilizing features to understand how suggestions might sound and integrating these suggestions into my text, made me feel more actively involved in the writing process. It felt as if I were crafting the text myself.”

5.2.4 Potential for language learning. Fourteen participants anticipated a potential learning impact with the integrated features. This expectation appeared to stem from the belief that **increased interaction with the features could lead to learning experiences**. P15 highlighted that “engaging in the comparative analysis of suggestions’ various aspects could deepen understanding in English and promote thorough learning.” Similarly, P13 noted that “the diverse insights provided by each feature enriched my engagement with English.” Moreover, P22, majoring in English education, noted that the tool could particularly benefit language learners with low proficiency, as “the tool provides explanations for suggestions, enhancing accessibility to information needed for learning English.” P1 nominated Example Sentence for helpful in learning English: “Example Sentence, unlike other features that provide direct hints, allows for the indirect exploration of various sentences and fosters thoughtful consideration.”

5.2.5 Sentence Quality Improvements. To examine whether paraphrasing with PARASCOPE improved writing quality, we evaluated 112 sentence pairs (original vs. paraphrased) using both human judgments and automated metrics. Among 112 pairs, 100 pairs were paraphrased using information aids, while 12 were paraphrased without using information aids. As a result, we observed an **overall improvement in sentence quality after paraphrasing**. Figure 8 presents the human evaluation results of the pairwise comparison between NNESS’ original and paraphrased sentences. Most evaluators preferred paraphrased sentences regarding fluency and politeness ($p < .001$). Similarly, in the automatic evaluation, the number of grammatical errors in the original sentences ($M=0.42$, $SD=0.66$) decreased significantly after paraphrasing ($M=0.13$, $SD=0.37$). These findings suggest that PARASCOPE, which integrates information aids into the paraphrasing process, can help users select higher-quality suggestions. However, the small number of sentences paraphrased without aids ($N=12$) limits direct comparisons and warrants further investigation.

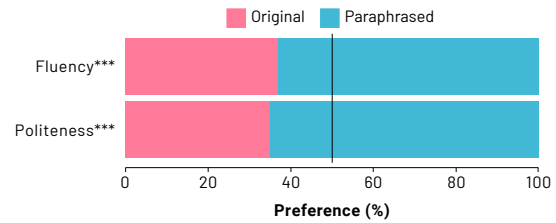


Figure 8: Human evaluation results of 112 paraphrased sentence pairs (original vs. selected paraphrase) composed in the user study. Among 336 comparisons, 65% of evaluators rated the paraphrases as more fluent than the original sentences, and 67% rated them as more polite. Overall, the paraphrased sentences showed a statistically significant improvement ($p < .001$).**

5.2.6 Summary of findings. Participants reported that multiple information aids improved decision-making confidence, with many using them in combination for verification. The integrated interface supported efficient access but occasionally caused cognitive overload. Some users experienced increased autonomy and ownership over their writing, while others saw potential for language learning through deeper engagement with features. Human and automated evaluations showed improved fluency and politeness in paraphrased sentences, suggesting benefits for NNESS’ writing quality.

5.3 PARASCOPE Feature Requests

We summarize key suggestions provided by participants for improving PARASCOPE’s user experience (RQ3). We derive findings from qualitatively analyzing user interview data.

5.3.1 Explainability in numeric measures. Ten participants emphasized the need for greater transparency and explainability in the AI Score feature. They wanted to understand “how the AI Score is calculated” (P15) and “why some suggestions receive such high scores” (P13). Similarly, P11 expressed a preference for AI Explanation over AI Score, stating: “I didn’t understand the rationale behind the AI Score—it didn’t make sense on its own. Instead, I preferred reading the AI Explanation, as it helped me accurately assess whether suggestions aligned with my criteria.” As such, P12 suggested “It would be helpful if the system explained the criteria used to calculate the scores and what those values represent.” A similar lack of transparency was noted for Frequency. P12 remarked: “It wasn’t clear what database the trends shown in Frequency were based on. For example, is this frequency higher because the word appears often in news articles? Knowing this would make the feature more helpful.”

5.3.2 Personalization of information aids. Six participants suggested providing options to personalize the interface by selectively displaying features they used most often. This was related to the section (about information overload) This suggestion was closely related to earlier concerns about information overload (see §5.2.2). Rather than displaying all information aids by default, participants preferred a more streamlined interface tailored to

their usage patterns. For instance, P16 suggested compressing the interface to focus on frequently used features: *“When using a tool repeatedly, people tend to stick to certain features. Instead of showing all features, it might be better to display only the ones I actively use.”* Participants also expressed interest in showing only frequently used features, with others available upon their needs. For instance, P15 suggested starting with AI Score as the default visible feature, with other aids revealed upon interaction: *“Initially, only the AI Score could be shown, and if I’m curious, I could click the score to see related aids like AI Translation or AI Explanation.”*

5.3.3 Interactive Suggestion Refinement. Participants often wished to edit suggestions while the interface was open, highlighting two primary use cases. First, seven participants expressed a desire to **make minor adjustments to their original text based on the suggestions provided.** For example, P17 explained: *“I didn’t want to completely rewrite my sentence; I aimed to keep the original structure intact while making minor adjustments to the words or expressions.”* Second, five participants **preferred creating new sentences by integrating elements from multiple suggestions.** For example, P9 shared an example of this process during the interview, where they started with the sentence ‘I figured out that there was a mistake’ and received suggestions including ‘I discovered that an error had been made’ and ‘I concluded that there was a miscalculation.’ They integrated parts of these suggestions, revising their original sentence to ‘I discovered that there was a miscalculation.’ In both cases, participants needed to type new sentences in the editor but found it challenging as the interface did not allow simultaneous editing and suggestion display.

5.3.4 Incorporating Original Text for Comparison. In our prototype, the original input text and its information aids were not displayed within the paraphrasing interface, following the design of existing tools [54, 112]. However, 11 participants **suggested incorporating features like AI Score and AI Translation for the original input, as they frequently referenced their original sentence during decision-making.** P6 explained: *“It would be more convenient if my original input was displayed alongside the suggestions for direct comparison. Paraphrasing provides four candidates, but the original sentence is essentially a fifth option. Including translations or scores for the original would make it easier to evaluate all options equally.”* This feedback indicates that treating the original sentence as an integral part of the decision-making process could enhance the usability of paraphrasing tools.

5.3.5 Summary of findings. Participants suggested four key improvements to PARASCOPE: (i) enhancing transparency in numeric features like AI Score and Frequency, (ii) enabling personalization by prioritizing frequently used aids, (iii) supporting flexible interaction with suggestions through in-place editing and combination, and (iv) displaying the original input with associated features for easier comparison. These suggestions highlight the need for more explainable, customizable, and interactive paraphrasing tools.

6 Discussion and Design Implications

With PARASCOPE, we studied how NNEs assess AI-generated paraphrases with information aids. Based on our findings, we explore

design implications for writing assistants to more effectively support NNEs and discuss broader impacts of the study.

6.1 Design Implications

In this section, we outline five design implications for writing systems tailored to NNEs, informed by our study findings.

6.1.1 Select information aids depending on user proficiency.

We observed that participants had different usage patterns in information aids according to their English proficiency: participants with lower English proficiency used more information aids and relied significantly more on AI Score than those with higher proficiency (§5.1.3). Based on the findings, we suggest that writing systems for NNEs should adapt the availability and presentation of information aids based on the user’s proficiency level. For example, systems could provide multiple features, including AI Score by default for lower-proficiency users. Conversely, for higher-proficiency users, the interface could simplify the display by emphasizing fewer features only.

In addition, writing tools could incorporate customization options, allowing users to set their own preferences for which information aids to display (§5.3.1). For instance, users could toggle specific features on or off based on their current needs or goals. Adapting to proficiency while allowing personalization of the tool interface could enhance usability and minimize cognitive overload by ensuring that only relevant features are presented.

6.1.2 Reveal information aids in stages.

We observed from the study that while participants generally preferred global features, they later resorted to local features when global features alone were insufficient for making decisions (§5.1.5). Considering this usage pattern, progressive disclosure could be an effective design strategy for managing the presentation of information aids. Progressive disclosure sequentially reveals information and functionalities, presenting only essential features initially while keeping more complex or less frequently used options accessible but hidden [19, 80, 102]. This approach reduces cognitive load, prevents information overload, and encourages efficient interaction with the interface [63, 73, 102]. In the context of our system, progressive disclosure could involve displaying global features by default while allowing users to explore local features on demand. For example, activating secondary displays through interface controls (e.g., clicking buttons or toggling options) could provide users with additional details without cluttering the primary interface.

6.1.3 Connect between information aids.

While integrating different information aids in a single interface was helpful in a more informed decision-making process—for example, by checking AI Explanation to understand why AI Score is high, and checking Example Sentence to understand why AI Explanation says the sentence is polite—participants had to make extra effort to make connections between such features, such as deciding which term to search in Example Sentence (§ 5.2.1). A more intuitive user interface design could effortlessly guide users through these information aids. For instance, AI Score might include an on-hover explanation or interactive icons detailing the reasoning behind certain scores. Moreover, visualizing attention scores [9] of AI-driven features, such as AI Score or AI Explanation, could help users easily find

key terms to delve deeper into. Attention scores measure the influence of different input tokens on a specific output token, and visualizing such influences (e.g., by highlighting each token with different opacity) could help users understand which input tokens significantly impact the model's output [4]. This support could be further enhanced by interaction designs that minimize user actions. For example, clicking the term with high relevance to the formality of a suggestion could automatically search the term in Example Sentence, streamlining the search process and reducing user effort.

6.1.4 *Make AI suggestions and information aids interactive.*

In this study, PARASCOPE provided static, non-editable suggestions and corresponding information aids, aligning with the design of existing tools [54, 71, 112]. However, participants preferred editable suggestions and dynamic features that could adapt in response to their suggestion edits (§5.3.2). This observation underlines the need to design paraphrasing tools that allow for a more flexible, interactive use of suggestions and information aids. We suggest that the suggestions provided within the interface be editable, so that users can modify suggestions within the interface. Previously static features like AI Score, AI Translation, and AI Explanation could be updated in response to the user's editing suggestions. For example, a user may take the increase in AI Score as the user changes the word in its suggestion, as the word becomes more appropriate. Similarly, while we implemented AI Explanation in a way that it only compares the original and paraphrased sentences, the future AI Explanation could be implemented so that it tracks the user edits to the suggestion and explains the effects of edits accordingly. Providing immediate feedback on modifications with AI-powered information aids like this could transform the tool into a 'what-if' analysis tool [111], where users can experiment with sentence modifications and immediately observe the implications on quality.

6.1.5 *Provide information aids for both original and paraphrased texts.* In our prototype, the original input text and its information aids were not displayed within the paraphrasing interface, following the design of existing tools [54, 112]. However, participants frequently considered their original sentences during decision-making, often comparing them to paraphrased suggestions and rejecting suggestions when they felt the original was sufficient. Participants expressed the need for information aids, such as AI Score and AI Translation, to be available for their original inputs to facilitate this comparison process (§5.3.4).

This finding suggests a design implication that information aids should encompass both original and paraphrased texts. Implementation of information aids could be adjusted accordingly; for example, our current metric, ParaScore [97], computes the AI Score of paraphrased text by measuring its semantic similarity and lexical divergence from the original, which is inapplicable to the original text itself. Instead, independent metrics, such as the metric for measuring fluency measurement [66], could offer a holistic view of both original and paraphrased texts.

6.2 Information Aids in Broader NNES Writing Tasks

Our study focused on designing information aids tailored to the specific needs of NNEs in email writing tasks (§3.3). However, the design of such information aids will likely vary depending on the requirements of different writing contexts. Exploring task-specific information aids offers a promising direction for future research, as distinct writing tasks emphasize different aspects of language.

As identified in our study, email writing emphasizes tone, formality, and appropriateness [107, 108]. In contrast, argumentative writing prioritizes clarity, conciseness, and logical coherence [71]. Future research could investigate the most effective types of information aids for supporting these priorities. Such aids might identify vague or redundant sentences, suggest more precise vocabulary, or evaluate the logical structure of arguments. They could also highlight weak transitions between ideas or provide templates for organizing persuasive content.

Another important area for exploration is addressing the challenges of working with LLM-generated content. Many LLM-based writing systems increasingly generate "watermarks"—formulaic patterns or predictable structures inherent in AI-generated text—to detect AI-generated contents [62]. Without adequate support, NNEs may struggle to identify and critically evaluate these watermarks. Information aids could help users navigate such challenges by visually flagging repetitive structures or suspicious phrases and providing actionable alternatives to improve fluency and originality. Such information aids could empower users to critically assess and refine AI-generated content, ensuring their writing remains both original and competitive.

Integrating task-specific information aids into writing systems raises several key research questions. How can these information aids be designed to address the demands of different tasks without overwhelming users? Are there universal aids that remain effective across various writing contexts? Addressing these questions could provide valuable insights into the design of adaptable and context-aware writing tools, ultimately tailored to the diverse needs of NNEs while maintaining usability and minimizing cognitive load.

6.3 Language Learning Effect of Information Aids

An intriguing direction for future research is to investigate whether information aids in writing tools can effectively promote language learning. Our study participants expressed optimism about the potential learning benefits of using information aids (§ 5.2.4). Future research could explore this by conducting long-term studies with NNEs students who consistently use the system, measuring the learning effects on language proficiency over time. Such research could provide valuable insights into the educational impact of integrating information aids into AI-driven writing systems.

However, findings from related work highlight a potential challenge: using LLMs can sometimes lead to over-reliance on AI, failing to improve users' language proficiency or even impact cognitive abilities [118]. This raises an interesting tension between the potential for language learning and the risk of fostering dependence on AI-generated suggestions. Investigating this balance could uncover

key design strategies for tools that encourage active learning while minimizing over-reliance.

6.4 Limitations

Our participant pool consisted exclusively of Korean individuals, which could limit the generalizability of our findings across different linguistic and cultural contexts. Participants were also primarily university students, narrowing the range of age and educational backgrounds represented. Future studies could involve a more diverse sample to examine how language proficiency and cultural variation influence engagement with AI-assisted writing tools. This work introduced PARASCOPE as a research artifact and examined its use through a mixed-method in-lab user study, observing real-time interactions and gathering both behavioral logs and reflective feedback in a controlled and task-oriented setting. While we sought to create a scenario as realistic as possible within laboratory constraints, in-lab studies may not fully capture the complexity of real-world writing contexts or longer-term usage patterns. Additionally, the scope of statistical analysis was limited by the relatively small number of participants. Future work could incorporate field deployments or longitudinal studies to evaluate system use over time and in more varied scenarios. Finally, the fixed interface layout—particularly the consistent ordering of information aid buttons—may have introduced ordering bias that influenced user preferences and interaction patterns. Future iterations could randomize or personalize feature placement to mitigate such effects and more accurately assess user behavior.

7 Conclusion

We investigated how NNEs use diverse information aids to assess and select AI paraphrase suggestions. By developing PARASCOPE, an AI paraphrasing assistant designed to integrate multiple information aids and collect user-system interaction data, we observed participants' paraphrasing workflows and preferences with information aids, revealing key patterns in how proficiency levels influence the use of information aids. While back-translation was the most frequently used aid, it was not sufficient on its own to drive decision-making, highlighting the importance of a combination of aids in supporting informed judgments. Our findings emphasize the potential of explainable AI paraphrasing tools to empower NNEs by enhancing their confidence, efficiency, and autonomy in writing tasks. However, careful consideration is needed to mitigate the risks of information overload when integrating multiple aids. Building on these insights, we propose actionable design implications for developing AI tools that effectively combine information aids, providing clear, contextualized explanations that meet the diverse needs of NNEs. These findings contribute to advancing explainable AI writing systems for NNEs and open avenues for broader applications in adaptive, user-centered writing support across various domains.

Acknowledgments

We thank our study participants for their valuable feedback on the design of PARASCOPE. We also thank the DIS 2025 ACs and reviewers for their thoughtful comments and suggestions. We disclose the use of generative AI tools in the process of writing this manuscript.

These tools were used solely for editing the authors' own text, and the authors ensured the final content was free from plagiarism, misrepresentation, fabrication, and falsification. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00337007).

References

- [1] Tazin Afrin, Omid Kashefi, Christopher Olshefski, Diane Litman, Rebecca Hwa, and Amanda Godley. 2021. Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 58, 13 pages. <https://doi.org/10.1145/3411764.3445683>
- [2] Mohamed Ally. 2005. Using learning theories to design instruction for mobile learning devices. , 5–8 pages.
- [3] Jose M. Alonso, A. Ramos-Soto, Ehud Reiter, and Kees van Deemter. 2017. An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* -, -, 1–6. <https://doi.org/10.1109/FUZZ-IEEE.2017.8015489>
- [4] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 412–421. <https://doi.org/10.18653/v1/D17-1042>
- [5] Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaña-Centellas, Kumar Paudel, Rachel Louise White, et al. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology* 21, 7 (2023), e3002184.
- [6] Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaña-Centellas, Kumar Paudel, Rachel Louise White, and Diogo Verissimo. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology* 21, 7 (07 2023), 1–27. <https://doi.org/10.1371/journal.pbio.3002184>
- [7] A. Ariyanti and Rinda Fitriana. 2017/10. EFL Students' Difficulties and Needs in Essay Writing. In *Proceedings of the International Conference on Teacher Training and Education 2017 (ICTTE 2017)*. Atlantis Press, -, 32–42. <https://doi.org/10.2991/iccte-17.2017.4>
- [8] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) (UIST '16). Association for Computing Machinery, New York, NY, USA, 603–608. <https://doi.org/10.1145/2984511.2984584>
- [9] Dmity Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). -. <https://api.semanticscholar.org/CorpusID:11212020>
- [10] Larry Beason. 2001. Ethos and Error: How Business People React to Errors. *College Composition and Communication* 53, 1 (2001), 33–64. <http://www.jstor.org/stable/359061>
- [11] Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics* 39, 3 (09 2013), 463–472. https://doi.org/10.1162/COLI_a_00166
- [12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., -, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [14] Jill Burstein and Magdalena Wolska. 2003. Toward Evaluation of Writing Style: Overly Repetitious Word Use. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Ann Copestake and Jan Hajič (Eds.). Association for Computational Linguistics, Budapest, Hungary, -. <https://aclanthology.org/E03-1003/>
- [15] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. <https://doi.org/10.1145/3411764.3445372>
- [16] Aoife Cahill, James Bruno, James Ramey, Gilmar Ayala Meneses, Ian Blood, Florencia Tolentino, Tamar Lavee, and Slava Andreyev. 2021. Supporting Spanish Writers using Automated Feedback. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Avi Sil and Xi Victoria Lin (Eds.). Association for Computational Linguistics, Online, 116–124. <https://doi.org/10.18653/v1/2021.naacl-demos.14>
- [17] Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2009. Common Sense Computing: From the Society of Mind to Digital Intuition and beyond. In *Biometric ID Management and Multimodal Communication*, Julian Fierrez, Javier Ortega-Garcia, Anna Esposito, Andrzej Drygajlo, and Marcos Faundez-Zanuy (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 252–259.
- [18] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management* 60, 1 (2023), 103111. <https://doi.org/10.1016/j.ipm.2022.103111>
- [19] John M Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (1984), 800–806.
- [20] Anna Uhl Chamot. 1987. A Study of Learning Strategies in Foreign Language Instruction. First Year Report. <https://api.semanticscholar.org/CorpusID:140882914>
- [21] Jim Chang and Jason Chang. 2015. WriteAhead2: Mining Lexical Grammar Patterns for Assisted Writing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Matt Gerber, Catherine Havasi, and Finley Lacatusu (Eds.). Association for Computational Linguistics, Denver, Colorado, 106–110. <https://doi.org/10.3115/v1/N15-3022>
- [22] David Chen and William Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). Association for Computational Linguistics, Portland, Oregon, USA, 190–200. <https://aclanthology.org/P11-1020/>
- [23] M-H Chen, S-T Huang, Jason S Chang, and H-C Liou. 2015. Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning* 28, 1 (2015), 22–40.
- [24] Yuexi Chen and Zhicheng Liu. 2024. WordDecipher: Enhancing Digital Workspace Communication with Explainable AI for Non-native English Speakers. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants* (Honolulu, HI, USA) (In2Writing '24). Association for Computing Machinery, New York, NY, USA, 7–10. <https://doi.org/10.1145/3690712.3690715>
- [25] Yuexi Chen, Yimin Xiao, Kazi Tasnim Zinat, Naomi Yamashita, Ge Gao, and Zhicheng Liu. 2025. Comparing Native and Non-native English Speakers' Behaviors in Collaborative Writing through Visual Analytics.
- [26] Ryuhaerang Choi, Taehan Kim, Subin Park, Jennifer G Kim, and Sung-Ju Lee. 2024. Private Yet Social: How LLM Chatbots Support and Challenge Eating Disorder Recovery. arXiv:2412.11656 [cs.HC] <https://arxiv.org/abs/2412.11656>
- [27] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML] <https://arxiv.org/abs/1706.03741>
- [28] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. <https://doi.org/10.1145/3491102.3501819>
- [29] Juliet Corbin and Anselm Strauss. 2008. Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory. <https://doi.org/10.4135/9781452230153>
- [30] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. <https://doi.org/10.1145/3526113.3545672>
- [31] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. <https://doi.org/10.1145/3544548.3580969>
- [32] Mark Davies. 2023. The Corpus of Contemporary American English (COCA). <https://www.english-corpora.org/coca/>.
- [33] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Alina Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3573–3590. <https://doi.org/10.18653/v1/2022.acl-long.250>
- [34] Muhammad Fareed, Almas Ashraf, and Muhammad Bilal. 2016. ESL learners' writing skills: Problems, factors and suggestions. *Journal of education and social sciences* 4, 2 (2016), 81–92.

- [35] John Flowerdew. 2007. The non-Anglophone scholar on the periphery of scholarly publication. *AILA Review* 20, 1 (2007), 14–27. <https://doi.org/10.1075/aila.20.04flo>
- [36] Liye Fu, Benjamin Newman, Maurice Jakesch, and Sarah Kreps. 2023. Comparing Sentence-Level Suggestions to Message-Level Suggestions in AI-Mediated Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 103, 13 pages. <https://doi.org/10.1145/3544548.3581351>
- [37] Michael Gamon. 2010. Using Mostly Native Data to Correct Errors in Learners' Writing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn (Eds.). Association for Computational Linguistics, Los Angeles, California, 163–171. <https://aclanthology.org/N10-1019/>
- [38] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) (DIS '22). Association for Computing Machinery, New York, NY, USA, 1002–1019. <https://doi.org/10.1145/3532106.3533533>
- [39] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing Using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) (DIS '22). Association for Computing Machinery, New York, NY, USA, 1002–1019. <https://doi.org/10.1145/3532106.3533533>
- [40] Gaëtanille Gilquin and Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1, 1 (2008), 41–61.
- [41] Google. 2023. Google Docs AutoCorrect. <https://docs.google.com/document/u/0/>.
- [42] Google. 2023. Google Ngram Viewer. <https://books.google.com/ngrams/>.
- [43] Google. 2023. Google Translate. <https://translate.google.com/>.
- [44] Grammarly. 2023. Grammarly: Free Online Writing Assistant. <https://www.grammarly.com>.
- [45] Shesen Guo and Ganzhou Zhang. 2007. Building a customised Google-based collocation collector to enhance language learning. *Br. J. Educ. Technol.* 38 (2007), 747–750. <https://api.semanticscholar.org/CorpusID:205566721>
- [46] Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring Methods for Generating Feedback Comments for Writing Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9719–9730. <https://doi.org/10.18653/v1/2021.emnlp-main.766>
- [47] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300809>
- [48] Md Naimul Hoque, Bhavya Ghai, and Niklas Elmqvist. 2022. DramatVis Persona: Visual Text Analytics for Identifying Social Biases in Creative Writing. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Virtual Event</city>, <country>Australia</country>, </conf-loc>) (DIS '22). Association for Computing Machinery, New York, NY, USA, 1260–1276. <https://doi.org/10.1145/3532106.3533526>
- [49] Chung-Chi Huang, Mei-Hua Chen, Ping-Che Yang, and Jason S. Chang. 2013. A Computer-Assisted Translation and Writing System. *ACM Transactions on Asian Language Information Processing* 12, 4, Article 15 (Oct. 2013), 20 pages. <https://doi.org/10.1145/2505984>
- [50] Jingshan Huang and Ming Tan. 2023. The role of ChatGPT in scientific communication: writing better scientific review articles. *American journal of cancer research* 13, 4 (2023), 1148.
- [51] Ju Chuan Huang. 2010. Publishing and learning writing for publication in English: Perspectives of NNES PhD students in science. *Journal of English for Academic Purposes* 9, 1 (2010), 33–44. <https://doi.org/10.1016/j.jeap.2009.10.001>
- [52] Yi-Ching Huang, Hao-Chuan Wang, and Jane Yung-jen Hsu. 2018. Feedback Orchestration: Structuring Feedback for Facilitating Reflection and Revision in Writing. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (<conf-loc>, <city>Jersey City</city>, <state>NJ</state>, <country>USA</country>, </conf-loc>) (CSCW '18 Companion). Association for Computing Machinery, New York, NY, USA, 257–260. <https://doi.org/10.1145/3272973.3274069>
- [53] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) (CHI '03). Association for Computing Machinery, New York, NY, USA, 17–24. <https://doi.org/10.1145/642611.642616>
- [54] Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. 2020. Langsmith: An Interactive Academic Text Revision System. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 216–226. <https://doi.org/10.18653/v1/2020.emnlp-demos.28>
- [55] Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jack Jamieson, and Kentaro Inui. 2023. Use of an AI-powered Rewriting Support Software in Context with Other Tools: A Study of Non-Native English Speakers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 45, 13 pages. <https://doi.org/10.1145/3586183.3606810>
- [56] Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the Human Revision Process in Scientific Writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9420–9435. <https://doi.org/10.18653/v1/2022.emnlp-main.641>
- [57] Matthew Jörke, Yasaman S. Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 902–918. <https://doi.org/10.1145/3581641.3584054>
- [58] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (April 2023), 29 pages. <https://doi.org/10.1145/3579592>
- [59] Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphor: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) (DIS '23). Association for Computing Machinery, New York, NY, USA, 115–135. <https://doi.org/10.1145/3563657.3595996>
- [60] Yewon Kim, Mina Lee, Donghui Kim, and Sung-Ju Lee. 2023. Towards Explainable AI Writing Assistants for Non-native English Speakers.
- [61] Tomi Kinnunen, Henri Leisma, Monika Machunik, Tuomo Kakkonen, and Jean-Luc LeBrun. 2012. SWAN - Scientific Writing Assistant: A Tool for Helping Scholars to Write Reader-Friendly Manuscripts. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Frédérique Segond (Ed.). Association for Computational Linguistics, Avignon, France, 20–24. <https://aclanthology.org/E12-2005/>
- [62] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2024. A Watermark for Large Language Models. arXiv:2301.10226 [cs.LG] <https://arxiv.org/abs/2301.10226>
- [63] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 2390–2395. <https://doi.org/10.1145/2858036.2858402>
- [64] Ola Knutsson, Teresa Cerratto Pargman, Kerstin Severinson Eklundh, and Stefan Westlund. 2007. Designing and developing a language environment for second language writers. *Computers & Education* 49, 4 (2007), 1122–1146. <https://doi.org/10.1016/j.compedu.2006.01.005>
- [65] Hiroe Kobayashi and Carol Rinnert. 1992. Effects of First Language on Second Language Writing: Translation versus Direct Composition†. *Language Learning* 42 (1992), 183–209. <https://api.semanticscholar.org/CorpusID:144621093>
- [66] Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 737–762. <https://doi.org/10.18653/v1/2020.emnlp-main.55>
- [67] Eka Kurniati and Rahmah Fithriani. 2022. Post-Graduate Students' Perceptions of Quillbot Utilization in English Academic Writing Class. *Journal of English Language Teaching and Linguistics* 7 (12 2022), 437. <https://doi.org/10.21462/jeltl.v7i3.852>
- [68] LanguageTool. 2023. LanguageTool - Online Grammar, Style & Spell Checker. <https://languagetool.org/>.
- [69] LanguageTool. 2023. LanguageTool API. <https://languagetool.org/http-api/>
- [70] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambagsans, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel

- Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1054, 35 pages. <https://doi.org/10.1145/3613904.3642697>
- [71] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. <https://doi.org/10.1145/3491102.3502030>
- [72] Weixin Liang, Mert Yuksekogul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- [73] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 2119–2128. <https://doi.org/10.1145/1518701.1519023>
- [74] Hajin Lim, Dan Cosley, and Susan R. Fussell. 2022. Understanding Cross-lingual Pragmatic Misunderstandings in Email Communication. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 129 (April 2022), 32 pages. <https://doi.org/10.1145/3512976>
- [75] Ting Liu, Ming Zhou, Jianfeng Gao, Endong Xun, and Changning Huang. 2000. PENS: A machine-aided English writing system for Chinese users. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. -, -, 529–536.
- [76] Ludwig. 2023. Ludwig • Find your English sentence. <https://ludwig.guru/>.
- [77] Ettore Mariotti, Jose M. Alonso, and Albert Gatt. 2020. Towards Harnessing Natural Language Generation to Explain Black-box Models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, Dublin, Ireland, 22–27. <https://aclanthology.org/2020.nl4xai-1.6/>
- [78] Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Towards Automated Document Revision: Grammatical Error Correction, Fluency Edits, and Beyond. arXiv:2205.11484 [cs.CL]
- [79] Ryo Nagata. 2019. Toward a Task of Feedback Comment Generation for Writing Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3206–3215. <https://doi.org/10.18653/v1/D19-1316>
- [80] Lloyd H. Nakatani and John A. Rohrlich. 1983. Soft machines: A philosophy of user-computer interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) (CHI '83). Association for Computing Machinery, New York, NY, USA, 19–23. <https://doi.org/10.1145/800045.801573>
- [81] Diane M Napolitano and Aanda Stent. 2009. TechWriter: An evolving system for writing assistance for advanced learners of English. *Calico Journal* 26, 3 (2009), 611–625.
- [82] Naver. 2023. Papago. <https://papago.naver.com/>.
- [83] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, -.
- [84] OpenAI. 2023. GPT-3.5 Turbo. <https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates>.
- [85] OpenAI. 2023. Introducing ChatGPT. <https://openai.com/blog/chatgpt>.
- [86] OpenAI. 2024. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>.
- [87] Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics* 24, 4 (2003), 492–518.
- [88] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] <https://arxiv.org/abs/2203.02155>
- [89] Taehyun Park, Edward Lank, Pascal Poupart, and Michael Terry. 2008. Is the sky pure today? AwkChecker: an assistive tool for detecting and correcting collocation errors. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) (UIST '08). Association for Computing Machinery, New York, NY, USA, 121–130. <https://doi.org/10.1145/1449715.1449736>
- [90] Charlene G Polio. 1997. Measures of linguistic accuracy in second language writing research. *Language learning* 47, 1 (1997), 101–143.
- [91] Prolific. 2024. Prolific | Quickly find research participants you can trust. <https://www.prolific.com/>.
- [92] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1051, 19 pages. <https://doi.org/10.1145/3613904.3642105>
- [93] Quillbot. 2023. Paraphrasing Tool - QuillBot AI. <https://quillbot.com>.
- [94] Marek Rei and Helen Yannakoudakis. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1181–1191. <https://doi.org/10.18653/v1/P16-1112>
- [95] Robert Ridley, Zhen Wu, Jianbing Zhang, Shujian Huang, and Xinyu Dai. 2023. Addressing Linguistic Bias through a Contrastive Analysis of Academic Writing in the NLP Domain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 16765–16779. <https://doi.org/10.18653/v1/2023.emnlp-main.1042>
- [96] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition (Virtual Event, Italy) (C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages. <https://doi.org/10.1145/3450741.3465253>
- [97] Lingfeng Shen, Lema Liu, Haiyun Jiang, and Shuming Shi. 2022. On the Evaluation Metrics for Paraphrase Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3178–3190. <https://doi.org/10.18653/v1/2022.emnlp-main.208>
- [98] Antonette Shibani, Ratnavel Rajalakshmi, Faerie Mattins, Srivarshan Selvaraj, and Simon Knight. 2023. Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing. In *Proceedings of the 16th International Conference on Educational Data Mining*, Mingyu Feng, Tanja Käpser, and Partha Talukdar (Eds.). International Educational Data Mining Society, Bengaluru, India, 183–193. <https://doi.org/10.5281/zenodo.8115695>
- [99] Momin N Siddiqui, Roy D Pea, and Hari Subramonyam. 2025. Script&Shift: A Layered Interface Paradigm for Integrating Content Development and Rhetorical Strategy with LLM Writing Assistants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 532, 19 pages. <https://doi.org/10.1145/3706598.3714119>
- [100] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 68 (Sept. 2023), 57 pages. <https://doi.org/10.1145/3511599>
- [101] Kacper Sokol and Peter Flach. 2018. Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (Stockholm, Sweden) (IJCAI'18)*. AAAI Press, -, 5785–5786.
- [102] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 29 (oct 2020), 32 pages. <https://doi.org/10.1145/3374218>
- [103] Anselm L. Strauss. 2017. *The discovery of grounded theory: Strategies for qualitative research*. Routledge, -.
- [104] Thesaurus.com. 2023. Thesaurus.com: Synonyms and Antonyms of Words. <https://www.thesaurus.com/>.
- [105] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971 (2023), -. <https://arxiv.org/abs/2302.13971>
- [106] DeepL Translate. 2023. DeepL Translate: The world's most accurate translator. <https://www.deepl.com/en/translator>.
- [107] Evangeline Marlos Varonis and Susan M Gass. 1985. Miscommunication in native/nonnative conversation. *Language in society* 14, 3 (1985), 327–343.
- [108] Jane A Vignovic and Lori Foster Thompson. 2010. Computer-mediated cross-cultural collaboration: Attributing communication errors to the person versus the situation. *Journal of Applied Psychology* 95, 2 (2010), 265.
- [109] Thiemo Wambganans, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022. Adaptive Empathy Learning Support in Peer Review Scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 227, 17 pages. <https://doi.org/10.1145/3491102.3517740>
- [110] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. LyriSys: An Interactive Support System for Writing Lyrics Based on Topic Transition. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (Limassol, Cyprus) (IUI '17)*. Association for Computing Machinery, New York, NY, USA, 559–563. <https://doi.org/10.1145/3025171.3025194>

- [111] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization & Computer Graphics* 26, 01 (jan 2020), 56–65. <https://doi.org/10.1109/TVCG.2019.2934619>
- [112] Wordtune. 2023. Wordtune - Rewrite Text in Seconds. <https://www.wordtune.com>.
- [113] Writefull. 2023. Writefull. <https://www.writefull.com/>.
- [114] Junjie Ye, Xuanning Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. arXiv:2303.10420 [cs.CL] <https://arxiv.org/abs/2303.10420>
- [115] Xing Yi, Jianfeng Gao, and William B. Dolan. 2008. A Web-based English Proofing System for English as a Second Language Users. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, -, -, -. <https://aclanthology.org/I08-2082/>
- [116] Seid Muhie Yimam and Chris Biemann. 2018. Demonstrating Par4Sem - A Semantic Writing Aid with Adaptive Paraphrasing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 48–53. <https://doi.org/10.18653/v1/D18-2009>
- [117] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. <https://doi.org/10.1145/3544548.3581388>
- [118] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.
- [119] Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A Web-based Revision Assistant for Argumentative Writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, John DeNero, Mark Finlayson, and Sravana Reddy (Eds.). Association for Computational Linguistics, San Diego, California, 37–41. <https://doi.org/10.18653/v1/N16-3008>
- [120] Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-Coach: Can Mixed Human-Written and Machine-Generated Text Be Detected?. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 409–436. <https://doi.org/10.18653/v1/2024.findings-naacl.29>
- [121] Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. 2023. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) (C&C '23). Association for Computing Machinery, New York, NY, USA, 368–370. <https://doi.org/10.1145/3591196.3596612>

A CEFR Rubrics

We list the rubrics we used for describing user English proficiency:

- A1 (Beginner): You can understand and use basic phrases and expressions. You can communicate in simple ways when people speak slowly to you.
- A2 (Elementary): You can participate in simple exchanges on familiar topics. You can understand and communicate routine information.
- B1 (Intermediate): You can communicate in situations and use simple language to communicate feelings, opinions, plans, and experiences.
- B2 (Upper Intermediate): You can communicate easily with native English speakers. You can understand and express some complex ideas and topics.
- C1 (Advanced): You can understand and use various languages. You can use English flexibly and effectively for social and academic purposes.
- C2 (Proficiency): You can understand almost everything you hear or read. You can communicate very fluently and precisely in complex situations.

B Technical Details

B.1 Paraphrase Generation

We leveraged GPT-3 [12] to generate multiple paraphrased suggestions. Below is the prompt used for generating paraphrases:

```
Generate four paraphrased variations for the given sentence(s) below.
###
Sentence(s): ask about
Paraphrased sentence(s):
inquire about
request details on
pose questions about
seek information about

Sentence(s): Nice to meet you.
Paraphrased sentence(s):
I'm glad to meet you.
Delighted to meet you.
Pleasure to make your acquaintance.
I am pleased to meet you.

Sentence(s): <ORG_TXT>
Paraphrased sentence(s):
```

The decoding parameters we used are:

- **Engine:** text-davinci-003
- **Max Tokens:** 250
- **Temperature:** 0.8
- **Top-p:** 0.9
- **Presence Penalty:** 0.5
- **Frequency Penalty:** 0.5

B.2 AI Explanation Generation

We used GPT-3.5 [84] to generate AI Explanation. Below is the prompt we used for generating AI Explanation:

Compare the paraphrased sentence(s) with the original sentence(s) regarding conveyed tone and appropriate use cases in less than 40 words.

###

Original: I'm sorry for the late submission of my assignment.
 Paraphrased: I apologize for the delay in submitting my assignment.
 Comparison: This sentence conveys a more formal tone. It could be used if the audiences are instructors, seniors, or anyone else where formal and respectful language is important.

Original: Nice to meet you.
 Paraphrased: Glad to meet you.
 Comparison: This sentence conveys the same tone using different wording. It could be used in either a professional context or a casual social encounter.

Original: decide to
 Paraphrased: resolve to
 Comparison: This phrase conveys a more determined and committed tone. It could be used when you want to emphasize your unwavering commitment to a decision.

Original: I am encountering considerable difficulty in comprehending the course material.
 Paraphrased: I'm really struggling to get a good grip on the course material.
 Comparison: This sentence conveys a more informal and open tone. It could be used when communicating with friends or peers.

Original: Could you please provide clarification on the third question in the assignment?
 Paraphrased: I was hoping you could help me understand the third question on the assignment.
 Comparison: This sentence maintains a polite and inquisitive tone. It could be used when communicating to your instructor or peers in an educational context.

Original: I'm looking forward to your lecture on Friday.
 Paraphrased: I am eagerly anticipating your lecture scheduled for this Friday.
 Comparison: This sentence has a more formal tone. It could be used when writing to academics, professionals, or in a formal setting where a higher level of vocabulary is expected.

Original: <ORG_TXT>
 Paraphrased: <PAR_TXT>
 Comparison:

The decoding parameters we used are:

- **Engine:** gpt-3.5-turbo
- **Max Tokens:** inf
- **Temperature:** 1
- **Top-p:** 1
- **Presence Penalty:** 0
- **Frequency Penalty:** 0

C User Study

C.1 Email Scenarios in the User Study

Table 3 shows the email scenarios we provided in the main study. The scenarios were created from the samples of the emails we received from the screening survey.

C.2 Thematic Analysis Codebook

Table 4 and Table 5 shows the codebook used for the thematic analysis of user experience on PARASCOPE.

Email Scenario	
1	Having attended a lecture on topic X, you are intrigued and wish to explore it further. You email the professor, seeking recommendations for related papers to expand your understanding. In your email, politely express your appreciation for the lecture, convey your enthusiasm for the subject, and request recommendations for relevant academic papers.
2	After reviewing your graded assignment, you notice an error in the calculation of your final score. You write an email to your professor explaining the issue to request a grade correction. In your email, politely explain the situation and express your concern about the potential impact on your grade. Show appreciation for their attention to the matter and your hope for a prompt resolution.
3	You have written a personal statement for your internship application. To refine your personal statement further, you've decided to reach out to your English professor and ask for their feedback to improve it. In the email, politely explain your purpose for sending the email, ask for their constructive feedback on the personal statement you attached, and appreciate their help.
4	You missed the class because you were sick and had to go to the hospital. You email your professor to ask if you can get an excused absence by presenting a medical record. When writing an email, politely explain your situation and show appreciation for your professor's understanding and consideration.
5	You are interested in auditing the course X at your university. You email the professor responsible for teaching course X to inquire about the possibility of auditing their course. In your email, briefly introduce yourself and politely explain your intention to audit the course. Inquire about the professor's policy on auditing and convey your enthusiasm for the subject.
6	You are planning to study abroad as an exchange student in the upcoming semester. However, you missed the deadline for the dormitory application, resulting in no dormitory assignment. You want to reach out to the university to inquire if they allow for a late dormitory application. In the email, politely express your situation, apologize for the oversight, and ask if there is a solution.

Table 3: The six email scenarios we used in the Main Task. We created the scenarios from the email excerpts from the recruitment survey.

Table 4: Codebook summarizing the dimensions, codes, code descriptions, and example quotes from participants regarding the usage purpose and patterns of PARASCOPE

Dimensions and Codes	Code Description	Example Quote
<i>What are users' purposes of using information aids?</i>		
Preserve intended meaning	Ensure the suggestion retains the original meaning.	<i>"I checked whether the suggested sentence included everything necessary or if anything important was missing or distorted."</i>
Assess tone appropriateness	Ensure tone and nuance fit the context and audience.	<i>"I mainly used it to see if what I wrote was appropriate for the situation and the intended reader."</i>
Explore real-world usage	Examine how expressions are used in real-world contexts.	<i>"I used it to see how unfamiliar expressions are used in actual sentences."</i>
Verify common usage	Confirm whether a phrase is commonly used or outdated.	<i>"It looked natural to me, but I wanted to make sure it was something people commonly say nowadays—like checking if it's not an outdated phrase no longer in use."</i>
<i>How do users use information aids to evaluate suggestions?</i>		
Screen for best fit	Narrow down options by dismissing weak suggestions.	<i>"My general process was to first filter out bad ones, then pick the best among the remaining options."</i>
Validate initial preference	Use features to validate an initial judgment.	<i>"Among the four options, I initially liked one the most and thought I would go with it. As I explored the features one by one, they confirmed what I had sensed—that it was the most polite and formal, just as I had thought."</i>
Cross-check information aids	Use one feature to interpret or validate another.	<i>"When I only looked at the AI Score, I doubted its reliability and couldn't understand why it rated a suggestion the highest. But when I checked the AI Translation and Explanation, it made sense, and I could accept the reasoning behind the score."</i>
Break tie with information aids	Let features guide selection when unsure between options.	<i>"If I was still unsure which of two sentences to use, I thought, perhaps the one with the higher numerical value would be better."</i>

Table 5: Codebook summarizing the dimensions, codes, code descriptions, and example quotes from participants regarding the user experiences and suggestions for PARASCOPE

Dimensions and Codes	Code Description	Example Quote
<i>What are perceived impacts of information aids on user experience?</i>		
Informed decision-making	Having access to diverse information aids supported more confident and informed choices.	<i>“Having access to various types of information that either confirmed or refuted the suitability of a suggestion made me feel more reassured and confident in my choice.”</i>
Improved efficiency	Having diverse information aids integrated in one interface streamlined the suggestion selection process.	<i>“Typically, my writing process involves constantly switching between windows: composing, searching dictionaries, and consulting ChatGPT. This system consolidates these actions within one platform, significantly aiding my workflow.”</i>
Cognitive overload	The abundance of information aids sometimes caused mental fatigue and inefficiency.	<i>“I feel compelled to use every information aid, even for sentences I could easily move on from. I don’t think this process is efficient, but I am nonetheless convinced that it improves writing.”</i>
Increased autonomy	Interacting with information aids enhanced agency and authorship in the writing process.	<i>“Utilizing features to understand how suggestions might sound and integrating these suggestions into my text, made me feel more actively involved in the writing process. It felt as if I were crafting the text myself.”</i>
Potential language learning	Participants perceived interacting with information aids as opportunities for language learning.	<i>“Example Sentence, unlike other features that provide direct hints, allows for the indirect exploration of various sentences and fosters thoughtful consideration.”</i>
<i>What are the factors that influence users’ feature usage experience?</i>		
Simplicity of presentation	Participants preferred concise, direct information presentation format.	<i>“I needed to make quick decisions on suggestions, but since it’s presented in long paragraphs, it wasn’t easy to skim. If it were in bullet points, it would have been faster and more convenient.”</i>
Explainability	Participants expressed a need for explainability and transparency about how numeric features were derived.	<i>“It wasn’t clear what database the trends shown in Frequency were based on. For example, is this frequency higher because the word appears often in news articles? Knowing this would make the feature more helpful.”</i>
Personalization	Participants wished for an interface tailored to personalized usage patterns.	<i>“When using a tool repeatedly, people tend to stick to certain features. Instead of showing all features, it might be better to display only the ones I actively use.”</i>
Interactive refinement of suggestions	Participants wished to directly modify suggestions without leaving the interface or fully rewriting the text.	<i>“I didn’t want to completely rewrite my sentence; I aimed to keep the original structure intact while making minor adjustments to the words or expressions.”</i>
Incorporate original text for comparison	Participants found it difficult to evaluate suggestions without seeing their original input.	<i>“It would be more convenient if my original input was displayed alongside the suggestions for direct comparison. Paraphrasing provides four candidates, but the original sentence is essentially a fifth option. Including translations or scores for the original would make it easier to evaluate all options equally.”</i>