Use of an AI-powered Rewriting Support Software in Context with Other Tools: A Study of Non-Native English Speakers

Takumi Ito
t-ito@tohoku.ac.jp
Tohoku University and Langsmith Inc.
Japan

Naomi Yamashita
naomiy@acm.org
NTT Communication Science Labs
Japan

Tatsuki Kuribayashi
kuribayashi@tohoku.ac.jp
Tohoku University and Langsmith Inc.
Japan

Masatoshi Hidaka
hidaka@edgeintelligence.jp
Edge Intelligence Systems Inc.
Japan

Jun Suzuki
jun.suzuki@tohoku.ac.jp
Tohoku University and RIKEN
Japan

Ge Gao
gegao@umd.edu
University of Maryland
United States

Jack Jamieson
jack@jackjamieson.net
NTT Communication Science Labs
Japan

Kentaro Inui
inui@tohoku.ac.jp
Tohoku University and RIKEN
Japan

ABSTRACT

Academic writing in English can be challenging for non-native English speakers (NNESs). AI-powered rewriting tools can potentially improve NNESs' writing outcomes at a low cost. However, whether and how NNESs make valid assessments of the revisions provided by these algorithmic recommendations remains unclear. We report a study where NNESs leverage an AI-powered rewriting tool, Langsmith, to polish their drafted academic essays. We examined the participants' interactions with the tool via user studies and interviews. Our data reveal that most participants used Langsmith in combination with other tools, such as machine translation (MT), and those who used MT had different ways of understanding and evaluating Langsmith's suggestions than those who did not. Based on these findings, we assert that NNESs' quality assessment in AI-powered rewriting tools is influenced by the simultaneous use of multiple tools, offering valuable insights into the design of future rewriting tools for NNESs.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

writing support, rewriting tools, text suggestions, natural language processing

ACM Reference Format:
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 *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3586183.3606810

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10.
https://doi.org/10.1145/3586183.3606810

1 INTRODUCTION

English has become the world's lingua franca, bringing inevitable disadvantages to non-native English speakers (NNESs), especially those with low English proficiency, in numerous aspects of work and business. Academia is one such field where English dominates. NNES researchers may find it challenging to write an academic manuscript in English concisely, clearly, and fluently, without spelling or grammatical errors. Helping NNESs overcome such language barriers is important to achieve diversity and inclusion in academia [28].

AI-powered rewriting tools play an important role in diminishing these language barriers [9, 12]. The current designs of such tools typically involve machines suggesting potential revisions (e.g., correcting grammatical errors and improving wording) to a user-written draft sentence [22]. However, users must themselves determine whether such revisions are compatible with their writing goals, regardless of their English abilities. This raises several questions about NNESs in these situations, including why they accept or reject AI-provided revisions, and how they form an overall assessment of these AI-powered tools. This issue has become increasingly pertinent as more individuals are utilizing AI-powered tools, such as ChatGPT[1], to polish up their sentences. Despite the sophistication of such AI tools, users are ultimately responsible for deciding whether or not to adopt the suggested changes. Understanding how users (e.g., NNESs) interact with and assess AI-powered tools is an important step toward designing better human-AI collaborative writing systems. To date, the use of AI-powered rewriting tools has been studied extensively in human-computer interaction (HCI) [4, 29, 54]. Yet, little of this research has focused on NNESs.

[1] https://openai.com/blog/chatgpt

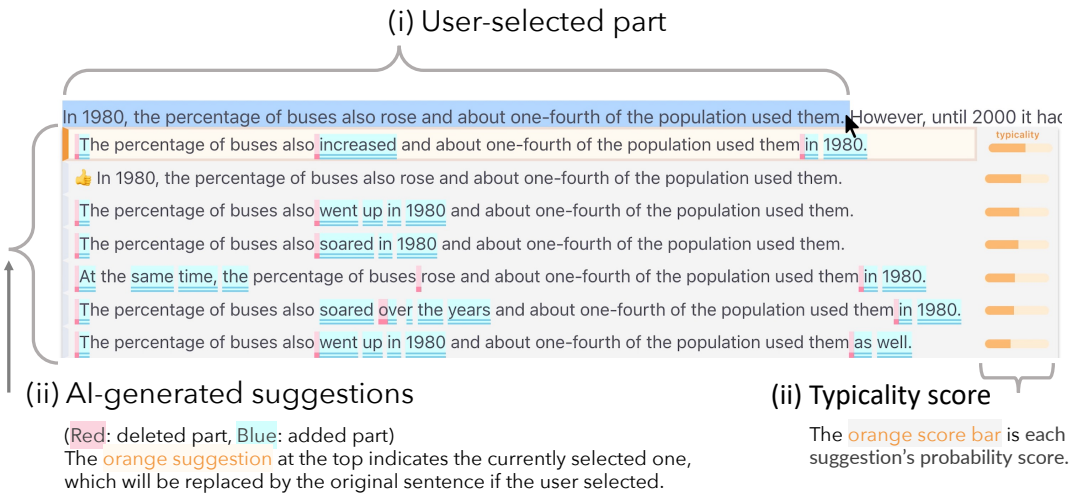

Figure 1: Screenshot of Langsmith. (i) The sentence selected by the user. (ii) The blue and red highlights indicate the addition and deletion of words or phrases in the user-selected sentence, respectively. (iii) The orange typicality bar indicating how natural a sentence is that is calculated using a language model.

Through a series of investigations using Langsmith [22] (Fig. 1), we studied how NNESs assess the validity of proposed revisions made by an AI-powered rewriting tool. Like other rewriting tools, Langsmith suggests potential revisions in terms of grammaticality, fluency, and style once the user inputs a draft sentence. Inspired by prior work [33] showing that NNESs uses a variety of resources, including search systems and dictionaries, to verify machine translation (MT) results, we conducted a preliminary investigation on resources that NNESs often use along with Langsmith. Our preliminary survey of NNES Langsmith users in Japan [2] revealed that many used MT in conjunction with this rewriting tool. This suggests that the use of MT is a common practice among NNES Langsmith users in Japan, and raises the question of how this might impact their assessment of the potential revisions offered by the rewriting tool. To explore this issue further, we conducted user studies and interviews. When analyzing the results, we observed differences in decision-making between those who did and did not use MT tools alongside Langsmith. Specifically, we investigated how members of these groups differed regarding whether and how they adopted the rewriting tool's suggestions, and explored factors influencing their overall assessment of the tool.

Our findings suggest that NNESs who used MT tend to face difficulty making appropriate selections on their own when assessing rewriting suggestions, and so they relied on other resources such as the score provided by the tool and back-translation. Furthermore, those NNESs who depended on other language support resources, such as MT, were more likely to lose confidence in the rewriting support tool when they discovered evident errors in the tool's suggestions. In contrast, NNESs who did not use MT in conjunction with the rewriting tool were more likely to view such errors as common and were less likely to be influenced in their assessment of the tool's effectiveness. There was some evidence that NNESs who relied on MT tended to possess lower English proficiency, while those who relied on their own abilities tended to have higher proficiency. These findings suggest that NNESs' perception and behavior towards an AI-based rewriting tool are not solely affected by the tool itself, but also by their simultaneous use of other tools such as MT. We provide valuable insights to the future design of rewriting tools for NNESs, emphasizing on the importance of studying these tools in conjunction with other language support resources and understanding the role of MT and other language support resources in shaping users' perceptions and behaviors when using rewriting support tools.

[2]Japan is one of the countries with low English proficiency according to the EF English proficiency index (www.ef.com/wwen/epi/regions/asia/japan/).

2 RELATED WORK

2.1 Writing environments for NNESs

English is the dominant language in numerous domains, including academia. NNESs may find themselves at a disadvantage due to a lack of knowledge of English grammar, collocations, phrases, and style [15, 20, 41, 44]. Huang [20] interviewed NNES Ph.D. students and found that many felt at a disadvantage due to their limited English proficiency. NNES students appear to draft the content of their planned writing in their native languages and then translate it into the target language during the writing process [5]. Thus, a lack of English skills can prolong writing times and, in some cases, lead to manuscript rejection despite reporting valuable research results [15, 20, 43, 44]. Although translation and editing services can be used to address this language barrier, this comes at a steep financial cost [44]. In addition to negative career impacts in 'publish or perish' academic environments, these barriers reduce the visibility of NNESs' research, which is detrimental to diversity and inclusion goals.

2.2 AI-powered rewriting tools

To overcome the above-mentioned barriers in writing and publishing activities, researchers have proposed and developed rewriting software specifically for academic papers written in English by NNESs [7, 10], which suggest more fluent sentences to their written text [13, 22, 23, 25, 37]. Many of these tools employ a user interface (UI) that suggests multiple alternatives to human-written drafts in their user interfaces [3, 22, 29, 54], and the human author then selects the one that best fits his or her intentions. Several UIs display scores alongside their suggestions, such as a prediction probability or confidence score [e.g., 16] in the hope that they will improve the users' performance in evaluating and adopting model outputs [14].

2.3 Interaction between NNESs and AI-powered writing tools

Although the performance of AI-powered writing support tools, including rewriting tools, is improving dramatically, their suggestions sometimes contain errors [24]. Moreover, AI-powered writing support tools are unable to provide users with reasons for their suggestions. Therefore, even if a probability score for each suggestion is presented, users must decide which suggestion to accept as appropriate based on information extrinsic to the system: often, a personal knowledge of English, which is inevitably lower among NNESs than among native speakers.

Studies indicated that NNESs have difficulty thinking and making judgments in English [5], and therefore tend to rely on MT. According to previous work [2, 30, 32], MT aids the NNESs' English-writing process by allowing them to write in their native language and ultimately helps them write better English than on their own.

Despite these findings indicating that NNESs tend to use their native language during their English-writing process and increasingly rely on MTs, research on human interaction with AI-powered writing support tools has mostly focused on native English speakers [4, 29, 54]. Among few exceptions, Buschek et al. [3] found that NNESs were more accepting of AI suggestions than native English speakers. However, the reason behind this result remains unclear, indicating a lack of understanding of how NNESs are affected by and use of the tools. Similarly, Ito et al. [22] have demonstrated that an AI-powered rewriting tool improves the English writing of Japanese students; however, how users assess and adopt rewriting suggestions has not been investigated. To address this gap in the existing research, we conducted a behavioral study on the users of an AI-powered writing tool.

3 RESEARCH TOOL: LANGSMITH

We adopted Langsmith [22], an AI-powered rewriting tool specifically tuned for academic English (Figure 1), as the focus of our research. The main users are NNES Japanese researchers and students.

While Langsmith offers several functions, including autocompletion, this study focuses on the rewrite function, which is its main feature, as indicated by the survey results of Ito et al. [22]. For each sentence that the user selects using the cursor, Langsmith suggests multiple rewritten options, as shown in Fig. 1. If the user selects part of a sentence, Langsmith intensively suggests rewritten options for that part.[3] Furthermore, Langsmith provides a "typicality score" for each rewrite candidate (orange bars shown in Fig. 1); the higher the score, the more likely the candidate is to be a good choice. This score is based on the generation probability of that sentence as calculated by the language model. A language model is often used to assess sentence fluency in NLP [27, 53]. In the present context, the typicality score is relative between suggestions, and the suggestions are ranked by the typicality score.

These functions are broadly the representative of other writing support tools that suggest rewritten alternatives to users' input [e.g., 3, 29, 54]. In particular, rewriting is a common feature in AI-powered writing support tools [8], such as wordtune[4] and quillbot[5]. Consequently, studying the ways in which users employ Langsmith provides findings that are relevant to other software as well. We chose to focus on Langsmith because, for this study, we were able to provide users with a modified version of Langsmith suitable for addressing our research questions. Additionally, investigating first-time users is advantageous in controlling other factors that could affect their impressions of the rewriting tool. Although users' perceptions of a rewriting tool may change over time, our study provides an initial step toward comprehending how users' engagement with a rewriting tool is shaped by simultaneously using complementary technologies.

4 PRELIMINARY STUDY

Our preliminary study is aimed at investigating what other tools people use in conjunction with Langsmith in their writing process. Using a survey, we asked about the tools the respondents usually used when writing papers in English and how they used Langsmith. This survey was sent to individuals on the Langsmith user mailing list and also posted on Twitter. We received responses from 39 Japanese adults (33 males and 6 females), whose average age was 32 (range: 22–54). They included thirteen faculty members, nine Master's students, nine Ph.D. students, one industry researcher, five public institution researchers, and two others. Survey results indicated that 85% of the respondents used MT tools such as DeepL and Google Translate in conjunction with Langsmith. A total of 75% used grammatical error correction tools like Grammarly and Trinka. More than 70% of the respondents also reported using other online/offline resources (e.g., Hyper Collocation and Power Thesaurus) to look up example sentences and alternative expressions. To the question of how they usually use Langsmith, exactly half of those surveyed answered that they used other editors to create English text and copied them into Langsmith as needed; under a third responded that they pasted MT output into Langsmith for editing; and five stated that they wrote sentences into Langsmith directly.

In sum, respondents typically used the rewriting tool in combination with other tools — mostly MT and other rewriting tools — while writing academic English. This result helps us formulate specific research questions and construct the design for our main study. We formed our research questions around how the use of

[3]Langsmith may rewrite areas outside of the selected range if the rewrite of the part requires another adjustment or if there are errors.
[4]https://wordtune.com/
[5]https://quillbot.com/

other tools affects NNESs to assess the appropriateness/validity of the rewriting suggestions. Because numerous users employed MT, we were particularly interested in understanding how this affects their use of rewriting tools and the writing process, and what factors influence the usefulness of the tools.

5 RESEARCH QUESTIONS

Our preliminary study and recent literature reviewed in Section 2 show that NNESs increasingly rely on MT tools. Thus, we pose the following research question:

RQ1: How do NNESs use MT alongside an AI-powered rewriting tool during their writing process?

Based on the answers to RQ1, we investigate how NNESs use AI-powered rewriting tools. When an AI-powered rewriting tool recommends word replacements, grammatical corrections, etc., it is not always possible for NNESs to make informed choices about whether these recommendations are worth adopting. Prior research has reported that NNESs use various strategies to verify MT output, such as using external resources (e.g., dictionaries, other MT software, and back-translation) or asking experts [33]. In the case of rewriting tools that provide their suggestions in English, however, it is unclear what strategies NNESs use to determine whether a suggestion should be adopted or rejected. Therefore, we ask

RQ2: How do NNESs decide whether or not to accept the suggestions of an AI-powered rewriting tool? What, if any, is the difference in the approach of NNESs who use MT and those who do not in assessing the suggestions?

Finally, we are interested in how users form an overall assessment of a rewriting tool, such as their perceptions of usefulness and reliability. This is an important issue because whether users continue to use a system depends on these perceptions [11]. We pay particular attention to what kind of writing suggestions influence NNESs assessments.

Further, we investigate differences between those who use MT and those who do not because we want to understand how rewriting tools are evaluated in relation to other tools, especially MT. Therefore, our next question is

RQ3: What factors are key to NNESs' overall assessment in a rewriting tool, and are there differences between those who use MT and those who do not regarding these factors?

6 MAIN STUDY

To explore the research questions presented in Section 5, we designed the main study, in which NNESs performed English-writing tasks using Langsmith's rewriting mode, followed by an online interview. To ensure that the focus of the main study remains solely on rewriting, we disabled features other than the rewriting feature.[6] Participants in the study were Japanese researchers and students recruited through a crowdsourcing platform. We collected data from screen recordings of them doing English writing tasks and transcriptions of the interviews. We then analyze their writing process and compare those who used MT to those who did not use MT. The detailed methodology is described in subsequent sections.

[6]Langsmith offers a feature wherein adding the special token '()' prompts the software to provide suggestions by replacing the token with a word or phrase. As this feature is not found in tools such as wordtune and quillbot, we disabled it for this study.

6.1 Procedure

A crowdsourcing platform, CrowdWorks[7] was used to recruit 24 NNESs, of whom 21 completed the writing tasks, as detailed in Section 6.3 below. After receiving a briefing on the study and signing consent forms, they were shown a video that explained how to use Langsmith. To familiarize themselves with this tool and the task format, they were asked to perform a brief (approximately 10-minute) practice writing task using Langsmith. Subsequently, they were asked to work on two writing tasks and make screen recordings during their writing process. The tasks were distributed to the participants via Google Docs, and we asked them to write their final answers on the same Google Docs files. The order of the two tasks was randomized among the participants. As our preliminary study indicates that NNESs use various tools for writing, we allowed the use of other tools besides Langsmith. However, the use of any such tools that could not be captured in screen recordings was forbidden. Notably, all participants had default access to the spelling- and grammar-correction functions of Google Docs.

Online semi-structured interviews were conducted within a week after both main tasks were completed. We randomly selected 15 participants and invited them for a follow-up interview, and 14 participated. The interview protocol was designed to encourage the interviewees to reflect on their writing process, including questions about their general practices when writing academic papers, use of writing-support tools, opinions and impressions of Langsmith, and how they assessed/selected the suggestions provided. All interviews were conducted in Japanese; they were audio-recorded, and lasted approximately one hour (range: 56–74 minutes). Compensation for participation in the main study was calculated based on the local pay rates for part-time work: participants who participated only in the writing task received 4,500 yen, and those who participated in both the writing task and the interview received 7,500 yen.

6.2 Writing tasks

We adapted all writing tasks from examples of the International English Language Testing System (IELTS) Academic Writing Task 1 posted on the website of iPassIELTS (an online IELTS course provider) with the company's permission. Each task comprised a bar graph and a table, which the participants were asked to describe.[8] The original IELTS Academic Writing Task 1 requires examinees to write a minimum of 150 words, and the estimated time for completing it is 20 minutes.[9] However, when we initially asked two NNES members of our laboratory team to complete a sample task, we found that a 20-minute limit left them very little time to polish their writing. Therefore, we provided a 30-min window for each writing task. However, we did not set a strict time limit, as our principal goal was not to assess the participants' English-writing ability, but rather to obtain a detailed picture of their English-writing process.

6.3 Participants

We recruited the participants for the main study via CrowdWorks, a Japanese crowdsourcing service. The participation criteria were:

[7]https://crowdworks.jp/

[8]https://www.ipassielts.com/ielts_training/study_plans_single/leisure-time and https://www.ipassielts.com/ielts_training/study_plans_single/ielts_task1_hotel_occupancy

[9]https://www.ielts.org/for-test-takers/test-format

Table 1: Demographics of participants in main study. ∗ indicates those who participated in the interview. CEFR is a language proficiency scale consisting of six levels: A1=beginners < A2 < B1 < B2 < C1 < C2= proficient. "Papers" shows the participants' number of English publications before the study. ✓indicates the participants who used MT in the main study.

	CEFR	Papers	Profession	Use of MT
P1∗	A2	0	Ph.D. stu.	✓
P2	B1	7+	faculty member	✓
P3∗	B2	0	undergraduate stu.	
P4∗	B2	5 − 6	faculty member	
P5∗	C1	0	undergraduate stu.	
P6∗	n/a	0	Ph.D. stu.	✓
P7	C2	7+	public institution res.	
P8∗	n/a	7+	public institution res.	✓
P9	B1	0	master's stu.	✓
P10∗	A2	0	master's stu.	✓
P11∗	B1	0	master's stu.	✓
P12∗	A2	3 − 4	undergraduate stu.	✓
P13∗	n/a	7+	faculty member	
P14∗	B1	1 − 2	Ph.D. stu.	✓
P15	A1	1 − 2	master's stu.	✓
P16∗	B1	0	master's stu.	✓
P17	A2	0	Ph.D. stu.	✓
P18∗	B2	0	Ph.D. stu.	✓
P19	B2	3 − 4	industry res.	✓
P20∗	n/a	1 − 2	Ph.D. stu.	
P21	A2	1 − 2	physician	✓

1) experience in writing academic English within the previous three years, and/or 2) a plan to write a paper in English within the following year. Of the 24 Japanese students and researchers who were recruited [10], three failed to complete the study due to technological issues, resulting in a final pool of 21 participants (13 males, 8 females) with a mean age of 31.8 (range: 20-47). None of them had prior experience using Langsmith. This is because we were interested in how their interactions with other tools might shape their initial impression of the tool (RQ3), which was made possible by focusing on novice users. Three were undergraduate students, four were Master's students, seven were Ph.D. students, one was an industry researcher, two were public institution researchers, three were faculty members, and one was a physician. Nine had never written a manuscript in English, but were intending to do so in the coming year, while four were rather experienced with more than seven English publications. Participants' English proficiency levels were reported as follows: one beginner (A1), five elementary (A2), four intermediate (B1), four upper-intermediate (B2), one advanced (C1), one professional(C2) on the Common European Framework of Reference for Languages (CEFR), and four did not report previous English proficiency testing. 14 participants (i.e., those marked with asterisks in Table 1) participated in the post-task interviews.

[10]All participants' first language was Japanese.

6.4 Measurement and analysis

We collected screen recording data from 21 participants and audio data from 14 interviews. Because we were interested in how participants used other tools alongside Langsmith, we first noted which tools they used during the writing tasks. 15 participants (71%) used MT, and 11 (52%) used web searches in addition to Langsmith. Four also used Grammarly, a grammatical error correction tool. Although Grammarly has a full-sentence rewriting feature, this feature is only available to paid users, and all four of the participants in question used the free plan.[11]

MT vs. noMT groups. Because MT was the most commonly used tool, particularly important for NNESs, we divided our participants into two groups, namely the *MT* and *noMT* groups. This allowed us to differentiate between the assessment of Langsmith's suggestions between those who used MT and those who did not. We compared these two groups' writing performances and analyzed differences in how their respective member sets assessed the suggestions made by Langsmith and formed impressions of it.

Writing performance. In understanding the differences in the writing processes between the *MT* and *noMT* groups, it is crucial to consider the participants' writing performance. We asked iPassIELTS to rate the participants' writing based on the indicators usually employed when providing course feedback.[12] Each participant's text was evaluated in four aspects: "Task achievement," "Coherence and cohesion," "Lexical resource," and "Grammatical range and accuracy." Each aspect consists of two scoring items, and all items were rated on the same four-point scale, i.e., 1 = satisfactory, 2 = good, 3 = very good, and 4 = excellent.

Video recordings. We used video recordings to address RQ1 and RQ2. To identify the tools used by participants along with Langsmith (RQ1), we reviewed the video recordings and counted the number of participants using each tool. For MT, we categorized its usage into two types: forward-translation (Japanese → English) and back-translation (English → Japanese). Forward-translation is commonly used to translate complex sentences into a foreign language, whereas back-translation is used to confirm the accuracy of MT output [33]. To address RQ2, we observed the video recordings and identified which tools were used for assessing the Langsmith suggestions. Importantly, the use of MT to validate Langsmith's suggestions was limited to back-translation, as Langsmith's suggestions were listed in English.

Interviews. The interviews added nuance to the findings from video analysis in RQ 1 and 2, shedding light on participants' common practices of using various resources including MT when writing in English, as well as their reasons for using each tool. During the interview, participants were shown clips from video recordings in which Langsmith was being used or switched to another tool, and then we asked how the suggestions were assessed, and why the tools were switched. Additionally, the interviews addressed RQ3, investigating participants' attitudes toward AI-based rewriting suggestions and identifying factors that influenced their assessment

[11]Plans and feature details of Grammarly: https://www.grammarly.com/plans (accessed on January 9, 2023)

[12]Sample feedback: https://www.ipassielts.com/images/uploads/Nuri_GDP_growth_web.pdf

Table 2: Comparison of writing performance between MT and noMT. TA=Task achievement; CC=Coherence and cohesion; LR=Lexical resource; GRA=Grammatical range and accuracy. Total is the sum of those criteria. Each value is the mean score of the participant's writings, and the value in parentheses is the standard deviation. The values in parentheses for groups are headcounts.

Group	TA	CC	LR	GRA	Total
MT (15)	2.5 (1.1)	6.2 (1.1)	6.7 (1.3)	7.5 (0.9)	22.9 (3.7)
noMT (6)	2.7 (0.5)	6.2 (1.7)	7.0 (1.5)	7.5 (0.6)	23.3 (4.0)

of the tool. Therefore, for RQ3, we asked, for example, "when did you use Langsmith and why did you use it at that time?", "what features were good or bad when you used the system?", and "would you like to continue to use Langsmith, and why?" In addition, since using Langsmith alongside other tools may affect the usefulness or value of Langsmith, we asked, for instance, "what are the forms of assistance that Langsmith can provide that other tools can't, and vice versa?"

The recorded interviews were transcribed using an automatic-transcription tool. The transcripts were then reviewed and corrected by the first author. We identified themes in the transcripts using an inductive approach [6]. Two authors separately analyzed one-third of the transcripts and sorted them into meaningful categories, while identifying relationships between the themes and looking for salient themes. The same two authors then iteratively checked and elaborated on their codes until agreement and saturation were reached. Then, the first author coded the rest of the transcripts.

7 FINDINGS

We report on the participants' writing performance and subsequently present our findings organized around our three research questions. When presenting quotations from participants, their ID numbers include an "-MT" or "-noMT" suffix according to whether that person used MT during the writing process.

Writing Performance. The average number of words produced in the two tasks was almost identical: 161 words (SD = 13.6, range: 131-186) for the bar-graph task and 165 words (SD = 17.1, range: 125-209) for the table task. Two participants, P10-MT and P12-MT, took more than 30 minutes on both of their tasks, yet failed to reach the 150-word goal on either. The participant's score for each of the four dimensions of writing performance was the sum of the relevant iPassIELTS-assigned item scores across both writing tasks. As indicated by the scores presented in Table 2, there were no significant differences between the *MT* and *noMT* groups in any writing-quality dimension.

7.1 NNESs' writing methods (RQ1)

This section describes the *noMT* and *MT* groups' respective writing methods — how they used Langsmith and MT in their writing process (RQ1) — based on 1) our observations of the participants' screen recordings, and 2) the interview data.

7.1.1 Usage of the MT during writing. MT was used by 15 of the 21 participants during writing tasks. Notably, all beginner (A1) and elementary English level participants (A2) used MT, while advanced (C1) and professional level (C2) participants did not use MT. Out of the 15 participants, 13 used forward translation to some degree. Six drafted the full text in Japanese and forward-translated it into English, and seven others drafted the text in Japanese for some parts and in English for others, and forward-translated the Japanese parts into English. Back-translation, on the other hand, was used by 10 participants, eight of whom also used forward-translation. In most cases, these participants would back-translate their English text into Japanese after refining it with Langsmith.

7.1.2 Writing methods.

noMT group. All participants in the *noMT* group created the draft themselves, entirely in English, and revised it with Langsmith. Some of them occasionally searched the Web using Japanese keywords, referred to Japanese-English dictionaries when encountering difficulty expressing certain concepts in English. They also consulted a thesaurus or an English-Japanese dictionary when a word or phrase they had written in English seemed inappropriate or unclear. As P4-noMT noted,

> *"I don't think about the text in Japanese when writing scientific papers."*

MT group. In contrast, all participants in the *MT* group revealed in their interviews that they used the MT technology regularly, i.e., not only for the writing tasks of this study. As P18-MT explained:

> *"I use machine translation almost all the time, except for expressions that come up many times or that I use often. So I think I rely on the machine translation 80% of the time when I write in English."*

Many (13/15) participants in the MT group created the drafts with forward-translation, and then revised them by repeating Langsmith and back-translation. They also occasionally edited the English text by themselves after back-translation. However, two (P1-MT and P6-MT) never edited the English text by themselves. Instead, those two participants rewrote the source Japanese text and repeated the process of forward-translation, Langsmith, and back-translation until a back-translation result (in Japanese) that they were comfortable with appeared. These participants seemed to follow a similar practice (i.e., repeat forward and backward translations) in their daily English academic writing:

> *"I usually draft some Japanese text, translate it into English using DeepL, and then translate it back into Japanese, again using DeepL. I check to see whether the Japanese is properly translated, and if it is not, I look at what is wrong and correct the sentences one by one."* (P1-MT)

7.1.3 Summary of findings (RQ1). NNES participants who did not use MT during the writing process rarely used tools other than Langsmith to revise their English text. However, those who used MT in the writing process often wrote sentences in their first language (Japanese) and translated them into English. After refining the English with Langsmith, they further back-translated it and evaluated its validity. These results show that NNESs who use MT

in conjunction with Langsmith to write English texts make creative use of MT's forward- and back-translations in combination, before and after using the rewriting tool.

7.2 NNESs' assessment of rewriting suggestions (RQ2)

To explore NNESs' decision-making on whether/how they adopt Langsmith's suggestions (RQ2), we first examined video recordings to assess what other tools were used to verify Langsmith's output. Then, we conducted a thematic analysis of NNESs' interview quotes and identified their assessment strategies.

7.2.1 Usage of other tools to check Langsmith output. Table 3 lists the tools used by the participants to verify Langsmith's output. To explore NNESs' decision-making on whether/how they adopt Langsmith's suggestions, we first examined video recordings to assess what other tools were used to verify the suggestions. If the other tool was used after reviewing with Langsmith, we considered that to be an assessment of Langsmith's suggestion. In the *MT* group, 10/15 participants used MT for back-translation, 3/15 used web search, 2/15 used a grammatical error correction tool, and 4/15 used no tools. In the *noMT* group, on the other hand, 2/6 used web search, 1/6 used grammatical error correction tool, and 3/6 used no tools. This indicates that one of the main use of MT was to back-translate Langsmith's output, while noMT group members were more likely to assess the output for themselves.

7.2.2 Assessment strategies. From the interview analysis, we identified four strategies by which the NNESs assessed Langsmith's writing suggestions. These were by consulting 1) Langsmith's typicality score, 2) back-translation, 3) web search results, and 4) their own English knowledge. According to the interviewees, search engines were used to check the meaning or usage of unclear words. Table 4 lists the percentages of interviewees from *MT* and *noMT* groups who claimed to adopt each of these assessment strategies. In the *MT* group, 9/9 claimed their judgments were shaped by the typicality score, 6/9 by back-translation, 2/9 by web search, and 3/9 by their own English knowledge. In the *noMT* group, on the other hand, 3/5 claimed their judgments were informed by typicality score, 2/5 by web search, and 2/5 by their own English knowledge. Overall, members of both groups seemed to rely on the typicality scores provided by Langsmith. However, members of the *MT* group seemed to rely more on system-generated suggestions (by the typicality scores provided by Langsmith or back-translations) than members of the *noMT* group. Specific insights gained from the interviews with members of the *noMT* and *MT* groups are reported below.

noMT group. Members of the *noMT* group generally appeared to decide whether or not to accept Langsmith's suggestions based on their own preferences. Although some referred to the system-provided typicality score, they did not seem to regard it as particularly important. As P13-noMT put it, *"I noticed there was a chart on the right side [i.e., typicality score], but I didn't pay much attention to it."* They referred to the typicality scores only when they were not sure about their own decisions. As P3-noMT commented:

> *"I could usually narrow it down to about one or two sentences, because even if I got many sentences, some of them were a bit different from what I really wanted to say. In the end, when it comes down to about two choices and I cannot decide which one to choose, I would refer to the credibility section on the right-hand side."*

Members of the *noMT* group reported both advantages and disadvantages of using the typicality score to assess Langsmith's suggestions. Some members of the *noMT* group reported that the typicality score allowed them to avoid having to rely on MT, and that they could keep their thoughts in English while writing, which was an advantage. As P3-noMT stated,

> *"I think it's more efficient to keep it in English when revising something written in English."*

However, P20-noMT reported a negative effect of using the typicality score, that it discouraged deeply engaging with the text itself.

> *"In a way, it seemed as if the level of my revision was becoming increasingly shallow. At first, I was comparing not only the typicality of the text with my original text, but by the end of the project, I felt like I was unconsciously or without thinking looking for text with a high typicality score."*

MT group. Members of the *MT* group tended to regard the typicality scores as more important than their *no-MT* group counterparts. Some *MT* members simply adopted the system proposal with the highest typicality score without checking it. For example, P1-MT stated that he was unsure of his English proficiency and thought the machine judgment was better than his own. Some others expressed a belief that Langsmith's suggestions with smaller typicality scores were completely unworthy of consideration. P18-MT explained,

> *"I only looked at the top three or four. I didn't look at the bottom of the suggestions because their typicality bars were so small that I didn't think I needed to look at them."*

However, this is not to suggest that *MT* group members completely trusted Langsmith's typicality scores. Rather, they relied on them due to a perceived lack of any other means of assessment.

> *"I can't judge whether it sounds fluent or not because I'm not a native speaker. I had my doubts about whether the sentence was really fluent, but I selected it."* (P16-MT)

Another characteristic strategy adopted by *MT* group members was back-translation using MT. All participants who stated that they ever used back-translation also back-translated their task text after using Langsmith. Mostly, they said they did this to ensure that Langsmith's output did not contain evident errors or evidently missing information, as it was more efficient for them to check it in Japanese. P10-MT noted,

> *"I tried it in DeepL first, and if something looked strange, I checked the English text myself."*

Some participants also guessed the quality of English sentences provided by Langsmith from the quality of back-translated Japanese. They believed that if their MT could translate system-produced English into error-free Japanese, the English would also be error-free.

Table 3: Tools used to check suggestions, identified from the video recordings. The values are headcounts.

	MT (Back-translation)	Web Search	Grammar Checker	Nothing
MT (15)	10	3	2	4
noMT (6)	-	2	1	3

Table 4: Suggestion assessment strategies identified via thematic analysis. The values are headcounts.

	Typicality	Back-translation	Web Search	Own proficiency
MT (9)	9	6	2	3
noMT (5)	3	-	2	2

7.2.3 Summary of RQ2 findings. To summarize, the participants' use of various tools for determining Langsmith's output was consistent with previous research findings that indicated NNESs rely on other resources to check MT results Liebling et al. [33]. Furthermore, *MT* group members tended to assign importance to Langsmith's typicality scores and were likely to use back-translation to make judgments.

7.3 NNESs' overall assessment of the rewriting tool (RQ3)

To address RQ3, we conducted a thematic analysis of the interview quotes and identified the factors shaping their overall tool assessment. "Perceived quality of suggestions" refers to statements where users assessed Langsmith's suggestions to be good and helpful. In particular, participants referred to times when they believed the best suggestion was the one with the highest typicality score. Similarly, "variety of suggestions" refers to cases where participants expressed that being presented with multiple, diverse suggestions was helpful to their writing.

By contrast, negative influences on the overall tool assessment included instances when participants identified "obvious errors in suggestions," such as incorrect changes to proper nouns or the removal of important content. In addition, "negative assessments of typicality score" pertained to cases where participants indicated that the typicality score did not aid in assessing the quality of the suggestions. These and other findings are discussed in more detail below.

noMT group. All interviewees ($n = 5$) in the *noMT* group mentioned "perceived quality of suggestions" and "variety of suggestions" as positive factors. For example, P20-noMT said, *"I thought it was a nice tool because it changed words and sentences for the better."* Some expressed appreciation for the fact that they could select and mix system-generated expressions according to their own preferences. As P5-noMT explained,

> *"I combined the first one with the third one and so on. I also kept what I wanted to keep from my writing, but used some of Langsmith's suggestions that I thought were better. I liked the fact that I could combine multiple suggestions."* (P5-noMT)

On the other hand, as negative factors, 1/5 of interviewees in the *noMT* group mentioned "obvious errors in suggestions". In addition, 2/5 of interviewees mentioned "negative assessments of typicality score." In particular, both of these participants referred to times when their own text appeared at the top of the suggestions, i.e., when they do not get the suggestions with higher typicality than their own sentence.

> *"When my writing reaches the highest typicality score, I wonder if there is anything more to do."* (P20-noMT)

This was described as unsatisfying because they were not content with their own sentence, and expected Langsmith to provide a suggestion with a higher typicality score.

However, interestingly, a few participants in the *noMT* group expressed positive assessment in Langsmith, even when it contained obvious errors or made low-quality suggestions. For example, P4-noMT expressed so much confidence in Langsmith that when they discovered errors, they attributed them to their own writing rather than blaming the software:

> *"My experience was that when I typed in longer sentences, I had the impression that Langsmith returned suggestions that didn't make much sense, so I figured that my sentence was too long to make sense in the first place."* (P4-noMT)

These quotes indicate that participants who were able to produce English sentences on their own were often able to evaluate and take advantage of the various suggestions generated by Langsmith.

MT group. 4/9 of the interviewees in the *MT* group mentioned "perceived quality of suggestions" as a positive factor, such as P11-MT, who said, "I thought it was a great tool because it was so accurate." In addition, P18-MT described feeling reassured when their own writing had a high typicality score:

> *"It was a relief when I received a high score from Langsmith for a text that I thought was good and that I had written or machine-translated myself."*

8/9 in this group mentioned "variety of suggestions" as a positive factor. However, some commented that it was difficult to select appropriate suggestions and/or to modify them as needed:

> *"If there was a big change, I wondered whether the sentence is weird, which gives me a chance to think about where to fix it. But since I am unable to fix it,*

> *I'm unsure of whether I'm making the right choice."* (P12-MT)

Thus, as described in Section 7.2, *MT* group participants turned to typicality scores to help them select among Langsmith's various suggestions. Although many participants in both groups (5/5 in *noMT* and 8/9 in *MT*) indicated that they appreciated having multiple suggestions, the role of those suggestions seemed to vary between groups. While the nuanced interpretation and combination of multiple suggestions were demonstrated by the *noMT* group, some *MT* group members said they simply appreciated multiple suggestions as a backup in case they did not prefer the top suggestion. As P16-MT put it,

> *"If there were only one suggestion and I was told that it was the best one, I would not be able to do anything about it, even if it looked weird. Langsmith gave me about seven suggestions, so I could look at them from the top down, and if a suggestion seemed wrong, I could choose the next one. I think it was a good feature."*

8/9 of the interviewees in the *MT* group also voiced problems with clear discrepancies between Langsmith's suggestions and the text they had prepared ("obvious errors in suggestions"), which was rarely noted by their *noMT* group counterparts. For example, P8-MT said, "[After using Langsmith,] there are parts that disappear, and it is difficult to re-check and correct them, so I had to be careful when using the tool." In particular, some of those who used back-translation commented that they could not identify errors when examining Langsmith's suggestions themselves. However, they described noticing errors via back-translation, i.e., when they appeared in Japanese, and said this caused them to have a negative impression of Langsmith's suggestions.

> *"At first, I fixed each sentence with Langsmith. I thought, 'Oh, that's good work,' and then put it into DeepL and translated it into Japanese. But then I noticed something was missing [because of Langsmith]."* (P1-MT)

Furthermore, 4/9 of *MT* group interviewees mentioned "negative assessments of typicality score." As discussed in Section 7.2, all interviewees in the *MT* group stated that they assessed the suggestions by the typicality score. However, as also discussed in Section 7.2, this does not suggest that they completely believed the suggestion or its score; rather, they used the suggestion with caution. This caution seemed to be associated with the user's feeling that they generally could not assess the suggestions on their own. P10-MT said,

> *"I don't know how far I could go to modify it, and in the end I'm not even sure which one was the best. I thought it (Langsmith) was nice because it made drastic revisions, but on the other hand, I felt that I was not sure about the high typicality score."*

For the typicality score to be useful in assessing a suggestion's quality, users themselves need to be able to assess the suggestion. Otherwise, the user may not be convinced or judge the suggestions properly, but may instead be confused or just follow the score.

7.3.1 Summary of RQ3 findings. We identified four factors that affected the overall assessment of the rewriting tool. Comparing the *MT* and *noMT* groups in terms of these factors, we found several differences. "Perceived quality of suggestions" was mentioned by all interviewees in the *noMT* group (5/5) but less often by interviewees in the *MT* group (4/9). Both groups mentioned the "diversity of suggestions," but there were differences in the reasons for appreciating diversity. In addition, many of the interviewees in the *MT* group mentioned Langsmith's obvious errors, and this was often found following back-translation. These suggest that those in the *MT* group have higher expectations for the accuracy of the tool than those in the *noMT* group.

8 DESIGN IMPLICATIONS

8.1 NNESs' writing strategies and evaluation within the dynamics of multiple tools

Existing research makes conflicting arguments and suggestions about NNES's use of MT in academic writing. While some studies recommend that NNESs not draft in their native languages nor use MT when writing papers [49, 50], recent studies [2, 30, 32, 47] claim that the use of MT can improve NNESs' writing and also provides educational benefits [31].

Our findings show that many of our participants created their initial drafts in their native language. They further seem to imply that the use of MT is widespread and highly trusted by Japanese NNESs, even in academic writing. We have drawn a few implications from this finding, which we outline below.

Most prior studies on user interactions with AI-powered (re)writing tools [3, 29] based their analyses on tool logs. However, this may not be appropriate for understanding the writing behavior of NNESs, given the variety of their tool-use behaviors that we observed. For example, our findings show that the text input of rewriting tools may not necessarily be drafted by the user, but rather the output of other tools (e.g., MT). Therefore, we encourage future developers and evaluators of writing-support systems for NNESs to adopt research designs that comprehensively observe the writing process.

8.2 Provide comprehensive support

As described above, when working with Langsmith to produce academic English writing, participating Japanese NNESs relied on several other tools and frequently switched among various applications and websites. In particular, eight participants in the MT group switched back and forth between MT and Langsmith. However, continuous switching between applications can increase one's workload and make it more difficult to focus [42]. Thus, evaluating AI-generated suggestions using additional software may be more demanding than optimal.

Consequently, editors that integrate frequently combined writing support tools may help NNESs write more effectively and with greater focus. One way to achieve this goal is to develop an integrated, all-in-one writing environment like Word or Google Docs. In fact, this direction is accelerating with the addition of writing support tools for Google Docs that use language models.[13] However, the development cost of such an integrated editor comes is high.

Another direction is to develop plug-ins for existing editors and browsers. This possibility can help users build a comprehensive

[13] https://blog.google/technology/ai/ai-developers-google-cloud-workspace/

writing environment by selecting the extensions they want to use. The more modest ambition of this direction also has the advantage of providing a call to action for smaller entities (e.g., researchers) that can create plug-ins but do not have the resources to build an integrated environment. To mitigate the risk of conflicts that are common in plug-in ecosystems [34], it will be valuable to establish technical standards for writing support tools [e.g., 19].

Furthermore, for comprehensive support, it is not enough to simply provide multiple functions simultaneously. Rather, the way in which multiple functions are integrated is an important design challenge to be addressed in the future. Appropriately combining multiple functions is expected to reduce user workload and facilitate appropriate decision-making on the part of the user. In the following sections, we discuss detailed design challenges suggested by our findings.

8.3 Back-translation by MT tools should warn about erroneous input

NNESs who faced challenges in evaluating the tool's suggestions appeared to utilize back-translation to assess the efficacy of the suggestions. Back-translation has actually been hailed as an effective strategy for MT-output assessment [33, 38]. For example, recent work on MT for outbound translations — i.e., ones in which the translation is into a language unknown to the MT user — reported that back-translation increased user confidence [57]. However, such use of MT may not always be effective. In fact, a majority of MT research has focused on improving translation robustness, even when NNESs make some errors in their original text [1]. This means that MT (including back-translation) can often correctly translate texts containing errors, e.g., by translating a phrase like "I has pen" into a grammatically correct translation. Such automatic correction of the original text suggests that use of MT is inappropriate to evaluate the correctness of the original text [50]. Ultimately, evaluating AI suggestions is not a main application of MT and is not well-studied. Using back-translation for this purpose may require new approaches, including improvements in aspects other than robustness. For example, if there are errors in the source English text, MT may need to warn the users instead of smoothing over the errors to translate it into Japanese. More simply, by making sure to apply a grammatical error correction before applying a back-translation, and having the user check the results before applying the back-translation, errors are less likely to be missed. Further, to help NNESs correctly assess the appropriateness of English text, metrics for text quality evaluation such as Langsmith's typicality score are important.

8.4 Prevent users from overly trusting machine scores

In addition to the usage of back-translation, numerous participants in our MT group selected system-generated English sentences based on their typicality scores, and some ignored sentences that were assigned low scores. One member in the *noMT* group also reported subconscious pressure to act according to typicality scores, which made his own elaboration more shallow. Although these numerical indicators provide a useful basis for NNESs' judgments, they also appear to have become an obstacle to NNESs checking the AI-powered tool's suggestions for themselves. Specifically, although "typicality" is a useful metric for writing text that matches the writing style in the training corpus, it may discourage users from selecting less typical suggestions, regardless of their quality for conveying the author's intended message. This finding encourages further research on the explanations and feedback for NNESs. For example, providing feedback in natural language on the reasons for corrections [39] or actual examples of phrases and expressions [26] might be worthwhile. How to integrate such techniques into rewriting tools, and how displays can help NNESs assess suggestions is a future design challenge.

8.5 Produce a variety of suggestions

When assessing Langsmith, many participants mentioned diverse output as a positive aspect of the tool. In particular, many participants in the *noMT* group combined multiple suggestions to produce a final sentence. Furthermore, *MT* group members indicated that they liked having multiple suggestions to choose from. However, participants in the *MT* group also faced difficulty assessing suggestions, indicating that diversity and quality must be balanced to support NNES users.

Langsmith uses an algorithm called diverse beam search [48] to achieve diverse outputs by adding noise to the probability distribution of the model's outputs. Due to the nature of the algorithm, high noise strength produces a variety of outputs, but also increases the probability that some suggestions are less appropriate than others. Several studies have been conducted on other algorithms to achieve diverse outputs, but many of them have also yielded slower or degraded performance [21, 36]. Evaluating diversity is not technically easy, and few NLP tasks have added diversity to their evaluation criteria [46]. Our findings highlight the need for further research on the improvement of algorithms and evaluation of models when such diversity is a goal.

It is also possible to generate diversity by using multiple rewriting models or tools instead of just one. However, simply increasing diversity and showing more suggestions will likely burden the user's assessment of the outputs, or, even worse, cause some suggestions to be ignored. Displaying suggestions in various formats is another design challenge. As an example, instead of always showing multiple outputs, it might be more appropriate to display additional examples only when the model confidence level is low or the user indicates dissatisfaction.

8.6 Implications for other generative AI tools

Recently, generative AI models are increasing in sophistication and ubiquity, including large language models (LLMs), such as ChatGPT. Several of the implications discussed above regarding rewriting tools are also relevant to generative AI models in a broad perspective.

Although generative AI tools have rapidly improved in performance, they still sometimes generate erroneous results that are "nonsensical or unfaithful to the provided source content" (often called "hallucinations") [24]. Due to the increased fluency of generated texts, it has become more difficult to detect hallucinations or other non-optimal suggestions. In many cases, generative AI systems are used precisely because users lack sufficient subject

expertise to generate content themselves, meaning that these users are also likely to struggle to effectively evaluate the generated content. While we demonstrated such a phenomenon in the context of NNESs evaluating English writing suggestions, the same also applies in common use cases such as asking questions and asking for information about unfamiliar topics. Thus, it is important to research new ways to help people evaluate content produced by generative AIs.

One of the most common tools for helping people evaluate AI-generated responses is quantitative scores, indicating the confidence of the model or the quality using automatic evaluation metrics in each suggestion [51, 55, 56]. However, even though users in the present study regarded the typicality score as useful, some felt it pushed them toward shallow interpretations, rather than considering the text deeply. Further, this sort of automatic evaluation metric often relies on a ground-truth against which generated content can be compared, which may not be readily accessible or interpretable in many applications [35, 45]. In that situation, many in our study relied on other AI tools, such as MT, as we have discussed. NNESs' integration of MT with Langsmith demonstrates how complementary AI systems may be integrated to improve the comprehensibility of each. Specifically, MT was not used to provide a quantitative score or a definitive rating but rather to provide additional context to help users understand AI-generated text. Researchers have demonstrated that innovative interfaces can help users understand the mechanisms and outputs of AI systems [18, 40, 51], demonstrating an important role for HCI and UIST researchers. Thus, the development of interfaces that integrate multiple AI systems to provide comprehensive support is important not only for rewriting tools, but also for other generative AI systems in which comprehensibility is a pressing challenge.

9 LIMITATIONS AND FUTURE DIRECTIONS

Our study has several limitations that should be the focus of future research.

First, the participants in our main study had never previously used Langsmith, and their usual writing practices may be different from the ones observed in our study. Furthermore, the participants were not allowed to use paper dictionaries and any other tools that could not appear on-screen (including secondary digital devices). This may have created further deviation from their common writing practice. Therefore, we hope to conduct future research over a longer period and in more realistic settings.

Second, we analyzed NNESs' behavior based on their writing process, specifically by dividing them into two groups according to whether they used MT. This allowed us to examine the impact of MT use on the rewriting process. There was some evidence that, compared to NNES with higher English proficiency, those with lower proficiency begin their writing less frequently in a second language [5] and employ MT more often [47]. Meanwhile, other HCI scholars have found that NNES may leverage MT for various English tasks regardless of their proficiency level [17, 52]. Taken together, this literature encourages future studies to examine the latent variables that shape an NNES's choice or strategy of MT use. Furthermore, since we targeted only Japanese researchers and students, there are limitations with respect to generalization; for example, there may be MT errors specific to the Japanese language. A more comprehensive study of NNESs would necessarily involve participants with a wider array of linguistic, cultural, and occupational backgrounds.

Finally, this is a case study based on Langsmith alone. While we believe that our findings will guide the future development of AI-powered writing-support systems for NNESs, research focused on other NLP models and interfaces could contradict our findings. Therefore, we emphasize the need for ongoing research as this technology develops.

10 CONCLUSION

We examined how one AI-powered writing-support tool, Langsmith, was used and perceived by NNES Japanese researchers in the context of English-language paper writing. We first investigated what other tools are used for this purpose and found that many participants supplemented their use of rewriting tools with MT. Further, we conducted a user study and interviews with these researchers to understand how they assessed Langsmith's rewriting suggestions and what factors they use to form an overall assessment of the tool.

Our results suggest that the NNESs who used MT tended to rely on sources of information other than their personal English proficiency to evaluate the English output of an AI-powered rewriting support tool. Probably as a consequence, we observed that they tended to discover errors in the rewriting tool's suggestions at a relatively late stage in their writing tasks, and that their overall assessment of the tool plummeted upon noticing evident errors.

In summary, the results of this study imply that interaction between NNESs with relatively low English proficiency and the focal AI-powered writing tool may be restricted by a language barrier. While AI-powered writing tools may help them become aware of new words and expressions, reliable verification processes remain time-consuming and expensive. Moreover, the study revealed how NNESs use multiple tools to compensate for their limited English skills when assessing AI-powered rewriting suggestions. This finding highlights the fact that users' perceptions and behaviors when using AI-based tools are influenced not only by the tools themselves, but also by the simultaneous use of other tools. Our study suggests that to benefit NNESs, a multilingual clue set should be provided by the interface of one tool or through an integrated workflow involving multiple tools. We hope our results will motivate both the HCI and NLP research communities to take up the challenge of developing writing-support tools specifically for NNESs.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful comments and the participants in this study. We are grateful to Professor Takeo Igarashi for valuable feedback. We also appreciate Yoriko Isobe for her assistance with this study. This work was supported by JSPS KAKENHI Grant Numbers JP22H00524 and JP21J14152.

REFERENCES

[1] Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural Machine Translation of Text from Non-Native Speakers. In *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 3070–3080. https://doi.org/10.18653/v1/N19-1311

[2] Nora Aranberri. 2020. With or without you? Effects of using machine translation to write flash fiction in the foreign language. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Lisboa, Portugal, 165–174. https://aclanthology.org/2020.eamt-1.18

[3] 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

[4] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[5] Andrew D. Cohen and Amanda Brooks-Carson. 2001. Research on Direct versus Translated Writing: Students' Strategies and Their Results. *The Modern Language Journal* 85, 2 (2001), 169–188. https://doi.org/10.1111/0026-7902.00103

[6] Juliet Corbin and Anselm Strauss. 2014. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications. https://books.google.co.jp/books?id=hZ6kBQAAQBAJ

[7] Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Nancy, France, 242–249. https://aclanthology.org/W11-2838

[8] Robert Dale and Jette Viethen. 2021. The automated writing assistance landscape in 2021. *Natural Language Engineering* 27, 4 (2021), 511–518. https://doi.org/10.1017/S1351324921000164

[9] 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

[10] Vidas Daudaravičius. 2015. Automated Evaluation of Scientific Writing: AESW Shared Task Proposal. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado, 56–63. https://doi.org/10.3115/v1/W15-0607

[11] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[12] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 96–108. https://doi.org/10.18653/v1/2022.in2writing-1.14

[13] 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)*. Association for Computational Linguistics, Dublin, Ireland, 3573–3590. https://doi.org/10.18653/v1/2022.acl-long.250

[14] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[15] John Flowerdew. 2007. The non-Anglophone scholar on the periphery of scholarly publication. *AILA review* 20, 1 (2007), 14–27.

[16] The Allen Institute for Artificial Intelligence. 2020. AllenNLP - Demo, Language Modeling. https://demo.allennlp.org/next-token-lm

[17] Ge Gao, Jian Zheng, Eun Kyoung Choe, and Naomi Yamashita. 2022. Taking a Language Detour: How International Migrants Speaking a Minority Language Seek COVID-Related Information in Their Host Countries. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 542 (nov 2022), 32 pages. https://doi.org/10.1145/3555600

[18] Maliheh Ghajargar, Jeffrey Bardzell, Alison Smith Renner, Peter Gall Krogh, Kristina Höök, David Cuartielles, Laurens Boer, and Mikael Wiberg. 2021. From "Explainable AI" to "Graspable AI". In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction* (Salzburg, Austria) *(TEI '21)*. Association for Computing Machinery, New York, NY, USA, Article 69, 4 pages. https://doi.org/10.1145/3430524.3442704

[19] Masato Hagiwara, Takumi Ito, Tatsuki Kuribayashi, Jun Suzuki, and Kentaro Inui. 2019. TEASPN: Framework and Protocol for Integrated Writing Assistance Environments. 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): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 229–234. https://doi.org/10.18653/v1/D19-3039

[20] 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

[21] Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of Diverse Decoding Methods from Conditional Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3752–3762. https://doi.org/10.18653/v1/P19-1365

[22] 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*. Association for Computational Linguistics, Online, 216–226. https://doi.org/10.18653/v1/2020.emnlp-demos.28 The Langsmith editor is online available at: https://editor.langsmith.co.jp/.

[23] Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the Rough: Generating Fluent Sentences from Early-Stage Drafts for Academic Writing Assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 40–53. https://doi.org/10.18653/v1/W19-8606

[24] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. https://doi.org/10.1145/3571730

[25] 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*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9420–9435. https://aclanthology.org/2022.emnlp-main.641

[26] Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for Language Learners Using Example-Based Grammatical Error Correction. https://doi.org/10.48550/ARXIV.2203.07085

[27] Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-Level Fluency Evaluation: References Help, But Can Be Spared!. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 313–323. https://doi.org/10.18653/v1/K18-1031

[28] Rassim Khelifa, Tatsuya Amano, and Martin A. Nuñez. 2022. A solution for breaking the language barrier. *Trends in Ecology & Evolution* 37, 2 (2022), 109–112. https://doi.org/10.1016/j.tree.2021.11.003

[29] 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

[30] Sangmin-Michelle Lee. 2020. The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning* 33, 3 (2020), 157–175. https://doi.org/10.1080/09588221.2018.1553186 arXiv:https://doi.org/10.1080/09588221.2018.1553186

[31] Sangmin-Michelle Lee. 2021. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning* 0, 0 (2021), 1–23. https://doi.org/10.1080/09588221.2021.1901745 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/09588221.2021.1901745

[32] Sangmin-Michelle Lee and Neil Briggs. 2021. Effects of using machine translation to mediate the revision process of Korean university students' academic writing. *ReCALL* 33, 1 (2021), 18–33. https://doi.org/10.1017/S0958344020000191

[33] Daniel J. Liebling, Katherine Heller, Margaret Mitchell, Mark Díaz, Michal Lahav, Niloufar Salehi, Samantha Robertson, Samy Bengio, Timnit Gebru, and Wesley Deng (Eds.). 2021. *Three Directions for the Design of Human-Centered Machine Translation*.

[34] Igor Lima, Jeanderson Cândido, and Marcelo d'Amorim. 2020. Practical detection of CMS plugin conflicts in large plugin sets. *Information and Software Technology* 118 (2020), 106212. https://doi.org/10.1016/j.infsof.2019.106212

[35] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6723–6737. https://doi.org/10.18653/v1/2022.acl-long.464

[36] Ruotian Luo and Gregory Shakhnarovich. 2020. Analysis of diversity-accuracy tradeoff in image captioning. https://doi.org/10.48550/ARXIV.2002.11848

[37] 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. https://doi.org/10.48550/ARXIV.2205.11484

[38] Mai Miyabe and Takashi Yoshino. 2009. Accuracy Evaluation of Sentences Translated to Intermediate Language in Back Translation. In *Proceedings of the 3rd International Universal Communication Symposium* (Tokyo, Japan) *(IUCS '09)*. Association for Computing Machinery, New York, NY, USA, 30–35. https://doi.org/10.1145/1667780.1667787

[39] Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared Task on Feedback Comment Generation for Language Learners. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 320–324. https://aclanthology.org/2021.inlg-1.35

[40] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe It or Not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3242587.3242666

[41] 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

[42] Jan Pilzer, Raphael Rosenast, André N. Meyer, Elaine M. Huang, and Thomas Fritz. 2020. Supporting Software Developers' Focused Work on Window-Based Desktops. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376285

[43] Stephen Politzer-Ahles, Teresa Girolamo, and Samantha Ghali. 2020. Preliminary evidence of linguistic bias in academic reviewing. *Journal of English for Academic Purposes* 47 (2020), 100895. https://doi.org/10.1016/j.jeap.2020.100895

[44] Valeria Ramírez-Castañeda. 2020. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. *PloS one* 15, 9 (2020), e0238372.

[45] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.* 55, 2, Article 26 (jan 2022), 39 pages. https://doi.org/10.1145/3485766

[46] Guy Tevet and Jonathan Berant. 2021. Evaluating the Evaluation of Diversity in Natural Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 326–346. https://doi.org/10.18653/v1/2021.eacl-main.25

[47] Shu-Chiao Tsai. 2020. Chinese students' perceptions of using Google Translate as a translingual CALL tool in EFL writing. *Computer Assisted Language Learning* 0, 0 (2020), 1–23. https://doi.org/10.1080/09588221.2020.1799412

[48] Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse Beam Search for Improved Description of Complex Scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 7371–7379. https://ojs.aaai.org/index.php/AAAI/article/view/12340

[49] Adrian Wallwork. 2016. *English for writing research papers*. Springer.

[50] Adrian Wallwork and Anna Southern. 2020. *100 Tips to Avoid Mistakes in Academic Writing and Presenting*. Springer.

[51] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 402–412. https://doi.org/10.1145/3397481.3450656

[52] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in Establishing Common Ground in Multiparty Groups Using Machine Translation. 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, 679–688. https://doi.org/10.1145/1518701.1518807

[53] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Text Style Transfer using Language Models as Discriminators. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., 7298–7309. https://proceedings.neurips.cc/paper/2018/file/398475c83b47075e8897a083e97eb9f0-Paper.pdf

[54] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[55] Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. ChrEnTranslate: Cherokee-English Machine Translation Demo with Quality Estimation and Corrective Feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 272–279. https://doi.org/10.18653/v1/2021.acl-demo.33

[56] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[57] Vilém Zouhar, Michal Novák, Matúš Žilinec, Ondřej Bojar, Mateo Obregón, Robin L. Hill, Frédéric Blain, Marina Fomicheva, Lucia Specia, and Lisa Yankovskaya. 2021. Backtranslation Feedback Improves User Confidence in MT, Not Quality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 151–161. https://doi.org/10.18653/v1/2021.naacl-main.14