

Sustaining Human Agency, Attending to Its Cost: An Investigation into Generative AI Design for Non-Native Speakers' Language Use

Yimin Xiao
College of Information
University of Maryland
College Park, Maryland, USA
yxiao@umd.edu

Cartor Hancock
UMIACS
University of Maryland
College Park, Maryland, USA
cartorh@terpmail.umd.edu

Sweta Agrawal
Instituto de Telecomunicações
Lisbon, Portugal
swetaagrawal20@gmail.com

Nikita Mehandru
School of Information
University of California, Berkeley
Berkeley, California, USA
nmehandru@berkeley.edu

Niloufar Salehi
School of Information
University of California, Berkeley
Berkeley, California, USA
nsalehi@berkeley.edu

Marine Carpuat
Department of Computer Science
University of Maryland
College Park, Maryland, USA
marine@umd.edu

Ge Gao
College of Information
University of Maryland
College Park, Maryland, USA
gegao@umd.edu

Abstract

AI systems and tools today can generate human-like expressions on behalf of people. It raises the crucial question about how to sustain human agency in AI-mediated communication. We investigated this question in the context of machine translation (MT) assisted conversations. Our participants included 45 dyads. Each dyad consisted of one new immigrant in the United States, who leveraged MT for English information seeking as a non-native speaker, and one local native speaker, who acted as the information provider. Non-native speakers could influence the English production of their message in one of three ways: labeling the quality of MT outputs, regular post-editing without additional hints, or augmented post-editing with LLM-generated hints. Our data revealed a greater exercise of non-native speakers' agency under the two post-editing conditions. This benefit, however, came at a significant cost to the dyadic-level communication performance. We derived insights for MT and other generative AI design from our findings.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing**; Collaborative and social computing theory, concepts and paradigms.

Keywords

Agency, Machine translation, AI-mediated communication, Non-native speakers

ACM Reference Format:

Yimin Xiao, Cartor Hancock, Sweta Agrawal, Nikita Mehandru, Niloufar Salehi, Marine Carpuat, and Ge Gao. 2025. Sustaining Human Agency, Attending to Its Cost: An Investigation into Generative AI Design for Non-Native Speakers' Language Use. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3706598.3713626>

1 INTRODUCTION

As it is increasingly feasible for AI-powered systems and tools to generate natural-sounding sentences and expressions, the question of how much humans can, and need to, exercise their own agency in communication deserves more attention than ever. Recent research in the HCI community has contributed valuable insights to this discussion. Specifically, existing work has cautioned against the mindless adoption of AI-generated materials (e.g., [45, 62, 68, 84, 105]). It has also demonstrated cases where people strike a good balance between “benefiting from AI usage” and “sustaining their agency during AI usage” by closely monitoring and evaluating AI outputs before using them (e.g., [38, 39, 67, 85]).

In the current research, we pay attention to machine translation (MT), the AI system that promises to mitigate language gaps in conversations. Prior work with non-native speakers of a society's dominant language (e.g., English) suggests that these individuals, often being new immigrants to a country, sometimes rely on MT to perform conversations with local native speakers [1, 40, 69]. However, the use of MT is often accompanied by a loss of human agency: non-native speakers can effectively formulate thoughts and put them into words in their native language, but the representation of their meanings in English is entirely controlled by MT. The human-MT interface today often provides non-native speakers with limited opportunities to influence MT outputs [24, 41, 87, 102, 106].



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
<https://doi.org/10.1145/3706598.3713626>

Inspired by the agency-resource connection discussed in Bandura's social cognitive theory [10, 13], we explore whether non-native speakers can leverage their own language resources in English - that is, the English vocabulary they have acquired to comprehend and produce English content - to influence MT outputs and, hence, better sustain their agency in MT-mediated conversations with native speakers. More specifically, we suspect that, given the unique composition of non-native speakers' English resources, future human-MT interfaces could be designed to facilitate the co-production of English messages by a non-native speaker and the MT in more than one way. We conducted an experimental study to compare the effects of three possible designs of this human-MT interface: a) the labeling interface, where non-native speakers can leverage a subset of their English resources (e.g., passive vocabulary) to mark their perceived quality of MT outputs for a conversation partner to view, b) the regular post-editing interface, where non-native speakers can leverage the full set of their English resources (e.g., passive and active vocabulary) to revise the MT outputs before relaying them to a communication partner, and c) the augmented post-editing interface, where non-native speakers can access large language model (LLM) generated hints as an additional resource to facilitate their revision of MT outputs. The comparison across these interface conditions allowed us to explore possible ways to sustain the non-native speakers' agency while remaining cautious about the associated costs.

The rest of this paper details the rationale, method, data analysis, and results of our research. In particular, our research indicated that, although all three human-MT interface designs could help sustain the non-native speakers' agency at different extents, the rise of their agency under the two post-editing interface conditions came at a greater cost to people's communication process and outcome. Compared to those assigned to the labeling interface condition, dyads under the two post-editing interface conditions exchanged more information in their conversations; however, these conversations were characterized by lower depth and resulted in less alignment established between interlocutors. These results shed light on future human-MT interaction design that can facilitate non-native speakers in leveraging their (constrained) English resources to sustain agency, as well as managing the trade-offs between agency and other essential aspects of communication.

2 RELATED WORK

In this section, we review previous literature that informs the hypotheses (Hs) and research questions (RQs) guiding our work. We begin with Bandura's theory of human agency, which constitutes an effective device for understanding AI-mediated communication (Section 2.1). Next, we focus on the use case of MT, considering the "agency loss" experienced by those who often rely on translation to navigate information-seeking conversations in their non-native language (Section 2.2). Building upon this literature, we investigate ways to sustain non-native speakers' agency in MT-assisted communication. Our point of departure is a comparison between three possible conditions of human-MT interfaces. They all have the potential to protect non-native speakers' agency in a conversational setting, but likely to varying extents and with different associated costs (Section 2.3).

2.1 Bandura's Theory and Its Implications for Agency in AI-Mediated Communication

Agency considers a person's capability to influence the achievement of desired outcomes [13]. Among the bulk of social science theories tackling this concept (e.g., [6, 32, 89, 93]), arguably the most influential one was proposed by Bandura [10, 13]. His social cognitive theory emphasizes that people are not merely onlookers of what happens to them but active contributors to the world they live in [11]. Rather than passively reacting to the environment, individuals demonstrate personal agency by evaluating opportunities and constraints and planning actions accordingly [9, 71].

At the heart of Bandura's conceptualization lies the connection between personal agency and the resources an individual can utilize - such as knowledge and materials within their own access - to take purposeful actions toward their goal [8, 10, 13]. A lack of such resources forces people to be "wholly determined" by situational influences [13]. Alternatively, the person may turn to other individuals or groups who possess the resources to act on their behalf, which places the self in "a vulnerable position resting on the competencies and favors of others [8, 12]".

The agency-resource connection provides an inspiring theoretical lens for studying AI-mediated communication. In particular, a rich body of recent HCI research has cautioned that the rise of generative AI today may challenge the preservation of users' agency (e.g., [38, 45, 62, 63, 68]). Much of this challenge, according to Bandura's theory, stems from AI's increasing ability to generate human-like resources, such as natural-sounding sentences and expressions, that are essential for task completion. For instance, Fu et al. conducted an experiment with 120 participants who completed email writing with the help of AI. Participants in this study reported a greater loss of personal agency when the AI drafted the entire email for them, rather than providing specific suggestions upon the user's requests and for their review [38]. Kim et al. interviewed 24 individuals who engaged in AI-assisted personal journaling over a 10-day period. While all interviewees expressed enthusiasm for this practice, they voiced concerns that the blunt adoption of AI-generated language would prevent people from documenting genuine thoughts and feelings in their journals [62]. Kobiella et al. conducted a longitudinal study with knowledge workers to understand their use of AI for content creation and information discovery. People reported a diminished sense of control when AI's contribution to the task overshadowed the person's own effort [63]. Gero et al. and Li et al. examined the agentic dynamics between human and AI in various cases of essay writing. Participants in both studies perceived their language use to be less authentic as they used AI assistance in less reflective manners [45, 68].

Together, the above literature underscores the necessity for humans to utilize their own resources toward task completion. In AI-mediated communication, good practices of technology design should allow or even encourage people to monitor, evaluate, and influence AI-generated resources, rather than have the AI act on their behalf.

2.2 Non-Native Speakers' Agency in the Use Case of Machine Translation

Our current research explores ways to sustain human agency in the use case of machine translation (MT), the AI system developed to bridge language gaps in communication. This choice situates our empirical discussion about human agency in a unique setting. Specifically, prior HCI research on MT-assisted conversations often involved individuals who aim to navigate information exchanges in their non-native language, such as immigrants with imperfect fluency in a society's dominant language but need to communicate their information needs with local native speakers (e.g., [1, 2, 14, 16, 29, 43, 69]). Numerous studies have demonstrated the potential of MT in assisting non-native speakers' message production in the target language (e.g., [1, 2, 14, 69, 70]); that said, the default setup of human-MT interface always compels non-native speakers to accept the system's translation as-is, even at times when people are not fully satisfied with the translation. Their personal agency in MT-assisted conversations is noticeably limited.

The agency-resource connection, as conceptualized by Bandura, points to a direction for better preserving non-native speakers' agency: non-native speakers should be allowed to influence MT outputs with the language resources they own; accordingly, the human-MT interface could be designed to enable this influence. While existing HCI literature has rarely discussed such interface design for the agency of non-native speakers, we notice relevant examples in the field of translation studies (e.g., [21, 48, 64]). In that context, professional translators employ MT as a productivity accelerator. They let MT take the first pass at processing the source text, then review the MT's output and post-edit it by leveraging their own language knowledge. As O'Brien's recent article notes, the post-editing process enables professional translators to speed up their workflow while maintaining "their intended production [79]." To further support professional translators in performing post-editing, several HCI projects have developed and tested tools tailored to this user group's needs. Recent examples include Green and colleagues' work on predictive translation memory [46, 47], Coppers et al.'s development of Intellingo [27], and Herbig and coauthors' system for multimodal translation editing [53, 54].

So, is post-editing also an effective way to sustain non-native speakers' agency in the use of MT? Our literature review suggests that the answer may not be straightforward. Unlike professional translators who are fluent bilinguals in both the source and the target languages, non-native speakers possess constrained resources in the target language. This constraint has been extensively evidenced in communication and psycholinguistics research on the "recognition and production gap (e.g., [44])." On the one hand, being a non-native speaker often means the person has passive vocabulary to sense expression problems in the target language, allowing for their self-initiated attempts to repair those problems [25, 28, 33, 56]. On the other hand, this person may or may not own sufficient active vocabulary to complete the repair [25, 33, 51, 66, 104]. Previous studies have found that, in order to improve the quality of writing in the target language, non-native-speaking writers sometimes consult model texts to enhance their awareness of a greater variety of expressions (e.g., [34, 44]). Other research has also documented instances where non-native speaking employees seek suggestions

from native speaking colleagues to refine their oral presentations in the working language (e.g., [7, 77]).

In summary, prior literature implies that non-native speakers have the language resources to influence MT outputs and, thereby, better maintain their agency in MT-assisted conversations. However, the composition of these resources differs from that of fluent bilinguals. Non-native speakers are generally more effective in utilizing a subset of their knowledge in the target language (e.g., passive vocabulary) for recognition, such as assessing a given expression as they encounter it. They tend to be less effective in utilizing their full range of knowledge in that language (e.g., passive and active vocabulary) for production, such as generating new expressions from scratch. These insights prompt us to explore alternative forms of human-MT interface design that enable non-native speakers to activate their language resources in a stepwise manner, allowing for varying levels of influence on MT outputs.

2.3 Interface Design for Non-Native Speakers' Influence on Machine Translation Outputs

We considered three forms of human-MT interfaces to facilitate information exchange between non-native and native speakers. These interfaces enable non-native speakers to influence MT outputs by utilizing their language resources across distinct scopes. One type is a *labeling* interface, where non-native speakers can use icon clicking to indicate their recognition of each MT output's quality without directly editing them. Another type is a *regular post-editing* interface, similar to those used by professional translators who can edit translation outputs without additional hints. The third type is an *augmented post-editing* interface, where non-native speakers can access additional hints, such as expressions recommended by an LLM, to scaffold their post-editing of MT outputs.

By introducing these three interfaces, we do not claim to have exhausted all possible designs for enabling non-native speakers to exert influence while using MT. Rather, these interfaces represent different systemic infrastructures that allow a person to act with varying levels of their resources along a spectrum. Non-native speakers are expected to utilize a smaller scope of their language resources in the labeling condition (i.e., passive vocabulary for recognition tasks only), the full scope of their resources in the regular post-editing condition (i.e., passive and active vocabulary for both recognition and production), and additional resources in the augmented condition (i.e., passive and active vocabulary with additional hints to better activate these resources). As the first line of inquiry in our investigation, we ask: How will non-native speakers' exercise of agency vary across these interfaces? Based on the agency-resource connection, we further hypothesize that non-native speakers would exercise a greater sense of personal agency in MT-assisted conversations when they could utilize a larger volume of resources to influence MT outputs:

H1a. *Non-native speakers will exercise a greater sense of agency in MT-assisted conversations when the task interface enables them to perform regular post-editing of MT outputs, compared to labeling their quality assessment of the outputs.*

H1b. *Non-native speakers will exercise a greater sense of agency in MT-assisted conversations when the task interface enables them to*

perform augmented post-editing of MT outputs with additional hints, compared to regular post-editing.

Moreover, recent discussions on AI-mediated communication have cautioned that enhancing human agency may come with associated costs [15, 98]. In the use case of MT, existing HCI research has been paying close attention to people's performance, in terms of their communication process and/or outcomes, and workload as two productivity-centered measures (e.g., [42, 102, 106, 108]). While we value non-native speakers' agency in MT-mediated conversations, we believe it is equally important to be mindful of any potential trade-offs between this and the productivity aspects of communication. We ask the following questions in our current research context:

RQ1a. *How will the dyadic level communication process, such as the breadth and depth of information exchanges, differ when non-native speakers can influence MT outputs through labeling, regular post-editing, or augmented post-editing?*

RQ1b. *How will the dyadic level communication outcome, such as the alignment between information seekers and providers, differ when non-native speakers can influence MT outputs through labeling, regular post-editing, or augmented post-editing?*

RQ2. *How will the task workload experienced by communicants differ when non-native speakers can influence MT outputs through labeling, regular post-editing, or augmented post-editing?*

3 METHOD

We designed a 3×2 between-subject experiment with 90 participants to investigate our hypotheses and RQs. This sample size was confirmed through a power analysis, and it aligned well with the practices of prior HCI studies involving dyadic or triadic participants performing communication tasks (e.g., [23, 30, 81, 82, 91, 107]). Half of our participants were Chinese new immigrants to the United States. They were all interested in housing selection tips and suggestions but lacked sufficient language proficiency to seek this information in English. The remaining participants were local residents in the United States, who had extensive experience in selecting local housing. They spoke English as their sole native language.

We paired participants into 45 dyads. Each dyad consisted of one non-native speaker and one native speaker of English. They performed text-based information-seeking conversations focused on a housing selection task that we developed. The text communication tool used in our experiment was embedded with an MT module, enabling non-native speakers to compose messages expressing their thoughts and questions in Mandarin before receiving an automated translation in English. The non-native speaker in each dyad was provided with one of the three human-MT interfaces: labeling interface, regular post-editing interface, or augmented post-editing interface. This experimental design allowed us to understand how the involvement of a non-native speaker's language resources would affect their agency in MT-assisted conversation, as the level of resource involvement varied according to the human-MT interfaces people used. Below, we provide the full details of our method.

3.1 Participants

We recruited all non-native speaking participants by distributing online and physical flyers within multiple Chinese new immigrant

communities in the United States. Among the 45 individuals who signed up and participated in this research, 29 were self-identified females and 16 were males. Their average age was 27 (S.E. = .73). According to their self-reports on 7-point scales, these participants had a medium-level English proficiency ($M = 4.56$, S.E. = .14; 1 = limited proficiency, 7 = full proficiency). They held limited existing knowledge about housing selection in the United States ($M = 2.18$, S.E. = .16; 1 = no knowledge at all, 7 = extensive knowledge). They had some experience of communicating with local native speakers in daily work and life settings ($M = 5.47$, S.E. = .20; 1 = no experience at all, 7 = extensive experience). They also had some experience with language technology, such as translation and English text editing tools ($M = 4.31$, S.E. = .27; 1 = no experience at all, 7 = extensive experience).

Native speakers in our sample were recruited from Prolific and other public social media platforms to ensure scheduling availability with non-native speakers. Among these 45 individuals, 31 were self-identified females and 14 were males. Their average age was 42.69 (S.E. = 1.79). They all spoke English as their native language and with full proficiency. According to their self-reports on a 7-point scale, these participants had sufficient knowledge about housing selection in the United States ($M = 5.69$, S.E. = .16; 1 = no knowledge at all, 7 = extensive knowledge). Their average experience of communicating with new immigrants to the country and with language technology was 3.71 (S.E. = .28) and 3.07 (S.E. = .25), respectively (1 = no experience at all, 7 = extensive experience).

All participants were compensated at an hourly rate of \$20.

3.2 Task Design

3.2.1 The housing selection scenario. We developed an information seeking task within the housing selection scenario. New immigrants often face challenges in navigating essential information exchanges with local people in this context. This issue has been well-documented in prior research on and with immigrant populations [16, 22, 94]. To protect participants' privacy and ensure fair comparison across dyads in data analysis, we provided all participants with pseudo profiles that we prepared, rather than engaging in exchanges based on their personal information in real life.

Our task design followed the established protocol of a "hidden profile task [92]." Specifically, the task requires one non-native speaking participant to act the role of an information seeker, describing their housing needs and asking questions about the comparison between different housing options. In response, one native speaking participant, acting as the information provider, offered suggestions based on the housing information available to them and their knowledge from prior life experiences. Each party in this task was provided with a package of information that was hidden from the other party. This setup mimics the real-world information gap between communicants: non-native speaking participants had full access to descriptions about one pseudo information seeker's family situation, whereas native speaking participants had full access to details about two housing options to discuss. It also allowed us to control for individual variations by situating all dyads within the same conversational context. Non-native speakers received their package of information in Mandarin, and native speakers received theirs in English.

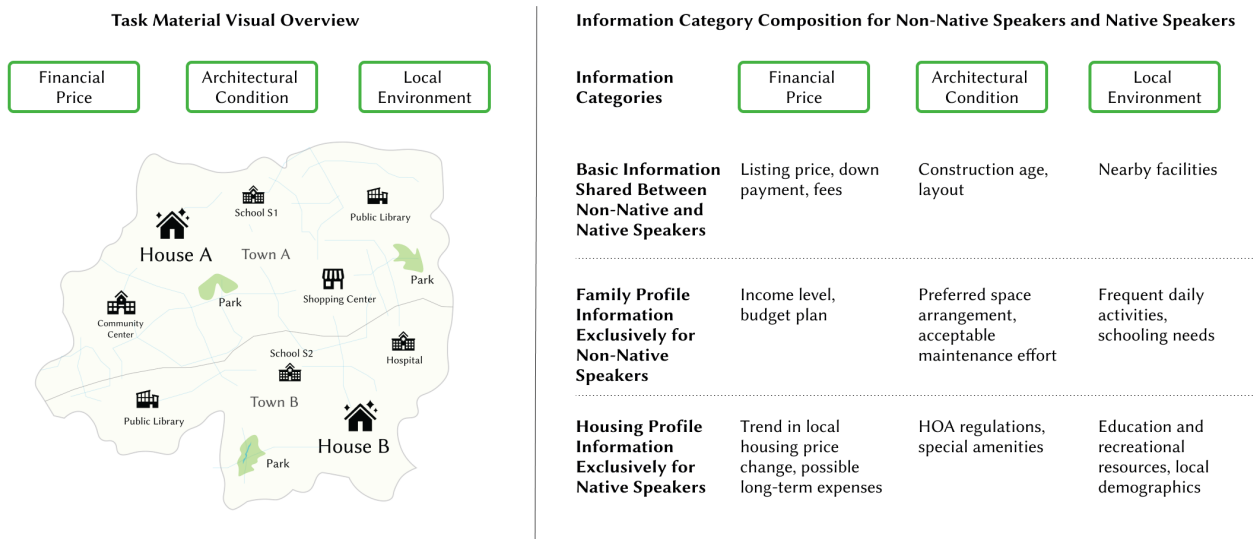


Figure 1: Composition of information in the task materials and its distribution between non-native speaking information seekers and native speaking information providers. As a participant clicks on each tab and housing options on the map view, the corresponding information will appear in a pop-up.

During the task conversation, participants shared information that their task partners would not have known otherwise. The ultimate goal of this information exchange was to bring both parties to a mutual understanding of the key considerations in making housing selections for the immigrant family’s needs. Figure 1 summarizes the types of information available to participants under the immigrant family’s profile and each housing option’s profile, depending on their task role (i.e., information seeker or provider).

3.2.2 Task procedure. We paired participants in dyads based on their shared availability. Prior to the information-seeking task, each participant attended an online tutorial, getting themselves familiar with our task scenario and the task materials. We invited all non-native speaking participants to test how their message production would work with the human-MT interface designed for their task condition. Similarly, all native speaking participants were invited to preview how a task partner would use the MT and the translated message they would get. These steps helped ensure that our manipulation of the human-MT interface functioned as intended.

During the information seeking task, dyads of participants completed their conversations about housing selection online. Each conversation session lasted 30 minutes and was performed entirely through text messages. The logs were recorded on our system server. Immediately after the conversation, the task interface directed participants to a short Qualtrics survey, which collected people’s numerical ratings of their conversation experiences across multiple aspects. Upon completing this survey, participants were provided with a Zoom link. They met the researcher for a brief session to receive instructions on compensation.

3.3 System Setup

We developed an in-house communication tool for convenient recording and storage of all conversation logs. This tool features a front-end design that is identical to commonly available instant messaging applications. For the MT module integrated into this tool, we used an open-source Chinese-English translation model trained on the OPUS dataset, released by Helsinki-NLP [100]. The model achieves a BLEU score of 36.1 with a brevity penalty of 0.948 on the Tatoeba-test.zho.eng, indicating reasonable translation quality.

Figure 2 illustrates the human-MT interface used by non-native speaking participants in each of the three task conditions, along with the corresponding view that native-speaking participants experienced when reading their task partners’ messages. All participants viewed their side of the conversation history in a language setting that matched their own language profile: native speakers would view all messages in English as the sole language that makes sense to them, while non-native speakers would always see each message displayed in two languages (e.g., the translated version and the source version). This design choice aligned with recommendations from previous work on MT-assisted conversations (e.g., [102, 108]), considering that non-native speakers often prefer to view a translated message alongside its source version. It allows people to make the most of their (imbalanced) knowledge across two languages.

3.4 Measurements

We employed three groups of measures in response to our hypotheses and RQs: participants’ subjective evaluation of the non-native speaking information seeker’s agency (H1a, H1b), various objective measures of the dyad’s information exchange performance (RQ1a, RQ1b), and the perceived workload of the MT-assisted information-seeking task (RQ2).

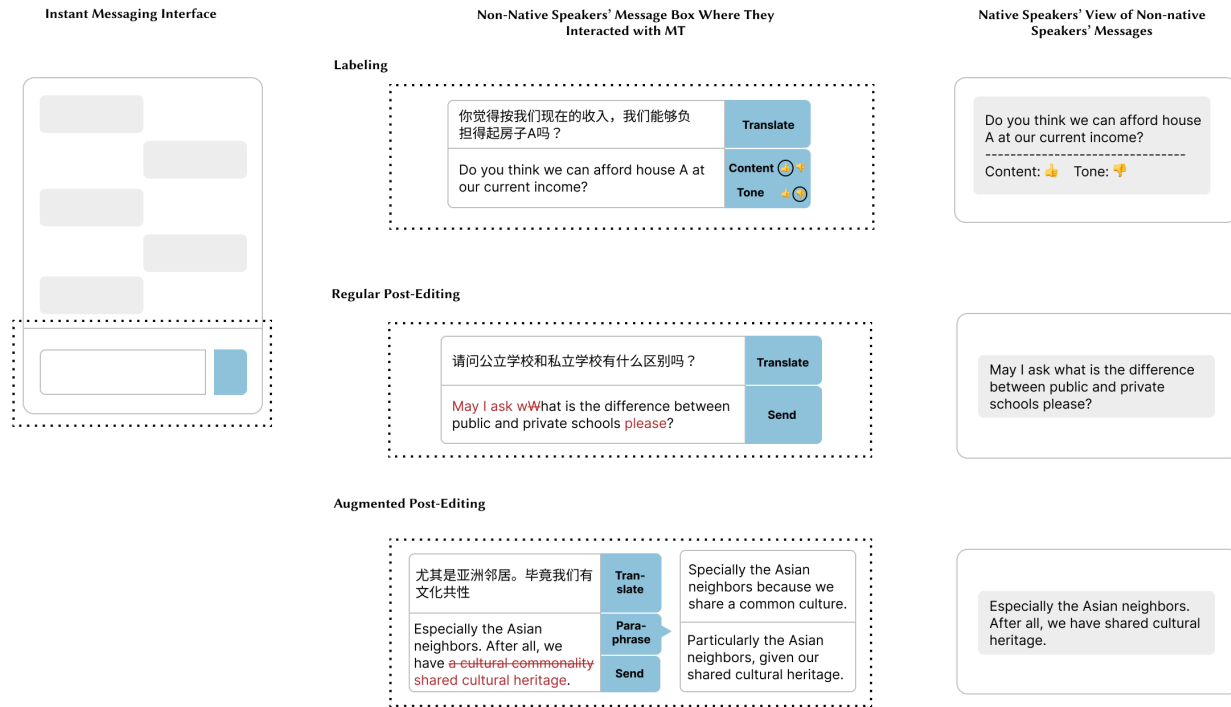


Figure 2: We manipulated how non-native speakers could influence their English production in the message box. In the labeling interface condition, non-native speakers could click on the thumbs-up and thumbs-down buttons to indicate their evaluation of the translation. Their native-speaking partners received both the English message and the evaluations. In the regular post-editing interface condition, participants could edit the original MT output in the English text box. In the augmented post-editing interface condition, participants could access two paraphrases of the initial MT output and edit it in the English text box. Their native speaking partners received the final version of English messages in the two post-editing conditions.

3.4.1 Non-native speakers' agency. We used 7-point scales to collect each non-native speaker's self-ratings and their native speaking partner's ratings on their personal agency demonstrated in the task context. Survey items under this measurement were adopted from the Sense of Agency Scale [99] (e.g., "[The item version rated by non-native speakers] I was in control of what my messages would convey to my partner," or "[The item version rated by native speakers] My partner was in control of what their messages would convey to me;" Cronbach's $\alpha = .80$). A higher average rating indicated a higher level of agency perceived by the participant.

3.4.2 Dyadic information exchange process. The process of information exchange is commonly examined by its breadth (i.e., the number of unique information pieces exchanged in the conversation) and depth (i.e., the number of elaborations made on each unique information piece being exchanged). To apply these two measures in our task context, we manually coded all conversation logs recorded on the system server. Specifically, we adopted Clark's classic framework, which categorized all messages in a conversation into two types [26]: content-related messages that attempt to carry out the official task (e.g., the exchange of the immigrant family's housing needs or a housing property's characteristics) and coordination-related messages that do not contribute new content

but are used to facilitate the conversation (e.g., "oh, okay" as a backchannel response). After that, we focused on each content-related message, categorizing it under one of the two groups: the initiation of a new information exchange, or the elaboration on an already-initiated exchange. Table 2 detailed all the message categories used in this analysis, along with examples from our participants' conversation logs. Two research assistants, blind to the rationale behind our study design, began by reviewing the entire corpus and independently coding a subset of all message turns. The inter-coder reliability achieved at the end of this step was already high, with Cohen's Kappa $\kappa = .82$. Following this, the two assistants discussed their discrepancies, clarified their coding rules, and proceeded to code the remaining message turns. They repeated the calibration steps until they arrived at unified coding results for the entire corpus. Based on the coded corpus, we calculated the total number of initiation messages in each task conversation (i.e., the breadth), with each initiation message representing one unique information piece being exchanged. We also calculated the average number of elaboration messages per unique information piece (i.e., the depth), as each elaboration message further deepened the discussion surrounding its initial status.

Table 1: Message categories and examples based on the information exchange functions they serve

Category and Description	Example Messages Within One Information Exchange
Content: Initiation of a new information exchange	"House A and B are four rooms, but one is a single floor, one is a third floor. Which one would be easier." [Non-native speaker]
Content: Elaboration on an already-initiated exchange	"That depends on how you like the stairs. Does anyone in your family have mobility issues?" [Native speaker]
	"The kids are 6 years old and 8 years old. I'm afraid they'll get hurt." [Non-native speaker]
	"So it's better on the single floor?" [Non-native speaker]
	"That's a valid concern. When my children were that age we had a single story home but we still had to maneuver stairs to get into and out of the home. I'm not sure about either one of these homes as far as that is concerned" [Native speaker]
Coordination	"Okay, I get it." [Non-native speaker]

3.4.3 Dyadic information exchange outcome. Prior research on information exchange often considers the alignment between information seekers and information providers as a key outcome at the end of the exchange. We adopted this criterion to evaluate the dyadic outcome in our task context. Specifically, we included one question in the post-task survey, asking all participants to rank a shared list of five items according to their importance to the immigrant family's housing selection. All five items on the list were derived from the task materials that participants had reviewed and centered their conversations around, including the anticipated financial expenses, the architectural condition of the house, the nearby services and amenities, the networking opportunities within the local community, and the educational resources available for children in the family. We computed the Rank-Biased Overlap (RBO) score to assess the alignment between the rankings provided by participants in the same dyad [103]. This similarity score ranges between 0 and 1, with a higher value indicating greater alignment and, hence, better dyadic information exchange outcome.

3.4.4 Workload. We used 7-point scales to collect each participant's self-ratings on the workload they experienced in the task context. Survey items under this measurement were adopted from the NASA TLX scale [50] (e.g., "How much mental burden (e.g., thinking, searching, deciding) did you experience during the entire conversation?"; Cronbach's $\alpha = .66$). A higher average rating indicated a higher level of workload perceived by the participant.

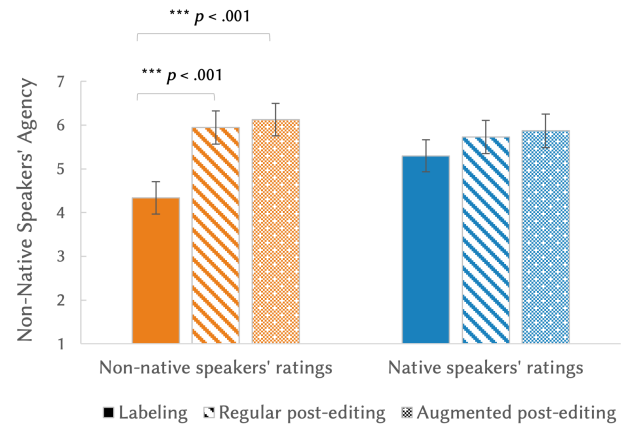
4 RESULTS

We conducted a series of ANOVAs to examine the effect of the human-MT interface design on several factors at the center of our interest: the preservation of non-native speakers' agency in MT-assisted conversations (H1a, H1b), dyadic performance as revealed in participants' information exchange process (RQ1a) and outcome

(RQ1b), and the perceived task workload (RQ2). We applied the Bonferroni correction to all pairwise comparisons. The remainder of this section details the model we used in each analysis and the relevant results.

4.1 Non-Native Speakers' Agency

Our H1a and H1b predicted that the three human-MT interfaces would all help non-native speakers sustain their personal agency in MT-assisted conversation, but the level would vary across conditions. We performed a 3×2 Mixed Model ANOVA to test the hypotheses. The results of this analysis supported H1a and rejected H1b (Figure 3).

**Figure 3: Non-native speakers' agency by the human-MT interface condition and the rater's language background**

Specifically, we set the dependent variable of this analysis to be each participant's rating of the non-native speaker's agency exercised in the task conversation. One independent variable was the human-MT interface condition: labeling interface, regular post-editing interface, or augmented post-editing interface. A second independent variable was the rater's language background: non-native speaker or native speaker. Participants were nested within the dyads. The Satterthwaite's approximation was applied to estimate the degrees of freedom, which often generated non-integer values. Control variables in this analysis included each person's demographic information (e.g., age, gender), existing knowledge about housing selection in the United States, prior experience with cross-lingual communication with people of the other residence status, and prior experience in using language technology.

The results indicated a significant main effect of the human-MT interface condition: $F [2, 41.43] = 8.17, p < .001$. The main effect of language background was not significant: $F [1, 72.68] = 0.09, p = .76$. The interaction effect between these two factors was marginally significant: $F [2, 40.19] = 3.19, p = .05$.

We used pairwise comparisons to examine ratings provided by the non-native speakers themselves and the native speakers working with them, respectively. In support of H1a, non-native speakers rated themselves as having a higher level of agency in the regular post-editing interface condition ($M = 5.95, S.E. = .38$) than in the labeling interface condition: $F [1, 77.67] = 14.99, p < .001$. They also reported a higher level of agency in the augmented post-editing interface condition ($M = 6.13, S.E. = .37$) than in the labeling interface condition ($M = 4.34, S.E. = .37$): $F [1, 77.82] = 18.43, p < .001$. That said, there was no significant difference between the two post-editing conditions, rejecting H1b. Meanwhile, the native speakers provided similar ratings to the non-native speakers' agency across the augmented post-editing interface condition ($M = 5.87, S.E. = .38$), regular post-editing interface condition ($M = 5.73, S.E. = .38$), and labeling interface condition ($M = 5.30, S.E. = .37$).

In summary, the above results suggested that the non-native speakers' self-perceived agency tended to vary according to the influence they could make on MT outputs: the agency was higher in the two post-editing interface conditions, compared to the labeling interface condition. This difference, however, was not observed in ratings provided by their native speaking partners.

4.2 Dyadic Information Exchange Process

We wondered how the adoption of the three human-MT interfaces would affect participants' dyadic information exchange process (RQ1a). In the context of this research, we examined the process by analyzing the breadth and depth of our participants' information exchange. Overall, our analyses showed that both the breadth and the depth of information exchange varied according to the human-MT interface condition, but these variations occurred in two opposite directions (Figure 4).

4.2.1 Breadth of information exchange. We used a one-way ANOVA to compare the total initiations of unique information pieces exchanged by dyads across three human-MT interface conditions. The dependent variable was the total number of unique information exchanges initiated per dyad. The independent variable was

the human-MT interface condition: labeling interface, regular post-editing interface, or augmented post-editing interface. Control variables were the same as those used for agency. The results indicated a significant main effect of the human-MT interface condition: $F [2, 42] = 6.32, p < .005$. Specifically, dyads of participants initiated a lower total number of unique information exchanges under the labeling interface condition ($M = 6.20, S.E. = .56$) than under the regular post-editing interface condition ($M = 7.93, S.E. = .56$; $F [1, 42] = 4.75, p < .05$) or the augmented post-editing interface condition ($M = 9.00, S.E. = .56$; $F [1, 42] = 12.39, p < .01$). There was no significant difference between the two post-editing interface conditions.

To further unpack this result, we also compared the total number of unique information pieces that the non-native speakers and the native speakers had initiated in their conversations, respectively. The analysis was performed with a 3×2 Mixed Model ANOVA (Figure 5), where the dependent variable was the total number of unique information exchanges initiated per person. The rest of the model setup was the same as that used for agency. The results indicated a significant main effect of the human-MT interface condition: $F [2, 40.46] = 6.92, p < .005$. The main effect of language background was significant as well: $F [1, 78.87] = 10.81, p < .005$. There was no significant interaction effect between these two factors: $F [2, 41.46] = 0.87, p = .43$.

Pairwise comparisons revealed that non-native speakers contributed more to the initiation of unique information exchanges under the augmented post-editing interface condition ($M = 6.12, S.E. = .56$) than in the regular post-editing interface condition ($M = 5.04, S.E. = .57$; $F [1, 74.07] = 3.42, p = .05$) or the labeling interface condition ($M = 4.35, S.E. = .56$; $F [1, 74.94] = 8.34, p = .005$). In contrast, native speakers' contributions to the initiation of unique information exchanges appeared similar across the augmented post-editing interface condition ($M = 2.89, S.E. = .57$), the regular post-editing interface condition ($M = 3.00, S.E. = .57$), and the labeling condition ($M = 2.84, S.E. = .56$).

4.2.2 Depth of information exchange. We used another one-way ANOVA to compare the average number of elaborations in dyadic conversations across three human-MT interface conditions. The dependent variable was the average number of elaborations generated per dyad. The rest of the model setup was the same as that used for dyadic breadth of information exchange. The results indicated a significant main effect of the human-MT interface condition: $F [2, 42] = 30.09, p < .05$. Specifically, dyads of participants made a larger average number of elaborations under the labeling interface condition ($M = 4.10, S.E. = .53$) than under the regular post-editing interface condition (Mean = 5.10, $S.E. = .53$; $F [1, 42] = 1.98, p = .05$) or the augmented post-editing interface condition ($M = 6.11, S.E. = .53$; $F [1, 42] = 7.23, p = .01$). There was no significant difference between the two post editing conditions.

To gain further insights, we compared the average number of elaborations that non-native speakers and native speakers had added to their conversations, respectively. The analysis was performed with a 3×2 Mixed Model ANOVA (Figure 6), where the dependent variable was the average number of elaborations generated per person. The rest of the model setup was the same as that used for agency. The results indicated no significant main effect of the human-MT interface condition ($F [2, 41.77] = 1.34, p = .27$), nor

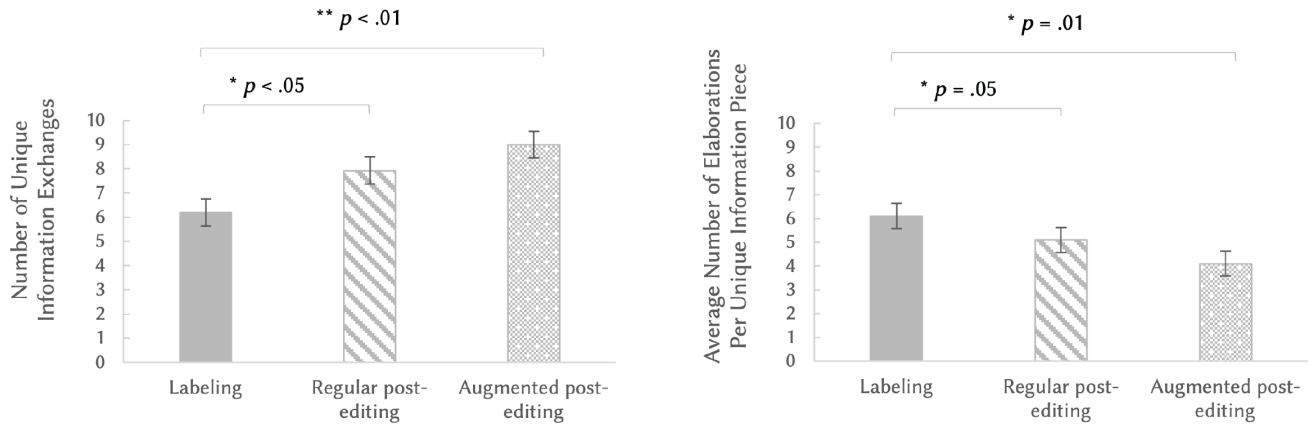


Figure 4: Dyadic information exchange process as in the breadth (left) and depth (right) by the human-MT interface condition

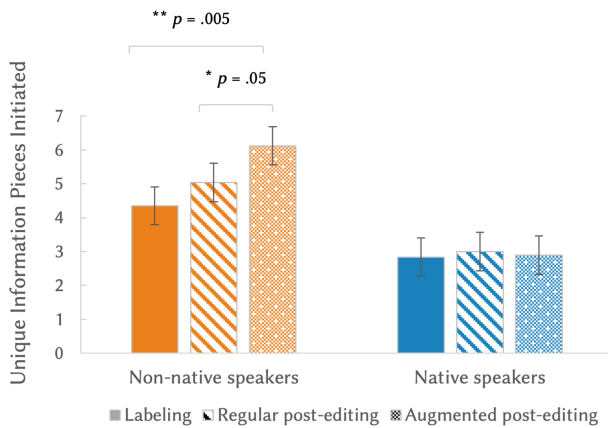


Figure 5: Unique information pieces initiated by the human-MT interface condition and individual language background

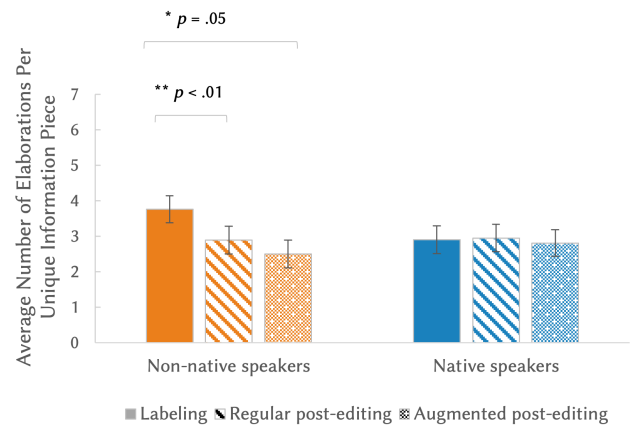


Figure 6: Elaborations added by the human-MT interface condition and individual language background

of the language background ($F [1, 53.65] = 0.13, p = .72$). However, there was a significant interaction effect between these two factors: $F [2, 38.35] = 4.29, p < .05$.

Pairwise comparisons revealed that non-native speakers contributed more to the generation of elaborations under the labeling interface condition ($M = 3.76, S.E. = .38$) than in the regular post-editing interface condition ($M = 2.89, S.E. = .39; F [1, 61.61] = 7.34, p < .01$) or the augmented post-editing interface condition ($M = 2.50, S.E. = .39; F [1, 61.28] = 3.56, p = .05$). In contrast, native speakers' contributions to the elaborations turned out to be equal across the augmented post-editing interface condition ($M = 2.81, S.E. = .39$), the regular post-editing condition ($M = 2.95, S.E. = .39$), and the labeling condition ($M = 2.90, S.E. = .38$).

In short, the above results suggested that the dyadic information exchange process varied according to the influence non-native speakers had on MT outputs: conversations under the labeling interface condition featured the highest level of depth but the lowest level of breadth. No significant differences were found between

the two post-editing interface conditions. More interestingly, non-native speakers appeared to act as the primary drivers of those observed differences, while native speakers performed similarly across all three interface conditions.

4.3 Dyadic Information Exchange Outcome

Our RQ1b asked how the adoption of the three human-MT interfaces would affect participants' dyadic information exchange outcome. To answer this question, we conducted a one-way ANOVA to compare the dyadic ranking alignment scores across three human-MT interface conditions (Figure 7). The dependent variable was the level of alignment between the two participants within a dyad. The rest of the model setup was the same as that of the dyadic breadth of information exchange. The results indicated a significant main effect of the human-MT interface condition. Specifically, dyads achieved a greater level of alignment when their information exchange took place under the labeling interface condition (Mean =

.57 S.E. = .06) compared to the regular post-editing interface condition (Mean = .34 S.E. = .06; $F [1, 42] = 6.74, p = .01$) or the augmented post-editing interface condition (Mean = .33 S.E. = .06; $F [1, 42] = 7.55, p < .01$). There was no significant difference between the two post-editing interface conditions.

In short, the dyadic task outcome, in terms of alignment achieved between two parties of the information exchange, was best under the labeling interface condition.

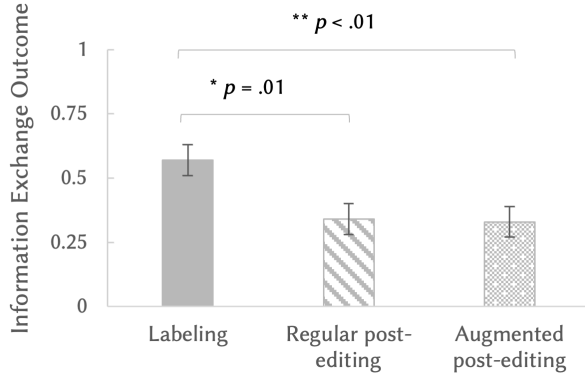


Figure 7: Dyadic information exchange outcome by the human-MT interface condition

4.4 Workload

Our RQ2 asked how the adoption of the three human-MT interfaces would affect participants' perceived workload. We performed a 3×2 Mixed Model ANOVA to answer this question. The model setup was the same as that used for agency, only except that the dependent variable became each participant's self-rating of their workload during the task. The results demonstrated no difference in workload across the three interface conditions ($F [2, 41.32] = 0.44, p = .65$), nor between participants with different language backgrounds ($F [1, 77.41] = 0.97, p = .33$). The interaction effect was not significant either ($F [2, 41.24] = 0.20, p = .82$).

That is, the perceived workload was consistent across all three interface conditions and for all participants.

5 DISCUSSION

To recap, the conventional design of human-MT interface today typically follows a straightforward two-box layout: non-native speakers of a target language (e.g., English) type in their source message in one text box; they receive and may preview the translated version of that message in another box before sending. Such an interface design generally does not allow users to exercise their agency by making further influence on the MT outputs. Our current work investigated ways to sustain non-native speakers' agency in MT-mediated conversations. We explored three potential designs for the human-MT interface. Each interface allowed non-native speakers to leverage their resources to a distinct extent, thereby influencing the translated messages sent to their task partners.

In the remainder of the Discussion section, we reflect on how our findings relate to existing literature in HCI and beyond. These reflections focus on three key aspects of our results: the agency-resource connection (Section 5.1), the associated costs of sustaining non-native speakers' agency (Section 5.2), and the comparable levels of agency observed under two post-editing conditions (Section 5.3). We conclude this section by presenting design insights derived from these reflections (Section 5.4).

5.1 Operationalizing Agency Preservation Through the Agency-Resource Connection

We speculated three conditions of human-MT interface for satisfying non-native speakers' agentic needs: the labeling condition, the regular post-editing condition, and the augmented post-editing condition. In each case, the average self-ratings given by non-native speakers confirmed that resource involvement would effectively anchor personal agency in MT-mediated conversations. Participants reported preserving medium to high levels of their personal agency when working with all three tested interfaces. Across conditions, the comparison of non-native speakers' self-ratings revealed how personal agency would vary based on the scope of resources they leveraged to influence MT outputs. Participants reported significantly lower agency when they utilized only a target portion of their language resources to label the quality of MT outputs, compared to when they used a broader set of their language resources for direct editing.

The above findings highlight the unique value of adopting Bandura's theoretical lens to better sustain non-native speakers' agency. In particular, Bandura's theory underscores the importance of understanding personal agency within the context of resource involvement. Beyond understanding, it also calls for adjusting the agency an individual can exercise by manipulating the resources they can access [8, 10, 12, 13]. Such a perspective not only provides HCI researchers with conceptual insights into the behavioral foundation of personal agency but also suggests actionable strategies for moderating agency through system interface design.

Inspired by the agency-resource connection, we suspect that the regular post-editing interface, commonly used to preserve the agency of fluent bilinguals (e.g., professional translators; [21, 48, 64, 79]), may not be the sole reasonable option to sustain non-native speakers' agency. We specified and compared three human-MT interfaces that enabled non-native speakers to utilize different scopes of their language resources in a progressive manner. The labeling interface addresses non-native speakers' agentic needs by leveraging their passive English vocabulary – a limited but reliable subset of the person's language resources – for recognizing or evaluating the quality of MT outputs. The augmented post-editing interface, conversely, served the same needs by activating more of the non-native speakers' active English vocabulary to facilitate their direct editing of MT outputs. Together, these interfaces expand our view on how to preserve a non-native speaker's agency, taking into account the sharper divide between passive and active vocabulary in the composition of these individuals' language resources in English.

Notably, while we adopted Bandura's agency-resource connection to guide the current investigation of agency, our intention

has never been to diminish the importance of other relevant theories. In fact, the concept of agency has been widely discussed in numerous previous HCI projects. According to Bennett et al.'s recent review [15], a substantial body of this prior work draws on self-determination theory to support their interpretations of agency. Other relevant frameworks include Suzy Killmister's four dimensions of self-governance [60, 61], Frankfurt's hierarchical notion of freedom [36, 37], and many others (e.g., [31, 32, 88, 97]). Most of these theories emphasize the importance of an individual's self-identity, personal values, intrinsic motivations, and desires in shaping their view of agency. Compared to them, Bandura's theory focuses more on an individual's exercise of agency [8, 10, 12, 13], often through the use of their resources as well as the task environment that contextualizes the access to and utilization of such resources. This comparison makes Bandura's framework particularly suitable for the purposes of our research, which centers on ideating and operationalizing human-MT interface design for non-native speakers' agency preservation.

5.2 Attending to Associated Costs of Non-Native Speakers' Agency

The results of our study demonstrated the potential of all three tested human-MT interfaces in supporting non-native speakers' agency. This naturally raises the question: Does one of these interfaces perform significantly better or worse than the others? Should we, as researchers and designers, recommend that future MT-mediated systems adopt or avoid any of the particular interfaces tested here? Reflecting on these questions, it becomes clear that we must first define what qualifies as a "better" or "worse" human-MT interface.

If we consider the preservation of non-native speakers' agency as the sole criterion, we might conclude that the labeling interface is less effective than the two post-editing interfaces. While such an interface design supported participants' agency, it did so only to a moderate degree. However, this conclusion overlooks other critical aspects of dyadic communication, especially when such communication serves high-stakes purposes, such as facilitating essential information seeking for immigrants.

To gain a more comprehensive understanding, we analyzed additional measures to explore how dyadic information exchange performance varied across the different human-MT interfaces provided to non-native speakers. Our findings revealed that the variation in agency had interesting associated effects. In the labeling condition, dyads exchanged fewer unique pieces of information compared to other conditions. However, they engaged in deeper discussions about each piece of information. Our data indicated that non-native speakers were the primary drivers of these in-depth discussions. It also revealed more successful alignment between the two speakers by the end of their information exchange.

Thus, when we adopt a holistic perspective that considers multiple qualifiers for successful communication, important strengths of the labeling interface become evident. The boosts in non-native speakers' agency observed under the two post-editing conditions came at significant costs. One possible explanation for these costs is that utilizing active English vocabulary to edit MT outputs depleted the attention participants might otherwise have allocated to

probing further into their discussions with task partners. Similar tensions have also been reported in recent literature on AI-mediated communication (e.g., [20, 39, 62]), and they may be even more pronounced for non-native speakers, whose language resources are characterized by the "recognition and production gap [44]."

Given these reflections, we believe that our study contributes an empirical understanding of the costs associated with supporting non-native speakers' agency. It complements existing HCI literature by addressing a critical shortfall: while there is broad consensus on the value of supporting agency, prior research often fails to clarify how agency relates to other aspects of users' experiences within specific task environments [15].

Returning to the question of what constitutes a "better" or "worse" human-MT interface, we find no straightforward answer due to the associated costs of agency, as detailed above. When the priority is to understand the detailed reasoning behind an immigrant's information needs, the labeling interface is a good choice: it maintains a degree of personal agency in language use for non-native speakers and facilitates in-depth information exchange between the immigrant and the local person. However, when the priority shifts to understanding the landscape of an immigrant's information needs without delving into details during a single conversation, the two post-editing interfaces may be more effective.

We believe it is important to raise awareness, within the research community as well as among MT users, of the potential trade-offs between agency and other aspects of communication. Equipped with this understanding, people can make informed decisions about human-MT interface designs that align with their priorities in specific contexts.

5.3 Making Sense of the Comparable Levels of Agency Under Two Post-Editing Conditions

The comparison of non-native speakers' agency between two post-editing conditions did not support our H2b. While the LLM-generated paraphrases were designed to activate more of the active English vocabulary of non-native speakers and scaffold their post-editing, this did not result in higher agency in the augmented condition, compared to the regular condition.

We uncovered additional evidence from the data to make sense of the above result. We examined all conversation logs from the augmented post-editing condition to verify whether non-native speakers in our study have utilized these language resources. Our data showed that, on average, participants consulted the paraphrases in 30.3% (S.E. = 9.1%) of the messages they sent. They adopted words and expressions from these paraphrases in 19.6% (S.E. = 6.4%) of their messages. Among all messages edited in the augmented post-editing condition, an average of 62.7% (S.E. = 10.1%) of the edited tokens were adopted from the LLM-generated paraphrases. This data confirms that the language resources hinted via the augmented interface did facilitate our participants' use of their active English vocabulary during MT output editing.

Thus, rather than suspecting a lack of engagement with the LLM-generated resources or a failure in establishing the agency-resource connection, we believe a "ceiling effect" provides a more reasonable explanation for our findings. That is, non-native speakers' agency in the regular post-editing condition was already high (i.e., an average

score of 5.95 out of 7), leaving limited room for another significant increase in the augmented condition.

Moreover, some measures analyzed in our study highlighted how the augmented post-editing interface benefited our participants, although this benefit was not directly related to agency. Specifically, we observed that non-native speakers in the augmented post-editing condition initiated more unique information exchanges with their task partners compared to how they behaved in other conditions. This finding, along with those discussed in Section 5.2, underscores the necessity of remaining mindful of associated costs and benefits when valuing personal agency.

5.4 Deriving Insights for Sustaining Non-Native Speakers' Agency Through System Design

Building on the critical aspects of the results discussed above, we outline three key directions for sustaining the agency of non-native speakers. These directions extend beyond the specific task context of this study. The first two focus on future human-MT interfaces to support conversational interactions, while the third envisions how such interfaces could facilitate innovative data collection methods for training translation algorithms.

5.4.1 Labeling as a low-cost and valuable feature for conversational interfaces. Our analysis underscores the potential of the labeling interface to support high-quality dyadic communication while preserving the personal agency of non-native speakers at a moderate level. Notably, similar labeling functions have already been integrated into several commercial MT systems. For instance, Google Translate and DeepL both encourage users to evaluate the translations they receive by clicking a thumbs-up or thumbs-down label. While these interfaces are convenient to use, they are not designed for real-time conversational contexts. The labels are primarily utilized by developers to improve translation services in the long term, rather than enabling users to influence MT outputs or communicate their inputs to conversation partners during a real-time interaction.

Some recent studies on MT usage have implemented labeling interfaces, similar to the one discussed in our work, for participants to flag low-quality translations encountered in MT-mediated instant message exchanges [87]. However, the focus of that research was not on examining personal agency during communication but rather on capturing human perceptions of translation quality.

Building on these prior efforts, our work emphasizes the broader value of incorporating the labeling feature into text-based communication tools equipped with MT modules. This feature is especially helpful for non-native speakers with low proficiency in the target language. It provides them with a feasible mechanism to effectively engage with a subset of their language resources (e.g., passive vocabulary in the target language), thereby exercising their influence on the messages sent to conversational partners.

5.4.2 Promoting awareness of the costs associated with enhanced agency. Besides incorporating the labeling function, our study results suggest other, more straightforward ways to enhance non-native speakers' agency in MT-mediated conversations. For instance, the interface could let the translated text in the message box remain editable until the communicant clicks the "send" button. This feature would be easy to implement and holds promising

potential to encourage non-native speakers to post-edit their messages. However, many instant messaging tools today, such as the messenger module on all versions of iPhones, have yet to consider this feature.

A substantial part of our results revealed costs that may be associated with enhanced agency in post-editing conditions. Further analysis pointed to reasons for these costs: while non-native speakers may attempt to utilize their active English vocabulary to edit MT outputs, their limited English proficiency makes this process effort-intensive. Consequently, post-editing appeared to distract these individuals from engaging in in-depth information exchanges with their conversational partners. To address this constraint, we envision that future system designs could encourage non-native speakers to post-edit MT outputs selectively, focusing on moments where high-stakes miscommunication is likely to occur. This guidance on prioritizing post-editing efforts may help users balance the benefits and costs of using a post-editing interface.

Recent advancement in LLMs provides the technical feasibility to generate the described guidance. In particular, several research projects have demonstrated the success of using LLMs to identify selected discourse acts [78, 86, 95, 96], track various aspects of the conversation dynamics over time [57, 72], and detect interaction patterns that reveal miscommunications [35, 55, 58, 83, 109]. These techniques could be applied to highlight moments where post-editing is most crucial for communicants.

5.4.3 Building customized MT systems based on first-person account data. The labeling and post-editing interfaces explored in this study open up new technical opportunities to train customized MT algorithms that reflect an individual's word choices and language styles, thereby sustaining their agency in a proactive manner. They hold the promise of offering complementary information to that used in today's MT system-building practices. The latter often relies on training data generated by third-party annotators (e.g., crowd workers who review the text corpus resulting from others' conversations), without incorporating first-person perspectives contextualized to the actual communicant [4].

More specifically, labels generated by non-native speakers during their real-time communication could serve as reinforcement learning feedback to personalize MT models [3, 65, 74, 76, 90]. Developers could also train a classifier to predict the same person's preferences in follow-up conversations over time [110]. Post-edits collected through the human-MT interfaces tested in our study could serve as in-context learning examples, aiding the fine-tuning of translation models [5, 17, 18, 75, 80] and correcting tacit errors deemed important by a given communicant [59, 101]. More collaborative research between HCI and NLP experts may emerge as more data from the first person account becomes available.

6 LIMITATIONS AND FUTURE DIRECTIONS

As one of the first HCI studies to examine non-native speakers' agency in MT-mediated conversations, our work was conditioned by the specific methodological choices and task settings described in this paper. Below, we reflect on the composition of participants, the three human-MT interfaces, and the communication modalities examined in this work. We detail our thoughts on the generalizability of our findings, as well as opportunities for future research.

6.1 The Composition of Participants

We investigated ways to sustain non-native speakers' agency in the context of essential information-seeking interactions between Mandarin-speaking immigrants (as seekers) and English-speaking local residents in the United States (as providers). This choice was motivated partly by the real-world language challenges we observed through communal work with this immigrant group and partly by the linguistic differences between Mandarin and English, which are of empirical significance for MT research. For example, the syntactic and pragmatic distance between Mandarin and English is much greater than that between most Romance languages and English. In the former case, it becomes even more critical to encourage non-native speakers' exercise of their personal agency in monitoring, evaluating, and influencing MT outputs.

If we were to switch to other immigrant groups speaking different native languages, as long as they have imperfect proficiency in English, the "recognition and production gap" in their use of English language resources would persist. Thus, our proposal to leverage the agency-resource connection as a conceptual foundation for preserving these individuals' agency would remain applicable.

In addition to native language type, English fluency level constitutes another important aspect of a non-native speaker's linguistic profile. Participants in our sample exhibited a medium level of English fluency, with small individual differences in this measure. The broader population of non-native speakers includes individuals with both lower and higher English fluency than our participants. In such cases, we believe that the overall design approach of enabling resource involvement at varying scopes would still help sustain people's agency in MT-mediated conversations. The more constraints a non-native speaker faces in utilizing their active vocabulary, the more likely they are to favor and benefit from the labeling interface.

Alternatively, these individuals could be provided with an augmented post-editing interface, where additional hints are offered to help activate more of their own language resources in English. A potential issue with this second setup is that it may be cognitively taxing for individuals to process the hints provided via the interface. It could raise more serious concerns regarding the associated costs discussed in our study, such as the reduced quality of the dyadic information exchange. Given these considerations, we suspect that the labeling interface can support agency preservation across a broader range of non-native speakers, whereas the post-editing interfaces may not offer the same level of accessibility or benefit.

6.2 The Three Human-MT Interfaces

Inspired by Bandura's agency-resource connection, we speculated on three human-MT interface designs. Our empirical comparisons across these interfaces showcased two pathways for calibrating a person's resource involvement to exercise agency. One pathway engages a smaller-than-full scope of a person's existing resource pool, as evidenced by our analysis and discussion of the labeling condition. The other provides additional hints to scaffold the activation of a person's full resources, as demonstrated in the augmented post-editing condition. They highlight two generalizable approaches for designing human-MT interfaces that meet the agentic needs of non-native speakers.

For the first pathway, in particular, we do not yet have well-developed ideas for other interface designs that also enable the involvement of a non-native speaker's passive vocabulary but do not follow the format tested in our study. We anticipate that readers of our paper may perceive the labeling condition as somewhat different from the two post-editing conditions: in the labeling condition, each message received by native speakers consisted of MT-generated textual content and labels provided by non-native speakers. In the two post-editing conditions, each message received by native speakers consisted of the MT-generated content and any edits made by non-native speakers; however, these two components appeared more visually integrated.

We were less concerned about this difference for two reasons. First, all native speakers in our study were made aware that their conversational partners were adding influence, through labeling or editing, to those translated messages before sending them (see Section 3.2.2 for details). Second, the agency ratings collected from native speakers showed no significant difference across conditions, suggesting that native speakers perceived their conversational partners' agency at an equal level, regardless of the human-MT interface used by those partners. For non-native speakers, even after assigning more visually prominent labels to MT outputs, they still rated their agency lower in the labeling condition than in other conditions. Such results align with what Bandura's theory would predict, indicating the comparative benefits and limitations of using the labeling interface to support people's agentic needs.

For the second pathway, we envision that numerous alternative interface designs could be explored beyond the augmented interface. For instance, the number of paraphrases and the format of LLM-generated language resources could both be adjusted. The interface could also provide clues to assist non-native speakers' critical assessment and adoption of AI-generated resources, such as translation quality feedback (e.g., [73]) and structured prompts to aid people's critical thinking process (e.g., [19, 52]).

6.3 The Communication Modality

Our current work focuses on text-based conversations. This setup captures the translation functions used in available instant messaging applications, such as Slack, WeChat, and iMessage. The text-based modality enables individuals to interact with MT outputs and seek paraphrases for each written message. It also allows the researcher to log and track the entire conversation. That said, these benefits may not fully extend to speech-based translations, where additional challenges for non-native speakers' language use often occur (e.g., [49]). Future work should explore how to sustain human agency in cases involving variations in participant composition, human-MT interfaces, and communication modalities beyond those examined in our study. The research materials and findings from our current study pave the way for this exciting area of exploration.

7 CONCLUSION

The current paper presents an experimental investigation involving 45 dyads, aimed at exploring the design of human-machine translation (MT) interfaces to sustain non-native speakers' agency in MT-assisted conversations. Each dyad was tasked with exchanging information in a given real-life scenario. In this scenario, one

non-native speaker, acting as an immigrant information seeker, sought housing selection tips and suggestions for their family in the United States. One native English speaker, acting as a local information provider, engaged in the conversation with the non-native speaker to assist with their needs. Inspired by Bandura's notion of the agency-resource connection, we proposed three different human-MT interfaces and provided each dyad with one of them. Interestingly, we found that the labeling interface – the one designed to help non-native speakers maintain a moderate level of agency – facilitated dyadic information exchange with better performance. The two post-editing interfaces could sustain a higher level of non-native speakers' agency but at the cost of dyadic performance. These results highlight the complexity of human agency in MT-assisted conversations. In particular, they underscore the critical trade-off between personal agency and collective performance in non-native speakers' use of MT. Based on these findings, we outline potential directions for future human-MT interface design that can preserve non-native speakers' agency while also guiding people to remain mindful of the associated costs.

Acknowledgments

This research was supported by the National Science Foundation, under grants #2147292 and #2229885. We thank the reviewers for their constructive feedback on this work. We are also grateful to Ruipu Hu, Connie Siebold, Siyi Zhu, Elana Blinder, Yongle Zhang, Jian Zheng, Victoria Chang, and Yuhang Zhou for their insights during the early stages of our system development and research design. Last but not least, we thank all participants who engaged in our study and shared their task experiences with us.

References

- [1] Adinawa Adjagbodjou and Geoff Kaufman. 2024. Envisioning support-centered technologies for language practice and use: needs and design opportunities for immigrant English language learners (ELLs). In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24)*. 1–15.
- [2] Denise Adkins and Heather Moulaison Sandy. 2020. Information behavior and ICT use of Latina immigrants to the U.S. Midwest. *Information Processing & Management* 57, 3 (May 2020), 102072.
- [3] Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M Guerreiro, and Andre Martins. 2024. Modeling user preferences with automatic metrics: creating a high-quality preference dataset for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14503–14519.
- [4] Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14491–14502.
- [5] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8857–8873.
- [6] Laura M Ahearn. 2001. Language and agency. *Annual review of anthropology* 30, 1 (2001), 109–137.
- [7] Nathalie Aichhorn and Jonas Puck. 2017. “I just don't feel comfortable speaking English”: foreign language anxiety as a catalyst for spoken-language barriers in MNCs. *International Business Review* 26, 4 (Aug. 2017), 749–763.
- [8] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American Psychologist* 37, 2 (1982), 122–147.
- [9] Albert Bandura. 1991. Social Cognitive Theory of Moral Thought and Action. In *Handbook of Moral Behavior and Development*. Psychology Press.
- [10] Albert Bandura. 1999. Social cognitive theory: An agentic perspective. *Asian Journal of Social Psychology* 2, 1 (1999), 21–41.
- [11] Albert Bandura. 2001. Social cognitive theory: An agentic perspective. *Annual review of psychology* 52, 1 (2001), 1–26.
- [12] Albert Bandura. 2003. On the psychosocial impact and mechanisms of spiritual modeling. *The International Journal for the Psychology of Religion* 13, 3 (2003).
- [13] Albert Bandura. 2006. Toward a psychology of human agency. *Perspectives on psychological science* 1, 2 (2006), 164–180.
- [14] Luis Fernando Baron, Moriah Neils, and Ricardo Gomez. 2014. Crossing new borders: computers, mobile phones, transportation, and English language among Hispanic day laborers in Seattle, Washington. *Journal of the Association for Information Science and Technology* 65, 1 (2014), 98–108.
- [15] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI understand human agency and autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18.
- [16] Paola Beretta, Elham Sayyad Abdi, and Christine Bruce. 2018. Immigrants' information experiences: an informed social inclusion framework. *Journal of the Australian Library and Information Association* 67, 4 (2018), 373–393.
- [17] Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the WMT 2022 shared task on automatic post-editing. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 109–117.
- [18] Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (Eds.). Association for Computational Linguistics, Singapore, 672–681.
- [19] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
- [20] 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 (CHI'21)*. 1–13.
- [21] Patrick Cadwell, Sheila Castillo, Sharon O'Brien, and Linda Mitchell. 2016. Human factors in machine translation and post-editing among institutional translators. *Translation Spaces. A multidisciplinary, multimedia, and multilingual journal of translation* 5, 2 (2016), 222–243.
- [22] Nadia Caidi, Danielle Allard, and Lisa Quirke. 2010. Information practices of immigrants. *Annual review of information science and technology* 44, 1 (2010), 491–531.
- [23] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
- [24] Mei-Ling Chen, Naomi Yamashita, and Hao-Chuan Wang. 2018. Beyond Lingua Franca: system-facilitated language switching diversifies participation in multi-party multilingual communication. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22.
- [25] Ann E Chun, Richard R Day, N Ann Chenoweth, and Stuart Luppescu. 1982. Errors, interaction, and correction: a study of native-normative conversations. *Tesol Quarterly* 16, 4 (1982), 537–547.
- [26] Herbert H Clark. 1996. *Using language*. Cambridge university press.
- [27] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: an intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–13.
- [28] Richard R Day, N Ann Chenoweth, Ann E Chun, and Stuart Luppescu. 1984. Corrective feedback in native-nonnative discourse. *Language learning* 34, 2 (1984), 19–45.
- [29] Jia Tina Du. 2023. Understanding the information journeys of late-life migrants to inform support design: information seeking driven by a major life transition. *Information Processing & Management* 60, 2 (2023), 103172.
- [30] Wen Duan, Naomi Yamashita, and Susan R Fussell. 2019. Increasing native speakers' awareness of the need to slow down in multilingual conversations using a real-time speech speedometer. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
- [31] Gerald Dworkin. 1988. *The theory and practice of autonomy*. Vol. 20. Cambridge University Press.
- [32] Mustafa Emirbayer and Ann Mische. 1998. What is agency? *American journal of sociology* 103, 4 (1998), 962–1023.
- [33] Marisol Fernández-García and Asunción Martínez-Arbelaiz. 2014. Native speaker–non-native speaker study abroad conversations: do they provide feedback and opportunities for pushed output? *System* 42 (2014), 93–104.
- [34] Dana R Ferris, Hsiang Liu, Aparna Sinha, and Manuel Senna. 2013. Written corrective feedback for individual L2 writers. *Journal of second language writing* 22, 3 (2013), 307–329.
- [35] Sarah E Finch, Ellie S Paek, and Jinho D Choi. 2023. Leveraging large language models for automated dialogue analysis. *arXiv preprint arXiv:2309.06490* (2023).

- [36] Harry Frankfurt. 2001. Freedom of the Will and the Concept of a Person. In *Agency And Responsibility*. Routledge, 77–91.
- [37] Harry Frankfurt. 2003. Alternate possibilities and moral responsibility. In *Moral responsibility and alternative possibilities*. Routledge, 17–25.
- [38] 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 (CHI'23)*. 1–13.
- [39] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From text to self: users' perception of AIMC tools on interpersonal communication and self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [40] Ge Gao and Susan R Fussell. 2017. A kaleidoscope of languages: When and how non-Native English speakers shift between English and their Native language during multilingual teamwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 760–772.
- [41] Ge Gao, Bin Xu, Dan Cosley, and Susan R. Fussell. 2014. How beliefs about the presence of machine translation impact multilingual collaborations. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, Baltimore Maryland USA, 1549–1560.
- [42] Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two Is better than one: improving multilingual collaboration by giving two machine translation outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, Vancouver BC Canada, 852–863.
- [43] 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. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.
- [44] Susan Gass. 1983. The development of L2 intuitions. *Tesol Quarterly* 17, 2 (1983), 273–291.
- [45] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)*. 1–15.
- [46] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014. Predictive translation memory: a mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST'14)*. 177–187.
- [47] Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 439–448.
- [48] Ana Guerberoof Arenas. 2013. What do professional translators think about post-editing. *JoSTrans The journal of specialised translation* 19 (2013), 75–95.
- [49] Kotaro Hara and Shamsi T Iqbal. 2015. Effect of machine translation in interlingual conversation: lessons from a formative study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3473–3482.
- [50] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
- [51] Ari MJ Hautasaari, Naomi Yamashita, and Ge Gao. 2014. "Maybe it was a joke" emotion detection in text-only communication by non-native english speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3715–3724.
- [52] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To err is AI! Debugging as an intervention to facilitate appropriate reliance on AI systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. 98–105.
- [53] Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: a multi-modal interface for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 1691–1702.
- [54] Nico Herbig, Santanu Pal, Josef van Genabith, and Antonio Krüger. 2019. Multi-modal approaches for post-editing machine translation. In *Proceedings of the 2019 CHI conference on human factors in computing systems, CHI'19*. 1–11.
- [55] Ryuichiro Higashimaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 3146–3150.
- [56] Yuri Hosoda. 2000. Other-repair in Japanese conversations between nonnative and native speakers. *Issues in Applied Linguistics* 11, 1 (2000).
- [57] Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Julie Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. How did we get here? Summarizing conversation dynamics. *arXiv preprint arXiv:2404.19007* (2024).
- [58] Yova Kementchedjieva and Anders Søgaard. 2021. Dynamic forecasting of conversation derailment. *arXiv preprint arXiv:2110.05111* (2021).
- [59] Dayeon Ki and Marine Carpuat. 2024. Guiding Large Language Models to post-edit machine translation with error annotations. 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, 4253–4273.
- [60] Suzy Killmister. 2013. Autonomy and false beliefs. *Philosophical Studies* 164, 2 (2013), 513–531.
- [61] Suzy Killmister. 2017. *Taking the measure of autonomy: a four-dimensional theory of self-governance*. Routledge.
- [62] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. Diary-Mate: understanding user perceptions and experience in human-AI collaboration for personal journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, (CHI'24)*. 1–15.
- [63] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. "If the machine is as good as me, then what use am I?"—how the use of ChatGPT changes young professionals' perception of productivity and accomplishment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, (CHI'24)*. 1–16.
- [64] Maarit Koponen. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* 25, 2 (2016), 131–148.
- [65] Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheuer, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberoof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (Eds.). European Association for Machine Translation, Lisboa, Portugal, 135–144.
- [66] Salla Kurhila. 2001. Correction in talk between native and non-native speaker. *Journal of Pragmatics* 33, 7 (2001), 1083–1110.
- [67] 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 (CHI'22)*. 1–19.
- [68] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
- [69] Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet needs and opportunities for mobile translation AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*. 1–13.
- [70] Jessica Lingel. 2011. Information tactics of immigrants in urban environments. *Information Research* 16 (2011).
- [71] Edwin A. Locke and Gary P. Latham. 1990. *A theory of goal setting & task performance*. Prentice-Hall, Inc, Englewood Cliffs, NJ, US.
- [72] Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. Dialogue-trm: exploring the intra- and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637* (2020).
- [73] Nikita Mehndru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician Detection of Clinical Harm in Machine Translation: Quality Estimation Aids in Reliance and Backtranslation Identifies Critical Errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). 11633–11647.
- [74] Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 312–318.
- [75] Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Tampere, Finland, 227–237.
- [76] Jason Naradowsky, Xuan Zhang, and Kevin Duh. 2020. Machine translation system selection from bandit feedback. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Michael Denkowski and Christian Federmann (Eds.). Association for Machine Translation in the Americas, Virtual, 50–63.
- [77] Tsedal B Neeley. 2013. Language matters: status loss and achieved status distinctions in global organizations. *Organization Science* 24, 2 (2013), 476–497.
- [78] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.
- [79] Sharon O'Brien. 2022. How to deal with errors in machine translation: post-editing. In *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*. Zenodo, 105–120.
- [80] Daniela Occhipinti, Michele Marchi, Irene Mondella, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2024. Fine-tuning with HED-IT: the impact of human post-editing for dialogical language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre

- Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11892–11907.
- [81] So Yeon Park and Sang Won Lee. 2023. Why “why”? The importance of communicating rationales for edits in collaborative writing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'23)*. 1–25.
- [82] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. In-body experiences: embodiment, control, and trust in robot-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. 1921–1930.
- [83] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEditIT: text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). 5274–5291.
- [84] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan “Michael” Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. ABScribe: rapid exploration & organization of multiple writing variations in human-AI co-writing tasks using Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'24)*.
- [85] Jeba Rezwana and Mary Lou Maher. 2023. Designing creative AI partners with COFI: a framework for modeling interaction in human-AI co-creative systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (sep 2023), 28 pages.
- [86] Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. , 172–180 pages.
- [87] Samantha Robertson and Mark Díaz. 2022. Understanding and being understood: user strategies for identifying and recovering from mistranslations in machine translation-mediated chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2223–2238.
- [88] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.
- [89] Carol D. Ryff. 1989. Beyond ponce de leon and life satisfaction: new directions in quest of successful ageing. *International Journal of Behavioral Development* 12, 1 (March 1989), 35–55.
- [90] Avneesh Saluja, Ian Lane, and Ying Zhang. 2012. Machine translation with binary feedback: a large-margin approach. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*. Association for Machine Translation in the Americas, San Diego, California, USA.
- [91] Eike Schneiders, Christopher Fourie, Stanley Celestin, Julie Shah, and Malte Jung. 2024. Understanding entrainment in human groups: optimising human-robot collaboration from lessons Learned during human-human collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'24)*. 1–13.
- [92] Stefan Schulz-Hardt, Felix C Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. 2006. Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology* 91, 6 (2006), 1080.
- [93] Shalom Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology* 53 (Sept. 1987), 550–562.
- [94] Snunith Shoham and Sarah Kaufman Strauss. 2008. Immigrants’ information needs: their role in the absorption process. *Information Research* 13, 4 (2008), 13–4.
- [95] Aina Gari Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring lexico-semantic alignment in debates with contextualized word representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SiCon 2023)*. Association for Computational Linguistics, 50–63.
- [96] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26, 3 (2000), 339–373.
- [97] Roger J Sullivan. 1989. *Immanuel Kant’s moral theory*. Cambridge University Press.
- [98] S Shyam Sundar. 2020. Rise of machine agency: a framework for studying the psychology of human-AI interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88.
- [99] Adam Tapal, Ela Oren, Reuven Dar, and Baruch Eitam. 2017. The sense of agency scale: a measure of consciously perceived control over one’s mind, body, and the immediate environment. *Frontiers in psychology* 8 (2017), 1552.
- [100] Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — building open translation services for the World. In *Proceedings of the 22nd Annual Conferenc of the European Association for Machine Translation (EAMT)*. Lisbon, Portugal.
- [101] Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pomal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: a multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15222–15239.
- [102] Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. Machine translation vs. common language: effects on idea exchange in cross-lingual groups. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 935–944.
- [103] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.
- [104] Jean Wong. 2000. Delayed next turn repair initiation in native/non-native speaker English conversation. *Applied linguistics* 21, 2 (2000), 244–267.
- [105] Yimin Xiao, Yuewen Chen, Naomi Yamashita, Yuexi Chen, Zhicheng Liu, and Ge Gao. 2024. (Dis)placed contributions: uncovering hidden hurdles to collaborative writing involving non-native speakers, native speakers, and AI-powered editing tools. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–31.
- [106] Naomi Yamashita and Toru Ishida. 2006. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW'06)*. 515–524.
- [107] Seraphina Yong, Leo Cui, Evan Suma Rosenberg, and Svetlana Yarosh. 2024. A change of scenery: transformative insights from retrospective VR embodied perspective-taking of conflict with a close other. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'24)*. 1–18.
- [108] Yongle Zhang, Dennis Asamoah Owusu, Marine Carpuat, and Ge Gao. 2022. Facilitating global team meetings between language-based subgroups: when and how can machine translation help? *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.
- [109] Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. 2024. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). 4455–4476.
- [110] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

APPENDIX A. GPT PROMPT FOR GENERATING PARAPHRASES IN THE AUGMENTED POST-EDITING CONDITION

Generate two paraphrases for this message by following the instructions:

The message is about two houses that the person is considering purchasing;

Use natural English expressions and sentence structure in your paraphrases;

Adjust the message’s wording to make one paraphrase more casual and one more formal;

The length of the paraphrase should be within a 5-word difference from the original message.