Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat

Samantha Robertson*
samantha_robertson@berkeley.edu
University of California, Berkeley
Berkeley, California, USA

Mark Díaz
markdiaz@google.com
Google Research
New York, New York, USA

ABSTRACT

Machine translation (MT) is now widely and freely available, and has the potential to greatly improve cross-lingual communication. In order to use MT reliably and safely, end users must be able to assess the quality of system outputs and determine how much they can rely on them to guide their decisions and actions. However, it can be difficult for users to detect and recover from mistranslations due to limited language skills. In this work we collected 19 MT-mediated role-play conversations in housing and employment scenarios, and conducted in-depth interviews to understand how users identify and recover from translation errors. Participants communicated using four language pairs: English, and one of Spanish, Farsi, Igbo, or Tagalog. We conducted qualitative analysis to understand user challenges in light of limited system transparency, strategies for recovery, and the kinds of translation errors that proved more or less difficult for users to overcome. We found that users broadly lacked relevant and helpful information to guide their assessments of translation quality. Instances where a user erroneously thought they had understood a translation correctly were rare but held the potential for serious consequences in the real world. Finally, inaccurate and disfluent translations had social consequences for participants, because it was difficult to discern when a disfluent message was reflective of the other person's intentions, or an artifact of imperfect MT. We draw on theories of grounding and repair in communication to contextualize these findings, and propose design implications for explainable AI (XAI) researchers, MT researchers, as well as collaboration among them to support transparency and explainability in MT. These directions include handling typos and non-standard grammar common in interpersonal communication, making MT in interfaces more visible to help users evaluate errors, supporting collaborative repair of conversation breakdowns, and communicating model strengths and weaknesses to users.

KEYWORDS

machine translation, human-AI interaction, computer-mediated communication, explainable machine learning

*Research conducted while an intern at Google Research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3534638

ACM Reference Format:
Samantha Robertson and Mark Díaz. 2022. Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3531146.3534638

1 INTRODUCTION

To use machine learning (ML) systems reliably and safely, end users must be able to assess the quality of system outputs and determine their reliability to guide decisions and actions. This is challenging when users lack information about how the system works, or its strengths and limitations. Machine translation (MT) is one ML system that has the potential to improve cross-lingual communication. However, limited language skills in either the source or the target language makes it difficult for users to determine when the model is wrong and recover from the error.

MT can help speakers of minority languages within a given society communicate with others and access resources. For example, 9% of people living in the United States have limited English proficiency [2], which can make it more difficult for them to access critical resources including housing [25], employment [2], and healthcare [66]. While MT has the potential to help, unexpected and undetected errors can cause confusion, frustration, and embarrassment [37]. When the stakes of an interaction are higher, the consequences can be far worse; instances have been recorded when MT systems produced harmful or threatening language from benign source inputs and vice versa, with grave consequences when used by police or content moderators [7, 58, 62]. For speakers of low-resource languages, these problems stand to be more frequent due to weaker machine translation support [34].

Conversational communication is an important use case for MT [37], and casual text-based communication mediated by MT is only likely to become more widespread as MT features are embedded into messaging apps[1] and social media sites.[2] Prior research has shown that people can have successful conversations across languages using imperfect MT [29, 69], but it is unclear why users are able to identify and recover from some errors, while they are misled by

[1]Examples include Microsoft Translator's integration into SwiftKeys (https://web.archive.org/web/20210329164205/https://support.swiftkey.com/hc/en-us/articles/360001314546-How-to-use-Microsoft-Translator-with-your-Microsoft-SwiftKey-Keyboard) and the Google Pixel 6 Live Translate feature (https://web.archive.org/web/20211021223350/https://www.xda-developers.com/pixel-6-live-translate-messages-captions/)

[2]Facebook (https://www.facebook.com/help/509936952489634) and Twitter (https://help.twitter.com/en/using-twitter/translate-tweets) offer users the option to translate content using MT.

others. As MT systems improve and produce increasingly fluent translations, it is especially important to understand when users are likely to be misled and how systems might intervene to promote reliable use of MT.

To this end, we collected 19 MT-mediated dyadic text conversations and in-depth debrief interviews. During the conversations, participants role played high-stakes employment and housing scenarios. In each conversation pair, one participant wrote in English, while the other participant, who was bilingual with English, wrote in one of Spanish, Persian/Farsi, Igbo, or Filipino/Tagalog. During the conversation, participants annotated confusing translations, and in the debrief interview we showed participants their conversation transcript along with source messages and their machine translations. This allowed us to document not only users' perception of conversation quality as it unfolded, but also to identify instances of misunderstanding and unnoticed miscommunication. We conducted qualitative analysis of the conversation and interview transcripts to understand user challenges, strategies for recovery, and the kinds of translation errors that proved more or less difficult for users to overcome.

Our findings show that users have difficulty identifying translation errors, particularly when translations are fluent and might reasonably make sense in context. As a result, several participants were unaware they had misunderstood parts of their conversation until the debrief interview. Uncaught mistranslations have the potential for serious harm in real world contexts; for example, if critical information is unknowingly misunderstood, or if erroneously rude translations are attributed to a person. In addition, participants' strategies for repair often hindered achieving common understanding. Finally, some users tried to avoid translation errors by adjusting their writing style and choices, but this proved difficult to achieve in practice and risked negatively impacting the social dynamics of the conversation.

Drawing from theories of grounding and repair in communication as well as prior work in explainable AI (XAI) and FAccT, we identify promising paths forward to support users in identifying, recovering from, and avoiding translation errors. We highlight opportunities for interface design, model development, and interdisciplinary collaborations that bridge natural language processing (NLP), explainable AI (XAI), and human-computer interaction (HCI).

2 RELATED WORK

Prior work has shown that people face challenges using MT in conversational settings because it is difficult to assess the quality of individual translations and, when users can identify a low quality translation, it is difficult to efficiently repair communication. To provide relevant, helpful, and actionable user support, we need to understand how users assess translation quality, and when it is particularly difficult to identify and recover from translation errors. In this section we first connect our work to the field of explainable AI (XAI) and the broader FAccT community. Then, we review related literature studying MT in conversational settings.

2.1 Transparency, Explainability, and Trust in Machine Learning

In many ML systems it can be difficult for an end-user to discern whether an output is correct or reliable. In the HCI and FAccT communities, scholars have explored trustworthy design for ML models and AI as a whole [60, 63] as well how to calibrate user trust in individual system outputs [14, 35, 72]. Research on explainable AI cites supporting appropriate user trust as a core goal [21, 38, 49]. To use an AI system reliably, a user must consider both their trust in the general design and technical underpinnings of a technology to function as expected, while also considering individual scenarios or outputs that may be more or less reliable for their needs.

In this work, we engage primarily with user evaluations of MT reliability in conversational settings. Investigations of trust calibration in ML systems often focus on contexts in which the user, such as a domain expert, can rely on alternative assessments if trust is questioned, such as their own judgment [59]. Often people use MT because they need to understand a language they do not know and for which no one is available to interpret or translate [37]. Jacovi et al. [33] describe distrust in AI as a mode of mitigating risk, which must be present for trust or distrust to manifest [39], but it is unclear how MT users with limited language abilities assess and mitigate the risk of inadequate translations, or calibrate their trust in MT. For example, a pilot study by Martindale and Carpuat [40] suggests that users' trust in MT was more impacted by encounters with disfluent translations than with inadequate ones. We build on this work by understanding how users assess translation quality, what information they seek when they believe a translation is poor quality, and in what circumstances they are unable to identify poor quality translations. This understanding is key to identifying future directions for human-centered explainable AI [22] for MT.

2.2 Machine Translation-Mediated Communication

As their availability has increased, MT systems have become a convenient option for people who need to communicate across language barriers [37]. Researchers have conducted user studies to understand how MT impacts communication [15, 24, 29, 54, 68, 70], and to evaluate new interface designs for conversational MT systems [23, 43, 52, 67]. In these studies, participants typically engage in text-based or spoken conversations mediated by MT to complete a given task, for example, collaborative storytelling [29], idea brainstorming [23], and simple games designed to force participants to develop shared referring expressions [24, 52, 68, 70]. Some researchers have also studied MT in more realistic settings. For example, Shin et al. observed participants using MT over the course of a four week clinical role-play [54], and Calefato et al. conducted a controlled experiment to study the impacts of MT in software requirements meetings [15].

Researchers have analyzed MT-mediated communication through the lens of grounding theory, which frames communication as a collaborative process of establishing shared understanding [19]. In this model, contributing to a conversation involves both producing an utterance, and verifying that it has been sufficiently understood by the addressee(s) [11]. Yamashita and colleagues showed that it is challenging for people to maintain grounding in MT-mediated

conversation because it is difficult to know what parts of your utterance have been understood by the other person [68, 70]. These challenges are exacerbated by inconsistent and asymmetric translations, which make it difficult to maintain consistent referring expressions or make reference to an earlier part of a conversation [70].

Prior work has also found that as users interact with an MT system, they develop adaptive strategies like simplifying their language, repeating and rephrasing, and guessing the meaning of confusing translations [29, 54]. Even with these adaptations, errors can make communication frustrating, cognitively burdensome, and imprecise [15, 29, 37, 54]. Users may make incorrect guesses, may not know how to rephrase in a way that improves the translation, or they may not even realize a translation error has occurred [69]. These challenges are particularly concerning in higher stakes settings and settings with a power difference between communicators. For instance, Liebling et al. [37] describe the story of a woman who lost a job after migrating to the United States because she did not speak English and her employer found it too difficult to communicate with her via a mobile MT app. Our work builds on prior user studies with a novel focus on understanding *when and why users have difficulty identifying translation errors, and how these difficulties impact communication.* A second goal of this work is to investigate these challenges in high-stakes conversations that reflect real-world power differentials.

3 METHOD

To understand how users identify and recover from errors in MT, we collected 19 MT-mediated text-based conversations in three realistic role play scenarios across four language pairs: English-Spanish (5), English-Farsi (5), English-Tagalog (5), and English-Igbo (4). In this section we describe our user study and our approach to data analysis.

3.1 Participant recruitment

We recruited English-speaking participants and bilingual participants who knew both English and one of Farsi, Tagalog, Igbo, or Spanish through dscout, an online user research platform. Due to anticipated challenges recruiting participants fluent in low-resource languages, we also recruited from an active employee resource group of Farsi speakers at a large technology company. Prospective participants filled out a screener survey that asked for their reading and writing proficiency in English and their other language on a scale from 1 (not well at all) to 5 (very well).[3] We only recruited respondents who rated their reading and writing proficiency in both English and (for bilingual participants) their other language at least 4 out of 5.[4] For bilingual respondents, we asked which dialect of the language they knew, how they learned it, and how they use it in their life (see supplementary material). We used the answers to these questions to verify participants' experience with the language.

All but three bilingual participants rated their reading and writing proficiency in their non-English language a 5 out of 5; the other three rated their reading a 5 and their writing a 4. Five of the bilingual participants rated their English reading a 5 and writing a 4; one rated their English reading a 4 and writing a 3. All participants who used English in the task rated their reading and writing proficiency in English a 5 out of 5. 22 participants were men, 13 were women, and 1 was non-binary. Participants self-reported their race or ethnicity as White (11), Asian (6), Middle Eastern (5), Black or African-American (4), Hispanic or Latinx (4), Iranian (3), Black or African-American and Hispanic or Latinx (1), and South Asian (1); 3 did not report. The median age was 36.5 years, with a range of 22-60.[5] All participants were familiar with machine translation prior to the study, and most were infrequent, casual users. The 10 English-Farsi participants were all full-time employees at a technology company, and 2 of the remaining 28 participants worked in the technology industry. We refer to participants by the study session they participated in (e.g., I1 through I4 for the Igbo-English sessions), as well as the language they wrote and received messages in.

We selected language pairs across a range of low to high-resource languages in terms of NLP training data as well as diverse geographic origins. First, we selected English-Spanish as a high-resource language pair because it is highly relevant to real world use cases in the United States. Next, we curated a list of 32 languages that were supported by Google Translate and were spoken at home by at least 100,000 people in the United States,[6] excluding Western European languages, Chinese, Japanese, and Arabic, which receive relatively high MT research attention. From this list we selected Tagalog, Farsi, and Igbo based on geographic diversity and participant availability. According to Joshi et al.'s six-point (0-5) scale of language resources in NLP, Igbo is classified as a 1 (very little labelled training data available), Tagalog a 3, Farsi a 4, and Spanish a 5 (massive investment in data collection and model development)[7] [34].

3.2 Study procedure

Each study session involved two participants communicating via Google chat. We randomly assigned each pair to one of three role play scenarios developed to reflect realistic use cases for machine translation based on Liebling et al. [37]: a tenant-landlord discussion about building repairs; a cleaner and client discussing workplace safety; and a parent interviewing a prospective nanny (see supplementary material). To further reflect real-world social dynamics, we consistently assigned the role with relatively less social status in the United States (tenant, cleaner, or nanny) to the non-English language, and the role with relatively more social status (landlord, client, or parent) to English. We were conscious of the risk that this choice would reinforce stereotypical associations between immigrants in the U.S. and low wage care professions. However, given our goal to evaluate MT challenges in realistic settings, we decided

[3] A similar scale is used in the American Community Survey English-Ability question https://www.census.gov/content/dam/Census/library/working-papers/2015/demo/SEHSD-WP2015-18.pdf

[4] Note: there was one exception who we recruited before adding an English proficiency question to our recruitment survey. After participating in the study, this person reported their English writing proficiency a 3 in a post-survey, and thus does not meet the inclusion criteria. However, English was their primary language at their job as an engineer in a large technology company.

[5] See supplementary material for details on demographic data collection.

[6] https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

[7] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

that it was important to simulate situations in which people who are marginalized on the basis of their English proficiency may face discrimination. Before the conversation began, participants gave informed consent to participate, and we reminded them that they could end the task at any time. One author was virtually co-present with each of the participants throughout the session, and continuously monitored the chat.

Each scenario required participants to resolve a disagreement and schedule a time to meet. The conversation was mediated by a custom Google Apps Script bot that translated messages using the public Google Translate API.[8] In all but four sessions, the participant using English did not have proficiency in the non-English language, and the other participant had written and reading proficiency in both languages. In four of the Farsi-English sessions both participants knew English and Farsi. Participants could not see or communicate with each other apart from the chat interface, and they saw only their sent messages and the translations of their partner's messages (*not* the untranslated source messages) (Figure 1). Participants were asked to use any emoji to mark messages from their partner that were unclear or confusing. These emojis were not visible to the other person and served only as flags for the debrief interviews and data analysis. The conversation task was complete when the participants agreed on a time, or after approximately 20 minutes. The conversations ranged from 9 to 34 minutes (median 22), and 10 of the 19 pairs completed the task.

After the chat task was complete, we conducted a semi-structured debrief interview with each participant one-on-one over video call. These interviews were conducted in English and lasted 38 minutes on average (range: 16-62 minutes). We asked the participant to reflect on their conversation, discuss anything they found challenging or surprising, and whether they altered anything about the way they read or wrote messages because of the MT. Next, we showed participants a transcript of their conversation with the source text and machine translation of every turn. For each turn they received, we asked bilingual participants whether their understanding of the message had changed after seeing the source message. We also asked them to correct the translations where relevant.[9] For participants who wrote in English, we focused on messages they marked confusing and asked about their strategies to make sense of them. Interviews were recorded with consent and transcribed for analysis.

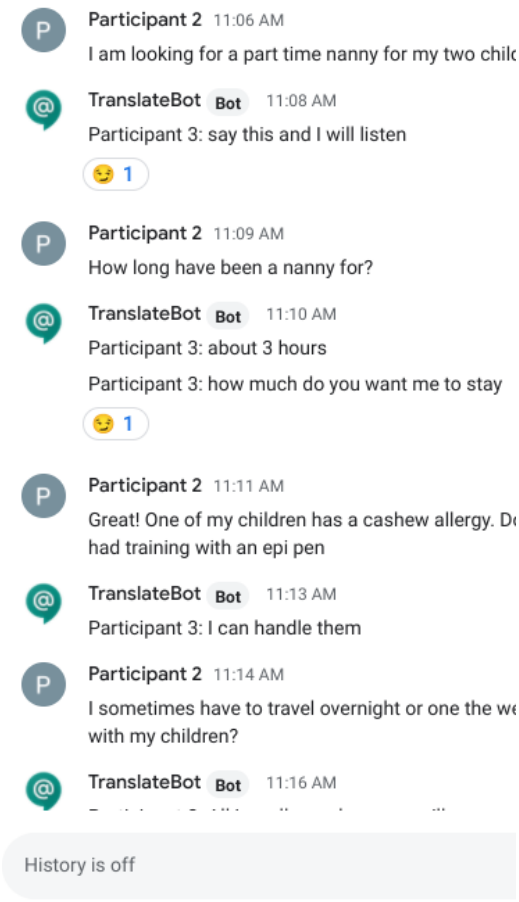

Figure 1: An extract of the user study interface from the perspective of I4, English. Complete image provided in supplementary material.

3.3 Data Analysis

Our dataset contained 19 conversation transcripts with confusion annotations and post-edits, and 38 debrief interview transcripts. First, we conducted inductive qualitative analysis on the interview transcripts [41]. The authors each independently conducted line-by-line open coding [18] on one interview transcript from each language pair. We then compared and discussed our codes. Next, we repeated this process on another set of one interview per language pair. At this point, we converged on a tentative code book, containing seven high-level codes including "error attribution," "uncaught mistranslation," "criteria for assessing quality," "error types," and "confusion strategies," each with up to seven subcodes describing specific examples (e.g., "confusion strategies: ignore confusing part" and "criteria for assessing quality: effort.") We then split the remaining interview transcripts and each continued this coding process on half of the data. We were in regular communication throughout this process, adding codes as necessary and resolving instances where we were unsure about our coding.

We then identified two phenomena of interest. First, we noticed that there were several instances where a participant thought they knew what their partner was trying to say, but had in fact misinterpreted an incorrect translation ("uncaught mistranslation"). Second, we noted several strategies that participants used when they were not sure about the meaning of a translation. We conducted deductive coding of the conversation transcripts. We read each conversation transcript in parallel with the associated interview transcripts and coded each conversation turn for: (a) recipient's

[8]We accessed the API via the Language service for Apps Script (https://developers.google.com/apps-script/reference/language/language-app) between July 30, 2021 and September 30, 2021.

[9]This process is called post-editing in the machine translation literature and is often a part of professional translation workflows [26]. One way to evaluate MT is to compare an MT-generated translation to a post-edited version using a string distance metric, e.g. Translation Error Rate (TER) [6].

understanding of the intended meaning (total / partial / none / bilingual), where "partial" referred to turns that the participant indicated they understood a portion but not all of a message, and "bilingual" referred to turns where the person explicitly relied on their knowledge of the other language to understand a literal translation; (b) response action (ignore / repeat / simplify / add detail / generalize / clarify / guess); and (c) uncaught mistranslation (yes / no). This coding process was non-exhaustive, because we relied heavily on what users verbalized in the debrief interview. We used these codes to connect our analysis across the interview transcripts and the conversation transcripts, and report counts of these codes as *rough* estimates of the frequency of different phenomena in the data. Although the method is non-exhaustive, triangulating participant behavior using both interview and chat transcripts allowed us to make sense of participant responses without interrupting the flow of live conversation.

The 19 conversations featured 628 turns and 938 sentences. In total, 228 turns were either marked confusing (in situ with an emoji) or post-edited. Including turns that we coded as "none" or "partial" understanding, or "uncaught mistranslation," there were a total of 236 turns that caused miscommunication or were post-edited. Throughout the paper, when referring to or reporting the intended meaning of messages in Spanish, Farsi, Tagalog, or Igbo, we use participants' post-edits, i.e., their own translations of their messages to English.

4 RESULTS

In this section we present our findings regarding how users identified errors, when they were unable to confidently identify issues, and how these challenges shaped conversations. In the following section we discuss the implications of these findings for the design of MT systems.

4.1 Fluency and dialogue flow are used as a (misleading) proxy for adequacy

Participants used the fluency of the translations they received and the logical flow of the dialogue as a proxy to judge translation quality. While fluency, dialogue flow, and adequacy were often correlated, this was not always true, leading to unidentified mistranslations.

Participants often referred to the fluency or flow of a conversation as evidence that the translations must have been accurate. Participants described conversations as *"smooth, easy"* (S2, Spanish), *"flowing"* (I4, Igbo), and *"pretty straightforward,"* (S1, English). Several participants felt very confident that the translations were accurate, even without verifying that the other person felt similarly or receiving any additional information about the quality of the translations. As S5, English put it, *"I would say like a hundred percent of the time. I was able to understand everything that the person was trying to say."*

However, this perception did not always align between conversation partners. The Igbo speaker in I1 indicated the conversation was *"smooth, there was no confusion,"* while his partner, who was using English, complained that she *"would be super frustrated"* if it had been a real life conversation. As shown in Figure 2, there was asymmetry in how often participants were confused by translations between the two directions of a language pair. This sometimes hindered repair if the participant receiving higher quality translations was confused by clarifying questions, not realizing how poorly their own messages were translated. This may reflect not only underlying asymmetries in quality, but also asymmetries in users' tolerance for errors. In the debrief interviews, Igbo and Farsi speakers, in particular, showed a higher tolerance for errors, informed by negative past MT experiences in their languages.

In some circumstances, participants were misled by messages that were both fluent and seemed to make sense in context, despite conveying the wrong meaning. In I4, The parent-participant explained to the nanny-participant that: *"One of my children has a cashew allergy. Do you have experience with taking care of children with allergies? If so have you had training with an epi pen,"* but the Igbo speaker interpreted the translation of this message to mean that the child was stubborn, and replied, *"ewerem ike ijikwa ha* [I can handle them]" The parent-participant took this as confirmation that the nanny-participant could handle a severe allergy and moved forward with the conversation. At the start of the debrief interview, the nanny-participant said, *"I understood everything,"* (I4, Igbo) but after seeing the conversation transcript realized he had not.

In the real world, a parent might make more effort to be certain that a prospective caretaker has fully understood their child's medical needs. Nevertheless, this is a powerful example of how harm could arise from translation errors that users are not able to identify. In another example T1, English received an untranslated sentence, *"Sige po* [Okay]," but the rest of the message was translated fluently, leading him to believe that it might refer to some kind of generational slang he was unfamiliar with. Our data indicates that fluent but inadequate translations are a particularly risky type of mistranslation, offering additional qualitative evidence in support of Martindale and Carpuat's findings that fluency has a greater impact than adequacy on people's perceptions of translation quality [40].

4.2 Difficulty attributing errors has social consequences

Participants were more skeptical of messages that seemed disfluent or out of place, but still lacked information to identify whether poor quality machine translation was to blame or if the other participant had said something they perceived to be odd or inappropriate. This uncertainty made it more difficult for participants to decide how best to resolve confusion and risks negative social consequences.

Issues with tone and formality were frequent across language pairs. One participant described messages he received as *"abrupt," "rude," "demanding,"* and even *"flirtatious"* (S6, Spanish). Other participants noticed instances where the other person seemed to be *"blaming"* (T5, Tagalog), or *"not respectful"* (T1, Tagalog). Recipients of these messages seemed to struggle to separate their judgment of the translation from their judgment of the other person.

The English-speaker in one Igbo-English session easily attributed errors to the machine translation when words weren't translated or when the English did not make any sense to him (e.g., when *"onwe ihemcho igwagi* [I have something to tell you]" (I3, Igbo) was translated to *"self-interest to talk"*).

> *"If [...] the words were English, but there were a few non-English words, then I assumed that the other person*

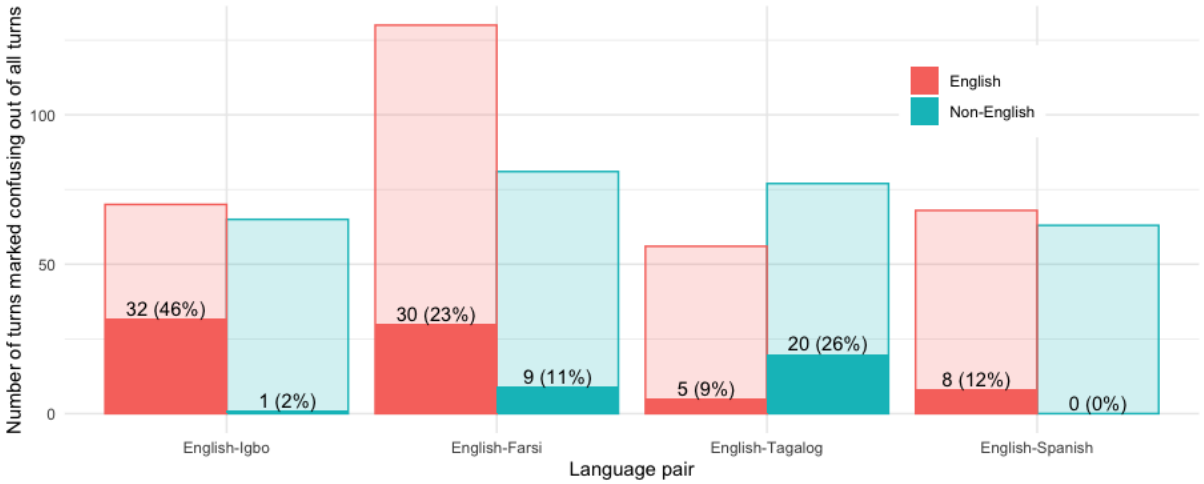

Figure 2: Participants using English were confused more than their partners when speaking with an Igbo, Farsi, or Spanish speaker, with Igbo to English translations annotated for confusion at the highest rate. In English-Tagalog sessions, Tagalog-speakers marked a higher proportion of messages they received as confusing.

> *typed a legible, totally legit [message] and that the translator had for some reason not worked on a few of those words."* (I3, English)

However, when more nuanced translation errors occurred later in the conversation, the same participant assumed the message reflected the other person's intent. The Igbo-speaker, playing a tenant, wrote *"biko ke mbe i ga kpota mmadu idozie uko ulo a* [Please, when can you get someone to fix this ceiling?]," but it translated to, *"Please hurry up and get someone to fix the ceiling."* In the debrief interview, I3, English expressed surprise by this wording, *"The please hurry up was, that was like, I don't expect that as a landlord, I guess."*

T1, English also found it difficult to distinguish MT errors. In reference to one translation he initially said, *"see, that's confusing to me, but I'm chalking that up to the person writing it rather than the translator,"* but later admitted that *"I'm not sure if that was the translator or the person writing it."* In I1, a straightforward clarifying question was completely mistranslated and was interpreted as rude. *"I felt like the person was like, you know, this person was tired of talking with me they just wanted me to go away"* (I1, Igbo).

These difficulties influence the social dynamic both by shaping people's perceptions of others, but also by shaping how people communicate themselves. People tend to mirror the language of the person (or agent) they are interacting with [10, 12, 28], and we saw that this remained the case even if participants weren't sure whether they were mirroring their partner or the MT.

> *"Now that I'm looking at it in English, it looks like it [the other person's messages] would just be as if I was chatting with a friend, but I guess a translator makes it formal, so I was responding more formal based on how [the messages] were translating."* (S2, Spanish)

Ultimately, participants struggled to distinguish MT errors from genuine interpersonal miscommunications, with potentially negative consequences for the conversation and interpersonal dynamic.

4.3 Guessing and ignoring errors can widen the understanding gap

When users were able to identify errors, they then had to determine how to move forward. Consistent with prior work [29, 67, 69], the most common strategy was to ignore parts they couldn't understand (Table 1). Guessing the meaning was also common. When participants ignored errors, they either responded to the parts they did understand, crafted a more general response, or changed the subject altogether. While these strategies frequently kept the conversation going, they sometimes created a false perception of mutual understanding.

If a participant believed they understood enough of a message to formulate an appropriate response, they often chose to ignore the part they did not understand.

> *"Maybe half of the sentences that translations were not correct, but because [the other person] for a few times said a few sentences and next to each other, I was able to understand what he means by understanding at least one or two of them."* (F1, English)

Participants also ignored mistranslations when they felt that the information was not critical to completing the task at hand, seeking *enough* understanding rather than perfect understanding.

> *"I had to [ignore an untranslated part] because I looked at the bigger picture and I said, okay. Well, whatever [the other participant] said there was kind of not relevant to what I was trying to solve."* (T1, English)

When messages seemed slightly off, participants were able to (often subconsciously) make meaning by guessing related terms that would allow a clearer interpretation of the meaning.

Strategy	N turns
Ignore	60
Guess	21
Clarify	20
Generalize	3
Repeat	3
Simplify	2

Table 1: Ignoring confusing parts of a message was the most popular strategy, followed by guessing the meaning or asking for clarification.

> *"When I go back and look at them, I think the fact that it was basically one word, that made me think that it was either typo or a translation thing, like the one where they said, "I don't have a CPR, but if you can pay for the training, you can take a class." Well, I just swapped out "you" for "I," meaning them, and that's what I figured was happening there."* (S3, English)

After interpreting a confusing message, participants generally responded as best they could to continue the conversation. Some were able to formulate a relevant response even with very little understanding of what their partner had said.

> *"[The translation] basically didn't make any sense. [...] That's why I just kind of went and answered with something in general. Like, "When can we start this?" [...] because I wasn't sure like this, it's telling me that their roofs are not going to be able to be used on Sundays and I'm like, I don't I don't know what to do with this information."* (T5, Tagalog)

If a response was sufficiently relevant, the sender of the original confusing message often accepted this as evidence they had been understood [20], not realizing that the other person may have made an incorrect guess or even ignored part of their message altogether. For example, I2, Igbo sent a message that said *"enwere m oge abalị* [Have a good time]," but was translated to *"I have a night time."* The recipient interpreted the message incorrectly, explaining, *"He's okay with nights. That's what that means to me. I don't think that's terribly off,"* (I2, English) and moved on. This misunderstanding was never caught and the Igbo speaker's well wishes were never received.

An accumulation of partially understood messages and vague responses made it difficult to have a specific conversation. For example, two participants playing the landlord role (I3, English and F2, English) understood that the tenant-participant had a leak, but found it difficult to ascertain its seriousness.

4.4 Avoiding errors is difficult even with conscious effort

While many participants had theories about how to produce the clearest translations, it was difficult to control translation quality in practice. First, it was difficult for participants to know which strategies worked. Moreover, even strategies that work in one case may fail in another, and can be difficult to maintain throughout a conversation.

The most common beliefs about MT among the participants were that it performs poorly on long sentences and complex sentence structures, that it is sensitive to spelling and grammar, and that it often translates idioms and metaphorical language literally, failing to convey their meaning. While these theories were largely consistent with the limitations of MT models, putting them into practice proved difficult. For example, two participants mentioned a trade-off between keeping messages simple and including sufficient detail.

> *"I was like, should I just like, you know, just give one word [response]? Like "no?" [...O]r should I say "no experience?" [...] I don't really know how I'll be able to respond to it for the person to understand."* (I2, Igbo)

More broadly, the MT model's limitations were at odds with realistic features of casual language. In addition to occasional typos, users rarely used complete punctuation, and the system frequently failed to translate messages at all when users omitted diacritics.[10] Participants also frequently used idiomatic and metaphorical language, even when they had intentionally tried to avoid it, leading to strange literal translations, e.g., "standing water" was translated into Farsi using the word to describe a person standing up.

An issue with attempts to simplify language is that it risks shaping and constraining human communication around the limitations of existing machine translation tools. For example, the English participant in F1 tried rephrasing a message several times in an effort to improve the translation. However, this process changed the tone of his message.

> *"So I repeated the same sentence a few times. [...] And every time I made it simpler and simpler because he wasn't understanding that I need the building cleaned by tonight. So for example, "I need this building cleaned by tomorrow," and very direct way of saying things, which I usually don't say. For example, if you were a real cleaner, I would say, "can you please clean the building for me," but when I was doing [the task], I told him, "I want this building cleaned by tomorrow." So I gave him very direct orders."* (F1, English)

Although the translation may have eventually conveyed the core meaning of the source text, the MT is indirectly shaping the interpersonal dynamic in an undesired way.

5 DISCUSSION

Two key principles of good user interface design are to prevent errors where possible, and when errors do occur, help users quickly recognize, diagnose, and recover from them [47]. Our findings identify important challenges that users face in identifying, recovering from, and preventing miscommunication due to translation errors in MT-mediated chat. We contextualize these findings in existing theories of computer-mediated communication and human-AI interaction to identify next steps for MT model development and user interface design that could improve the user experience in each of these areas. We end with a discussion of how systems could

[10] This was particularly a problem for Igbo-English because several of the Igbo-speaking participants did not know how to access diacritics on their computer's keyboard.

adapt support to different contexts to provide relevant and useful information without becoming intrusive.

5.1 Identifying and recovering from errors

It is difficult for MT users to identify translation errors without knowing both the source and target languages. Our findings echo concerns that this is especially difficult with current state of the art neural machine translation systems, which can produce very fluent translations that are not necessarily adequate [4, 40]. One risk of language models that produce seemingly fluent and coherent output is that people are inherently driven to make meaning from such outputs, regardless of how they were produced or whether they reflect any meaning or intent [4]. In this study, participants had confidence in their interpretation of apparently fluent and coherent translations, even when their interpretation did not match their partner's intent. Thus, there is a need for novel approaches that interrupt this process and help users identify and recover from translation errors.

5.1.1 Make MT more visible. Participants found it difficult to identify MT errors and frequently attributed system errors to their conversation partner. In real world scenarios, translations that are erroneously offensive or rude translations, or that fail to convey well wishes could alter how users are able to present themselves and jeopardize interpersonal relationships, with potentially serious consequences in cases where users are seeking employment or assistance [28]. MT-mediated communication has historically been designed to feel seamless and as close as possible to a chat with someone speaking the same language [52]. However, this seamlessness may actually make it more difficult for users to identify and attribute errors, and easier for them to forget that MT is in use. A 2014 study by Gao et al. found that users attributed errors *less* to their conversation partner when they believed the conversation was mediated by MT, compared to when they believed they were speaking to someone for whom English was a second language [24]. As MT becomes more fluent, it becomes less salient to users, possibly making them more likely to attribute errors to their partner even when they are initially aware of MT.

Future work could investigate how designs that make MT more visible (see, e.g., seamful design [17, 31]) or adopt alternative metaphors for MT, such as that of an agent or interpreter [52], could heighten users' awareness of MT, help users identify errors, and reduce their tendency to attribute MT errors to their conversation partner. This is aligned with approaches to explainable AI that seek to encourage more deliberate and critical thinking about model predictions before making a decision about whether to rely on them [13]. At the same time, increased visibility may not always be appropriate or desired. One challenge will be designing tools that help users rely on predictions appropriately without adding frustration. Making MT more visible may limit users who want to retain control over how and whether they share aspects of their identity, including their language abilities, which may be associated with stigmatized social categories [3] and which can be the basis for linguistic discrimination (e.g., [27]).

5.1.2 Warn users when errors occur. Helping users identify incorrect predictions is a challenge across machine learning domains [21, 38]. In MT, there has been sustained effort to develop quality estimation (QE) models, which predict the quality of a translation without comparison to a reference translation and could thus be used to warn users of low quality translations in real time [9, 16, 56, 57]. This prediction task has proved difficult, and it is not clear what kind of quality indicators (prediction targets) would be both feasible to predict and helpful and actionable for end-users [56]. One study by Miyabe and Yoshino suggests that it is difficult for users to apply numeric quality indicators to repair translation errors, particularly if those quality indicators could, themselves, be inaccurate [44]. One direction for future work is to focus QE and other translation-level information interventions on specific kinds of errors that are particularly difficult for users to identify. For example, our findings suggest that supporting users to identify fluent but inadequate translations, as well as errors that change the tone of a message should be a high priority for conversational MT. The dominant approach to QE has been supervised learning, which requires expensive labelled training data, favoring high-resource languages. Given that people using MT with low-resource languages are those most in need of support, QE methods that are effective for low-resource languages will be especially critical.

5.1.3 Support collaborative repair. Theories of repair in communication suggest that people prefer to identify and correct errors in their own messages before sending them, avoiding the need to expend collaborative effort on repair [50]. Prior research has proposed interfaces that encourage self-repair, for example, by showing users the back-translation of their message [42, 53, 68], or suggesting changes to improve the translation [43, 45, 52], but even with support this process is challenging for users who do not speak the target language. Repair costs are shaped by the medium of communication [19]; in MT-mediated communication, users' preference for self-repair may be much weaker because they are forced to guess if their self-repair attempt is likely to be successful. Future work could examine lower cost mechanisms for engaging in collaborative repair. For example, Hu et al. developed a system that allows two monolingual people who each know a different language to collaboratively produce high quality translations from one language to the other using MT [30]. Another possibility is to develop interactions that support repair without relying on MT. In this study, participants annotated messages with emojis to indicate to the research team that they found a message confusing, but the other participant could not see those annotations. One participant often received more confusing translations than the other (Fig. 2), but it was difficult to communicate that to the other person. Offering a specific annotation that both participants can see to indicate that an entire translation, or a portion of it, is unclear could enable lower cost repair activities. Encouraging collaborative repair could avoid disproportionately burdening one person in a conversation with identifying and resolving misunderstandings. Another possibility is to offer standard clarification utterances that have been professionally translated. Such phrases could ease the difficulty of communicating specific issues such as pointing out untranslated words or asking for a statement to be reworded.

5.2 Preventing errors

It is important to be able to identify and recover from errors when they happen, but it is even better when users can prevent those errors from happening in the first place. Existing MT systems offer little insight into model performance, despite widely varying performance across language pairs, and even directions within a language pair [46, 71]. Although MT developers are aware of systematic weaknesses (e.g., [5, 32, 51, 61]), this information is not conveyed to end-users. Instead, users must develop their own theories about MT's strengths and weaknesses through interacting with these systems over time. Theories based on interactions with MT in a particular language pair at a particular time may be misleading when applied to a different language pair or after updates to the model. Moreover, not acknowledging disparities in performance between languages with large investments and those with less support reinforces an expectation that speakers of lower-resource languages should accept poorer performance. Greater transparency into model performance across language pairs and on specific types of language could help users better adjust their expectations, calibrate their trust in the system, and learn how to minimize the risk of translation errors.

One path forward is to develop onboarding materials for new users to teach them about the system's capabilities and known failure modes [1, 14, 65]. While lengthy instructions may not be feasible across all use cases, visual and contextual indicators or warnings could be a first step toward onboarding nudges. Further research is needed to identify what would help users understand what the system can and cannot do, and then apply that understanding to avoid harmful translation errors. This engagement must be ongoing; when the MT model is updated and improved, users should be kept up to date with specific guidance about how to update their strategies for reliable use [1].

Systems could also use interactive teaching strategies [65] and provide reminders when a user tries to use the system in a way that is not supported. For example, human communication is rarely fluent and free of errors [11], but MT systems perform poorly on text with typos, abbreviations, grammatical errors, and other normal features of casual language. While telling users about this limitation upfront would be useful, even users who are explicitly aware of these limitations struggle to abide by them consistently, especially in text messaging where casual language is broadly accepted and expected. As MT models are integrated into messaging apps and social media, a priority should be to ensure they are robust to casual language. A complementary approach could be to interactively assist users to write in a way that is suited to current MT capabilities, from interventions that are straightforward with existing technology like spelling and grammar correction, to more sophisticated interventions like detecting and suggesting alternatives for idiomatic and metaphorical language. Certainly, such an approach would constrain how people are able to communicate when using MT. Over time, this could lead to changes in language use driven by the arbitrary constraints of MT models, especially if inputs to MT systems are then used as training data for future systems. However, with careful attention to this dynamic we can help users work within the limitations of existing MT systems, while simultaneously expanding system capabilities to reduce those constraints in the future.

5.3 Adapting support to the context of use

A consistent finding across study sessions was that people are tolerant of translation errors and can have successful conversations without perfect MT. Because this study involved a role play, participants may have been more accepting of partial understanding than they would be in real life. This is consistent with the idea in grounding theory that people's *grounding criterion*, or how much evidence a person needs that the other person has understood them before they move on with the conversation, changes not only with the medium of communication, but also with the purpose [19]. In different situations, we would expect users to hold MT to a different standard and adopt different strategies for assessing translation quality.

Accordingly, when designing and evaluating MT models and user interfaces, we should be accounting for the purpose of communication and evaluating how well the system serves that purpose. Our findings and the next steps we have proposed above, for example, are specific to MT-mediated text chat with a clear task or goal. Translation systems that adequately serve this purpose may be less effective for conversations with open-ended or creative goals, such as story-telling or getting to know someone, or communicating information that needs to be understood verbatim. Translation systems that use other modalities, such as speech-to-speech translation, also introduce different challenges. Hara and Iqbal [29] found that people using MT over video call also face challenges identifying and recovering from errors, and that users employ similar strategies to recover, like simplifying their language. However, the most useful interventions to improve communication may differ. For instance, visual and audio cues may make it easier for a user to identify misunderstanding, while text may be more conducive to identifying and correcting specific errors [29].

Users' purposes for MT can also shift over time, or even within a conversation. For example, a conversation between a parent and a prospective caretaker could easily shift between friendly chat and building rapport, to sharing critical information about a child's health condition. In high-stakes discussions, such as discussing allergies, people's tolerance for errors may be very low. In fact, studies of MT use in healthcare have found that patients and healthcare providers prefer phrase-based translation tools over open-ended MT because they are more reliable [48, 55, 64]. One path forward could be considering ways to smoothly integrate different types of translation support to match users' relative need for accuracy and flexibility in different contexts. Ideally, MT systems would be designed for flexible use, offering more or less intrusive support based on the context and stakes.

On the other hand, given that users' criteria for assessing translation quality shift according to the context, users' perceptions of translation quality may be an inconsistent proxy for their actual understanding. Several prior studies that introduce new interface designs for MT rely on users' *perceptions* of clarity to evaluate the new system, but do not compare those perceptions against other measurements of translation quality (e.g., professional translators' evaluations) to determine whether users actually understood the

intended meaning. This makes it difficult to know whether these designs improved users' *actual* understanding and quality of communication, or whether they only improve *perceived* understanding and quality. By engaging bilingual participants in debrief interviews with the full conversation transcript and translations, we were able to compare users' perceptions of quality in situ to their understanding of their partner's intended meaning. The fact that we saw several instances of uncaught mistranslations, where a participant thought they understood a message until they saw the original source text, suggests a need to consider this gap more explicitly in future evaluations of systems designed to improve understanding in MT-mediated communication.

5.4 Limitations & Opportunities for future work

We faced several trade-offs in designing the study to resemble real-world high-stakes communication while remaining feasible. Here we identify drawbacks of our approach and discuss how they could be addressed in future work.

We recruited bilingual participants for two reasons: first, bilingual participants were able to compare the source messages and translations in the debrief interview, offering us insight into the difference between in situ *perceived* quality and *actual* quality of MT-mediated communication; second, it allowed us to conduct recruitment and debrief interviews in English. However, users in the real world are unlikely to be using MT to translate between languages they are fluent in, making our set-up less realistic. Further, bilingual participants were sometimes able to infer meaning from poor quality translations that would be difficult for someone who does not speak the source language to understand. For example, idiomatic or metaphorical language translated literally may be intelligible to a bilingual person because of an ability to backtranslate. We partially addressed this limitation in the Spanish, Igbo, and Tagalog sessions by having only one bilingual participant in each pair. Future work could improve on this further by recruiting only participants who have limited or no knowledge of their target language and hiring professional translators to assess translation quality.

We also faced issues with the limitations of the input devices that participants had available. Particularly in the Igbo sessions, some users could not access certain diacritics on their laptop. It is possible that this reflects realistic real-world use, but this is not something that we investigated. Future work could identify what kinds of input devices users might typically have access to when using MT with a specific language and replicate this in user studies.

Finally, the participants knew that they were role playing, so our study only partially replicates realistic high-stakes scenarios and power dynamics. Our insights could be further understood by observing MT-mediated interactions in the real world, for instance, drawing on ethnographic methods [36] or contextual inquiry [8]. Our choice of task prompts, and the choice to put an English speaker in the position of relative social power reflect our context as U.S.-based researchers, and would be complemented by future work in other cultural and linguistic contexts.

6 CONCLUSION

In this work we conducted a user study to explore how users evaluate translation quality and recover from translation errors in MT-mediated text conversations. 19 participant pairs engaged in an MT-mediated role-play conversation modeled after real-world, high-stakes scenarios in English and one of Spanish, Persian/Farsi, Igbo, or Filipino/Tagalog. Through analysis of debrief interviews, chat transcripts, and annotations of confusion provided by participants in situ, we demonstrate that users have difficulty identifying translation errors and validating their own understanding, particularly when translations are fluent, but inadequate. Often these difficulties were asymmetric within conversation pairs and participants were not always aware of their partner's difficulties, at times leading them to attribute MT errors to their partner. We build on existing scholarship in explainable AI (XAI), FAccT, and HCI to identify directions for interdisciplinary research and design to support users in identifying and recovering from MT errors.

ACKNOWLEDGMENTS

We would like to thank the study participants for their time and insights. This work builds on a collaboration with Wesley Hanwen Deng, Niloufar Salehi, Timnit Gebru, Margaret Mitchell, Katherine Heller, Daniel J. Liebling, Michal Lahav, and Samy Bengio. We are also grateful to Romina Stella, Ben Hutchinson, and the anonymous reviewers for helpful discussions and feedback.

Funding: This research was supported by Google.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Jeanne Batalova and Jie Zong. 2016. Language Diversity and English Proficiency in the United States. (nov 2016). Retrieved July 19, 2021 from https://www.migrationpolicy.org/article/language-diversity-and-english-proficiency-united-states-2015

[3] Nicole Baumgarten and Inke Du Bois. 2019. Linguistic discrimination and cultural diversity in social spaces. *Journal of Language and Discrimination* 3, 2 (2019), 85–91.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[5] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 257–267. https://doi.org/10.18653/v1/D16-1025

[6] L. Bentivogli, M. Cettolo, M. Federico, and C. Federmann. 2018. Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *Proceedings of International Conference on Spoken Language Translation (IWSLT '18)*. 62–69.

[7] Yotam Berger. 2017. Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'. *Haaretz* (Oct 2017). https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

[8] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[9] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland, 315–321. https://www.aclweb.org/anthology/C04-1046

[10] Holly P. Branigan, Martin John Pickering, Jamie Pearson, and Janet McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42 (2010), 2355–2368.

[11] Susan E. Brennan. 1998. *The Grounding Problem in Conversations With and Through Computers.* Lawrence Erlbaum, Hillsdale, NJ, 201–225.

[12] Susan E. Brennan and Justina O. Ohaeri. 1994. Effects of Message Style on Users' Attributions toward Agents. In *Conference Companion on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 281–282. https://doi.org/10.1145/259963.260492

[13] Zana Buçinca, 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[15] Fabio Calefato, Filippo Lanubile, Tayana Conte, and Rafael Prikladnicki. 2016. Assessing the impact of real-time machine translation on multilingual meetings in global software projects. *Empirical Software Engineering* 21, 3 (June 2016), 1002–1034. https://doi.org/10.1007/s10664-015-9372-x

[16] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada, 10–51. https://www.aclweb.org/anthology/W12-3102

[17] M. Chalmers, I. MacColl, and M. Bell. 2003. Seamful design: showing the seams in wearable computing. In *2003 IEE Eurowearable*. 11–16. https://doi.org/10.1049/ic:20030140

[18] Kathy Charmaz. 2014. *Constructing grounded theory: A practical guide through qualitative research* (2 ed.). SAGE Publications Ltd, London, United Kingdom.

[19] H. H. Clark and S. E. Brennan. 1991. *Grounding in communication.* American Psychological Association, 127–149. https://doi.org/10.1037/10096-006

[20] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22 (1986), 1–39. Issue 1. https://doi.org/10.1016/0010-0277(86)90010-7

[21] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[22] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466.

[23] Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, Paris, France, 449–458. https://doi.org/10.1145/2470654.2470719

[24] 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 (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1549–1560. https://doi.org/10.1145/2531602.2531702

[25] Edward Golding, Laurie Goodman, and Sarah Strochak. 2018. Is Limited English Proficiency a Barrier to Homeownership? (mar 2018). Retrieved July 19, 2021 from https://www.urban.org/sites/default/files/publication/97436/is_limited_english_proficiency_a_barrier_to_homeownership.pdf

[26] 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*. ACM Press, Honolulu, Hawaii, USA, 177–187. https://doi.org/10.1145/2642918.2647408

[27] Craig Hadley and Crystal Patil. 2009. Perceived discrimination among three groups of refugees resettled in the USA: associations with language, time in the USA, and continent of origin. *Journal of Immigrant and Minority Health* 11, 6 (2009), 505–512.

[28] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (Jan. 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022 _eprint: https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf.

[29] 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 (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 3473–3482. https://doi.org/10.1145/2702123.2702407

[30] Chang Hu, Benjamin B. Bederson, Philip Resnik, and Yakov Kronrod. 2011. MonoTrans2: a new human computation system to support monolingual translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, Vancouver, BC, Canada, 1133–1136. https://doi.org/10.1145/1978942.1979111

[31] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300508

[32] Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2486–2496. https://doi.org/10.18653/v1/D17-1263

[33] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[34] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[35] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.

[36] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 9 - Ethnography. In *Research Methods in Human Computer Interaction (Second Edition)* (second edition ed.), Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser (Eds.). Morgan Kaufmann, Boston, 229–261. https://doi.org/10.1016/B978-0-12-805390-4.00009-1

[37] 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*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376261

[38] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[39] Niklas Luhmann. 1979. Trust: A mechanism for the reduction of social complexity. *Trust and power: Two works by Niklas Luhmann* (1979), 1–103.

[40] Marianna Martindale and Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Association for Machine Translation in the Americas, Boston, MA, USA, 13–25. https://www.aclweb.org/anthology/W18-1803

[41] Sharan B Merriam and Associates. 2002. Introduction to qualitative research. In *Qualitative research in practice: Examples for discussion and analysis*. Jossey-Bass, Hoboken, NJ, USA, 1–17.

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

[43] Mai Miyabe and Takashi Yoshino. 2010. Influence of Detecting Inaccurate Messages in Real-Time Remote Text-Based Communication via Machine Translation. In *Proceedings of the 3rd International Conference on Intercultural Collaboration (ICIC '10)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1841853.1841863

[44] Mai Miyabe and Takashi Yoshino. 2011. Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?. In *Human-Computer Interaction. Interaction Techniques and Environments*, Julie A. Jacko (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–377.

[45] Mai Miyabe, Takashi Yoshino, and Tomohiro Shigenobu. 2008. Effects of Repair Support Agent for Accurate Multilingual Communication. In *PRICAI 2008: Trends in Artificial Intelligence*, Tu-Bao Ho and Zhi-Hua Zhou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1022–1027.

[46] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P.

Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

[47] Jakob Nielsen. 1994. 10 Usability Heuristics for User Interface Design. (April 1994). Retrieved December 19, 2021 from nngroup.com/articles/ten-usability-heuristics/

[48] Anita Panayiotou, Kerry Hwang, Sue Williams, Terence W H Chong, Dina LoGiudice, Betty Haralambous, Xiaoping Lin, Emiliano Zucchi, Monita Mascitti-Meuter, Anita M Y Goh, Emily You, and Frances Batchelor. 2020. The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing* 29, 17-18 (Sep 2020), 3516–3526. https://doi.org/10.1111/jocn.15390

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[50] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. http://www.jstor.org/stable/413107

[51] Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 376–382. https://aclanthology.org/E17-2060

[52] Chunqi Shi, Donghui Lin, and Toru Ishida. 2013. Agent Metaphor for Machine Translation Mediated Communication. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. Association for Computing Machinery, New York, NY, USA, 67–74. https://doi.org/10.1145/2449396.2449407

[53] Tomohiro Shigenobu. 2007. Evaluation and Usability of Back Translation for Intercultural Communication. In *Usability and Internationalization. Global and Local User Interfaces*, Nuray Aykin (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 259–265.

[54] JongHo Shin, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2013. Enabling effective design of multimodal interfaces for speech-to-speech translation system: An empirical study of longitudinal user behaviors over time and user strategies for coping with errors. *Computer Speech & Language* 27, 2 (2013), 554–571. https://doi.org/10.1016/j.csl.2012.02.001 Special Issue on Speech-speech translation.

[55] Hervé Spechbach, Ismahene Sonia Halimi Mallem, Johanna Gerlach, Nikolaos Tsourakis, and Pierrette Bouillon. 2017. Comparison of the quality of two speech translators in emergency settings : A case study with standardized Arabic speaking patients with abdominal pain. In *Proceedings of European Congress of Emergency Medicine (EUSEM 2017)*. Athens, Greece. https://archive-ouverte.unige.ch/unige:100812

[56] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online, 743–764. https://www.aclweb.org/anthology/2020.wmt-1.79

[57] Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics, Belgium, Brussels, 689–709. https://doi.org/10.18653/v1/W18-6451

[58] Steve Stecklow. 2018. Why Facebook is losing the war on hate speech in Myanmar. *Reuters* (Aug 2018). https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

[59] People + AI Research team. 2019. *Explainability + Trust*. https://pair.withgoogle.com/chapter/explainability-trust/

[60] Lauren Thornton, Bran Knowles, and Gordon Blair. 2021. Fifty Shades of Grey: In Praise of a Nuanced Approach Towards Trustworthy Design. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 64–76.

[61] Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 1063–1073. https://aclanthology.org/E17-1100

[62] Yeganeh Torbati. 2019. Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees. *Pro Publica* (September 2019). https://www.propublica.org/article/google-says-google-translate-cant-replace-human-translators-immigration-officials-have-used-it-to-vet-refugees

[63] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[64] Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health Surveill* 5, 1 (Jan 2019). https://doi.org/10.2196/11171

[65] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 448–459. https://doi.org/10.1145/3301275.3302290

[66] Elisabeth Wilson, Alice Hm Chen, Kevin Grumbach, Frances Wang, and Alicia Fernandez. 2005. Effects of limited English proficiency and physician language on health care comprehension. *Journal of General Internal Medicine* 20, 9 (2005), 800–6. https://doi.org/10.1111/j.1525-1497.2005.0174.x

[67] Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving Machine Translation by Showing Two Outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3743–3746. https://doi.org/10.1145/2556288.2557171

[68] Naomi Yamashita, Rieko Inaba, Hideaki Kuzuoka, and Toru Ishida. 2009. Difficulties in Establishing Common Ground in Multiparty Groups Using Machine Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/1518701.1518807

[69] Naomi Yamashita and Toru Ishida. 2006. Automatic Prediction of Misconceptions in Multilingual Computer-Mediated Communication. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06)*. Association for Computing Machinery, New York, NY, USA, 62–69. https://doi.org/10.1145/1111449.1111469

[70] 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*. ACM Press, Banff, Alberta, Canada, 515. https://doi.org/10.1145/1180875.1180955

[71] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148

[72] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

A SUPPLEMENTARY MATERIAL

A.1 Screener survey used to recruit bilingual Spanish and English speakers and English speakers who do not know Spanish (dscout)

(1) Are you comfortable having a written chat conversation in Spanish (i.e. writing and reading messages)?
- Yes
- No → Skip to Q7

(2) What dialect or variety of Spanish do you speak? (e.g. Mexican, Chilean, ...)
- Open ended, up to 140 characters

(3) How did you learn Spanish?
- Open ended, up to 140 characters

(4) What is your experience with using Spanish?
- Open ended, up to 140 characters

(5) How well do you READ in Spanish? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(6) How well do you WRITE in Spanish? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(7) How well do you READ in English? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(8) How well do you WRITE in English? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

A.2 Screener survey used to recruit bilingual speakers of English and a low-resource language (dscout)

(1) Are you comfortable having a written conversation over instant messaging (i.e. writing and reading messages) in one of these languages? If you know more than one of these language, please select the one you know the best or use most frequently.
- I do not know any of these languages → Knocked out
- Albanian
- Amharic
- Armenian
- Bengali
- Croatian
- Gujarati
- Haitian Creole
- Hebrew
- Hindi
- Hmong
- Igbo
- Khmer
- Korean
- Lao
- Malayalam
- Persian
- Punjabi
- Romanian
- Russian
- Serbian
- Swahili
- Tagalog
- Tamil
- Telugu
- Thai
- Turkish
- Ukrainian
- Urdu
- Vietnamese
- Yiddish
- Yoruba
- Zulu

(2) If this language features letters or characters not used in standard English, are you able to set up your computer so you can type in this language for an instant messaging chat?
- Yes
- No
- This language doesn't feature letters or characters not used in standard English.

(3) What is your experience with using this language?
- Open ended, up to 140 chars

(4) What dialect or variety of this language do you speak? (Leave blank if you're not sure)
- Open ended, up to 140 chars

(5) How well do you READ in this language? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(6) How well do you WRITE in this language? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(7) How well do you READ in English? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

(8) How well do you WRITE in English? (1 = Not well at all; 5 = Very well)
- Scale from 1 to 5

A.3 Screener survey used to recruit bilingual speakers of English and a low-resource language (shared internally at a large technology company)

Let us know what languages you speak (other than English).

You can fill this out for up to three languages, with the option to let us know if there are other languages you can read and write in.

During the study you will have a conversation over Google Chat, so please list languages you can TYPE in on one of your devices.

(1) Language (and dialect if applicable)
- Short answer text

(2) How well do you READ this language?
- 1 - Not well at all
- 2
- 3
- 4
- 5 - Very well

(3) How well do you WRITE in this language?
- 1 - Not well at all
- 2
- 3
- 4
- 5 - Very well

(4) Do you have another language to add? *(Shown up to 2 times)*
- Yes → Return to (1)
- No → End.

A.4 Post-session survey used to collect participant demographics (for English-Farsi participants only)

Note: dscout provided demographic information for participants recruited through their platform (including age, gender, education, employment status, job title, race and ethnicity, household income, and industry).

(1) How well do you READ in English?
- 1 - Not well at all
- 2
- 3

- 4
- 5 - Very well

(2) How well do you WRITE in English?
- 1 - Not well at all
- 2
- 3
- 4
- 5 - Very well

(3) Which of these best describes how often you use an automatic translation tool (e.g. Google Translate)?
- Never
- A few times a year
- About once a month
- Multiple times a month
- Multiple times a week
- Every day

(4) What is your age?
- 18-25
- 26-30
- 31-35
- 36-40
- 41-45
- 46-50
- 51-55
- 56-60
- 61-65
- 66-70
- 71-75
- 76-80
- 80+

(5) What is your gender?
- Short answer text

(6) What is your race and/or ethnicity?
- Short answer text

A.5 Screenshot of the study user interface

Figure 3 replicates Figure 1 with the full text shown.

B INSTRUCTIONS FOR PARTICIPANTS

Instructions
Welcome to the study! We are so excited to have you participate today. Before we get started, please read these instructions and let us know if you have any questions.
This study has two parts:

(1) Role play **conversation** over Google chat. (SEE YOUR ROLE ON THE NEXT PAGE)
(2) One-on-one follow-up interview over video call. We will ask you to open and edit a Google doc during the interview.

IMPORTANT During the chat portion, **use emoji reactions (any emoji is fine) to mark messages that you receive whenever you are not sure whether you have understood what your partner is saying.**
TO GET STARTED:
Go to chat.google.com and log in with these credentials:
USERNAME: [*Participant gmail account*]
PASSWORD: [*Password, randomly generated and reset after each session.*]
[*You should start the conversation/ by sending a direct message to TranslateBot./Your partner will start the conversation and you will see their message (translated) in the chat with TranslateBot.*]
Key points:

- Send messages in *[English/Spanish/Farsi/Tagalog/Igbo]* only.
- Stay in character. Do not share any personally identifiable information.
- Mark messages with an emoji reaction if you're not sure you understood what your partner is trying to communicate.
- Your partner's messages may come in slowly sometimes and you will not be able to see when they are typing, so please be patient when waiting for a response to your messages.
- The conversation is over when you schedule a time or after 30 minutes, whichever comes sooner. We will watch the conversation and confirm when you are done.
- If you have any questions, unmute and ask at any time.

YOUR ROLE IS ON THE NEXT PAGE! »>
[page break]
You will be using: *[English/Spanish/Farsi/Tagalog/Igbo]*
Your role: *[Cleaner/Tenant/Nanny/Real estate agent/Landlord/Parent]*
Role description: *[Relevant description (see below)]*
Your availability: (yellow shows times when you are available) [*One of the two calendars in Figure 4*]

B.1 Tenant

You live in a rental apartment that is old and poorly maintained. There's always something broken in your apartment, and it has started really interfering with your ability to work from home. You always let your landlord know when there are problems. They sometimes try to help, but recently they have been too slow to respond and you've had to fix things yourself.
You've just noticed a drip from the ceiling in your bathroom. You've placed a bucket underneath but the paint is starting to sag and you're worried about flooding. **Text your landlord to let them know about the leak and ask for help.** You expect them to organize and pay for all the work; this looks like it could be a big job.
Find a time for someone to come and fix the leak. The calendar below shows your availability:

B.2 Landlord

You are a landlord and you manage a small apartment building. One of your tenants is always texting you complaining about problems in their apartment. The wifi is too slow, or the washing machine is too loud, or the neighbors are smoking. You try your best to be responsive but you feel like sometimes they are too demanding.
You get a text from your tenant - there's another issue. **Find out what is going on and how serious it is.** If they need help, determine who would be the appropriate person to call (e.g. plumber, electrician, roofer). Your task is over when you either: **agree that no help is needed, or arrange a time for someone to come fix**

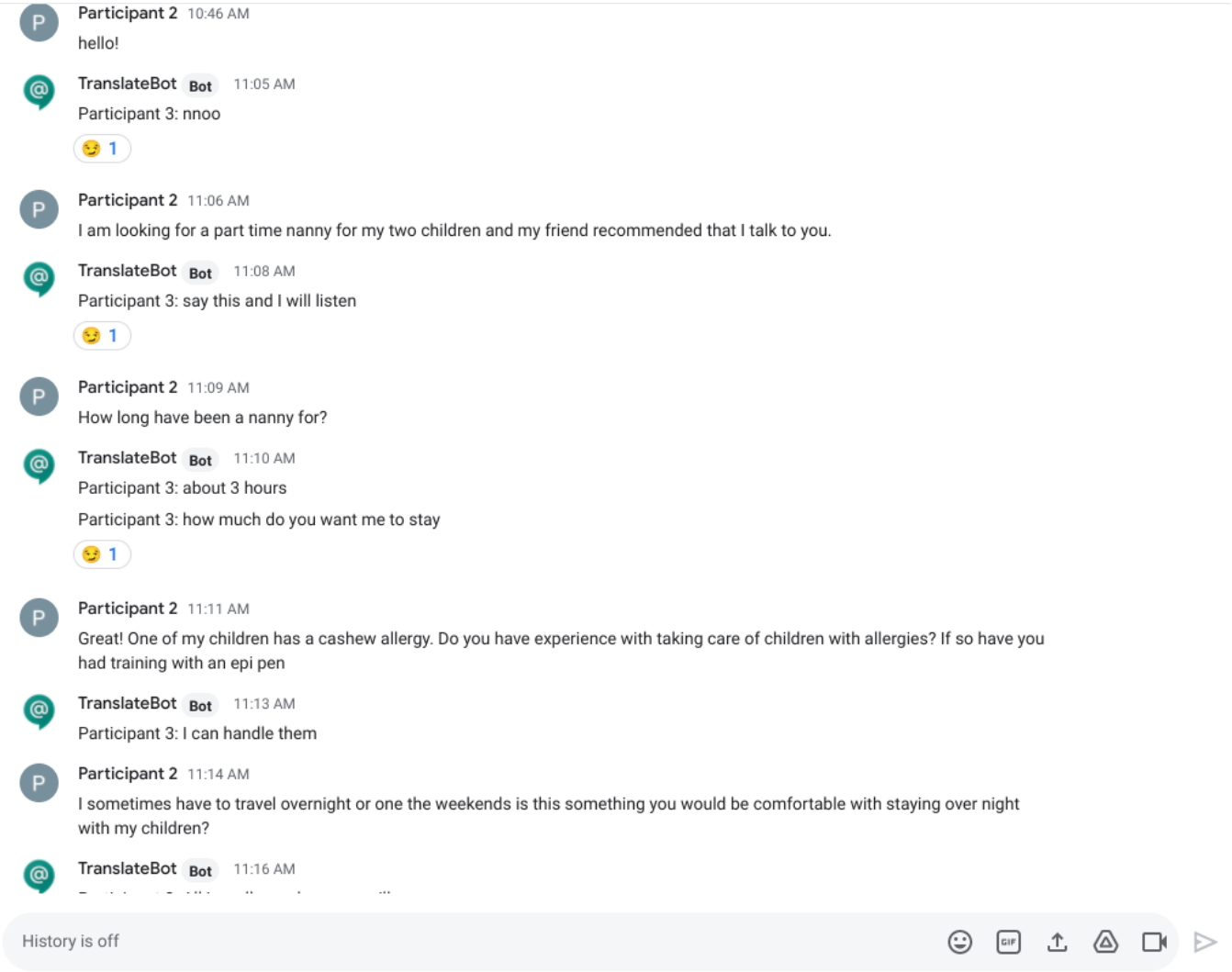

Figure 3: The user study interface from the perspective of the participant using English in session I4.

it. The calendar below shows when you or a plumber[11] is available to visit the apartment:

B.3 Nanny

You are a part-time nanny and you are looking for a new family to care for. A parent texts you wanting to find out more about your experience and availability. **Answer their questions and try to get the job.**
Some facts about you to help you answer the parent's questions:

- You have a flexible schedule, but you balance nannying with another job so you need advance notice before your shifts.
- You have a driver's license but no car.
- You don't have any specialized healthcare training (e.g. CPR), but if the parent pays for it you are happy to do a course.

[11]Participants pointed out that this was an error in the instructions, as it gave away that a plumber was needed. Future work using this protocol should correct this.

(Note: you don't have to get all of this info across, just use it if you need help answering the parent's question.)
If the parent asks any questions you don't know the answer to, feel free to get creative.
If the parent thinks you might be a good fit, they will ask you to come to their house for an interview. The calendar below shows your availability for the week:

B.4 Parent

You are looking to hire a part-time nanny. You have two children, a 6 month old and a 2 year old. A friend recommended someone and you would like to find out about their past nannying experience. Here are a few of your constraints:

- One of the twins has a severe allergy to cashew nuts, so you want to make sure the nanny has training in how to care for kids with allergies and how to use an EpiPen.

Figure 4: Each participant in a pair was shown one of these two calendars to indicate their availability. Each calendar has eight available two hour blocks, but only two overlap.

- You have a large dog, so the nanny needs to be comfortable with dogs.
- You sometimes work late at night or travel on the weekends, so they need to be okay with staying over at your house occasionally.
- Your older child needs to be driven to and picked up from a kids play group twice a week.

Text the nanny to find out whether they meet these needs. Once you have a sense of their experience, arrange a time for them to come to your house. The calendar below shows your availability for the week:

B.5 Cleaner

You work for a cleaning company and you've been assigned to a job. You normally do routine home cleaning, often for nice homes that are about to go on the market for sale.
You show up to find that the house is extremely run down. Worse, the walls are covered in mold in multiple rooms. You weren't warned about biohazards and you didn't bring any special equipment. You have severe asthma, and exposure to mold for long periods could make you very sick. You're nervous to confront the client because if they complain to your employer you could get in trouble and you can't afford to lose your job right now.
Text the client to explain the situation and let them know you cannot clean the house today. Negotiate with the client to find a solution that meets their needs and protects your health. Agree on a time by which the work can be finished. The calendar below shows your availability:

B.6 Real-estate agent

You are a real estate agent and you work for a large agency. You've been assigned to sell an old, run-down building for a very important client that needs to close the sale ASAP. You have a few potential buyers lined up for tomorrow and you've hired a professional cleaning company to come and clean up the place before they arrive. Someone arrives to start cleaning, but they look concerned. You're frustrated because you're under a lot of stress with this sale and the last thing you need is a delay right now. **Discuss and resolve the situation with the cleaner.** Agree on a time by which the work can be finished. The calendar below shows your constraints: