Bridging Fluency Disparity between Native and Nonnative Speakers in Multilingual Multiparty Collaboration Using a Clarification Agent

WEN DUAN[1], Cornell University, USA & NTT, Japan
NAOMI YAMASHITA, NTT, Japan
YOSHINARI SHIRAI, NTT, Japan
SUSAN R. FUSSELL, Cornell University, USA

Multiparty collaboration using a common language is often challenging for nonnative speakers (NNS). Conversation can move forward rapidly, with terms and references unfamiliar to NNS often going unexplained because NNS do not request clarification due to cognitive overload or face concerns. Language difficulties may further lead to NNS having a low level of participation in a conversation, which could be a loss for multilingual teams. To help NNS resolve potential confusions due to unfamiliar language use without risking face concerns, we created a conversation agent that asked clarification questions intended to help NNS follow and participate in multiparty conversations. We conducted a within-subjects laboratory experiment with 17 triads of 2 NS and 1 NNS, who performed a series of collaborative tasks under three conditions: a) no agent, b) a high-level agent that resembles a NNS with good command of English, and c) a low-level agent that resembles a NNS with poor English skills. Results suggest that NS made significantly more clarifications in both agent conditions than without an agent. In the high-level agent condition, NNS reported an increase in understanding after the agent's interruption and spoke significantly more. Further, NNS evaluated their communication competence in English highest in the low-level agent condition and lowest in the control condition. Our findings suggest several directions to improve the tool to better facilitate multilingual multiparty communication.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing** → Empirical studies in collaborative and social computing

KEYWORDS
Multilingual teams, Conversation agent, Nonnative speakers, Computer-mediated communication

ACM Reference format:
Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R. Fussell. 2021. Bridging Fluency Disparity between Native and Nonnative Speakers in Multilingual Multiparty Collaboration Using a Clarification Agent. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, No. CSCW2, Article 435, October 2021. ACM, New York, NY, USA. 33 pages. DOI: https://doi.org/10.1145/3479579.

1 INTRODUCTION

Collaborators across the world often use a common language such as English to communicate, which can create grounding problems due to differences in fluency. For instance, nonnative

2573-0142/2021/October - ART435 $15.00

https://doi.org/10.1145/3479579.

speakers (NNS) may give little feedback when they do not understand native speakers' (NS) messages, leading NS to believe their message has been understood when in reality it has not. In multiparty conversations, especially when NS predominate, the pace of conversation may be faster than NNS can comfortably follow [14] and include idiomatic language and cultural references that are unfamiliar to NNS [66]. The fast pace leaves little room for NNS to pinpoint confusions or formulate clarification requests when needed. Additionally, NNS report feeling embarrassed to interrupt the group conversation [35]. Consequently, NNS may continue to have difficulty understanding and participating in multiparty conversations. This low level of NNS participation may also lead NS to form erroneous beliefs about their NNS partners' attitudes and personality. At the same time, NNS may feel that their NS partners are uncaring and insensitive by dominating the conversation, resulting in a cycle of negative impressions formed about each other that disrupts collaboration [31].

Previous work has explored various tools to address this challenge, such as helping NS make their messages clearer for NNS (e.g. [14][49]) or helping NNS understand NS messages better by providing them with automated transcripts [25] or translations [23]. These approaches have shown the potential of language tools in facilitating NNS understanding. However, most such tools provide support throughout a conversation, rather than at the moment potential confusions arise, creating additional cognitive load. To make potential points of confusion transparent for NS, Gao et al. [26] displayed how NNS were using transcripts and a bilingual dictionary, which did improve task-related communication. However, this approach lacks privacy, and NNS sometimes reported feeling that it was face-threatening to have their confusions made obvious to their NS teammates.

We propose an alternative approach that has the potential to provide NS clarifications when needed by NNS partners without privacy concerns. Our strategy is to transfer the face threat and the formulation effort of requesting clarifications to a conversation agent that asks these clarification questions "on behalf of NNS". We developed a high-level conversation agent that resembles someone with good command of English who interrupted the NS and requested clarification when detecting a difficult word. We intended the agent to set the norm and demonstrate the value of asking questions to NNS, such that NNS might model the question-asking behavior when needed. We sought to explore the trade-offs between a high-level agent whose questions could match NNS confusions but requires sophisticated programming, and a low-cost agent that simply demonstrates to NNS they are welcome to ask any questions (even dumb ones). To do this, we developed another type of conversation agent that resembles someone with poor English skills (low-level agent) who interrupted the NS and requested clarification when detecting an easy word.

Using a laboratory experiment, we examined the effect of the two types of clarification agent on NNS understanding, participation, and their motivation to ask clarification questions through NS response to the agent. Seventeen triads of 2 NS and 1 NNS performed three discussion tasks under each of the three conditions: a) no-agent (control), b) high-level agent, c) low-level agent. Tasks and conditions were counterbalanced across three trials. We found that NS made significantly more clarifications with the agent (regardless of type) than without the agent. NNS reported a significant increase in understanding after the high-level agent's interruption and had higher levels of active participation in the high-level agent condition. We discuss implications of these findings in greater detail and how they can inform the design of agent-supported multilingual multiparty collaboration.

2 RELATED WORK

We start with reviewing the grounding challenges in multiparty collaboration between NS and NNS, then we describe how conversation agent has the potential to address the grounding challenges.

2.1 Grounding Challenges in Multiparty Collaboration between NS and NNS

Grounding is the collaborative process by which conversation partners establish that a message is understood as intended. For a message to be sufficiently grounded, both the speaker and the listener need to coordinate on the process and content [8]. The listener needs to provide positive evidence when grounding has occurred (e.g., a backchannel acknowledgement such as "uh huh", or a relevant next turn) and negative evidence when grounding has not occurred (e.g., a request to repeat or clarify). Depending on whether the evidence is taken by the speaker as understanding, nonunderstanding, or misunderstanding, the speaker might decide to initiate a new utterance, provide clarification or repetition per the listener's request, or repair their message, and further assess the listener's understanding, until both parties mutually believe that the utterance is understood sufficiently well for the current purpose [8][9].

The disparity in language fluency between NS and NNS may create grounding challenges for multilingual communication. For instance, research suggests that NNS give fewer responses in communication with NS [35][58], likely due to difficulty in processing NS messages [56] as well as in formulating their own messages [52][56]. These challenges are often exacerbated in multiparty conversations where NS outnumber NNS, as the discussion can move forward rapidly [14][64], leaving no room for NNS to cut in even if they do want to request clarification. Further, such conversation often involves idiomatic language, higher level vocabulary and cultural references only shared, understood and thus easily grounded among NS [66]. The grounding success between NS can create the illusion that communication is successful among all members when it's not [31][36], leaving NNS further behind. The more NNS participants are left behind, the more difficult it can become for them to pinpoint the source of confusion to request clarification. Furthermore, for reasons such as maintaining face, NNS (especially East Asians) often refrain from interrupting the conversation for clarification [30][36] or to signal NS to slow down [14][51]. In some cases, NNS may even send confusing backchannels such as "yeah", that lead NS to think they are following the conversation when they are not [62].

Researchers in the CSCW community have explored various support tools to facilitate grounding between NS and NNS. To provide NS evidence of NNS status of understanding, Gao and colleagues [25][26] implemented a system to show NS how NNS are using real-time automated transcripts and a bilingual dictionary. As such, NS are made aware of whether their message is understood by NNS, and of the words that are likely to elude them and thus need clarification. This technique made NS more sensitive to understanding problems encountered by NNS and adjust their speech accordingly. However, NS adjustment did not match what the NNS needed. For instance, NS rephrased and clarified words that NNS already looked up in the dictionary. Duan and colleagues [14] developed a Speech Speedometer to encourage NS to speak at a rate slow enough for NNS follow along. But their results suggest that NNS understanding did not significantly improve because NS had difficulty slowing down their speech even when they were motivated to do so, and unknown words remained unknown no matter how slowly a NS spoke. Echenique and colleagues [15] found that providing video and text cues enabled NNS to cross reference and access the common ground already established between NS. However, this

approach comes with the cost of additional workload, which left little room for NNS to participate in real-time conversation. Therefore, a tool that can display NNS status of understanding without imposing extra cognitive load on them would be ideal.

While direct display of NNS understanding is neither possible nor appropriate for reasons such as privacy concerns, an approximate of their potential confusion can nevertheless be utilized to signal to NS that they need to clarify a message. Along this line, Duan and colleagues [15] developed a clarification agent that asked NS members to explain words flagged as unclear by speech recognition accuracy. Unfortunately, the timing of their agent's interruptions and the requested words did not match NNS points of confusion and did not improve NNS understanding or collaboration quality. In our work, we develop an interruption mechanism that relies on a pretested dictionary so that our high-level agent only interrupts and requests clarification when detecting a difficult word that matches NNS potential confusion.

2.2 Agent-Facilitated Group Conversation

A host of studies in HCI and HRI fields have demonstrated that individuals apply human categories such as gender [46], ethnicity [44], and in/out-group membership [43], to computers and non-human agents. People also apply social rules to computer agents, including politeness [45], and reciprocation [22]. There is also evidence that people are attracted to computers that demonstrate similar "personalities" as themselves [46]. In essence, individuals' interaction with computers and conversation agents is fundamentally social [47]. Based on this previous work, we propose that using an intervention agent could be a promising approach for addressing the challenges that previous multilingual tools have faced, namely, providing language aid without imposing additional cognitive load on NNS [15][25] or causing any privacy and face concerns [26].

Conversation agents have been used to support human-human multiparty collaboration by performing social roles (e.g. [34][55]) such as mediating conflicts and facilitating team functioning, primarily in the form of chatbots. For instance, Kim and colleagues [34] implemented a chatbot (GroupfeedBot) into a text-based group chat to facilitate group discussions by managing the discussion time and encouraging members to participate more evenly (i.e. prompting lurkers to speak up). Qualitative feedback from small groups suggested that GroupfeedBot induced efficient discussion and that members perceived the chatbot as a group member. Furthermore, quantitative results from medium-sized groups suggest that GroupfeedBot promoted more equal participation among group members, encouraged opinion diversity and increased the effectiveness of group communication.

The chatbot studies suggest that an agent-based approach is promising for improving conversational dynamics in multilingual groups. However, text-based communication is different from video-mediated communication with respect to media affordances [8], and these differences could affect how well the agent can influence conversational turn-taking [61], the disruptiveness of the agent intervention, and so forth. In addition, the linguistic and cultural backgrounds of group members [56] and the group as a whole [62] could interact with media affordances to shape communication patterns. Since most such chatbots were developed for and evaluated using linguistically homogenous groups, it is unknown how these factors will play out in agent-facilitated group conversations. This suggests a need to further explore the dynamics of agent-facilitated group conversation in video-mediated platforms and in multilingual multicultural contexts, which this study seeks to do.

3 THE CURRENT STUDY

Our review of prior work on multilingual multiparty tools highlights several challenges in designing such tools: to encourage NS to make clarifications that NNS need, to improve NNS understanding without imposing extra cognitive load, and to enable NNS to participate in the ongoing conversation in real time. In this section, we discuss the design rationale behind our clarification agent and how it has the potential to address these challenges.

The agent serves to detect and ask about difficult words in the conversation that match NNS need for clarification. While there are other dimensions of speech (e.g. speech rate) that also affect NNS understanding, "lack of vocabulary" was identified as the most prominent problem hindering NNS comprehension in multilingual communication [5]. Additionally, Duan and colleagues' work [14] suggests that for their Speech Speedometer to work effectively and to match NNS needs, the recommended speech rate has to take into consideration word characteristics, such that difficult words are uttered more slowly than easy words. Drawing on these findings and design insights, we intend for our tool to address word familiarity to NNS while holding other speech dimensions (e.g. speech rate, accent) relatively constant.

We compare two types of agent, one that requests clarification on difficult words using grammatically correct English (high-level), and the other that asks about basic words using grammatically incorrect English (low-level). We seek to explore the trade-offs between the high-level agent whose questions could match NNS confusions but requires sophisticated programming, and the low-level agent that is relatively low-cost. With the low-level agent, we intend for it to set the norm and demonstrate the value of asking questions (even dumb ones), such that NNS might feel welcome to ask any questions.

3.1 NS Clarification

Since people tend to apply social norms of human relationships when interacting with computer agents [22][43][44][45], we suppose that NS will treat clarification questions from an agent in a way similar to how they treat these questions from human conversation partners. In human-human conversations, when the speaker is requested by the listener to clarify a message, the speaker will provide clarification until both parties mutually believe that the message is understood as intended [9][54]. In cross-lingual contexts, Li and colleagues [35][36] demonstrated that more clarification requests led to more pieces of information clarified. These suggest that the more clarification requests in the conversation, the more clarifications there will be. We therefore predict:

H1: NS will provide more clarifications in the high-level and the low-level agent conditions than in the no-agent control condition.

We are curious about:

RQ1: How will the high- and low-level agent affect NS clarifications respectively?

3.2 NNS Understanding

The clarifications NS make in response to an agent may in turn improve NNS understanding even though they did not request the clarification themselves. In fact, studies have observed that in multiparty conversations, people who did not overtly engage in grounding could build shared common ground as a result of another participant's response (e.g., a clarification request) [18]. We therefore predict:

H2: NNS will report higher levels of understanding in the high-level and the low-level agent conditions than in the control condition.

While the high-level agent might ask about words that match NNS need for clarification and thus directly improve NNS understanding, we are curious about how NS responses to the low-level agent and clarifications of easy words will affect NNS understanding. We therefore ask:

RQ2: How will the high- and low-level agent affect NNS levels of understanding?

3.3 NNS Participation

Much research on multilingual communication has shown that understanding can greatly impact NNS ability to contribute and their actual participation in multiparty collaboration (e.g., [31][35][52][56]). For instance, He et al. [31] showed that NNS attribute their low level of participation to language difficulties such as not being able to follow. Understandably, when foreign language processing takes up much of NNS cognitive load, there's not much room left for formulation [56]. If NS response to the agent helps improve NNS understanding, they might also be able to allocate some of their cognitive resources to formulating and expressing their own opinions and increase their participation. We therefore hypothesize:

H3: NNS level of participation will be higher in the two agent conditions than in the control condition.

Because we do not know how the two types of agent will affect NNS understanding, and in turn their participation, we therefore ask:

RQ3: How will the high- and low-level agent affect NNS level of participation?

3.4 Effects of Different Types of Agent on NNS and NS

The fluency level exhibited by the two types of agent might trigger different social psychological processes such as social comparison [21] in NNS. For instance, research has shown that NNS are more motivated to compare themselves with less linguistically competent others [41]. Those who made downward social comparisons with a fictional other who was said to have poorer English skills reported a better self-concept, produced more words, and even asked more questions [7]. Similarly, in Duan and colleagues' clarification agent study NNS reported being less embarrassed to ask their own questions following an agent asking dumb questions, and they felt more confident and relaxed after the agent showing language incompetence [15]. On the other hand, NNS might compare themselves with the high-level agent and view it as a high reference point that they cannot reach. This upward comparison might result in diminished self-evaluation [4][27]. We therefore hypothesize:

H4: NNS will report a more positive self-evaluation in the low-level agent condition followed by control condition, followed by the high-level agent condition.

Low self-evaluation often causes anxiety in communicating in a foreign language [42][50]. Studies have suggested that even advanced-level NNS can experience anxiety and apprehension when communicating in a foreign language [60], and this anxiety is not alleviated in computer-mediated communication [1]. We hypothesize:

H5: NNS will report lower foreign language anxiety in the low-level agent condition followed by control condition followed by high-level agent condition.

Furthermore, different types of agent might have different effect on NNS's motivation to ask questions. Even though we intend for the low-level agent to demonstrate to NNS that they are welcome to ask any (even dumb) questions, and that anecdotal evidence [15] suggests that this

might happen, it is also possible that NNS might not want to model the inappropriate behavior exhibited by an inferior other. Instead, they might compare themselves to the high-level agent who they might find better off than themselves in terms of fluency and courage and be motivated to emulate the better other [4][41][51] (e.g. to ask their own questions). However, it is yet unclear how NNS will perceive different agents and whether/how they will compare themselves to different agents. We therefore ask:

RQ4: How will different types of agent affect NNS motivation to request clarification?

Lastly, answering the agent's questions might impose extra cognitive load for NS. Once their train of thought is interrupted, it may take some cognitive effort for NS to resume where they left off. Additionally, the effort required to clarify on a difficult word for the high-level agent and a basic word for the low-level agent might be different for NS. For NNS, even though we intend for the high-level agent to reduce their workload, there might be a possibility that the reduced workload could be offset by their increased participation. We therefore ask:

RQ5: How will the two types of agent affect NS and NNS workload?

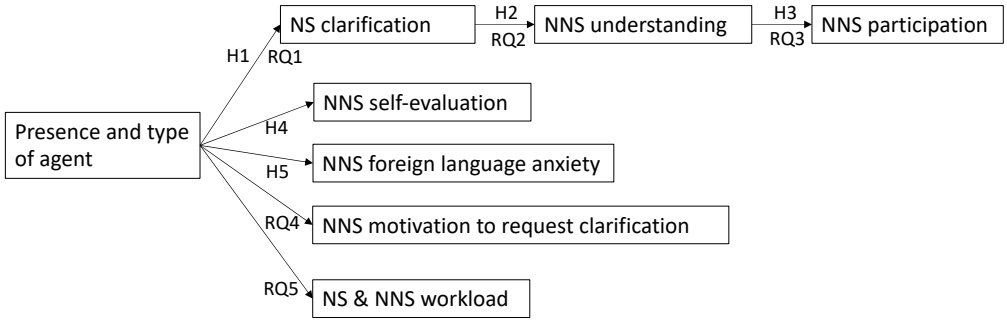

Figure 1. Diagram Mapping the Hypotheses and Research Questions.

Figure 1 summarizes our hypotheses and research questions for better clarity. We hypothesized that the presence of agent will increase NS clarification, which in turn will increase NNS understanding and participation. We are curious about whether and how different types of agent will affect these as well as NNS self-evaluation, foreign language anxiety, motivation to request clarification, and workload for both NS and NNS.

4 METHODS

To test our hypotheses and explore the research questions, we conducted a laboratory experiment in which we manipulated the presence and type of agent using a within-subjects design. We treated each group of participants as our unit of manipulation and had each group undergo all three conditions in random order. We chose a within-subjects design for two reasons: a) comparing the discussions of the same group of participants reduces variance from individual difference such as talkativeness, speech rate, NNS language fluency, etc.; and b) it allows us to elicit participants' comparisons of experiences in different conditions. Seventeen triads consisting of 2 NS and 1 NNS participated in 3 rounds of discussions performing a series of legislative dilemma tasks using a video conferencing tool. For each round, they performed the discussion under a different experiment condition: a) without agent (control), b) with a high-level agent, c) with a low-level agent. The tasks and experiment conditions were counterbalanced to avoid

ordering effects. Note that this is not a full counterbalance (1 group short for a full counterbalance), as we originally recruited 21 groups but ended up having only 17 three-person groups due to one or more participant's cancellation.

4.1 Participants.

A total of 51 participants were recruited through a third-party international personnel agency or social media. Among them, 34 (13 female) were native English speakers who grew up and received their education in an English-speaking country, including the USA, UK, Canada, and Australia. They had lived in Japan for 0.25 to 30 years (M=8.62, SD= 9.17) and had frequent communication with nonnative English speakers in their daily life (M=6.00, SD=1.23, on a scale from 1=never to 7=very often). Their mean age was 36.03 (SD=11.83). The remaining 17 participants (9 female) were native-born Japanese who spoke English as a second language. Among them, 10 had never lived in an English-speaking country, and 7 had lived in one for no longer than 2 years. They reported to have limited interaction with native English speakers in their daily life (M=2.59, SD=1.42, on a scale from 1=never to 7=very often). Their mean age was 22.06 (SD=2.16).

4.2 Materials.

4.2.1 Task

We chose a task for the study based on the following criteria: a) the conversation topic was sophisticated enough that NS would use advanced vocabulary in order to get their opinions across, which simulates many multiparty multilingual collaborations [20]; and b) the task must allow all participants to discuss freely with no assigned roles or prescribed order of speaking.

As a result, we adapted a Legislative Dilemma Task [40] into three tasks, each on a different issue, plus a mini-task for warmup. In the tasks, the participants were asked to imagine that they were UNESCO (United Nations Educational, Scientific and Cultural Organization) representatives who were allocating $1.8 billion in funding among five competing projects. The funding constraints only allowed for supporting two projects, with their first choice fully funded at $1 billion and their second choice partially funded at $0.8 billion. For each trial, participants chose from five (three for the mini-task for warmup) candidate projects on the same issue – culture, health or education. They first made their own choices of two projects, then communicated their individual decision and reasoning with the group, and upon completion of discussion, came to an agreement as a group.

Our creation of the projects was inspired by the UNESCO ongoing projects on their website. We adapted them such that none of the projects were said to be located in Japan or countries where NS participants were likely to come from, in order to reduce potential bias. To foster discussion, we included at least one project likely to cause disagreement between NS and Japanese NNS [39][40] in each set of projects after pilot testing. Sample project candidates included: strengthening minority language publishing in East Africa (culture), helping youth cope with mental health issues arising from traumatic experience of war and humanitarian emergencies (health), and supporting the development of female technical talents in Latin America (education).

4.2.2 Surveys

Participants completed three surveys for the study. First, they were given a *pre-experiment survey* before the warmup, in which they provided demographic information, frequency of communication with NS/NNS, and for NNS in particular, a pre-experiment foreign language

anxiety level. Second, they completed a *post-trial survey* at the end of each trial, which included questions about participants' workload, level of understanding, performance self-evaluation, and willingness to ask questions.

In addition, we used a *retrospective survey* to assess the moment-by-moment status of the participants' conversation experience after all three discussions were finished. In the retrospective survey, the participants watched twelve 10-second video clips from the recordings of their group discussion, four per trial, and answered questions about their level of comprehension. To gain an in-depth understanding of whether the agent interrupted at times when NNS needed and to what extent helped NNS to follow the conversation, we selected the time points based on the following principles: For high-level and low-level agent conditions, we selected 10 seconds before and 10 seconds after the agent's first and last interruption for that trial. For control condition, we selected 10 seconds before and 10 seconds after a high-level keyword was uttered by any NS for the first and last time during the trial, or in other words, where the high-level agent would have interrupted. This gave us four 10-second videoclips per trial/condition, and 12 data points for assessment of their moment-by-moment understanding.

All survey data except the retrospective survey was collected using an online survey tool. The retrospective survey was pen-and-paper. Survey questions were displayed in English with Japanese translation.

4.2.3 Interviews

We developed an open-ended interview protocol that asked participants to compare their communication experience across conditions and their impression about different agents. Specifically, how they thought about the agent's questions, how they decided the way to respond to the agent, how they gauged each other's understanding during the conversation and how they thought of the conversation flow. We conducted the interviews with participants in their native language.

4.3 Software and Equipment.

4.3.1 Speech recognition tool and interruption mechanism

We used real time speech recognition developed by IBM Watson for detecting keywords from the pre-set range. The range was determined using a preliminary study described as follows. To develop a list of words that NNS were likely to be unfamiliar with, we used a word bank from https://www.wordfrequency.info/intro.asp which documents the frequency of use based on the Corpus of Contemporary American English. Each word has a rank index that indicates how frequently it is used in different contexts (e.g., newspaper, spoken English, fiction, etc.) and in general. The lower the rank of a word, the more frequently it is used.

To find out whether frequency of use maps well onto NNS unfamiliarity with the word, we did a preliminary study with a random sample of 38 NNS who were students in a US university. We took the range from Rank 10,000 and above as high-level words and Rank 10 to Rank 200 as low-level words. We implemented them into a survey where 10 words from each of these two ranges were randomly displayed to the 38 NNS. They were asked to what extent they were familiar with the word or found the word difficult. Results of Independent T-tests suggested that NNS were much less familiar with the words from high-level range (M=4.11, SD=.15 on a scale from 1=extremely unfamiliar, 7= extremely familiar) than with those from low-level range (M=6.79,

SD=.05, t(74)=-16.80, p<.001), and found the ones from high-level range (M=3.61, SD=.16 on a scale from 1=extremely easy, 7=extremely difficult) much more difficult than those from the low-level range (M=1.19, SD=.06, t(74)=14.52, p<.001). Note that although the high-level range is not perfect for the purpose of detecting words that NNS are likely to be unfamiliar with (4.11 out of 7, 7=extremely familiar), there is a trade-off between setting the range too narrow (too infrequent) such that it may end up not helping NNS at all because words from that range would be much less likely to occur in conversation with NS. Further, the resulting range of difficult words determined by advanced NNS studying in the US can be applied to the experiment taking place in Japan, in that words difficult for advanced NNS are only more likely to be found difficult for intermediate level NNS who have limited overseas experience. Therefore, the range determined by the preliminary study is more effective in detecting words that are difficult for less advanced NNS participating in the experiment.

To design two types of agent that embody characteristics of a fluent speaker or a less fluent speaker, the first author and two undergraduate research assistants (1 NS, 1 NNS) first brainstormed a series of clarification request formats that each of these two types of speakers would typically and distinctively use. Using a vignette study [53], we tested how individuals perceived the language competence of a speaker who uttered the different clarification questions we presented in the survey. Results gave us a grammatically incorrect form for the low-level agent: "What mean …?", which was rated the least competent in English (M=2.78, SD=.09 on a scale of 1=extremely low, 7=extremely high); and a more polite and sophisticated form for the high-level agent: "Would you mind explaining the word … for me please?", which was rated the most competent in English (M=5.47, SD=.09).

Regarding interruption frequency, prior research shows that too many clarifications made too frequently may run the risk of sounding patronizing (e.g. foreigner talk [23], p.141) and of losing the listener's attention and interest, and thus might hinder understanding. To minimize such counterproductive effect, we experimentally set the number and frequency of agent clarification requests at no more than 3 times per 15 minutes conversation and no more than once for two minutes.

4.3.2 Videoconferencing interface

We implemented the agent feature into the three-way videoconferencing interface by adding a fourth "participant" whose image was an animated female cartoon character matched with a young woman's voice. To draw participants' attention, the agent is designed to wave her hand as she starts to ask for permission to interrupt; and moves her mouth as she speaks. The frame of her window also flashes blue as a way to indicate that she's speaking (see Figure 2).

Figure 2: Videoconferencing Interface with the Clarification Agent

4.3.3 Equipment setup

We used three Dell Inspiron laptops with Intel Core i5 CPU, 16GB memory, and 15.6-inch screens for the videoconferencing setup. Participants were asked to wear headsets with a microphone to communicate with each other and receive instructions from the experimenter. We screen recorded participants' screen activities including their verbal and nonverbal behaviors over the video-mediated conversations using Bandicam. Three camcorders were located behind each participant to video and audio record the entire experiment including interviews.

4.4 Procedure

Each experiment lasted about 2.5 hours. Upon arrival, participants (2 NS and 1 NNS) were led to separate soundproof rooms. The main experimenter gave general introduction and instructions in English over audio. NNS were assisted by a native Japanese-speaking experimenter who gave clarifications in Japanese.

After signing a consent form, participants completed the *pre-experiment survey*. Then they were asked to do a mini-task for warmup. For both the warmup and 3 real trials, they made their individual decision first, and then started free discussion to reach agreement. The experimenter stopped each trial after 15 minutes (10 minutes for warmup). After each of the three trials, they completed a *post-trial survey*.

Upon completion of all three trials and post-trial surveys, participants were asked to watch twelve 10-second videoclips and answer questions on the *retrospective survey*. At the end, we conducted semi-structured interviews with each participant individually in their native language.

The following figure (Figure 3) illustrates the experiment conditions and procedure.

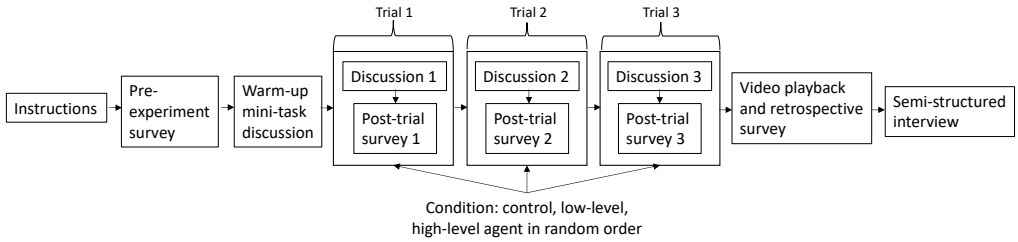

Figure 3. Diagram Illustrating the Experiment Conditions (in random order) and Procedure.

4.5 Conversation Coding

Before coding the conversation, two undergraduate research assistants were asked to proof and correct the automated transcripts based on the video recordings of the conversations. To keep our unit of coding consistent, they were asked to maintain the structure of the automated transcripts such that each row is considered an automatically transcribed utterance for our coding unit.

We adapted Carletta and colleagues' coding scheme [5] developed to analyze spontaneous task-oriented conversations. According to this coding scheme, an utterance is either coded as an "initiation" or "response" at the highest level, which matches our definition of active versus passive participation. We defined "clarification" as an utterance that made a previously expressed idea, statement or sentence less confused and more comprehensible.

The first author pilot coded randomly selected 6 conversations using the definition, and then trained two undergraduate research assistants who were blind to the purpose of the study to code the same 6 conversations. They discussed and resolved disagreements. The trained undergrad research assistants went on to code the rest of the conversations. A total of 8274 utterances were coded. Intercoder reliability based on 12% of the data was satisfactory (Cohen's Kappa = .86).

4.6 Measures

4.6.1 NS clarification

NS clarification was measured by calculating the number of words for utterances coded as "clarify". Following recommended data transformation procedure (e.g. [11][19]), we took the log10 of word count, to improve the normality of the data, before performing the statistical analyses.

4.6.2 NNS understanding

NNS overall understanding was measured in the post-task survey after each discussion trial using a 4-item 7-point Likert scale (1=strongly disagree, 7=strongly agree). Items included "I understood exactly what my partners were saying", "I had difficulties understanding my partners' opinions" (reverse coded), etc. They were averaged to form a reliable scale (Cronbach's alpha = .86).

For moment-by-moment understanding, we asked participants in the retrospective survey to "indicate their level of understanding for this time period" on a 7-point scale (1=completely lost, 7=all of it). Moment-by-moment understanding was measured before and after the agent's interruption, or before and after a supposed interruption when the agent "would have interrupted" in the control condition where a high-level word was detected.

4.6.3 Motivation to request clarification

We developed a 7-item scale to measure this construct. Participants rated on a 7-point scale (1=strongly disagree, 7=strongly agree) for statements such as "I felt reluctant to cut my partners' speech while they were talking" (reverse coded), "I felt embarrassed to ask my partners to explain" (reverse coded), "I felt comfortable to reveal that I didn't understand something", etc. The seven items formed a reliable scale (Cronbach's alpha = .80) and were averaged to form one measure for motivation to requestion clarification.

4.6.4 Overall and active participation

Participants' overall participation was measured by the total number of words spoken in each trial. Participants' active participation was measured by the total number words for utterances classified as "initiation" in each trial. Following recommended data transformation procedure (e.g. [11][19]), we truncated outliers to the mean +2.5SD and took the log10 of word count, to improve the normality of the data, before performing the statistical analyses.

4.6.5 Self-evaluation of communication competence

We measured participants' self-evaluation of their own communication competence on two dimensions: understanding others' opinions and expressing own opinions in English. Two items such as "my ability in understanding my partners' opinions was" (1= terrible, 7=excellent) were included. They were averaged to form a reliable measure (Cronbach's alpha = .90).

4.6.6 NNS foreign language anxiety

We measured NNS Foreign Language Anxiety using an adapted Foreign Language State Anxiety Questionnaire (FLSAQ) [1] designed to measure L2 learner's state anxiety following an interactive task, which served our purpose. We removed items that involved anxiety in classroom settings. Participants responded on a 7-point Likert scale (1=strongly disagree, 7=strongly agree) to the 8 statements including "I got flustered when my partners communicated things I did not understand", "I started to panic when I had to speak without preparation". The items formed a reliable scale (Cronbach's alpha = .90) and were averaged to form a Foreign Language Anxiety measure and loaded onto two subscales: Speaking Anxiety (Cronbach's alpha = .82) and Understanding Anxiety (.90).

4.6.7 Workload

Workload was measured using NASA-TLX [29], a widely used tool for measuring subjective task workload. We adapted the scale by dropping the "physical" dimension out of the 6 dimensions (mental, physical and temporal demand, performance, effort and frustration) since our task did not involve much physical activity. Participants responded on a 7-point Likert scale (1=extremely low, 7=extremely high). The items formed a reliable scale (Cronbach's alpha = .77) and were averaged to form a workload measure.

4.7 Data Analyses

4.7.1 Semi-structured interviews

All the interviews were conducted in the participant's native language, and were audio recorded for transcription and translation. The second author, who is a native Japanese speaker, conducted all the interviews with NNS. The Japanese transcripts were translated into English by professional translators. We employed a grounded theory approach [27] and did open coding while examining

through the data in the initial phase. A loose array of open codes was generated from the interviews and was sorted using affinity diagraming, through which high-level themes and relationships between the themes were identified. We then went through an iterative process of axial coding where we grouped together related open codes and refined them.

4.7.2 Retrospective survey

Retrospective survey data on paper was digitalized by the first author immediately after each experiment. Except for Group 16 where participants did not finish the retrospective survey due to an unforeseen natural disaster, all data was converted into SPSS files for analysis.

5 QUANTITATIVE RESULTS

In this section, we present the statistical results of a series of Mixed Models ANOVAs on our key measures that provide answers to our hypotheses and research questions. First, we report effect of the agent on NS clarifications, NNS understanding and participation. Then, we report the effects of different types of agent on NNS self-evaluation of their communication competence, NNS foreign language anxiety, and NNS motivation to request clarification. Lastly, we report the effects of different types of agent on the workload for both NS and NNS.

5.1 Effects of the Agent on NS Clarification

To test H1 and answer RQ1, we conducted a 3 (Condition) by 3 (Trial) Mixed Models ANOVA on the log10 of NS clarification word count. Consistent with H1, there was a significant main effect of condition ($F [2, 66.98] = 126.01$, $p < .001$). Pairwise comparisons showed that NS made more clarifications with the high-level agent ($M=1.98$, $SE=.09$) and the low-level agent ($M = 1.96$, $SE = .09$) than without the agent ($M= .52$, $SE = .09$, $p < .001$). But there was no significant difference in the amount of clarification NS made with different types of agent ($p = .83$). There was also no significant main effect of trial ($F [2, 63.93] = 1.31$, $p = .28$) and no significant interaction between condition and trial ($F [4, 92.47] = .72$, $p = .58$, see Figure 4). These results support H1 that predicts NS will make more clarifications with than without agent. With regard to RQ1, there is no significant difference in the number of NS clarifications between the low- and the high-level agent.

5.2 Effects of NS Clarifications on NNS Understanding

To test H2 and explore RQ2, we conducted a 3 (Condition) by 3 (Trial) Mixed Models ANOVA on the NNS *overall understanding*. No significant main effect of condition ($F [2, 95.98] = .22$, $p = .80$) or trial ($F [2, 84.20] = 1.28$, $p = .28$) was found. Neither was there a significant interaction effect of condition and trial ($F [4, 121.98] = 1.64$, $p = .17$).

However, a 3 (Condition) by 2 (before or after agent interruption) Mixed Models ANOVA on the NNS moment-by-moment understanding suggested that, even though there was no significant main effect of condition ($F [2, 155.63] = 1.58$, $p =.21$). There was a significant main effect of agent (supposed) interruption ($F [1, 111.36] = 9.81$, $p <.01$). NNS reported better understanding after the agent (supposed) interruption ($M = 5.05$, $SE = .23$) than before the (supposed) interruption ($M = 4.58$, $SE = .23$). There was also a significant interaction effect between condition and agent interruption ($F [2, 117.25] = 4.50$, $p = .01$; see Figure 5). A simple main effect analysis showed that at the moment before the agent interruption, NNS reported lowest understanding with the high-level agent ($M=3.74$, $SE=.29$) than with the low-level agent ($M=4.81$, $SE=.29$) or without the agent

(M = 4.96, SE = .29). Whereas at the moment after the agent interruption, there was no significant difference in NNS reported understanding across all conditions (p = .89). This suggested a significant increase of understanding following the high-level agent's interruption, which supports H2.

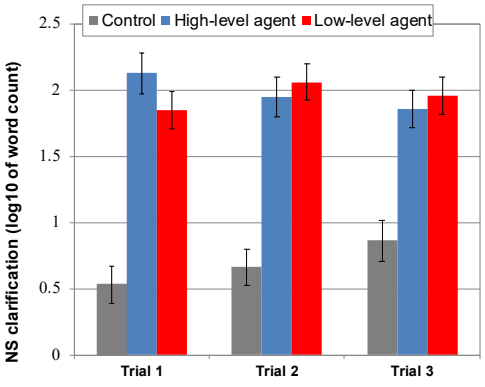

Figure 4. Mean NS Clarification (log10 of word count) by Condition for 3 trials (error bars represent standard error of the mean).

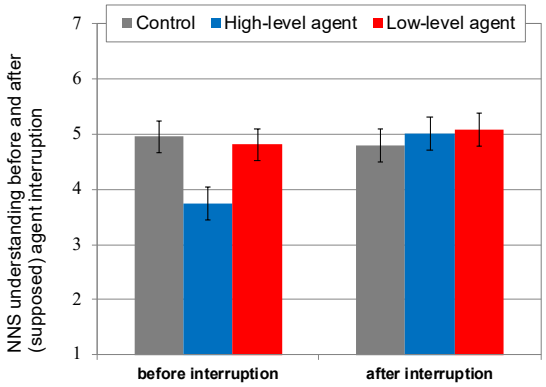

Figure 5 Mean NNS Moment-by-Moment Understanding on a scale of 1 to 7, by Condition for before and after (supposed) agent interruption (error bars represent standard errors of the mean).

5.3 NNS Overall and Active Participation

To test H3 and explore RQ3, we conducted a 3 (Condition) by 3 (Trial) by 2 (NS vs NNS) Mixed Models ANOVA on the participants' *overall participation*. There was a significant main effect of speakers' native language (F [1, 47.40] = 92.63, $p < .01$). NS (M = 2.88, SE = .03) had significantly higher participation than NNS (M=2.34, SE=.05). There was a trend towards a significant main effect of condition (F [2, 96.20] = 2.95, $p = .06$). Overall participation was higher in the high-level agent condition (M=2.63, SE=.03) and the low-level agent condition (M=2.61, SE=.03) than in control (M=2.58, SE=.03). There was also a trend towards a significant interaction between condition and speakers' native language (F [2, 96.20] = 2.81, $p = .07$). For NS, there was no significant difference across conditions with respect to their participation (ps >.77). For NNS, their overall participation was lower in the control condition (M=2.27, SE=.05) than in the low-level agent condition (M=2.35, SE=.05) and high-level agent condition (M=2.38, SE=.05, ps <.04). There was no significant main effect of trial (F [2, 83.41] = .11, $p = .90$).

A 3 (Condition) by 3 (Trial) by 2 (NS vs NNS) Mixed Models ANOVA on the *active participation* suggested a significant main effect of condition (F [2, 95.02] = 17.46, $p<.01$). Pairwise comparisons suggested that active participation was higher in the high-level agent condition (M=2.45, SE=.08) than in either control (M=2.12, SE=.08, $p <.01$) or low-level agent condition (M=2.24, SE=.07, p<.01). There was also a significant interaction effect between condition and speakers' native language (F [2, 95.02] = 24.10, p < .01; see Figure 4). For NS, there was no significant difference across conditions with respect to their participation (ps > .37). For NNS, pairwise comparisons suggested that NNS active participation was highest in the high-level agent condition (M=2.19, SE=.13), followed by low-level agent condition (M=1.72, SE=.13, $p <.01$), followed by control (M=1.46, SE=.13, $p = .01$). There was no significant main effect of trial (F [2, 82.47] = 1.58, $p =.10$). These results support H3: NNS participation (both overall and active participation) was

significantly higher in the agent conditions than in control (see Figure 6). With regard to RQ3, NNS active participation was significantly higher with the high-level agent than with the low-level agent.

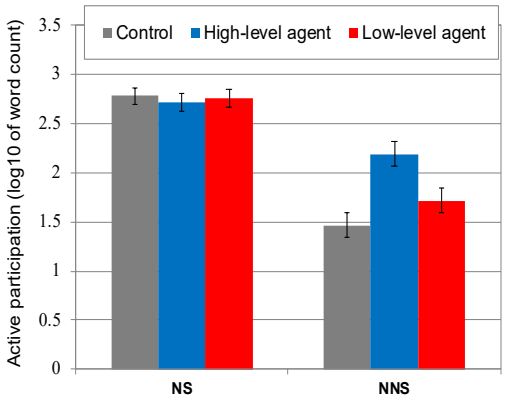

Figure 6. Mean Active Participation (in log10 word count), by Condition for NS and NNS (error bars represent standard errors of the mean).

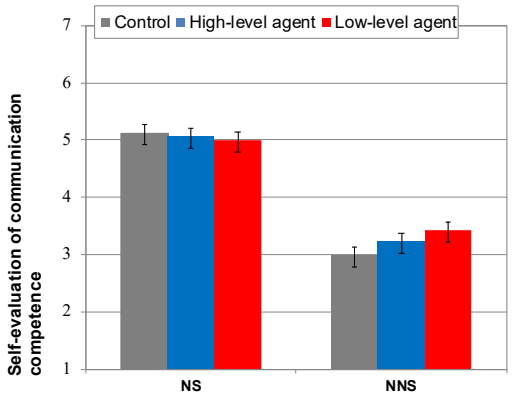

Figure 7. Mean Self-evaluation of Communication Competence on a scale of 1 to 7, by Condition for NS and NNS (error bars represent standard errors of the mean).

5.4 Effect of Types of Agent on Self-Evaluation of Communication Competence

To test H4, we conducted a 3 (Condition) by 3 (Trial) by 2 (NS vs NNS) Mixed Models ANOVA on participants' *self-evaluation of communication competence*. Results showed that there was a significant main effect of speakers' native language (F [1, 46.97] = 68.98, $p < .01$, see Figure 5. NS (M=5.06, SE=.13) reported higher self-evaluation than NNS (M=3.21, SE=.18). There was no significant main effect of condition (F [2, 95.82] = 1.37, $p = .26$) or trial (F [2, 82.96] = 1.41, $p = .25$). However, there was a significant interaction effect of condition by speakers' native language (F [2, 95.82] = 4.16, $p = .02$). Pairwise comparisons showed that NNS self-evaluation was significantly higher with the low-level agent (M=3.42, SE=.20) than in the control condition (M=2.98, SE=.20, $p = .01$). There was no significant difference in NNS self-evaluation between high- and low-level agent conditions ($p = .63$) or between control and high-level agent ($p = .56$). There was also no significant difference in self-evaluation across all conditions for NS ($ps > .30$). The results partially support H4 that NNS reported a more positive self-evaluation with the low-level agent than without agent (see Figure 7).

5.5 Effect of Types of Agent on NNS Foreign Language Anxiety

To test H5, we conducted a 3 (Condition) by 3 (Trial) Mixed Models ANOVA on NNS *foreign language anxiety*. Results showed that there was a significant main effect of trial (F [2, 25.79] = 3.57, $p = .04$). NNS reported significantly higher foreign language anxiety in Trial 1 (M=4.32, SE=.33) than Trial 2 (M=3.85, SE=.33, $p = .02$) and Trial 3 (M=3.77, SE=.33). There was also a borderline significant main effect of condition (F [2, 63.93] = 1.31, $p = .07$, see Figure 8). Pairwise comparisons suggested that NNS experienced significantly higher foreign language anxiety with the high-level agent (M=4.24, SE= 33, p = .02) than with either low-level agent (M=3.95, SE= 33) or without the agent (M=3.76, SE= 33). There was no significant interaction between condition and trial (F [4, 29.74] = .35, $p = .85$). These results are contradictory to H5.

To further explore the contradictory results, we ran a 3 (Condition) by 3 (Trial) Mixed Models ANOVA on the subscales: *anxiety associated with understanding* and *anxiety associated with speaking*. Results suggested that there was no significant main effect of either condition (F [2, 29.12] = .91, p = .41) or trial (F [2, 25.71] = 2.16, p = .14), or interaction effect (F [4, 30.77] = .59, p = .68) for anxiety associated with understanding. But for anxiety associated with speaking, there was a significant main effect of trial (F [2, 25.79] = 3.87, p = .03). NNS reported significantly higher foreign language anxiety with respect to speaking in Trial 1 (M=4.34, SE=.31) than Trial 2 (M=3.89, SE=.31, p =.02) and Trial 3 (M=3.65, SE=.31). There was also a significant main effect of condition (F [2, 29.23] =3.22, p = .05). Pairwise comparisons suggested that NNS experienced significantly higher foreign language anxiety associated with speaking with the high-level agent (M=4.24, SE= 31) than without agent (M=3.72, SE= 31, p = .02). There was no significant difference in speaking anxiety between the high- and low-level agent (M=3.92, SE= 31, p = .36). There was also no significant interaction between trial and condition (F [4, 30.94] =.47, p = .76). In sum, NNS foreign language anxiety associated with understanding did not seem to differ across conditions, but that associated with speaking was higher with the high-level agent than without agent. These results partially support H5.

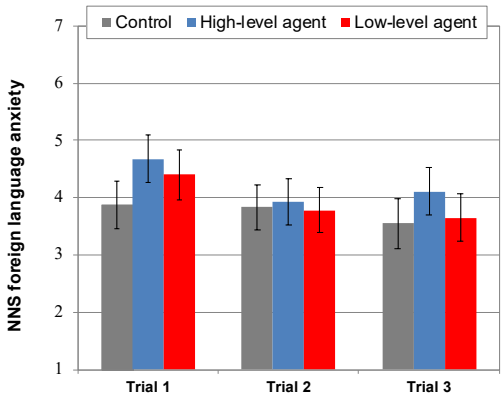

Figure 8. Mean NNS Foreign Language Anxiety on a scale of 1 to 7, by Condition for 3 Trials (error bars represent standard errors of the mean).

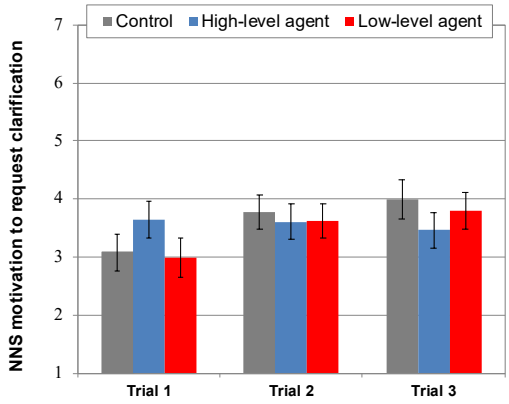

Figure 9. Mean NNS Motivation to Request Clarification on a scale of 1 to 7, by Condition for 3 Trials (error bars represent standard errors of the mean).

5.6 Effect of Types of Agent on NNS Motivation to Request Clarification

To answer RQ4, we conducted a 3 (Condition) by 3 (Trial) Mixed Models ANOVA on motivation to request clarification. There was a significant main effect of trial (F [2, 26.18] = 3.87, p = .03). Pairwise comparisons showed NNS reported lowest motivation to request clarification in Trial 1 (M=4.06, SE=.13) than in Trial 2 (M=4.32, SE=.13, p = .02) and Trial 3 (M=4.44, SE=.13, p = .01). But there was no significant main effect of condition (F [2, 28.64] = .18, p = .84) on NNS motivation to request clarification. There was a significant interaction effect of condition and trial (F [4, 40.54] = 2.69, p = .03). A simple main effect analysis revealed that in Trial 1, NNS reported higher motivation to request clarification if they were exposed to the high-level agent (M=3.65, SE=.31) compared to the low-level agent (M=2.98, SE=.34, p = .03) and control (M=3.08, SE=.32, p = .04); whereas in Trial 3, NNS reported lower motivation to request clarification with the high-level agent (M=3.46, SE=.31) compared to control (M=4.00, SE=.34, p =.05) and with the low-level agent

(M=3.80, SE=.32). In sum, NNS motivation to request clarification did not differ significantly across conditions overall but depended on the order of conditions. In Trial 1, NNS were more motivated to request clarification with the high-level agent than without agent or with the low-level agent (see Figure 9).

5.7 Effect of Types of Agent on NS and NNS Workload

To explore RQ5, we conducted a 3 (Condition) by 3 (Trial) by 2 (NS vs NNS) Mixed Models ANOVA on participants' workload. Results showed that there was a significant main effect of Speaker Type (F [1, 50.22] = 27.97, p < .01). NS (M=3.51, SE=.16) reported significantly lower workload than NNS (M=4.95, SE=.22). There was no significant main effect of condition (F [2, 95.82] = 1.37, p=.26) or trial (F [2, 82.96] = 1.41, p = .25). Nor was there any significant interaction effect between condition and trial (F [4, 120.24] = .54, p = .71).

	Measures	Control		Low-level agent		High-level agent		F	df	*p* (s)
		M	SE	M	SE	M	SE			
H1, RQ1	**NS clarification**	**0.52**	0.09	**1.96**	0.09	**1.98**	0.09	126.01	66.98	**<.001**
H2, RQ2	**NNS understanding (overall)**	4.43	0.27	4.21	0.27	4.33	0.27	0.22	95.98	0.80
	NNS understanding (moment-by-moment)							4.50	117.25	**0.01**
	before interruption	**4.96**	0.29	4.81	0.29	**3.74**	0.29			
	after interruption	4.80	0.29	5.09	0.29	5.02	0.29			
H3, RQ3	**Participation (overall)**							2.81	96.20	0.07
	NNS	**2.27**	0.05	2.35	0.05	**2.38**	0.05			
	NS	2.87	0.04	2.88	0.04	2.88	0.04			
	Participation (active)							24.10	95.02	**< 0.01**
	NNS	**1.46**	0.13	**1.72**	0.13	**2.19**	0.13			
	NS	2.78	0.09	2.75	0.09	2.72	0.09			
H4	**Self evaluation of communication competence**							4.16	95.82	**0.02**
	NNS	**2.98**	0.20	**3.42**	0.20	3.23	0.20			
	NS	5.12	0.14	5.01	0.14	5.06	0.14			
H5	**NNS foreign language anxiety**									
	associated with speaking	**3.72**	0.31	3.92	0.31	**4.24**	0.31	3.22	29.23	**0.02**
	associated with understanding	3.84	0.42	4.00	0.42	4.22	0.42	0.91	29.12	>.20
RQ4	**NNS motivation to request clarification**							2.69	40.54	**0.03**
	Trial 1	**3.08**	0.36	**2.98**	0.39	**3.65**	0.36			
	Trial 2	3.77	0.35	3.63	0.35	3.61	0.38			
	Trial 3	4.00	0.39	3.80	0.36	3.46	0.36			
RQ5	**Workload**							0.56	98.27	0.58
	NNS	4.74	0.27	5.13	0.27	4.99	0.27			
	NS	3.37	0.19	3.50	0.19	3.66	0.19			

Table 1. Summary of Results (bolded numbers indicate significance at p<.05 level)

6 QUALITATIVE RESULTS

In this section, we present the key themes emerged from our qualitative analysis of the interviews with NS and NNS by organizing them around our hypotheses and research questions.

6.1 Effect of Different Agent on NS Clarification

All NS participants perceived the difference between the high- and low-level agent. They distinguished the agents by calling the high-level agent the one asking "real questions" and "good questions" that "are related to the topic", and the low-level agent as the one asking "weird" questions that "are nowhere related to the topic". These different perceptions of the two agents and their questions affected how effectively NS clarified on the requested words and beyond.

6.1.1 Effect of the low-level agent on NS clarification

Despite responding to the low-level agent as requested in the conversations no matter what, in the interviews, NS noted the inappropriateness of the agent asking about the meanings of basic words that are irrelevant to a conversation discussing serious matters. They reported the challenge of explaining the basic vocabulary and expressed confusion about its purpose as opposed to the high-level agent.

More than half of the NS thought the low-level agent was inappropriate in a formal conversation in terms of the words it requested clarification on. For instance, one NS mentioned that it felt like a kid interrupting a serious conversation.

> (the low-level agent sounded) immature in between a serious conversation, like involving a kid. We're surprised by the questions. I thought it was too childish in between a serious conversation. It was nowhere related to the topic, we felt like it was not necessary to have interrupted. The word was too simple for us to have discussed in between a serious conversation. – 18A (NS)

Additionally, many NS (15 out of 34) found that simple words were difficult to explain. One even expressed insecurity in getting her point across to someone who does not even know basic words.

> I actually do find it difficult to explain really simple words, like the word "really". How do I explain that? So it makes me feel really insecure. Because immediately I feel like, Oh, no, if I can't explain this word, how am I even going to get my point across? – 21B (NS)

The grammatically incorrect use of English along with the basic word the low-level agent asked was inconsistent with NS perception of artificial intelligence and made them feel clueless about the purpose of the low-level agent. Whereas with the high-level agent, they immediately figured the agent was trying to help. For instance, one NS noted:

> Why would you ask me about a basic vocabulary word? If you're truly a computer, you should have an index of those terms in your vocabulary already… I was clueless about the purpose. But then you had the next agent to compare it to, then you could come to realize Oh there are people who do need help with things like that. If you only hear the first one, (you wonder) why is she talking like that? But the (one asked about) more difficult words is much more natural. You want to ask someone the meaning of something. – 20A (NS)

6.1.2 Effect of high-level agent on NS clarification

Awareness of NNS needs: Once NS realized that the agent was trying to help, they came to be aware of the disparity in language fluency in the group conversation with the NNS, and tried to bring down the level of difficulty not just for the agent but for NNS, too. For instance, NS would explain the word the agent asked for the sake of their NNS partners. One NS reflected:

> I think the agent was more helpful in the first one (high-level) because for example, I didn't know if C (NNS) understood these things, I was actually explaining back to C because I was worried that he wasn't being included in the conversation this much. – 17A (NS)

One NS explicitly noted that the high-level agent asked about a word that she actually had worried whether her NNS partner might have understood. However, she did not explain immediately after she said it. The agent's question allowed her to revisit it and clarify for the NNS.

> I did think maybe Ms.C (NNS) would maybe not have understood (the word "liquidate"). But the way she was engaging in the conversation, she didn't ask or she didn't seem like she didn't know. So I kind of let it go. So when the agent asks, I thought like oh okay, now it's a good opportunity to explain that word. – 21A (NS)

Voluntary and proactive clarification: The exposure to the high-level agent has encouraged NS to be mindful of the word they use and voluntarily provide explanation and dynamically adjust the level of difficulty. For instance, for concepts that NS realized could be unfamiliar to or too complicated for NNS, they would proactively offer explanations without the agent prompting them to clarify. As the following quote suggests,

> It encouraged us more to offer explanations or try to proactively think of… for example, if I came up to a certain word, or a more complicated or less known concept, (I'd) try and check and see if she (NNS) knew what was going on. – 16B (NS)

The high-level agent has also encouraged NS to proactively bring down the level of difficulty and actively check to make sure everyone in the conversation could understand.

> (The high-level agent) does remind us that when we are speaking in a group where not everyone has the same level of English ability, then we have to be more aware of the words we use… I tried to bring it down for C (NNS), because when she speaks, it seems like her English is not that good. So I want to make sure that she understands everything. – 21B (NS)

Some (5 out of 34) even noted that they continued to speak at the level that the high-level agent encouraged them to speak at, even in subsequent conversations without the agent or with a different agent. As the following quote illustrates:

> If you actually look at my notes, following the first conversation with the (high-level) agent I actually dumbed down my arguments. Because I was gonna say that we should "repatriate the historical objects", but now I put "send back" and that's what I actually use in the conversation instead, because I thought they'd be better for the conversation flow. The first conversation, I was trying to speak like it was actually UNESCO. And then the second conversation I was trying to speak more like I was speaking with nonnative speakers. – 17A (NS)

Despite their intention to bring down the level of difficulty for NNS, NS reported that it was not easy to do, especially when engaging in a formal meeting, where they automatically registered to using higher level vocabulary. For instance, one NS admitted:

> When talking about these kinds of things, I almost default to a higher level of English. When I'm talking about a casual thing, I would never use the word liquidation. But it's all about this context, I'm trying to be formal. I'm using the word liquidate. But then no word immediately comes to mind. And then I think, ah dilute, so immediately, I think, but dilute might actually be the same level of what is liquidate. So how is that going to help? – 21B (NS)

Alternative strategies to accommodate: Because of this difficulty mentioned above, NS used various other strategies to work around tuning down the vocabulary, while attending to, gauging and checking for NNS understanding. Continuing the last quote in the previous section, the NS (21B) reported that she employed alternative strategies such as rephrasing or giving examples to illustrate her point.

> So instead, I tried to rephrase the sentence completely. So instead of say "liquidate means", I rather try to be like, let me say why I did this. I tried to go around it so that they might understand something else. So throughout the rest of that interaction, I said "it will water down the culture". And then I wanted to give the whole example of Justin Bieber, so that, because an example like telling a story might encourage understanding if that person knows about that song, but then I realized that she might not even know the song. Man, she's like, even more confused. So then I went back to say "water down the culture", that seemed to work -21B (NS)

Similarly, many NS (26 out of 34) reported to have employed at least one of the techniques such as repetition, paraphrasing, providing context, examples or simply more information, until they felt the NNS partner understood. The following quote represents an effort to try a comprehensive array of clarification techniques:

> I will say something and then restate it using a similar phrase or expression, not the same word, but you're giving someone information. So for example, "antiquities were looted", many things were stolen and taken out, people came in and took them out. So, you make an expression, but then you're giving examples or extra details which reinforce and explain the situation. So I could tell that C (NNS), if he might not have caught the word from the expansion, he can understand the meaning of the situation. He was like (nodding)… So you repeat the same several times, maybe slightly different ways, or with expansion and examples so that people can understand what you're trying to get across – 19B (NS)

In some cases, NS were able to use a similar example in the context familiar to NNS to illustrate a complicated idea. For instance, one NS talked about her explaining minority language publishing to her NNS partner:

> Especially I know similar situations in Japan. So I tried to engage her (NNS). I tried to exchange probably what she knew about the Japanese… trying to maybe have her relate to a similar situation in Japan. – 16B (NS)

6.2 Effect of NS Clarification on NNS Understanding

This section presents NNS perspectives of how NS clarifications to different types of agent affected their understanding.

6.2.1 Effect of NS clarification to the low-level agent on NNS understanding

NNS understanding did not directly improve from NS response to the low-level agent. But some NNS (5 out of 17) were able to take advantage of the time that NS took to respond to the agent to catch up, process, and make sense of what had been said.

> When she (the low-level agent) stopped the conversation, I was pretty much at my limit, so it did give me some time to think things over for myself. I didn't come to understand

> anything because of the agent's questions directly, but because the conversation was forcibly stopped, I was able to put together the flow of the discussion up until that point, or at least as much of it as I had been able to make out… it bought me some time, so in that regard I was grateful to have the interruptions. It's not really something I could do for myself, to stop their discussion. – 2C (NNS)

Nevertheless, the fact that NS had to explain basic words that are irrelevant to the topic had made a few NNS feel overwhelmed by the disruption of the conversation flow, and has thrown them off by derailing the topic that they had worked hard to stay on track.

> It asked questions about vocabulary that I think it should have known. The native speakers spoke quite fast, so I only understood about half of what they said. (The low-level agent) Cutting in made me forget what was being said. ...I lost track of what I had been working hard to follow. The shock (bewilderment) due to the second (low-level) agent was too great. – 4C (NNS)

6.2.2 Effect of NS clarification to the high-level agent on NNS understanding

NS response to the high-level agent, on the other hand, were acknowledged by all NNS as being much more helpful, as it asked words they also did not catch or understand.

> I could sometimes understand the other two (NS) better depending on how they answered the agent's questions, so I thought that part was different from the first one (low-level). –18C (NNS)

Almost all (15 out of 17) NNS reported having the same question the high-level agent asked and expressed how timely the agent prompted NS for the clarifications that NNS needed. One made explicit comparison with the agent regarding their English levels.

> It asked the words that I missed, like the word "reverberating"… I was hoping NS would give more information for me to figure that out. Then the agent caught that word. It felt like we were speaking from the same level. It seemed quite smart, with English ability similar to mine. It seemed to be asking about special terminology and words I didn't know. –19C (NNS)

For fluent and less fluent NNS respectively, the extent to which the high-level agent was helpful for their understanding was illustrated in their reflections. Less fluent NNS reported being grateful and "greatly helped" by the high-level agent.

> … when I didn't understand something NS said due to my lack of vocabulary, the agent asked about that word, so I felt that that behavior helped my understanding quite a bit, it greatly increased my comprehension, so on that point it was very useful. –20C (NNS)
> …it was asking about words I had no idea about, so I felt very grateful. –21C (NNS)

Quite a few (7 out of 17) NNS felt the high-level agent served as a bridge between them and NS. Considering the fact that all 7 used the exact same metaphor, this concurrence is illuminating.

> when (the agent) asked about words that were rather difficult, NS explained rather thoroughly… I don't know with what purpose the agents were acting, but I felt that they were asking about things that I did not understand, so I felt the agent became a bridge for me between A-san and B-san. – 6C (NNS)

Fluent NNS acknowledged that NS responses to the high-level agent helped them understand better because they could not interrupt and ask for themselves even though they also did not know the word.

> The first agent asked questions about words I was curious about, words that I wanted to ask myself… (but) I wouldn't ask because it would take great courage to cut them (NS) off. – 10C (NNS)
> It also asked about an ambiguous word, which was helpful to me. I didn't know the word "indigenous." It was helpful that the agent asked. I prefer the agent asking about words that are a bit difficult because I myself couldn't ask. – 5C (NNS)

One of the reasons they would not ask for themselves even though they also did not know the word the high-level agent picked up was that they could infer the meaning from the context. In other words, not knowing the word may not have affected their overall understanding greatly. As one NNS noted:

> I think the agent asked question on words such as "forego". Although I could figure that word from its context, I did not understand its meaning as a word so it helped me understand it. Their (NS) answer confirmed my guess. So I could follow along with greater confidence. That's the way it is with words such as 'forego'. I think the agent was picking up words that people are not used to hearing. –11C (NNS)

6.3 Effect of Agent on NNS Motivation to Request Clarification and Participation

Ten out of the seventeen NNS mentioned that the high-level agent's question asking behavior made them feel that they could ask clarification questions as well. As one noted:

> The (high-level) agent asking about difficult words made me happy when I myself couldn't ask. It also made me feel that I can also ask questions if needed. – 5C (NNS)

However, only 4 NNS have taken the courage to ask their own questions. One NNS explicitly recalled how he modeled the agent's behavior and asked a question himself.

> The agent that helped me was the one that asked difficult words. Because of this, I tried asking a question one time. I tried asking about a word similar to how the agent asked, and it became a bit easy to ask. Probably because it was an agent, I thought, well, that's how it is. Certainly if you were to ask me if I could cut off a discussion like that, I'd think it's a bit difficult for me. – 10C (NNS)

Some (5 out of 17) expressed that in subsequent conversations following the agent asking questions, they started to take the initiative to inquire and prompt NS for more information.

> I didn't ask about the meaning of a specific word, but rather, I started asking questions like asking for a concrete example, asking why they think so if they have a different opinion than my own. I expressed my opinion concisely, and then asked if they had a different opinion. –19C (NNS)

In one case, a NNS braced up to cut off two NS for clarification as a precautionary act to prevent a more face-threatening situation.

> I wanted the agent to ask a question and interrupt when the discussion was heated up (between NS), but it didn't. So I had to do it myself. Because if they (NS) cannot reach

> agreement they will ask me "what do you think?" Then at that point, if I say I don't understand, it's going to be so bad. So before they came to me, I should ask and confess now. Better to ask now "before the wound gets deeper". So I confessed that I didn't understand and asked them to explain, even though I didn't want to. –17C (NNS)

NNS also reported how having "someone" of similar English level joining the conversation made it easier for them to speak up. A few (4 out of 17) mentioned that the agent interruptions have bought them time to prepare what they wanted to say and say it at the opportune timing created by the agent.

> The agent stopped the conversation in a personally helpful way. From the beginning, I was not able to keep up with the speed of the conversation. So when the agent cut in, I got some time to listen and figure out what would be the next topic of discussion and what the persons (NS) then were talking about. It wasn't a situation in which I particularly wanted to know the content of the agent's words, but instead I thought it was good that I could get a break from the conversation and think about its overall direction. And I had some leeway in being able to say what I wanted to say. – 6C (NNS)

6.4 Effect of Agent on NNS Self-Evaluation

Consistent with the quantitative findings on NNS self-evaluation, quite a few NNS (7 out of 17) made explicit comparison to the low-level agent and reported feeling better about themselves in terms of English ability. As one noted,

> The words the second (low-level) agent asked, I knew them already. It may be odd to compare myself with the agent, but I felt superior to the agent. –16C (NNS)

In one exception, the NNS admitted that knowing better than the low-level agent did not change his position as the least fluent person in the conversation, because he did not view the agent as a person but as a tool.

> I didn't understand a hundred percent, but the second (low-level) agent was like someone who couldn't understand even more… I didn't perceive it as a person in the first place. It was like a tool. So I recognized that there were three people and one tool. I couldn't change the fact that I was the least abled person. – 4C (NNS)

Contrary to this view, many NNS (8 out of 17) indicated in the interviews that they perceived the high-level agent as a person instead of a tool and reported feeling relieved to know they were not the only person who did not know the word the agent asked. As the quote below illustrates,

> I felt a sense of relief that I'm not the only one who can't understand the discussion. – 3C (talking about high-level agent)

Some (6 out of 17) reported a feeling of security to have had "another NNS" in the conversation and that the high-level agent made them feel more relaxed. NNS indicated that the sense of relaxation came from knowing that there's someone having similar English ability and difficulties as themselves.

> It gives you a sense of security when people in the same talk have the same level as you. – 6C (talking about high-level agent)
> I felt unconsciously that one more non-native speaker was there, and (it) gave me a sense of security. – 5C (talking about high-level agent)

Some (6 out of 17) also mentioned that seeing NS positive reaction to the agent regardless of type further made them let down their guard.

> It was impressive that the NS were very polite to the agent and explained the word appropriately, so that it took away my pressure. I felt happy and relieved. – 17C (NNS)
> Their (NS) attitude of answering to the agent's questions put me at ease. –14C (NNS)

7 DISCUSSION

In this study, we explored the effect of two types of clarification agent in multiparty virtual collaboration between NS and NNS: one that resembles a low-level NNS who requested clarification on easy words using broken English, and one that resembles a high-level NNS who requested clarification on difficult words using more polite and sophisticated English. Our findings indicated that high-level agent was preferred and yielded more desirable outcomes. Specifically, NS made many more clarifications with both types of agent than without agent. NNS understanding was directly improved after high-level agent interruption and they voluntarily spoke more in high-level agent condition. On the other hand, the low-level agent was not without merits. As predicted, NNS reported higher self-evaluation in the low-level agent condition than in the control condition. But the low-level agent did not significantly reduce NNS foreign language anxiety as we had predicted. In the rest of this section, we discuss these findings in light of qualitative results and previous literature.

7.1 NS Clarification and NNS Understanding

Both statistical and qualitative results suggested that agent (regardless of type)'s clarification requests successfully elicited clarification from NS. Moreover, NS continued to make clarifications in the subsequent trials after they have been exposed to the agent. This is evident in both NS reflection (e.g. 17A) and the statistics (Figure 4). NS who were randomly assigned to the control condition *after* both agent conditions made many more clarifications without the presence of an agent in comparison to NS who were randomly assigned to other condition orders (see Figure 4). NS uniformly preferred the high-level agent and employed various clarification strategies to explain not only what was asked (the meaning of a difficult word), but to get their ideas across someone whose level of English was indicated by their needing clarification of that difficult word. NS offered these clarifications not only for the sake of answering the agent, but also for their NNS partners. NS also went beyond to proactively provide explanation, examples, contexts familiar to NNS, in addition to leveling down the vocabulary without the agent prompting.

The effect of NS clarifications on NNS understanding, however, seemed to be limited only to the moments after the high-level agent interrupted, as NNS overall understanding assessed at the end of discussion was not significantly different across all conditions, and the moment-by-moment analysis suggested an increase in NNS understanding after the high-level agent's interruption. This indirectly demonstrates that the timing the high-level agent interrupted matched to when NNS actually needed. But the agent did not interrupt every time NNS needed it, which was part of the design to avoid frequent interruptions. It was highly likely that NS uttered difficult words incomprehensible by NNS when the agent was set to mute. As several of the NNS (e.g. 17C) recalled how they wished the agent would have asked about a specific word they did not know in addition to the words the agent already picked up on.

It is also worth noting that NS's voluntary clarifications (e.g. switching to lower-level vocabulary, providing examples, and so on without the agent prompting) seem to have gone unnoticed by NNS. At least in the interviews, not a single NNS have acknowledged any of these efforts their NS made as deliberate. Instead, NNS attributed their ease of understanding to the nature of the task. For instance, 17C (NNS) believed the discussion on culture was easier (while some NNS found it to be the hardest topic), where one of his NS partners (17A) later reported how she had substituted easy words for difficult words (e.g. "send back" for "repatriate") in that session. Anecdotal as this might seem, it alludes to a possibility that NS voluntary clarifications inspired by the high-level agent may have helped NNS understand better without NNS knowing that they were helped. In fact, Duan and colleagues [14] have reported similar findings with their Speech Speedometer, where NNS did not interpret their NS partners' slowing down as a deliberate effort made for themselves but still benefitted from this effort. While Duan and colleagues [14] had the urge to communicate NS accommodative effort to NNS, we view this lack of awareness on NNS part as a strength in such tools, in that many NNS prefer not to be explicit about their needs for help due to face concerns [36]. NS voluntary clarifications could pass as simply more information given, just as slowing speech could pass as natural speech variation, instead of deliberate accommodation. In a nutshell, NS voluntary clarifications and their direct responses to the high-level agent are both discreet ways to address potential language difficulties that NNS might have without their face being threatened. As such, the high-level agent has not only advanced the design of language tools for multilingual multiparty collaboration (e.g. [15]) by accurately identifying potential language difficulties for NNS in real time, but also addressed an important concern for face in such interactions that prior tools (e.g. [25][26]) have overlooked.

The function of the high-level agent is not limited to eliciting explanation of the meaning of a difficult word, which a dictionary would do. Rather, it improved NNS understanding by buying them time to put together overflowing information (e.g. 2C), to confirm their understanding (e.g. 11C) and to get prepared for upcoming information (e.g. 6C). These would not have been possible by simply providing a built-in dictionary as previous work has done [26], since looking up a word would increase NNS cognitive load and may cause them to miss important information in the ongoing conversation. The fact that in our study, both NS and NNS workload did not differ significantly across conditions suggests that at least the high-level agent has achieved what previous tools (e.g. [15][25][26]) have not – displaying an approximate of NNS status of understanding without imposing extra cognitive load.

7.2 NNS Participation and Other Social Psychological Processes

The statistical results suggested that NNS active participation was higher with the high-level agent than either with the low-level agent or without agent. This could be partially explained by the increase in NNS understanding after the high-level agent interruption. Ease in understanding may have led to a temporal liberation of cognitive resources [56] for NNS to allocate to formulating their own argument, or even to asking NS to clarify. As evidenced by the statistical result on NNS motivation to request clarification (see Section 5.3) as well as NNS self-reports (see Section 6.3), NNS took actions to ask their own questions or prompt NS for more information following the high-level agent setting the norm. We therefore speculate that some kind of social learning [1] or imitation processes may have occurred, at least with the high-level agent and for those NNS who have great self-efficacy [1]. Some NNS reported how they observed and recalled the agent asking questions, how they observed NS explaining to the agent patiently and thoroughly, and how they modeled the agent's question asking behavior. It is also true that not all

NNS believed they could interrupt and ask questions as the agent did. After all, those who have low self-efficacy are less likely to adopt the observed behavior [1]. To encourage NNS to ask more questions by setting a model, future work should also look at ways to enhance NNS self-efficacy, for example, through creating mastery experience [1]. These findings indicate a need for future work to explore whether and how social psychological processes such as social learning could occur between human and conversation agents.

With the low-level agent where NNS understanding did not significantly improve, NNS also had higher active participation than without agent (albeit not higher than with the high-level agent). One explanation could be: NS clarifications on simple words, though did not directly improve understanding, brought the conversation to a pause such that NNS could take advantage to sort out information that is otherwise overwhelming (e.g. 2C). This served just as the clarifications elicited by the high-level agent, to liberate NNS cognitive resources to allocate to speaking. Drawing on the statistical results of NNS self-evaluation, we offer an alternative explanation for NNS active participation with the low-level agent. In line with the downward social comparison proposition [21] and empirical findings (e.g. [7]), NNS evaluated the highest of their own communication ability in the low-level agent condition. This heightened self-evaluation may have led to more willingness to communicate in foreign language [38]. However, to empirically test this, one would need to set up a between-subjects experiment such that self-evaluation, willingness to communicate and actual participation are measured at sequential time points.

Additionally, for some NNS, the presence of the agent (regardless of type) made "it (feel) more like a fair game with NS" (20C). The feeling of having "another person speaking at the same level" (6C) may have rebalanced the power imbalance induced by the disparity and asymmetry in language proficiency between NS and NNS [32][59]. This power re-balance with the presence of the agent in turn may have essentially altered the social dynamic for NS and NNS in their conversation, which was identified by Macintyre et al. [37] as another factor influencing NNS willingness to communicate.

7.3 Trade-offs between the High- and Low-Level Agent

It is without doubt that the high-level agent outperformed the low-level agent in improving NNS understanding and active participation at least for video-mediated communication. But the low-level agent did not require a wordlist with difficulty index. It worked just as an agent picking up on random words and it also improved NNS active participation. Is it worth implementing a sophisticated high-level agent that eats up much system memory for video-mediated meetings for all kinds of multilingual teams? Or a less sophisticated agent without the word index would do just as well? Further, the word index used in the study might not match the need of NNS of every level and could become outdated as new Internet slang evolves every day. Is it worth the customization and the maintenance of an ever-evolving word index? We examine the trade-offs between the high- versus low-level agent with respect to the type of task that the agent is intended to support.

We used a negotiation task for our study where the primary purpose was to convince others with compelling arguments. NS participants "default to higher-level English" (21B) to sound more persuasive. The conversation could fill with big words that likely elude NNS. Without fully understanding NS perspective, it would be hard for NNS to dispute or build on others' arguments.

In such cases, insufficient understanding prevents NNS from full participation and engagement. Our study suggests that the high-level agent may work the best for such cases as it can address low participation through improving NNS understanding. On the other hand, for tasks such as idea generation and sharing, where NNS participation does not hinge on their understanding of NS opinions, the low-level agent or an agent that does not require sophisticated programming, one that asks random or generic questions would be enough to create a scenario for social comparison to take place, which in turn might encourage NNS to contribute more as a result of heightened self-concept [51] and increased willingness to communicate [38].

8 DESIGN IMPLICATIONS

8.1 A Clarification Companion for NS

Our findings suggest that despite their awareness to avoid using big words, NS had difficulties to level down vocabulary especially in formal meetings discussing important matters. One way to address this difficulty is for the system to offer NS suggestions of synonyms or alternative expressions in a discreet manner. For instance, when the agent picks up on a difficult word and requests clarification, the system already has pulled out alternative words, phrases and expressions from a wide range of sources and suggested to NS on their screen. Each alternative can have an index number/ranking indicating its level of difficulty for NNS, which could be proxied by frequency of use or other crowdsourcing methods (i.e. crowdsourcing frequency of look-ups by NNS, their familiarity with the expression, etc. as suggested in [16]). The NS could choose an alternative expression based on their assessment of their NNS partner's level. Or even better, with the advancement of natural language processing and machine learning, the system could suggest the expression at the appropriate level for the NNS based on its automatic assessment of NNS English level within a few exchanges in the conversation.

8.2 Aligning Agent's Clarification Request and NNS Needs

To better align the agent's question and NNS confusion, the system could allow NNS to control the timing of the agent's interruption. Ideally, when the NNS clicks to indicate confusion, it sets the system to search what has been said a few seconds prior to the click, for words, expressions or even cultural references of high difficulty or low frequency of use. Alternatively, the system could detect NNS facial expression for a confused look such as frowning, as a signal to activate the agent. This will save NNS effort and cognitive resources to pinpoint their confusion.

8.3 Encouraging Clarification Request through Setting a More Appropriate Model

Our study suggested that the low-level agent achieved the purpose of encouraging NNS to voluntarily speak more. This was partly through demonstrating the value of asking any (even dumb) questions, since the NS explained with patience despite their thinking the low-level agent was inappropriate. To improve this, we suppose that an agent asking generic questions such as "could you please elaborate more on that?", "can you give an example?" at appropriate timing could achieve the same purpose of setting the norm of question-asking for NNS to model, while without sounding inappropriate to NS. This will not require a sophisticated word index. The timing of activating the agent could be determined by detecting NS speech rate. Fast speech rate may indicate a need to clarify, as NNS are often left behind when NS speak too quickly [14].

9 LIMITATIONS AND FUTURE WORK

The findings of this study need to be interpreted with at least two limitations in mind. We discuss the limitations regarding our choice of task and the representativeness of our sample. We also lay out directions for future work to explore.

First, we deliberately chose the task to simulate global collaboration in formal settings in real world, where NS tend to use big words and complicated concepts to make arguments. While our two types of agent did achieve the goal to make NS clarify more, they did find it hard to explain complicated concepts and to tune down the level of word they used when discussing global issues. Future research may explore the effect of task and type of group interaction on the ease of leveling down vocabulary using different types of agent. For instance, if participants were to perform a task where the team performance hinges on how well members understand each other and the amount of information shared among members (e.g. problem-solving or hidden profile task) instead of advocating one's own opinion and persuasion, how might a clarification agent affect the communication.

Second, our sample of NS participants was not representative of an average NS living in their home country. For instance, as a NNS (of Japanese) living in a foreign country themselves, they may empathize more with NNS of English than an average NS. Additionally, a majority of them have lived in Japan for a long period of time and are familiar with Japanese culture. They are used to and are understanding about NNS not asking questions even when they do not follow. These people might have been more sensitive to the purpose of the clarification agent than an average NS. This limitation leaves room for future research to explore how an average NS living in their home country would react to the clarification agent.

10 CONCLUSION

There has been an increasing interest in tackling the grounding difficulty arising from the language disparity between NS and NNS in multilingual teams. In this study, we sought to address the challenge that CSCW researchers has faced in designing such tools – to provide the aid as-needed without posing extra cognitive load and face threat on NNS. Our approach transfers the face threat and the formulation effort of requesting clarifications to a conversation agent that asks clarification questions "on behalf of NNS". We conducted an experiment in which 2 NS and 1 NNS collaborated over videoconferencing on a series of negotiation tasks under three conditions (with the high-level clarification agent that asks NS to provide clarifications of difficult words, vs. the low-level agent that asks NS to provide clarifications of easy words vs. without agent). We found that with the high-level agent, NS provided the most clarifications that improved NNS understanding, which in turn led to NNS more active participation. Our findings suggest the potential of a conversation agent to support multilingual collaboration through sociopsychological processes that, to our knowledge, have not been examined in the contexts of agent-supported human interaction or human-agent interaction. Our work opens up a space for investigation into the nature and the dynamics of human-agent interaction on a deeper level and extends the CASA (Computers-as-Social-Actors) paradigm [47] by inquiring if and under what circumstances do human perceive conversational agents as proxies of their own kind such that the much complicated cognitive, behavioral and social processes such as social learning and social comparison might take place in human-agent communicative events, and that propositions of

such social psychological theories might apply. With these inquiries, we hope to inspire more investigation into the nature and the dynamics of such human-agent interactions.

ACKNOWLEDGMENTS

We thank all our participants for their time and patience, and our undergraduate research assistants Mind Apivessa and Jintana Cunningham for their graphic design. We also thank NTT software developer Kobayashi-san for making things work.

REFERENCES

[1] Albert Bandura. 1977. *Social Learning Theory*. Englewood Cliffs, NJ: Prentice-Hall.

[2] Albert Bandura. 1989. Human Agency in Social Cognitive Theory. *American Psychologist.* 44, 9: 1175–1184. doi:10.1037/0003-066X.44.9.1175.

[3] Melissa Baralt and Laura Gurzynski-Weiss. 2011. Comparing learners' state anxiety during task-based interaction in computer-mediated and face-to-face communication. *Language Teaching Research* 15, 2: 201–229. https://doi.org/10.1177/0265532210388717

[4] Hart Blanton. 2013. Evaluating the self in the context of another: The three-selves model of social comparison assimilation and contrast. *Cognitive social psychology*. Psychology Press: 79–91.

[5] Xun Cao, Naomi Yamashita, Toru Ishida. 2016. Investigating the impact of automated transcripts on non-native speakers' listening comprehension. *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo Japan, Oct. 2016), 121–128. https://doi.org/10.1145/2993148.2993161

[6] Jean Carletta,, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, Association for Computational Linguistics.

[7] Patrick Chambres and Pierre -Jean Marescaux. 1998. Fictitious social position of competence, and performance in a foreign-language interaction situation: An experimental approach. *European Journal of Psychology of Education* 13, 3: 411–430.

[8] Herbert H. Clark and Susan E. Brennan. 2001. Grounding in communication. In L.B. Resnick, J.M. Levine & S.D. Teasley (Eds.) *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington, DC: American Psychological Association.

[9] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, *13*(2), 259–294. DOI: https://doi.org/10.1016/0364-0213(89)90008-6.

[10] Sunny Consolvo, David W. McDonald, and James A. Landay. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. *Proceedings of SIGCHI Conference on Human Factors in Computing Systems . ACM.* https://doi.org/10.1145/1518701.1518766

[11] Peter. J. Costa, (2014). Truncated outlier filtering. *Journal of Biopharmaceutical Statistics, 24*(5), 1115-1129. https://doi.org/10.1080/10543406.2014.926366

[12] Pino Cutrone. 2014. A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners. *Intercultural Pragmatics, 11*(1): 83–120. https://doi.org/10.1515/ip-2014-0004

[13] Joan Morris DiMicco, Anna Pandolfo, and Walter Bender. 2004. Influencing group participation with a shared display. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work (CSCW '04).* ACM, New York, NY, USA, 614-623. DOI: http://dx.doi.org/10.1145/1031607.1031713

[14] Wen Duan, Naomi Yamashita, and Susan R. Fussell. 2019. Increasing Native Speakers' Awareness of the Need to Slow Down in Multilingual Conversations Using a Real-Time Speech Speedometer. *Proc. ACM Hum.-Comput. Interact. 3, CSCW*, Article 171 (November 2019), 25 pages. DOI: https://doi.org/10.1145/3359273

[15] Wen Duan, Naomi Yamashita, Sun Young Hwang, and Susan R. Fussell. 2018. "Let Me Ask Them to Clarify If You Don' t Want To"— A Clarification Agent for Nonnative Speakers. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*: 1–6. https://doi.org/10.1145/3170427.3188600

[16] Wen Duan and Susan R. Fussell. 2021. Understanding and Identifying Design Opportunities for Facilitating Humorous Interactions in Multilingual Multicultural Contexts. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan, May 2021), 1–6. https://doi.org/10.1145/3411763.3451668

[17] Echenique, A. et al. 2014. Effects of video and text support on grounding in multilingual multiparty audio conferencing. In *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology - CABS '14* (Kyoto, Japan, 2014), 73–81. https://doi.org/10.1145/2631488.2631497

[18] Arash Eshghi. and Patrick G.T. Healey. 2016. Collective Contexts in Conversation: Grounding by Proxy. *Cognitive Science*. 40, 2 (Mar. 2016), 299–324. DOI: https://doi.org/10.1111/cogs.12225.

[19] Russell H. Fazio. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Review of personality and social psychology: Vol. 11. Research methods in personality and social psychology* (pp. 74–97). Newbury Park, CA: Sage.

[20] Alan J. Feely and Anne-Wil Harzing. 2003. Language management in multinational companies. *Cross-Cultural Management: An International Journal:* 10, 2: 37–52. https://doi.org/10.1108/13527600310797586

[21] Leon Festinger. 1954. A theory of social comparison processes. *Human Relations*, 7: 117–140. DOI: 10.1177/001872675400700202

[22] BJ Fogg and Clifford Nass. 1997. How Users Reciprocate to Computers: An experiment that demonstrates behavior change. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '97*. https://doi.org/10.1145/1120212.1120419

[23] Cindy Gallois, Tania Ogay, and Howard Giles. 2005. Communication accommodation theory: A look back and a look ahead. In *W.B. Gudykunst (Ed.) Theorizing about Intercultural Communication*: 121-148. Thousand Oaks: Sage.

[24] Ge Gao, Bin Xu, David Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One : Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. *CSCW '15 Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*: 852–863. https://doi.org/10.1145/2675133.2675197

[25] Ge Gao, Naomi Yamashita, Ari Hautasaari, Andy Echenique, and Susan R. Fussell. 2014. Effects of public vs. private automated transcripts on multiparty communication between native and non-native English speakers. *Proceedings of the 2014 ACM Conference on Human Factors in Computing Systems*: 843–852. https://doi.org/10.1145/2556288.2557303

[26] Ge Gao, Naomi Yamashita, Ari Hautasaari, and Susan R. Fussell. 2015. Improving multilingual collaboration by displaying how non-native speakers use automated transcripts and bilingual dictionaries. In *Proc. CHI 2015*, 3463–3472. https://doi.org/10.1145/2702123.2702498

[27] J. P. Gerber, Ladd Wheeler, and Jerry Suls. 2018. A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin* 144, 2: 177–197. https://doi.org/10.1037/bul0000127

[28] Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: strategies for qualitative research*. Aldine Publishing Company, Hawthorne, NY.

[29] Sandra G. Hart and Lowell E. Staveland. 1998. Development of NASA-TLX: Results of empirical and theoretical research. *Advances in Psychology*, *52*: 139–183.

[30] Anne-Wil Harzing and Alan J. Feely. 2008. The language barrier and its implications for HQ-subsidiary relationships. *Cross Cultural Management: An International Journal, 15*(1): 49–61. https://doi.org/10.1108/13527600810848827

[31] Helen Ai He, Naomi Yamashita, Chat Wacharamanotham, Andrea B. Horn, Jenny Schmid, and Elaine M. Huang. 2017. Two Sides to Every Story: Mitigating Intercultural Conflict through Automated Feedback and Shared Self-Reflections in Global Virtual Teams. *PACM Human-Computer Interaction CSCW*, Article 51 (December 2017) 1: 21 pages. DOI: https://doi.org/10.1145/3134686

[32] Pamela J. Hinds, Tsedal B. Neeley, and Catherine Durnell Cramton. 2014. Language as a lightning rod: Power contests, emotion regulation, and subgroup dynamics in global teams. *Journal of International Business Studies, 45*(5): 536–561. https://doi.org/10.1057/jibs.2013.62

[33] Isbister, K. et al. 2000. Helper agent: designing an assistant for human-human interaction in a virtual meeting space. *Proceedings of the SIGCHI conference on Human factors in computing systems CHI '00* (The Hague, The Netherlands, 2000), 57–64. https://doi.org/10.1145/332040.332407

[34] Kim, S. et al. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, Apr. 2020), 1–13. https://doi.org/10.1145/3313831.3376785

[35] Han Z. Li. 1999. Grounding and information communication in intercultural and intracultural dyadic discourse. *Discourse Processes* 28, 3: 195–215. https://doi.org/10.1080/01638539909545081

[36] Han Z. Li, Young-ok Yum, Robin Yates, Laura Aguilera, Ying Mao, and Yue Zheng. 2005. Interruption and Involvement

in Discourse: Can Intercultural Interlocutors be Trained? *Journal of Intercultural Communication Research, 34*(4): 233–254.

[37] Peter D. Macintyre, Richard Clément, Zoltán Dörnyei, and Kimberly A. Noels. 1998. Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation. *The Modern Language Journal* 82, 4: 545–562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

[38] Peter D. Macintyre. 2007. Willingness to Communicate in the Second Language: Understanding the Decision to Speak as a Volitional Process. *The Modern Language Journal* 91, 4: 564–576. https://doi.org/10.1111/j.1540-4781.2007.00623.x

[39] Joseph Edward McGrath. 1984. *Groups: interaction and performance*. Prentice-Hall, Englewood Cliffs, NJ.

[40] Brian E. Mennecke, Joseph S. Valacich, and Bradley C. Wheeler. 2000. The Effects of Media and Task on User Performance: A Test of the Task-Media Fit Hypothesis. *Group Decision and Negotiation, 9*(6): 507-529. https://doi.org/10.1023/A:1008770106779

[41] Sarah Mercer. 2011. *Towards an understanding of language learner self-concept*. Springer, Dordrecht.

[42] Nicole Mills, Frank Pajares, and Carol Herron. 2006. A Reevaluation of the Role of Anxiety: Self-Efficacy, Anxiety, and Their Relation to Reading and Listening Proficiency. *Foreign Language Annals* 39, 2: 276–295. https://doi.org/10.1111/j.1944-9720.2006.tb02266.x

[43] Clifford Nass, B.J. Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6: 669–678. https://doi.org/10.1006/ijhc.1996.0073

[44] Clifford Nass and Eun-Ju Lee. 1998. Does the ethnicity of a computer agent matter? An experimental comparison of human-computer interaction and computer-mediated communication. *Proceedings of the 1998 Workshop on Embodied Conversational Characters*.

[45] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. *Journal of Applied Social Psychology* 29, 5: 1093–1109. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x

[46] Clifford Nass, Youngme Moon, B. J. Fogg, Byron Reeves, and Chris Dryer. 1995. Can computer personalities be human personalities? *Conference companion on Human factors in computing systems - CHI '95*, ACM Press, 228–229. https://doi.org/10.1006/ijhc.1995.1042

[47] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '94*, ACM.

[48] Tsedal B. Neeley, Pamela J. Hinds, and Catherine D. Cramton. 2012. The (Un)Hidden Turmoil of Language in Global Collaboration. *Organizational Dynamics* 41, 3: 236–244. DOI:10.1016/j.orgdyn.2012.03.008.

[49] Mei-hua Pan, Naomi Yamashita, and Hao-chuan Wang. 2017. Task Rebalancing : Improving Multilingual Communication with Native Speakers - Generated Highlights on Automated Transcripts. *Proc. CSCW 2017*: 310–321. https://doi.org/10.1145/2998181.2998304

[50] Aveni Valerie A. Pellegrino. 2005. *Study Abroad and Second Language Use: Constructing the Self*. Cambridge University Press.

[51] Damian J. Rivers. 2011. Evaluating the self and the other: Imagined intercultural contact within a 'native-speaker' dependent foreign language context. *International Journal of Intercultural Relations* 35, 6: 842–852. https://doi.org/10.1016/j.ijintrel.2011.08.003

[52] Pamela Rogerson-Revell. 2008. Participation and performance in international business meetings. *English for Specific Purposes 27*, 3: 338–360. https://doi.org/10.1016/j.esp.2008.02.003

[53] Peter Henry Rossi, & Steven L. Nock. (1982). *Measuring social judgments: The factorial survey approach*. SAGE publications, Incorporated.

[54] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language, 50*(4), 696. DOI: https://doi.org/10.2307/412243.

[55] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond Dyadic Interactions: Considering Chatbots as Community Members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 450. https://doi.org/10.1145/3290605.3300680

[56] Leslie D. Setlock, and Susan R. Fussell. 2010. What's it worth to you?: the costs and affordances of CMC tools to asian and american users. *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10* (Savannah, Georgia, USA, 2010), 341-350. https://doi.org/10.1145/1718918.1718979

[57] Yohtaro Takano and A. Noda. 1993. A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology, 24*, 4: 445–462. https://doi.org/10.1177/0022022193244005

[58] Yohtaro Takano and Akiko Noda. 1995. Interlanguage Dissimilarity Enhances the Decline of Thinking Ability During Foreign Language Processing. *Language Learning* 45, 4: 657–681. https://doi.org/10.1111/j.1467-1770.1995.tb00457.x.

[59] Miyuki Takino. 2017. Power in International Business Communication and Linguistic Competence: Analyzing the Experiences of Nonnative Business People Who Use English as a Business Lingua Franca (BELF). *International Journal of Business Communication*: DOI: 10.1177/232948841771422.

[60] Zsuzsa Tóth. 2011. Foreign language anxiety and advanced EFL learners: an interview study. *WoPaLP*, *5*(19).

[61] Traeger, M.L. et al. 2020. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*. (Mar. 2020), 201910402. DOI: https://doi.org/10.1073/pnas.1910402117.

[62] Hao-Chuan Wang and Susan R. Fussell. 2010. Groups in groups: conversational similarity in online multicultural multiparty brainstorming. *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10* (Savannah, Georgia, USA, 2010), 351-360. https://doi.org/10.1145/1718918.1718980.

[63] Jean. Wong. 2000. The token "yeah" in nonnative speaker English conversation. *Research on Language and Social Interaction 33*, 1: 39–67. https://doi.org/10.1207/S15327973RLSI3301_2.

[64] Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving machine translation by showing two outputs. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 3743–3746. https://doi.org/10.1145/2556288.2557171

[65] Naomi Yamashita, Andy Echenique, Toru Ishida, and Ari Hautasaari. 2013. Lost in Transmittance: How Transmission Lag Enhances and Deteriorates Multilingual Collaboration. *ACM Conference on Computer Supported Cooperative Work Social Computing Social Computing*: 923–934. https://doi.org/10.1145/2441776.2441881

[66] Chien Wen Yuan, Leslie D. Setlock, Dan Cosley, and Susan R. Fussell. 2013. Understanding Informal Communication in Multilingual Contexts. *In CSCW'13*: 909-922. https://doi.org/10.1145/2441776.2441880

Received January 2021; revised April 2021; accepted July 2021.