# Task Rebalancing: Improving Multilingual Communication with Native Speakers-Generated Highlights on Automated Transcripts

**Mei-Hua Pan[1], Naomi Yamashita[3], Hao-Chuan Wang[1,2]**

[1]Department of Computer Science
[2]Institute of Information Systems and Applications
National Tsing Hua University
101, Sec.2, Kuang-Fu Rd. Hsinchu 300, Taiwan
s103062509@m103.nthu.edu.tw
haochuan@cs.nthu.edu.tw

[3]NTT Communication Science Labs
2-4 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, Japan
naomiy@acm.org

## ABSTRACT

In multilingual communication through a common language among both native speakers (NS) and non-native speakers (NNS), NNS may encounter problems in comprehending the messages of NS or following conversations. Even though automated speech recognition (ASR) transcripts provide support to NNS, such transcripts may contain errors and impose the need to simultaneously listen and read. To reduce this burden, we propose adding another channel (i.e., highlighting) through which NS can help NNS by highlighting the critical parts of transcripts, thus making them more useful to NNS. In a laboratory study involving 14 triads (two NS and one NNS in each triad), participants engaged in collaborative discussions under two conditions: audio conferencing plus ASR transcripts with and without the highlighting function. NS showed various motivations to perform the extra task of highlighting. The highlighting efforts helped NS themselves focus on the discussion and enhanced their task performance while increasing the clarity and comfort perceived by NNS during communication. Having NS generating highlights can benefit both NS and NNS, but in different ways. We discuss the implications for research and design of multilingual collaborative work.

### Author Keywords

Automated speech recognition (ASR); multilingual communication; social annotation; highlighting tools; real-time transcripts

### ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

*CSCW '17, February 25-March 01, 2017, Portland, OR, USA*
© 2017 ACM. ISBN 978-1-4503-4335-0/17/03□$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998304

## INTRODUCTION

Since many organizations operate in distributed and international ways, multilingual communication is becoming more common, where people choose a common language (*lingual franca*), generally English, for collaboration [2, 19]. However, due to the lack of language proficiency, non-native speakers (NNS) often face significant difficulties during multilingual communication. For example, they are often overwhelmed with such multiple parallel processes as phonetic analysis, parsing ongoing conversations, and intensive thinking, all of which are typically accompanied by internal speech in their native language [6, 20, 22]. In conversations dominated by native speakers (NS) as a majority, NNS are left behind as discussions advance rapidly [15, 27].

Researchers in the HCI/CSCW field have developed technologies to assist NNS in multilingual communication. Among them, transcripts generated by automated speech recognition (ASR) technologies have been proven to be useful for improving the comprehension of NNS during audio conferencing [18]. However, research also shows that ASR transcripts impose extra burdens on NNS. These burdens increase as the transcript error rate increases or during delays in showing the transcripts. When the error rate and delay exceed a certain threshold, the ASR transcripts become simply a source of burden with little value to the NNS [28].

To avoid distracting NNS by ASR transcript errors/delays and to more efficiently use ASR transcripts, we propose adding another channel (i.e., highlighting) through which NS can provide help to NNS by highlighting the important parts of ASR transcripts. This idea frees NNS from following every word of a transcript; NNS do not have to give attention to the entire transcript if highlighting emphasizes its critical parts. Since NS do the highlighting, we assume the information is reliable and useful for NNS. It is important to note that the task loads of NS and NNS in conventional multilingual group communication are strongly unbalanced [4]. Even though NNS often suffer from difficulties due to language barriers, NS effortlessly handle the conversations [19, 20]. Thus, ideally, our additional channel will reduce the communication work of NNS by exploiting the underutilized cognitive resources of NS while improving the overall group communication quality.

We conducted a laboratory experiment to examine this idea. 14 triads of two NS and one NNS engaged in two collaborative discussions under different conditions: audio conferencing plus ASR transcripts with and without the highlighting function. In the *with-highlighting* condition, the two NS individually highlighted key points of ongoing conversations, simultaneously allowing NNS to see the highlighted sentences.

From a combination of quantitative and qualitative analyses of the experiment's data, we determined that the highlighting interface indeed increased the burden on NS, but at the same time it improved the task performance (remembering the agreements reached during the discussion) of the NS themselves. As expected, highlighting by the NS helped the NNS in multilingual communication by clarifying and simplifying the messages, and it also created a more relaxing cross-lingual interaction.

## BACKGROUND

In this section, we first introduce previous literature on the difficulties faced by NNS when interacting with NS as well as technologies intended to support NNS in multilingual conversations. We then focus on task rebalancing in the context of multilingual communication and discuss our design decisions (i.e., highlighting key discussion parts in automated transcripts by NS). Finally, we describe how the transcripts highlighted by NS can improve overall group communication quality.

### Difficulties of NNS in Multilingual Communication

Multilingual groups generally adopt a common language through which to communicate, requiring some members to communicate in a non-native language [2, 19]. Even though the diversity of a group's members might enhance its productivity, language barriers can also hinder effective group collaboration [24]. NNS face a number of difficulties when communicating in a second language, which might decrease the efficiency of group communication [27, 29]. During communication with NS, NNS are at a significant disadvantage due to their much higher cognitive load. They require more processing power and time to understand and follow streams of speech, not to mention generating proper and timely responses [19, 20]. Particularly in conversations held with a NS majority, discussions can move rapidly, leaving NNS far behind [15, 27].

### Current Technological Support

To support NNS in multilingual group conversations, researchers have explored various ways to improve their comprehension and ability to contribute. One line of such research explores the use of machine translation technology. For example, Wang et al. showed that machine translation facilitates NNS productivity by allowing them to express themselves in their native languages [24]. Furthermore, Gao et al. suggested highlighting keywords in machine-translated outputs so that NNS could focus on the important parts of the message without getting confused by translation errors [3]. Their experimental study showed that highlighting keywords on machine-translated sentences was indeed useful in enhancing NNS's comprehension.

Researchers have also explored automated speech recognition (ASR) technology as a means to support NNS in multilingual conversations. According to Pan et al., automatically generated transcripts improve NNS's understanding in one-way non-interactive scenarios if the transcript has little delay and few errors [18, 28]. For real-time audio conferencing, previous work showed that sharing automated transcripts among multilingual group members improved group communication quality. NS consciously spoke more clearly to reduce the error rate of the automated transcript, improving the communication quality as perceived by the NNS [4]. Displaying how NNS use automated transcripts in audio conferencing also improved the grounding of multilingual communication between NS and NNS, since this made the former aware of the problems encountered by the latter [1, 5].

While highlighted machine translations and automated transcripts are two useful methods for supporting the comprehension of NNS in multilingual communication, to date no work has investigated the effects of integrating the two approaches, i.e., using highlighted ASR-generated transcripts as communication support.

### Highlighting Automated Transcripts

Even though automated transcripts provide many benefits to multilingual conversations, they also add burdens and increase the loading of NNS. First, automated transcripts often contain errors. When recognition errors are present, NNS may require extra effort to understand the conversation, or misinterpretations might be caused [4]. Second, due to time delays between an utterance and the availability of the corresponding transcript, reading transcripts during a conversation requires multitasking, which can be cognitively demanding. Things could get worse when a native speaker speaks too fast or when the conversational content is dense, forcing the NNS partner to read a long transcript in real-time to glean any benefit from it.

To control the NNS task load and make ASR-generated transcripts more helpful, we adopt the technique of keyword highlighting [3, 25]. Highlighting is also a common strategy when people read, especially in academic or scientific contexts. From paper to digital media, people benefit from highlighting annotations for processing information of various kinds and genres [17, 21, 25].

We assume that highlighting critical words or utterances on automated transcripts would guide the attention of NNS to them, reducing the effort required compared to processing the entire transcript. Highlighting can also be used to filter ASR errors, reducing possible misreading or misinterpretation.

### Rebalancing NS/NNS Tasks with Highlighting

Numerous previous studies have identified the positive effects of highlighting key parts on cognitive information processing tasks such as reading, comprehension, and interaction [23, 25]. However, in implementing this transcript highlighting in multilingual communication, we must answer two critical questions: (1) Who is going to make the highlighting annotations? and (2) Would people be motivated to voluntarily make them?

In multilingual communication, the work of NS and NNS participants is unbalanced due the inherent gap of language proficiency. Therefore, we propose giving NS participants the responsibility of highlighting ASR-generated transcripts as a way to rebalance the workload of language processing. Consequently, we need to explore whether NS participants are motivated to voluntarily highlight transcripts when a highlighting tool is available to them.

We also need to explore whether key sentences highlighted by NS would help NNS to skim or to ignore the unimportant parts of the automated transcript in order to reduce their work in communication.

## RESEARCH QUESTION AND HYPOTHESES

In this study, we introduced the highlighting function to NS and explained how to highlight the transcripts. However, we did not force them to use it—it was up to the NS to use it during the study's communication tasks. Therefore, we asked the following research question:

*RQ1: Would NS be motivated to highlight transcripts? If so, what's the motivation?*

We also examined the impact of highlighting on multilingual communication. We first examined our main assumption: highlighting rebalances NS and NNS workloads in multilingual communication.

If NS take the lead, read the transcript, and decide which part is important or whether the group has agreed on a certain topic, they need to process more information than they are generally faced with. Since this situation increases their workload, they will probably engage more in the tasks and remember more commonalities when the highlighting tool is available.

On the other hand, for the NNS, when highlights are available, they can reduce their effort, compared to reading the entire automated transcript, and more easily or/and confidently understand the key points of the ongoing conversation. Such "extra" cognitive resources can be diverted to improve task performance. Based on the above line of thought, we hypothesized as follows:

*H1a. The NS workload will increase when the highlighting tool is available.*

*H1b. The NNS workload will decrease when the highlighting tool is available.*

*H2a. NS will remember more agreements reached during discussion when the highlighting tool is available.*

*H2b. NNS will remember more agreements reached during discussion when the highlighting tool is available.*

In terms of group communication quality, highlighting might serve as an additional backchannel for NS to emphasize their points and for NNS to confirm whether they focused on the correct parts of conversations. If NS highlight the critical parts of transcripts to explicitly indicate the points on which they believe all group members agree, we assume that the discussion would be less ambiguous, less confusing, and thus smoother. To summarize, we hypothesize:

*H3a. Participants will perceive better group communication quality in terms of comfort.*

*H3b. Participants will perceive better group communication quality in terms of message clarity.*

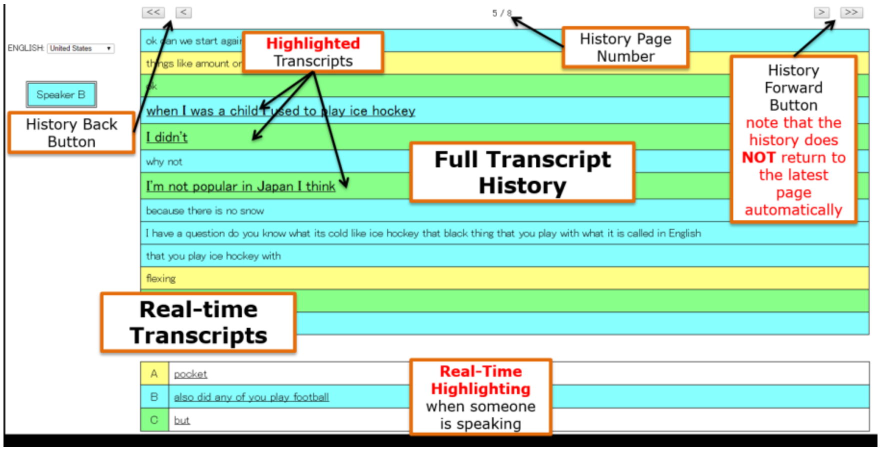
## METHOD

To answer the research question and test the hypotheses, we conducted a within-subject experiment that compared ASR-transcript-based multilingual communication support with and without the highlighting feature. 14 triads participated in a commonality-finding task. Each triad consisted of two native English speakers and one Japanese non-native English speaker. The participants were asked to use English as the common language in their audio conferencing discussions.

Participants were required to talk about their background and experience on a given topic: childhood memories or adulthood experience. They were also asked to identify as many common experiences that are shared by two or more members of the group as possible. There were two conditions: *with-highlighting* and *without-highlighting* (baseline). In both conditions, all of the triad participants could see the ASR transcript. In the *with-highlighting* condition, NS could highlight sentences by clicking on them, and the highlights were made visible to the NNS (Figure 1). Each group went through two sessions and completed two commonality-finding tasks. The orders of the conditions and discussion topics were counterbalanced.

### Participants

We recruited 42 participants (28 native English speakers and 14 non-native English speakers) using a personnel recruiting company. The study was advertised to participants as an exploration of designing new technologies for supporting group work. Participants were randomly assigned to triads of two NS and one NNS. There were 14 triads in total.



ok can we start agai

things like amount or

ok

when I was a child I used to play ice hockey

I didn't

why not

I'm not popular in Japan I think

because there is no snow

I have a question do you know what its cold like ice hockey that black thing that you play with what it is called in English

that you play ice hockey with

flexing

A pocket

B also did any of you play football

C but

**Figure 1. Transcript interface for all participants. Underlined sentences displayed in real-time transcript area indicate who is talking. Utterances of different speakers are distinguished by different background colors. When NS clicks on a sentence, it becomes enlarged and underlined.**

The 28 native English speaker participants (10 females) lived in Japan during the study. They all grew up and received the vast majority of their education in English-speaking countries. Their mean age was 40.89 (SD = 12.03). They reported little experience using audio conferencing software such as Skype for multiparty communication (M = 3.36, SD = 1.91 on a 7-point Likert scale from 1 = never to 7 = very often). They also reported little experience with ASR technology (M = 1.79, SD = 0.92 on a 7-point Likert scale from 1 = never to 7 = very often). They were all relatively familiar with communicating with non-native English speakers (M = 6.57, SD = 0.84 on a 7-point Likert scale from 1 = never to 7 = very often).

The remaining 14 participants were native Japanese speakers who grew up and received the vast majority of their education in Japan. Their mean age was 24.5 (SD = 6.17). The mean of their TOEIC English proficiency test (Test of English for International Communication) scores was 806.25 (SD = 58.55, min = 730), indicating that they were relatively highly proficient in English. However, the participants self-reported themselves as not being fluent English speakers (M = 4, SD = 1.18 on a 7-point Likert scale from 1 = not fluent at all to 7 = very fluent). They reported little experience with audio conferencing tools such as Skype (M = 2.71, SD = 1.54 on a 7-point Likert scale from 1 = never to 7 = very often). They also reported little experience using ASR technology (M = 2, SD = 1.52 on a 7-point Likert scale from 1 = never to 7 = very often).

### Materials

*Task*. Building shared knowledge or common ground is central to team collaboration and group work. Thus it is a common, universal need for members of an international workgroup to engage in communication aimed at establishing mutual understanding [1, 12, 14]. Finding commonalities is one kind of task in such a scenario. In this study, we designed a commonality-finding task in which we asked participants working in the same group to discover shared experiences or memories. We designed the task in such a way that (1) each member could contribute to the conversation freely, (2) members could participate with no *a priori* knowledge, and (3) members needed to build common ground.

We asked participants to discover as many commonalities as possible in 15 minutes. We also told them that their group performance would be evaluated by the number of commonalities remembered by the group's members. Therefore, to achieve high performance, all of the members in the group had to talk openly about themselves and move on quickly so that they could find as many commonalities as possible during the discussion. We gave the participants two topics to discuss, one in each session: *childhood memories* (before high school) and *adulthood experiences* (after high school). The order of the topics was counterbalanced. After each discussion, we asked them to write down the commonalities they remembered.

The task requires the members to build common ground on a number of things. This is a common feature in global meetings where team members discuss and reach agreement on multiple issues in a single meeting. However, when the team composition is majority NS and minority NNS, the conversation can move forward rapidly between NSs while NNS are left behind, missing important points that were discussed during the meeting. Such a situation can be especially challenging for NNS, who need to understand and follow the conversation closely to remember the commonalities that were grounded during the conversation. Indeed, previous literature shows that NNS often easily forget the things they have heard because they become overburdened by multitasking [6]. For example, their memory space often gets occupied by processing language and contents [20]. In our study, we were interested in evoking such a situation and testing the feasibility of using NS-generated annotations to offload NNS' cognitive burden.

*Survey.* Before the experiment sessions started, participants completed an online survey of their demographic information (age, gender, nationality, native language). The survey also asked them to rate on 7-point Likert scales their English fluency, familiarity with talking to NS/NNS speakers, experience with audio conferencing tools, and experience with ASR transcription tools.

After each session, participants filled out another online survey, which consisted of a manipulation check (whether they noticed the highlighting tool), NASA Task Load Index [8], adapted questions from Liu et al.'s Quality of Communication Experience (QCE) scale [16], and questions about their attitude toward our highlighting function.

*Answer Sheet.* After each session, participants wrote down all of the commonalities they found during the previous discussion. First, they only wrote them down from memory independently. Note that the list of commonalities serves as a rough measure of each member's comprehension and burden level. Then they checked the conversation's transcript and submitted another answer sheet that included those commonalities they had forgotten.

*Interview.* At the end of the experiment, we held open-ended interviews with the participants in their native languages. We asked for their opinions of the ASR transcript interface and the highlighting tool and whether the presence of the tools changed how they spoke during the study.

### Software and Equipment

*Speech Recognition Tool.* In this experiment, we use Google Web Speech API as the ASR service to automatically transcribe participants' speech into text in real-time.

*Transcript Interface*. While the participants were talking, the transcript generated by the ASR service was displayed on the interface we designed for this study (see Figure 1). The interface had two main components: full transcript history and real-time transcripts. In the highlighting condition, NS could click on both of the transcripts (history and real-time) to highlight the sentences they considered important.

### Procedure

Participants were assigned to triads consisting of two NS and one NNS and directed separate rooms. The experimenter introduced the study and explained the task and the transcript interface. For NS, we explained how to highlight the transcripts and some possible usage – for later reference when filling into the answer sheets after the conversation, for referring them later in the conversation, for helping NNS follow the conversation. However, we did not force them to use it. The NSs were told that the highlighting was an optional function and that they could decide themselves whether to use it or not. For NNS, we explained the mechanism of the highlight function.

After all the triad participants had finished the demographic survey, they put on headset microphones. The experimenter gave instructions about the task by audio conferencing during the rest of the experiment.

Depending on the pre-planned counterbalanced order, participants started with either the *with-highlighting* or the *without-highlighting* condition. After a 15-minute discussion, participants wrote down the commonalities they remembered. Then they checked the conversation history to submit another answer sheet that included the commonalities they had forgotten. After finishing the experiment's two sessions, the experimenter conducted a semi-structured interview with all of the participants.

## MEASURES

We evaluated our highlighting tool with two types of measures: participants' perceived experience of work and collaboration (subjective) and objective measures of their work performance.

### Manipulation Checks

*Language Proficiency of group members.* We used a single-choice question to examine each participant's perception of the language proficiency of their partners. The question asked them to identify whether they talked to two native English speakers, one non-native and one native English speaker, or two non-native speakers.

*Accessibility to Highlighting Function.* We used a single-choice question to examine the participant's perception of accessibility to the highlighting function. The question asked them to identify whether they discussed under automated transcripts with or without highlighting function.

### Measurement of NS' Motivation

We measured the possible motivations for the NS to make highlights with three 7-point Likert subscales: *I highlighted parts of conversation because (1) I thought I needed to remember them, (2) I thought I would need to refer to them later in the conversation, (3) I thought my non-native partner would find them useful.* The first and second items are motivations to highlight for themselves, while the third item is a motivation to highlight for the NNS. In the post-task interview, we further asked NS if they had any other motivation to highlight text segments.

### Measurement of Perceptions and Experience

*Workload.* We adopted five questions (mental demand, temporal demand, performance, effort, frustration) from NASA-TLX [8] to measure the workload of the participants during each session. The questions were highly reliable (Crobach's α = .85) and were averaged to provide an overall workload score.

*Group Communication Quality.* We composed a measure of the group communication quality (GCQ) from two subscales: clarity and comfort. We slightly modified the original question to match the scenario of this experiment. The clarity subscale consisted of four items (α = .82), and the comfort subscale also consisted of four items (α = .90).

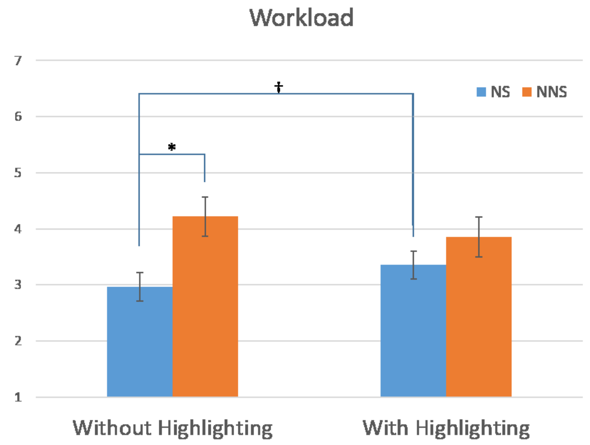


**Figure 2. Workload by condition for NS and NNS (error bars represent standard errors and mean)**
*** *p*<.05, †=.07**

The questions in each subscale were averaged to provide an overall score for that scale.

### Performance Measures

*Highlighting Percentage.* To understand whether the NS were motivated to provide highlights for the NNS, we calculated the highlighting percentage as follows:

*Highlighting Percentage = #highlighted utterances / #all utterances.*

*All Identified Commonalities.* To determine the number of commonalities found by each group, we put up a non-redundant set of commonalities that a group's members identified during the discussion.

*Recalled Commonalities.* We evaluated how well the participants remembered the group commonalities they found by calculating the ratio of the number of commonalities they remembered to the number of all commonalities found by the group:

*Recalled commonalities = #remembered commonalities / #all identified commonalities.*

## RESULTS

To test our hypotheses, we conducted a 2 (condition: with-highlighting vs. without-highlighting) × 2 (language background: NS vs. NNS) mixed-model ANOVA.

### Manipulation Checks

Our manipulation checks on the perception of group members' language proficiency and the accessibility to the highlighting function showed that both manipulations were successful. All participants could precisely perceive the language proficiency of their partners (100%) and their accessibility to the highlighting function (100%).

### Motivation of NS

To answer RQ1, we asked why the NS made highlights. The NS reported that the greatest motivation for making highlights is to use their highlights to help their non-native partners (M = 5.11, SD = 1.9). We found moderate or limited motivation to highlight for themselves, either in considering the possibility of referring to the key points in the subsequent conversation (M = 4.18, SD = 1.83) or for their personal needs of remembering (M = 3.82, SD = 1.85).

We further calculated the percentage of the transcript content highlighted by the NS to determine the empirical usage of the highlighting tool. As expected, some NS highlighted more content than others. The mean percentage of content highlighting by NS was 16.41% (SD = 8.53%, max = 32.68%, min = 4%).

To test the actual effects of highlighting on NNS, we excluded groups whose highlighting percentage was lower than the bottom quartile (9.01%) among all groups. 11 groups remained, and we included only them in the following analysis.

| Motivation | Mean | SD |
|---|---|---|
| For NNS | 5.11 | 1.9 |
| For reference | 4.18 | 1.83 |
| For remembering | 3.82 | 1.85 |

**Table 1 Mean and standard variance of NS' motivation to make highlights**

### Task Rebalancing

The first set of our hypotheses addresses our main assumption: NS highlighting rebalances workloads in multilingual communication. We hypothesized that the NS workload would increase while the NNS workload would decrease, closing the gap between the two parties. We also hypothesized that both the NS and the NNS would have better task performance, i.e., remembering more commonalities.

*Workload.* We conducted 2×2 mixed-model ANOVAs on the workload scores (Figure 2). There was a significant interaction effect between the highlighting function and the participants' language background (*F*[1, 31] = 4.20, *p* <.05). The main effect of the highlighting function was not significant (*F<1, n.s.*), but the main effect of the participants' language background was significant (*F*[1, 21] = 6.47, *p* <.05). As predicted, NS workloads were marginally higher in the highlighting condition (M = 3.35, SD = 1.00) than in the baseline condition (M = 2.96, SD = 0.98): *F*[1, 31] = 3.38, *p*=.07). The NNS workload remained about the same in the highlighting condition (M = 3.85, SD = 1.494) and in the baseline condition (M = 4.22, SD = 1.437), although there was a trend for NNS workloads to be reduced by highlighting: *F* [1, 31] = 1.46, *p*=0.23.

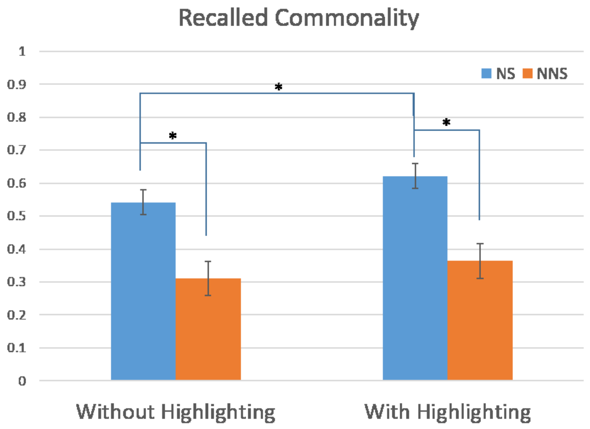


**Figure 3. Recalled commonalities by condition for NS and NNS (error bars represent standard errors). * *p*<.05**

Consequently, H1a was partially supported but H1b was not. Note that we also found that language background had a main effect on the workload in the baseline condition (*F* [1, 44.12] = 8.43, *p* < .01), but no effect on the workload in the highlighting condition *(F* [1, 44.12] = 1.34, *p* = 0.253). NS highlighting on the transcripts rebalanced the workload between NS and NNS, equalizing the effort taken by the two sides.

*Recalled Commonality.* We conducted 2×2 mixed-model ANOVAs on the recalled commonalities (Figure 3). The main effects of the highlighting function (*F* [1, 29.24] = 5.94, *p*<.05) and the language background of the participants (*F* [1, 29.89] = 17.47, *p*<.001) were all significant. The NS significantly recalled more commonalities in the highlighting condition (M = 0.61, SD = 0.17) than in the baseline condition (M = 0.54, SD = 0.20): *F* [1, 29.51] = 6.31, *p* = .01). The NNS also tended to recall more commonalities, but the difference between the highlighting condition (M = 0.36, SD = 0.115) and the baseline condition (M = 0.31, SD = 0.148) was not significant: *F* [1, 29.11] = 1.43, *p* = .24. H2a was supported, but H2b was not.

### Group Communication Quality

The second set of our hypotheses states that the highlighting function will improve the participants' perceived quality of communication. We conducted 2×2 mixed-model ANOVAs on each of the subscales of the main GCQ scale: clarity and comfort.

*Group Communication Quality-Comfort.* There was a significant interaction effect between the highlighting function and the participants' language background on the comfort subscale (*F* [1, 31] =4.20, *p* < .05) (Figure 4). The highlighting function and language background had no main effects on the participants' perceived comfort scores. We found that NNS's perceived GCQ-comfort score was marginally higher in the highlighting condition (M = 5.48, SD = 0.93) than in the baseline condition (M = 5.10, SD = 0.87): *F* [1, 31] = 3.42, *p*=.07). In terms of interaction effect, we found that language background had an effect on comfort scores in the baseline without-highlighting condition (*F* [1, 42.8] = 6.91, *p* < .05) but no effect on the scores in the with-highlighting condition (*F* <1, *n.s.*).

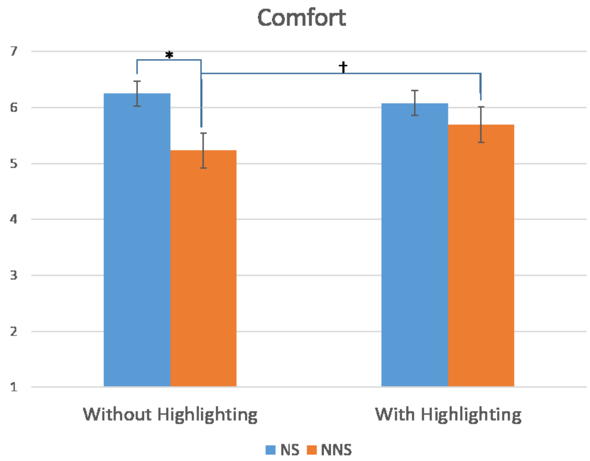


**Figure 4. GCQ-Comfort by condition for NS and NNS (error bars represent standard errors). * *p*<.05, †=.07**

*Group Communication Quality-Clarity.* The main effects of the highlighting function (*F* [1, 31] = 6.31, *p* < .05) and the language background of the participants (*F* [1, 31] = 12.40, *p* < .01) on their perceived GCQ-clarity scores were all significant (Figure 5). The NNS perceived GCQ-clarity scores were higher in the highlighting condition (M = 5.5, SD = 1.04) than in the baseline condition (M = 4.90, SD = 0.84): *F* [1, 31] = 6.66, *p*< .05.

### Correlations between Performance and Perceptions

To understand the relationship between task performance (recalled commonalities) and the participants' perceptions and experiences, we further examined pairwise correlations among the dependent variables.

As shown in Table 2, there were significant positive correlations between the recalled commonalities and the perceived group communication quality scales, including comfort and clarity. We also found a significant negative correlation between workload and perceived group communication quality scales, indicating that when people perceived a higher workload, their perception of the group communication quality also suffered.

## DISCUSSION

In summary, our study's results suggest that the NS were motivated to voluntarily highlight key parts of the conversation for their non-native partners. Our results also suggest that automated transcripts with highlighting tools can rebalance the workloads of NS and NNS. Even though asking NS to highlight transcripts added to their workload,

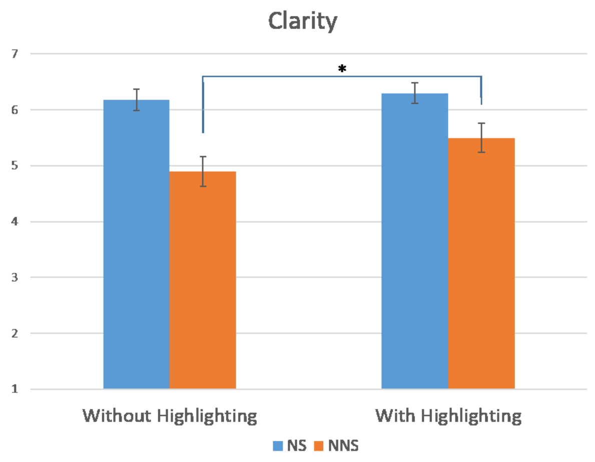


**Figure 5. GCQ-Clarity by condition for NS and NNS (error bars represent standard errors) * *p*<.05**

they also benefited from such extra effort and performed the task better (more recalled commonalities). Moreover, their highlights improved the quality of the clarity and comfort as perceived by the NNS.

### Motivation of NS to Highlight

As discussed in the background section, while NS can freely highlight utterances on the transcripts, this doesn't mean that they will be motivated to do so.

From our results, we found that NS participants were motivated to make highlights. NS participants seemed to highlight messages especially when they thought the highlights were helpful for their NNS partners. During the interviews, NS participants confirmed this again with:

> *"Actually, (I) mostly highlighted for the non-native speaker, just to help out, but then also for me to remember later, but mostly for the non-native speaker, just in case he needs help or because native speaker we have no problems but you know maybe.." (NS-12)*

> *"I highlighted most commonalities. I highlighted for both, you know, her [NNS] and me, and it would be useful to look at this one, we have to make a list of commonalities." (NS-19)*

> *"I highlighted 50% for myself and 50% for NNS. I sometime would scroll back to underscore." (NS-07)*

This finding provides insight for new ways of using textual highlighting. Our study shows that textual highlighting can serve as a backchannel for team members in a meeting to help other members catch up with the ongoing conversation. Meanwhile, previous works on textual highlighting have mainly focused on the cognitive properties of highlights, which may help readers process and organize textual information [17, 26]. Although recent works have tested the advantages of using textual highlighting in a conversational context, the highlights were used either for themselves (e.g., for note taking [11]) or for overcoming the disadvantages of the communication media (e.g., overcoming the errors generated by machine translation [3]). No work has explored highlighting's possible use as collaborative support for supporting others in a conversation.

From our results, we also found that few NS were not motivated to highlight, and some NS gave up highlighting during the conversation. A possible reason may be because their workload had increased. We discuss this further in the next subsection.

### Rebalanced Workloads

From our results, the NS workload increased, which is consistent with previous studies such as Kalnikaitė et al. [11]. They found that during automated-transcript-aided meetings, people perceived the highest workload while using highlighting as a note-taking tool.

Multitasking is one plausible reason why generating highlights during the conversations increased the NS workload. Some participants reported that they tried to highlight both for the NNS and for themselves and to remember the commonalities at the same time, which was still difficult:

> *"In the beginning, it was easy to highlight; as the discussion went on, I was trying to come up with ideas, then I gave up highlighting."* [NS-17]

Another possible reason is that the automated transcript contained errors, so that the NS had to choose which sentence to highlight:

> *"Errors made it difficult to choose, sometimes [I] wouldn't do highlighting because of the error."* [NS-17]

Although NNS' workload wasn't significantly reduced by the NS' highlights, we still found an interesting phenomenon: The difference in the workload between the NS and NNS evaporated in the highlighting condition, indicating that both were sharing a similar level of loading.

While previous works on multilingual group collaboration have successfully improved overall task performance or the quality of group communication, those approaches appeared to impose extra burden on NNS [5, 9]. One possible reason is that the NNS in those studies had to adapt themselves to the support technologies in order to take advantage of them. Meanwhile, textual highlights generated by the NS in our study did not seem to require an extra burden for NNS. Although reading transcripts perhaps increased their workload [4, 5], the highlighted text may have helped them reduce the amount of information they had to handle in the conversation.

### How NS-generated Highlights Helped NNS

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Recalled commonalities | - | | | |
| 2. Workload | -.28* | - | | |
| 3. GCQ-Clarity | .30* | -.39** | - | |
| 4. GCQ-Comfort | .28* | -.64** | .50** | - |

*$p < .05$, **$p<.01$

**Table 2. Correlation among recalled commonalities, workload, GCQ-Clarity, and GCQ-Comfort**

Our results showed that perceived group communication quality of the NNS improved in terms of clarity and comfort. One possible explanation is that NNS confirmed their understanding of the consensus by comparing what they thought and what was highlighted by the NS in the highlighting condition.

> *"In the first condition [with the highlighting function], I thought highlighting was not necessary. But in the second condition, I realized that I was relying on the highlights.*

> *"In the first condition, I could confirm my understanding by looking at the highlights - whether we found things in common or not. But in the second condition, I was often not sure if we actually reached a consensus or not. Like, was that a commonality?...and then they quickly slipped off my head."* [NNS-2]

Another possible mechanism is that the NS focused more on the task and changed their way of speaking to perform the highlighting task. Such new ways of speaking were clearer to NNS.

> *"[In highlighting condition,] I would say, 'Okay, so we have this in common.' I just want to make sure that everyone has common ground."* [NS-15]

> *"'This is what we have in common,' I tried to summarize so that I could click on it."* [NS-16]

> *"The two native speakers focused more on the task in the second condition [with the highlighting condition]. There were less side talks. Maybe the highlights helped them concentrate on the task. NS were highlighting the things we found in common. They matched with my understanding."* [NNS-5]

Having the NS generating highlights on automated transcripts offers an ideal situation to NNS, where (1) NS and NNS can use highlights as a backchannel to non-verbally confirm and ground their intention and (2) NS' behaviors can be shaped to better accommodate what the NNS need, such as extra confirmation and summary.

While textual highlighting is a well-known approach to helping readers process and organize textual information, it is actually a unique approach for supporting NNS in multilingual group communication. Previous studies on multilingual group collaboration have provided NNS word/sentence level support that allows them to compensate for the missed parts of conversation by reading transcripts or translating words [5, 7, 10, 24]. Few works have aimed at helping NNS gain understanding in context. While some researchers tried to make use of automatic keyword highlighting of ASR transcripts in a multilingual group meeting, it turned out not to be useful for NNS because some of the highlighted words were unimportant [9]. Our study showed that textual highlighting by humans (NS) is useful for NNS and that highlighting during conversation is a feasible support method.

### Correlations

From our results, not surprisingly, there were significant positive correlations between recalled commonalities and perceived group communication quality scales, including comfort and clarity, indicating when people have a better experience in the conversation they also perform the task better. We also found that workload and perceived group communication quality scales had a significant negative correlation. It's not surprising to learn that when people perceived a higher workload, their perception of group communication quality would suffer.

However, it's unclear why there was a moderately negative correlation, rather than no correlation, between workload and recalled commonalities. This shows that our task rebalancing strategy may also need to consider the possible negative impact of work overload in future designs. More work is required to investigate the causal mechanisms between the variables.

## DESIGN IMPLICATIONS

As discussed in the previous section, our findings show that highlights generated by NS can actually benefit the group itself at the cost of the workload of NS. Previous research on supporting multilingual communication mostly used machines as aids [5, 10, 24]. In this study, we didn't simply use a machine to support the collaboration. Instead, we added NS as a new component of the support system. Our main assumption, as well as the core of our study's proposed idea of task rebalancing, is that the spare cognitive resources of NS can be shared to deal with the problems that NNS encounter. Therefore, we extended the space of design and research from the earlier focus of "supporting NNS' comprehension with machine aids" to "taking an angle of human computation [13] to provide support." Based on these findings, we proposed some design implications for future multilingual communication support systems.

### Social Annotations Shared among Group Members

In our study, NS participants reported a desire to see what the other NS participant working in the same group highlighted. We may need to consider increasing the transparency of highlights among NS participants so that they are aware of the status of highlighting and won't duplicate their efforts by highlighting the same parts. However, careful design considerations should be taken, since increasing the transparency of highlights among NS participants may increase the power of the majority (NS) and rule out the diverse perspectives from the minority (NNS).

Similarly, our current design doesn't allow NNS participants to contribute highlights. Even though allowing NNS to add highlights might risk increasing their workload, giving them a non-verbal tool for generating contributions might also improve their collaboration with their NS counterparts.

Allowing more public generation and sharing of highlights are interesting features for future designs based on the understanding we obtained from this study. It would also be interesting to investigate how such a human computation system would affect the group dynamics.

### Augmenting Multilingual Communication with Versatile Highlighting

One noteworthy item from our interviews is that our participants started to appropriate highlighting as a non-verbal backchannel for confirmation and summarization.

Therefore, we might convert the interface into a versatile annotation tool by providing different types of highlights. Each type could be color- or icon-coded and given a different yet publicly shared meaning, such as "important," "agreed on," or "question?" etc.

Since such highlights are shared among all group members, group agreements and disagreements become explicitly visible on the transcript, which might ease the burden of multilingual conversations.

### Empowering NNS through Redistribution of Work

In multilingual collaboration, NS often take the lead in discussions, relegating NNS to the role of followers [27]. Although NNS benefited from the automated transcripts highlighted by NS, NS also gained influence and power in discussions by deciding which part of the conversation to highlight. NNS passively consumed what NS had decided.

Therefore, it may be necessary to not just rebalance the task but also to redistribute types of work to NS and NNS by taking their respective language proficiency into consideration. For example, some NNS participants may encounter difficulty in expressing ideas but can comprehend others' ideas without problem. In this case, it may be useful to ask NNS to instead take the lead in highlighting the transcripts, since they can focus on reading others' ideas and deciding which parts to highlight as a way to influence the group's work.

## LIMITATIONS AND FUTURE WORK

There are some limitations to this study. First, our main assumption of task rebalancing relied on NS having motivation to perform extra work to help their NNS partners. The existence of NS motivation was evidenced in the study. However, as a lab experiment, participants only interacted with one another for 15 minutes. Further studies with relaxed time constraints in the lab and in the field would be helpful.

Second, we developed and employed a commonality-finding task for the laboratory study. While the underlying process of commonality finding is considered universal and fundamental to intercultural collaboration, testing the proposed approach of task rebalancing with NS-generated annotations on other collaborative tasks, such as group brainstorming and group decision making, would help generalize the current results.

Finally, although our proposed idea successfully included NS in the supportive system by adding an additional channel, the back-channel was unidirectional (from NS to NNS), making NNS passive in the system. Further study to facilitate NNS' active participation is required.

## CONCLUSION

We conducted a laboratory study to examine the idea of asking NS to highlight key parts of automated transcripts to rebalance the tasks that NS and NNS perform in automated transcript-supported multilingual communication. We found that NS indeed had motivation to voluntarily highlight the transcripts for NNS and that the added burden of NS could benefit both NS and NNS. We also found that highlighting improved NNS's perception of group communication quality and that NNS would use highlighting as an additional channel to confirm their understanding of the ongoing conversation. Accordingly, it's feasible to add NS-generated highlight annotations as a component in the design of supportive tools for multilingual communication. Our results suggest a new approach to improve multilingual communication through in-group task rebalancing that requires limited extra cost.

## ACKNOWLEDGMENTS

The work is supported in part by Ministry of Science and Technology, Taiwan (102-2221-E-007-073-MY3; 104-3115-E-007-001; 105-2628-E-007-004-MY2), Microsoft Research Asia and Google Faculty Research Award. We also thank anonymous reviewers for the constructive feedback, which help us improve the work vastly.

## REFERENCES

1. Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley, (Eds.). *Perspectives on Socially Shared Cognition* (pp. 127-149). APA Press.
2. Alan J. Feely and Anne-Wil Harzing. 2003. Language management in multinational companies. *Cross-Cultural Management: An Int'l Journal*, 10: 37–52.
3. Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 449-458.
4. Ge Gao, Naomi Yamashita, Ari M.J. 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. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 843-852.
5. Ge Gao, Naomi Yamashita, Ari M.J. Hautasaari, and Susan R. Fussell. 2015. Improving Multilingual Collaboration by Displaying How Non-native Speakers Use Automated Transcripts and Bilingual Dictionaries. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3463-3472.
6. Christine C.M. Goh. 2000. A cognitive perspective on language learner's listening comprehension problems. *System* 28, 1: 55-75.
7. Cheng-Hsien Han, Chi-Lan Yang, and Hao-Chuan Wang. 2014. Supporting second language reading with picture note-taking. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '14), 2245-2250.
8. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload*. Amsterdam: North Holland Press.
9. Ari Hautasaari and Naomi Yamashita. 2014. Catching up in audio conferences: highlighting keywords in ASR transcripts for non-native speakers. In *Proc. of the 5th ACM International Conference On Collaboration Across Boundaries: Culture, Distance & Technology*, 107-110.
10. Rieko Inaba, Yohei Murakami, Akiyo Nadamoto, Toru Ishida. Multilingual Communication Support Using the Language Grid. In *Intercultural Collaboration*. Springer, Berlin, 118-132.
11. Vaiva Kalnikaitė, Patrick Ehlen, and Steve Whittaker. 2012. Markup as you talk: establishing effective memory cues while still contributing to a meeting. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW '12), 349-358.8
12. Mardi Kidwell. 2000. Common ground in cross-cultural communication: Sequential and institutional contexts in front desk service encounters. *Issues of Applied Linguistics* 11, 1: 17-37.
13. Edith Law and Luis von Ahn. 2011. *Human Computation* (1st ed.). Morgan & Claypool Publishers.
14. Han Z. Li. 1999. Grounding and information communication in intercultural and intracultural dyadic discourse. *Discourse Processes* 28, 3: 195-215.
15. Han Z. Li, Young-ok Yum, Robin Yates, Laura Aguilera, Ying Mao, & Yue Zheng. 2005. Interruption and Involvement in Discourse: Can Intercultural Interlocutors be Trained? *Journal of Intercultural Communication Research* 34, 4: 233-254.
16. Leigh Anne Liu, Chei Hwee Chua and Günter Stahl. 2010. Quality of communication experience: definition, measurement, and implications for intercultural negotiations. *Journal of Applied Psychology* 95, 3: 469-487.
17. Catherine C. Marshall. 1997. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries* (DL '97), 131-140.
18. Yingxin Pan, Danning Jiang, Michael Picheny, and Yong Qin. 2009. Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 2353-2356.
19. Pamela Rogerson-Revell. 2008. Participation and performance in international business meetings. *English for Special Purposes* 27, 3: 338-360.
20. Yohtaro Takano and Akiko Noda. 1995. Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing. *Language Learning* 45, 4: 657-681.
21. Craig S. Tashman and W. Keith Edwards. 2011. Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 2927-2936.
22. Michael D. Tyler. 2001. Resource consumption as a function of topic knowledge in nonnative and native comprehension. *Language Learning* 51, 2: 257-280.
23. Eric Wallen, Jan L. Plass, and Roland Brünken. 2005. The function of annotations in the comprehension of scientific texts: Cognitive load effects and the impact of verbal ability. *Educational Technology Research and Development* 53, 3: 59-71.

24. Hao-Chuan Wang, Susan Fussell, and Dan Cosley. 2013. Machine translation vs. common language: effects on idea exchange in cross-lingual groups. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13), 935-944.

25. Joanna L. Wolfe. 2000. Effects of annotations on student readers and writers. In *Proceedings of the fifth ACM conference on Digital libraries* (DL '00), 19-26.

26. Jen-Her Wu, Yufei Yuan. 2003. Improving searching and reading performance: the effect of highlighting and text color coding. I*nformation & Management* 40, 7: 617-637.

27. Naomi Yamashita, Andy Echenique, Toru Ishida, and Ari Hautasaari. 2013. Lost in transmittance: how transmission lag enhances and deteriorates multilingual collaboration. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13), 923-934.

28. Lin Yao, Yingxin Pan, and Danning Jiang. 2011. Effects of automated transcription delay on non-native speakers' comprehension in real-time computer-mediated communication. In *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction - Volume Part I* (INTERACT'11), 207-214.

29. Chien Wen Yuan, Leslie D. Setlock, Dan Cosley, and Susan R. Fussell. 2013. Understanding informal communication in multilingual contexts. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13), 909-93.