Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs

Ge Gao[1], Bin Xu[2], David Hau[2], Zheng Yao[2], Dan Cosley[2], Susan R. Fussell[1,2]
[1]Department of Communication
[2]Department of Information Science
Cornell University
Ithaca NY 14850 USA
[gg365, bx55, dch229, zy87, drc44, sfussell] @cornell.edu

ABSTRACT
Machine translation (MT) creates both opportunities and challenges for multilingual collaboration: While MT enables collaborators to communicate via their native languages, it can introduce errors that make communication difficult. In the current paper, we examine whether displaying two alternative translations for each message will improve conversational grounding and task performance. We conducted a laboratory experiment in which monolingual native English speakers collaborated with bilingual native Mandarin speakers on a map navigation task. Each dyad performed the task in one of three communication conditions: MT with single output, MT with two outputs, and English as a common language. Dyads given two translations for each message communicated more efficiently, and performed better on the task, than dyads given one translation. Our findings show the value of providing multiple translations in multilingual collaboration, and suggest design features of future MT-based collaboration tools.

Author Keywords
Machine translation; Collaboration; Multilingual communication

ACM Classification Keywords
H.5.3 [Group and Organization Interface]: Computer-supported cooperative work

General Terms
Experimentation; Human Factors

INTRODUCTION
Many modern organizations actively seek opportunities to collaborate across national and language boundaries. As people work together at a global scale, a multilingual context in which collaborators communicate via "a cocktail of languages [17]" arises. Multilingual organizations often require that all members speak English as a common language [9], which puts heavy cognitive load and pressure onto non-native speakers [15, 36, 35]. Non-native speakers report language performance anxiety, status worries, and job insecurity due to lack of fluency in English [29]. Language differences can also make it difficult to establish trust and collaborate effectively in teams consisting of native and non-native speakers [37].

Figure 1. The nature of machine translation mediated communication leads to difficulties in detecting translation errors within each single translation output.

Machine translation (MT) provides an opportunity for multilingual collaborators to interact via their own native languages. Using MT might reduce the cognitive effort and social challenges associated with speaking in a non-native language. At the same time, MT mediated communication can poses new challenges. In particular, the meaning of the original message may be lost or distorted after being processed by MT algorithms [e.g., 43, 44], and it can sometimes be difficult to detect these problematic translations in real time.

In the scenario illustrated by Figure 1, for example, a native English speaker gets a MT translated message from a Chinese speaker, saying "I clear the table, next to my legs kitten asleep". This translated message is different from its original version in Chinese, because the MT algorithm confused "the table legs" with "my legs". Native English speakers, who typically do not know precisely how MT works, would have little chance of detecting this translation error. Such errors can accumulate over the course of a conversation, leading to serious misunderstandings.

CSCW '15, March 14-18, 2015, Vancouver, BC, Canada.

ACM 978-1-4503-2922-4/15/03…$15.00
http://dx.doi.org/10.1145/2675133.2675197

In this paper, we explore the value of one low cost way to help people detect errors and infer the intended meaning of messages: providing multiple translations from multiple MT tools [42]. In the above scenario, for example, the same message is translated differently by Google and Bing:

- Google translation output:
  I clear the table, next to my legs kitten asleep.
- Bing translation output:
  I was cleaning the tables, kittens fell asleep next to my table legs.

Comparing these two translations offers the message recipient additional resources to infer what the original Chinese message means. The consistent parts between the two outputs is information the receiver can be fairly confident about (e.g., that there is a table and a cat), while the inconsistent parts suggest possible translation errors that need to be clarified (e.g., *where* the cat is, whether the table is being "cleared" vs. "cleaned"). In research by Xu and colleagues [42], dyads consisting of English and Mandarin Chinese speakers reported that having two MT outputs rather than one helped them infer the intended meaning of messages in an unstructured chat context.

Building on this work, the current paper presents a laboratory study that explores how giving two MT outputs affects communication efficiency and task performance during multilingual collaboration. We first propose hypotheses based on literature about MT and multilingual collaboration. We test these hypotheses with a laboratory experiment in which monolingual native English speakers collaborated with bilingual native Mandarin speakers in dyads on a map navigation task. Dyads finished the task under one of three communication conditions: MT with a single output for each message, MT with two outputs for each message, and English as a common language. We compared communication efficiency and task performance between these conditions, along with self-evaluations of communication experience and workload.

In general, our data show the value of providing two MT outputs. Dyads given two MT outputs communicated more efficiently and performed better on the task than dyads given a single MT output. Having two MT outputs also did not increase cognitive workload over having a single MT output. This suggests that for areas such as MT where imperfect computer agents sometimes make mistakes, designing systems that explicitly account for these errors is likely to pay dividends.

RELATED WORK AND HYPOTHESES

Although MT was initially used primarily for written documents, [e.g., 19, 26, 33], research has recently begun to explore the use of MT to facilitate real time communication and collaboration among speakers of different native languages. A number of studies [e.g., 16, 22, 25, 41] show that MT has both benefits and costs for communication.

On the positive side, MT allows people communicate with little concern about their language fluency. Previous work on multilingual collaboration suggests that using MT may benefit team collaboration. Lim and Yang [25] found that English-Chinese speaking dyads worked together on a negotiation task performed better using MT vs. English as a common language. Wang and colleagues [41] found that Chinese speaking participants generated more ideas when using MT rather than English as a common language. Hautasaari [16] similarly found that using MT improved the performance of Japanese-English speaking dyads and also increased the use of socio-emotional messages.

On the negative side, however, MT can make it challenging for participants to ground their utterances, that is, to build shared knowledge and beliefs [4]. A key component of conversational grounding is the identification of objects and locations [21]. When a tourist asks directions from a local resident, for example, the two must work together to make sure every landmark is mutually understood. In what Clark and Wilkes-Gibbs [6] call the "basic exchange," a speaker refers to an entity (e.g., "the clock tower") and the listener indicates that he or she has understood (e.g., "got it"). Here, the message recipient provides positive evidence of understanding (e.g., acknowledgement) and no negative evidence of understanding (e.g., questions or requests for confirmation) [5].

Grounding in MT-Mediated Conversation

There is evidence that grounding of referents can be problematic in MT-mediated communication, even more so than when English is used as a common language. For example, very similar names for the identical object or landmark may be translated differently [43, 44]. The following example shows how such inconsistency could induce problems to grounding in MT-mediated communication. In this scenario, an English speaking tourist is looking for the place to rent a bicycle. A Chinese speaker gives direction like this:

> **Chinese message:**
> 图书馆正对着一个钟楼
> [The library faces a clock tower]
> 楼下有一个服务处可以租自行车
> [under the tower there is a service center where you can rent bikes]
>
> **Google translated English:**
> Libraries facing a bell tower
> downstairs there is a place you can rent a bike service

In this example, both "钟楼 (clock tower)" and "楼 (the tower)" refers to the same entity in the original Chinese message. When the message is translated into English, however, the word "楼 (the tower)" is mistranslated into "downstairs", which may impede the grounding of this referent.

Even when a term is translated consistently, it may still be wrong in ways that disrupt grounding, especially when languages are distant both structurally and lexically (e.g., Chinese and English) [3, 39]. The next example shows how grounding could be hurt under this occasion:

Chinese message:
向前走直到看见一个卖水果的
[Go ahead until you see a fruit stand]
再右转就到了
[then turn right you will arrive]

Google translated English:
Go forward until you see a fruit
and then turn right on to the

In this example, the word "卖水果的 (fruit stand)" is mistranslated into "fruit", which creates a potential grounding problem. Although similar difficulties of grounding could happen with every message and reference, message recipients have few clues to figure out how different a translated message is from its original version.

Under the assumption that it will be some time until machine translation algorithms are perfected, investigators have turned instead to a consideration of how current MT output can be supplemented to facilitate grounding. We focus on improving grounding in MT rather than grounding when English is used as a second language [e.g., 12, 24] because of the demonstrated value of composing messages in one's native language [41]. Google's translation interface as of mid-2014 allows people to look up alternate translations for individual words. This approach can improve grounding, but requires participants to actively request information. More automated approaches include back-translations to help speakers identify when a translation might be confusing to the recipient [28, 32], keyword highlighting to help people focus on the gist of a translated message [10], and enhancing the verbal message with semantically linked pictures [40].

The Current Study

The solutions outlined above all have additional software requirements beyond the machine translator, such as tools to implement keyword highlighting or picture selection. A simpler approach, the one we investigate in the current study, is to use a chat interface that provides outputs from several established MT algorithms (e.g., Google Translate and Bing) at the same time [42]. With two translation outputs instead of just one, people may be better able to figure out the intended meaning of a message. For example, the Bing translation for the example above,

Bing translated message:
Go straight until you see a fruit seller, then turn right into the

clarifies that it is a fruit seller not a fruit that is the probable turning point.

Providing participants with two translations may also help them identify otherwise undetectable translation errors. In our example, the shift from "on to" to "into" at the end of the translation flags a potential point of confusion (though neither makes it clear that the traveler has reached the destination). At the same time, when two translations have overlapping content (here, "go forward/straight", "fruit," "turn right"), people may feel more confident about their understanding of the message.

One previous study has compared two vs. one translation(s) in a conversational setting. Xu and colleagues [42] found that two translations increased participants' confidence that they had correctly understood partners' messages. However, they used an informal chat task and did not examine whether the interface affected conversational grounding or task performance. The current study explores these issues in the context of a well-defined collaboration task, the HCRC map task [8], which requires grounding to succeed.

For the reasons outlined above, we expect that providing two MT outputs will improve message comprehension. There are few clues to identify accurate vs. problematic parts of a single MT output, but with two outputs, people can analyze commonalities and differences to better identify the probable meaning of the message.

H1. People will perceive less difficulty in understanding messages with two MT outputs vs. one MT output.

This decreased difficulty in understanding should be reflected in how people ground their messages. When messages are easy to understand, we would expect more positive evidence of understanding, such as basic presentation-acceptance exchanges, and less negative evidence of understanding, such as requests for clarification. Thus, we hypothesize that:

H2a. People will show more positive evidence in grounding (e.g., acknowledgment) with two MT outputs vs. one MT output.

H2b. People will show less negative evidence in grounding (e.g., questions, requests for confirmation) with two MT outputs vs. one MT output.

Because successful grounding of referents is essential for route-giving tasks such as the HCRC map task [e.g., 8, 12], we hypothesize that improvements in grounding from two translations will result in better task performance:

H3. People will achieve better task performance with two MT outputs versus one MT output.

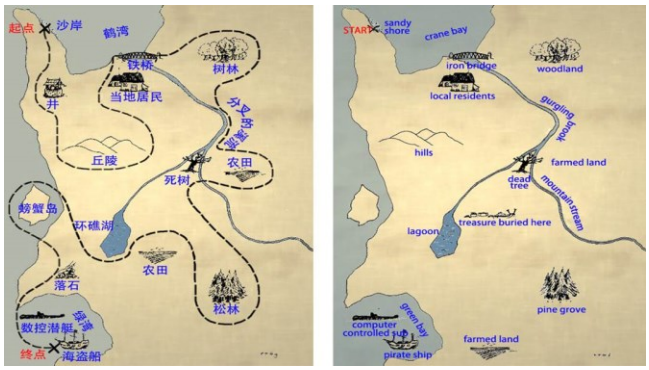

Figure 2. One set of maps used in the navigation task (the left one for the Chinese speaking instructor, the right one for the English speaking follower).

However, processing two translation outputs may require more cognitive effort than processing a single translation. Although Xu and colleagues [42] did not find this for informal chat conversations, we expect that in the context of a longer, more structured task, the overall accumulated reading effort will lead to higher perceived workload:

H4. People will experience higher workload with two MT outputs vs. one MT output.

Finally, given that people often use English as a common language in today's multilingual organizations [9], we want to compare two MT outputs to an English-only condition. While prior work suggests that grounding is more successful with English as a common language vs. MT with one output [43, 44], we wondered if the gap between MT and English as a common language would be reduced or eliminated when two MT outputs are provided,

RQ. How do the perceived difficulty in understanding messages, frequency of positive and negative evidence in grounding, task performance, and perceived workload differ between having two MT outputs versus English as a common language?

METHOD

Overview

We conducted a between-subjects laboratory experiment to examine the effects of two MT translation outputs on conversational grounding and collaborative performance. Dyads consisting of a monolingual native English speaker and a native Mandarin speaker who spoke English as second language performed the HCRC map navigation task [e.g., 1, 7, 38] under one of three communication conditions: MT with a single translation output for each message, MT with two translation outputs for each message, and English as a common language. In the first two conditions, both native English speaking participants and native Mandarin speaking participants typed and received messages in their native languages. In the third condition, participants typed and received messages in English. Post-task surveys, IM logs and task logs were used to measure message understanding, conversational grounding behavior, task performance, and cognitive workload.

Participants

We recruited 96 participants from a U.S. university. Half (N = 48, 22 female) were native English speakers who had grown up in the U.S. and had little or no knowledge of Mandarin. Their mean age was 21.5 years (*SD* = 2.70). They reported some previous experience with MT (M = 3.88, *SD* = 1.31 on a 7-point scale ranging from 1= never to 7 = very often). The other half (N = 48, 24 female) were native Mandarin speakers who had grown up in the People's Republic of China and been in the U.S. for two years or less. Their mean age was 26.38 years (*SD* = 5.66). They spoke English as second language with moderate fluency (M = 4.54, *SD* = 1.20 on a 7-point scale ranging from 1= not fluent at all to 7 = very fluent) and had some previous experience with MT (M = 3.38, *SD* = 1.72). Under all conditions, participants were assigned to dyads consisting of one native English speaker and one native Mandarin speaker.

Materials

Task. Dyads completed two HCRC map navigation tasks (Figure 2). Each task consisted of one instructor map and a paired follower map. The instructor map showed a prescribed route around a list of landmarks. The follower map contained similar but not identical landmarks and had no route shown. Each pair of HCRC maps has some differences in the depicted landmarks, adding challenges to the task and increasing the need for grounding work. Both pairs of maps have the same number of landmarks, and require similar steps to complete the navigation. We manually translated the labels on each map into Mandarin Chinese for the Chinese speakers.

Surveys. Participants completed an online pre-experiment questionnaire that collected their demographic information. During the formal experiment, participants were asked to fill out a short online survey after each task. This post-task survey consisted of seven 7-point Likert scale questions. One question is a manipulation check to confirm which communication condition participants were assigned to. Two questions pertained to participants' perceived difficulty of understanding received messages during the task. The other four questions asked about participants' workload while doing the task. Since participants need to work on two sets of maps, they finished two such surveys.

Software and Equipment

We used a custom chat interface with embedded MT that could show one or two MT outputs as appropriate for the experiment condition [42]. The interface consists of an instant messaging (IM) window on the left and a map window on the right, as shown in Figure 3.

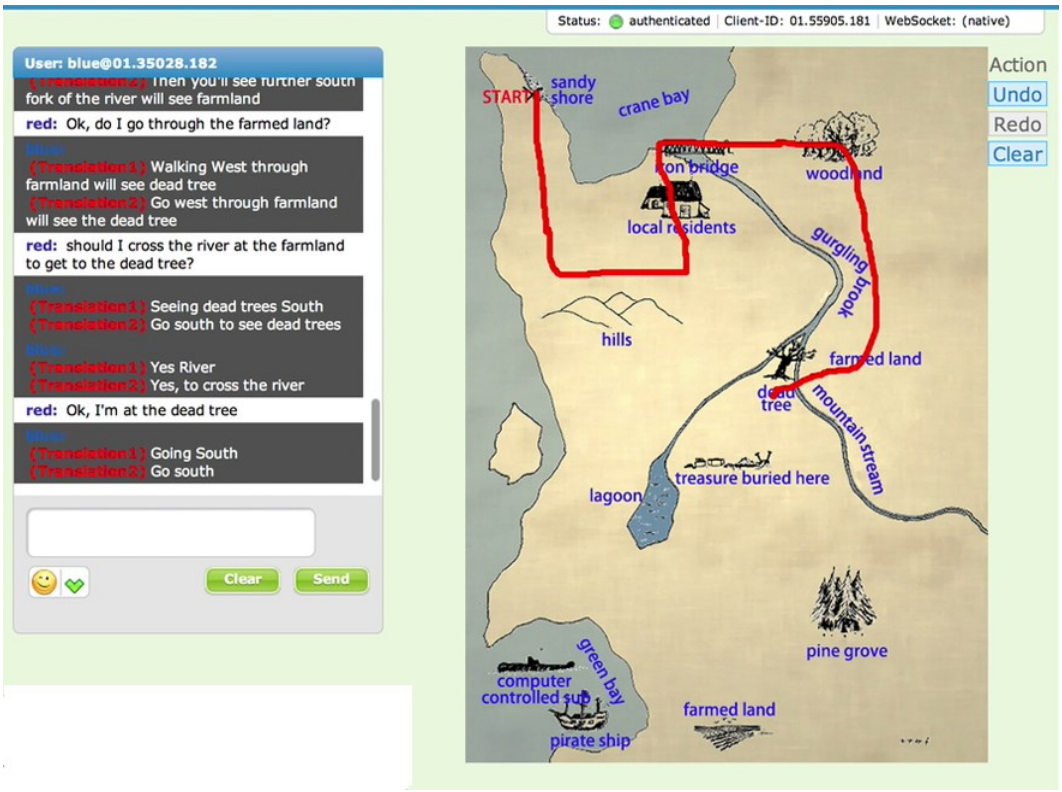

Figure 3. The chat interface as seen by an English speaking follower when getting two MT outputs for each message.

IM window. For the MT mediated conditions, participants typed messages in their native language and received messages translated into their native language. For pairs given two MT outputs, messages were processed by Google and Bing translations, chosen because they are competitive in quality but use distinct algorithms [45]. Pairs in the one MT output condition saw all messages translated by either Google or Bing, counterbalanced between dyads. Translations were always automatically generated and no preview or pre-editing was afforded by this interface. For the English condition, the MT module was absent and participants typed and received messages in English.

Map window. Instructors saw the map with both landmarks and route, while followers saw the map with only landmarks. The interface allowed followers to draw the route on their map following the instructor's guidance, but followers and instructors could only see their own map.

Equipment. Participants used Mac Pro laptops with 13.3 inch monitors and were separated from their partners by a barrier. They wore headphones to reduce distraction from outside noise.

Procedure

Pairs were seated at desks separated by a divider. The experimenter then introduced the study and explained the chat tool, MT module, and map navigation task. After that, the experimenter introduced the detailed procedure. Participants then practiced the chat tool and MT module for 5 minutes before starting the formal task.

In the formal task, dyads were randomly assigned to use one of three communication conditions: MT with one translation output, MT with two translation outputs, or English as a common language. They then performed two map tasks. For the first task, participants were randomly assigned as either an instructor or a follower. In the second task, participants switched roles with their partner and worked on the other map. Each task lasted for 15 minutes. The order of maps and roles were fully counterbalanced using a Latin Square design.

Participants answered a short survey to rate their communication experience and workload after each task. The order of maps and participant roles were fully counterbalanced using a Latin Square design.

Measures

We collected three types of measures of participants' communication and performance: participants' ratings on the post-task survey, coding of evidence in grounding based on the conversation corpus, and navigation performance based on the map route.

Post-Task Survey

Manipulation check. Communication condition was checked by a single choice question asking which language

Coding category	Definition	Example in the current corpus
Positive evidence in grounding		
Acknowledgment	A verbal response that shows a message is understood and accepted.	"Got it."
Negative evidence in grounding		
Align	Queries that check the partner's attention, agreement, or readiness.	"Is everything clear?"
Check	Queries that requests the partner to confirm some unsure information.	"So you want me to go around the bottom of the hills?"
Question-YN	Queries that take a yes or no answer but not counted as a Check or an Align.	"Can you see pelicans?"
Question-W	Any query not covered by the other three types of queries above.	"Where are the stairs?"

Table 1. Evidence of grounding in the current conversation corpus based on the HCRC dialogue structure coding manual.

medium (MT with single output vs. MT with two outputs vs. English as a common language) participants used.

Difficulty of understanding. Participants rated the perceived difficulty of understanding received messages on two 7-point Likert scales ("The unclear information in my partner's messages was distractive", "I had to think harder to understand the unclear information in my partner's messages", 1= strongly disagree, 7 = strongly agree). The questions formed a reliable scale (Cronbach's $\alpha = .86$) and were averaged to create a measure of difficulty of understanding.

Workload. Perceived workload was measured using four 7-point Likert scale adapted from the NASA Task Load Index (TLX [14]) ("Mental demand", "Temporal demand", "Effort", and "Frustration", 1 = low, 7 = high). The questions formed a reliable scale (Cronbach's $\alpha = .82$) and were averaged to create a measure of workload.

Evidence in Grounding

Coding on conversation corpus. We coded the corpus following the HCRC dialogue structure coding manual [2]. This coding scheme differentiates 13 types of conversational moves speakers may generate while doing the map navigation task. We asked two English-Mandarin bilingual speakers, blind to our hypotheses, to work independently and code all IM conversations line by line. Inter-coder agreement was good (Cohen's kappa = .83). The coders then discussed and resolved all disagreements.

Five conversational moves directly pertained to our hypotheses: acknowledgement, check, align, question-YN, and question-W (see Table 1). The first gives positive evidence in grounding, while the other four show negative evidence, i.e., an indication of a lack of understanding [5]. We calculated number of words (as in [13]) used for each type of evidence rather than turns because people use different strategies to break up sentences into turns.

Navigation Performance

Route accuracy. We scored each route the follower drew under the instructor's guidance using the system developed by Diamant and colleagues [7]. Dyads were given one point (1) every time they hit the correct landmark in the right order, but got no score (0) if they hit a wrong landmark or went to the landmark in the wrong order. Route accuracy was calculated as the proportion of correct landmarks.

We did not use time as a measure of task performance, since all groups used the whole session to accomplish the task.

RESULTS

We conducted 3 (communication condition: MT with one output vs. MT with two outputs vs. English as a common language) × 2 (native language: Mandarin vs. English) Mixed Model ANOVAs that took into account the fact that each participant provided two sets of measures. Participants were nested into pairs. The demographic variables (e.g., age and gender) and previous MT experience were set as control variables in all models. The order of trial (first vs. second trial), task (bay map vs. lake map), and role (guider vs. follower) were included as covariates.

Manipulation Check

The manipulation check indicated that all manipulations were successful. All participants identified the language medium they used correctly. The chat logs indicated that all participants used the correct language (English or Mandarin) for their assigned condition.

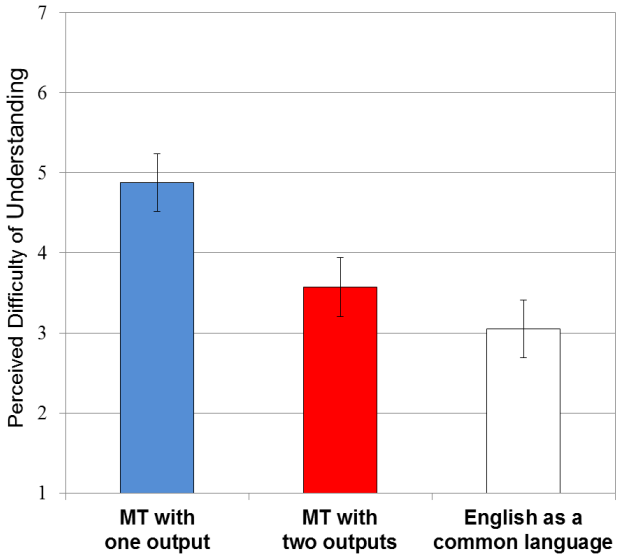

Figure 4. Mean difficulty of understanding by language medium on a scale of 1 (not difficult at all) to 7 (very difficult). (Error bars represent the standard errors of the mean.)

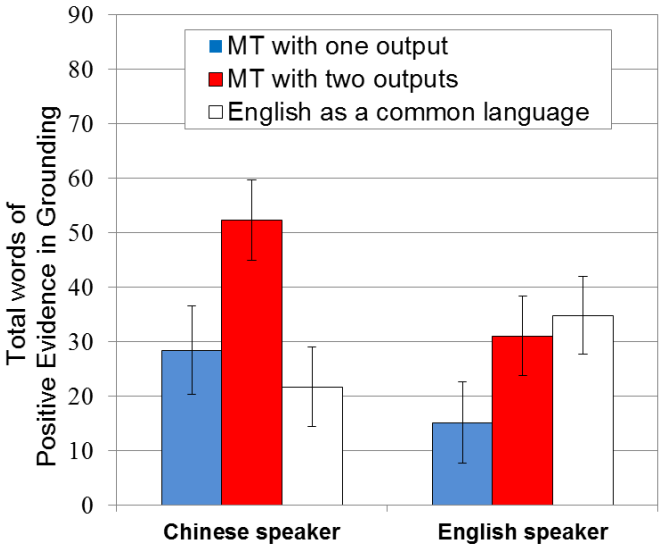

Figure 5. Mean words of positive evidence in grounding by communication condition and native language. (Error bars represent the standard errors of the mean.)

Difficulty of Understanding

H1 predicted that participants would perceive less difficulty in understanding messages when having two MT outputs rather than one. To test this hypothesis, we examined the effect of language medium on participants' rating on the difficulty of understanding.

The results support H1 (Figure 4). There was a significant effect of communication condition on self-reported difficulty of understanding ($F [2, 39] = 6.84$, $p < .01$). Pairwise comparisons indicated that participants perceived less difficulty with two MT outputs ($M = 3.57$, $SE = 0.37$) vs. one MT output ($M = 4.88$, $SE = 0.36$; $F [1, 39] = 5.70$, $p = .02$). They also perceived less difficulty with English ($M = 3.04$, $SE = 0.36$) vs. one MT output ($F [1, 39] = 13.16$, $p < .01$). However, there was no significant difference between MT with two outputs and English as a common language ($F [1, 39] <1$). These effects were not qualified by any further interaction effects. For both Chinese and English speaking participants, having two MT outputs helped them understand the received message better.

Evidence in Grounding

H2a and H2b predicted that participants would give more positive evidence (e.g., acknowledgement) and less negative evidence (e.g., questioning and requests for confirmation) of grounding when having two rather than one MT output. To test this hypothesis, we examined the effect of communication condition on the total number of words used for positive and negative evidence of grounding. Given that the map navigation task assigned asymmetric roles to collaborators, this analysis was restricted to messages produced by followers.

Positive Evidence in Grounding

In support of H2a, there was a significant main effect of communication condition on the total words of positive evidence of grounding ($F [2, 37] = 3.28$, $p = .05$; see Figure 5). Pairwise comparisons indicated that followers gave more acknowledgments when having two MT outputs ($M = 41.63$, $SE = 5.34$) vs. only one MT output ($M = 21.74$, $SE = 5.20$; $F [1, 37] = 6.41$, $p = .01$). Further, when participants used MT with two outputs vs. using English as a common language ($M = 28.20$, $SE = 5.06$), there was no difference on positive evidence given by followers: $F [1, 37] = 3.09$, $p = .09$.

We also found a significant interaction between communication condition and the follower's native language ($F [2, 37] = 3.38$, $p = .04$; see Figure 5). Native Mandarin speaking followers gave more words of acknowledgement when using MT with two outputs ($M = 52.25$, $SE = 7.34$) vs. MT with one output ($M = 28.38$, $SE = 8.27$) or English as a common language ($M = 21.64$, $SE = 7.29$). Native English speaking followers gave more words of acknowledgment when using MT with two outputs ($M = 31.01$, $SE = 7.26$) and English as a common language ($M = 34.76$, $SE = 7.07$) than when using MT with one output ($M = 15.10$, $SE = 7.54$).

Negative Evidence in Grounding

We examined the effect of communication condition on the total words of all queries (aligns, checks, question-YN, and question-W) given by the follower. As shown in Figure 6, the results partially support H2b. The main effect of communication condition was not significant ($F [2, 37] <1$). However, there was a significant interaction between communication condition and native language: $F [2, 37] =$

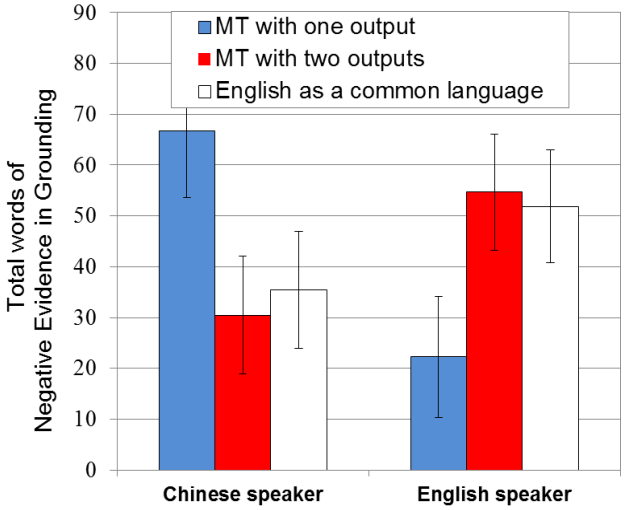

Figure 6. Mean words of negative evidence in grounding by language medium and native language. (Error bars represent the standard errors of the mean.)

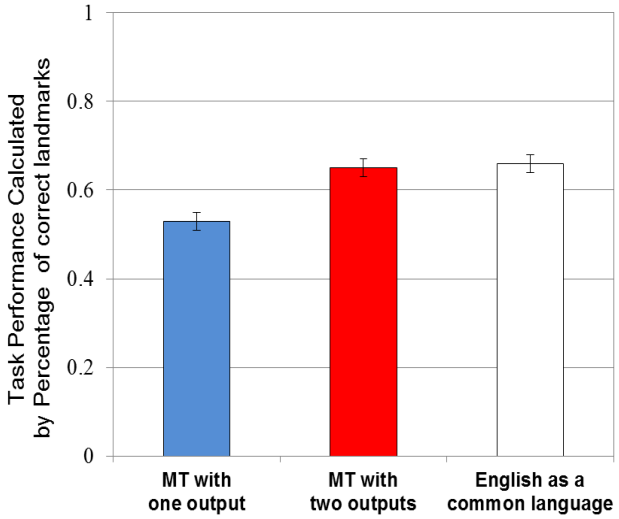

Figure 7. Mean task performance by language medium. (Error bars represent the standard errors of the mean.)

5.07, p = .01. The effect of communication condition was opposite for native Mandarin and English speakers. Native Mandarin speaking followers used more query words when using MT with one output (M = 66.70, SE = 13.04) than when using MT with two outputs (M = 30.45, SE = 11.60) or English as a common language (M = 35.42, SE = 11.49). Native English speaking followers used fewer query words when using MT with single outputs (M = 22.28, SE = 11.90) than when using MT with two outputs (M = 54.68, SE = 11.44) or English as a common language (M = 51.87, SE = 11.15).

Task Performance

H3 predicted that participants would perform better with two MT outputs rather than one. As shown in Figure 7, our results support H3. There was a significant effect of communication condition on route accuracy (F [2, 39] = 15.18, p < .01). Pairwise comparisons indicated that dyads having two MT outputs (M = 0.65, SE = 0.02) drew more accurate routes than dyads having only one MT output (M = 0.53, SE = 0.02; F [1, 39] = 19.70, p < .01). Dyads using English (M = 0.66, SE = 0.02) also drew more accurate routes than dyads having one MT output (F [1, 39] = 25.07, p < .01). Further, when dyads used MT with two outputs vs. using English as a common language, there was no difference in task performance (F [1, 39] <1).

Workload

H4 predicted that participants would experience higher workload when having two rather than one MT output. This hypothesis was not supported (see Figure 8). There was a significant effect of communication condition on workload (F [2, 39] = 4.54, p = .02) but pairwise comparisons indicated no difference in workload with two MT outputs (M = 4.37, SE = 0.30) versus one MT output (M = 4.72, SE = 0.29; F [1, 39] = 0.65, p = .42. Instead, participants using English (M = 3.55, SE = 0.28) reported lower workload than participants given one MT output (F [1, 39] = 8.58, p < .01) or two MT outputs (F [1, 39] = 3.77, p = .05). These effects were not qualified by any further interaction effects.

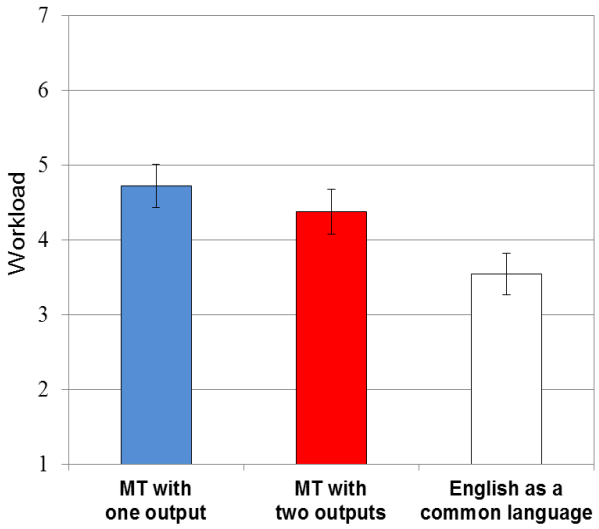

Figure 8. Mean workload by language medium on a scale of 1 (very low) to 7 (very high). (Error bars represent the standard errors of the mean.)

DISCUSSION

We examined conversational grounding, task performance, and cognitive workload as a function of whether participants received two MT outputs, one MT output, or used English as a common language. Our results suggest that two translation outputs improve grounding and

performance without increasing cognitive load.[1] We discuss each of these findings below and then suggest some possible implications for the design of tools to support multilingual communication.

Conversational Grounding

Consistent with Yamashita and colleagues [43, 44], our participants had more trouble grounding messages with a single MT output than they did when using English as a common language. Giving people two translations of each message strongly improved grounding. Participants reported that they could understand messages more easily with two rather than one, and this ease of grounding was evidenced in their greater use of acknowledgments. In fact, ease of grounding with two translations did not differ significantly from grounding using English as a common language.

Original message		Translation output(s) received by an English speaking follower
MT with one output		
Guider	到达一片农田	Reach a field
Follower	Treasure?	Treasure?
Guider	农田	Farmland
MT with two outputs		
Guider	下一个地方是另外一个农田	T1: Next place is another field T2: The next place is another farm
Follower	I am there	I am there

Table 2. Sample dialogues from the one MT output and two MT outputs conditions.

A possible mechanism behind these findings is that message recipients drew on the similarities and differences between the two outputs to infer the meaning of the original message. Table 2 shows an example of how two MT outputs may help people establish grounding efficiently. In each dialogue, a Mandarin speaking instructor is trying to guide an English speaking follower to a farm on the map. When using MT with one output, the word "农田 (farmland)" was translated into "field". The follower had to ask a clarifying question, because a field with hidden treasure was also shown on the map. The dyad with two translations grounded the instructions more easily because the "field" mentioned in T1 was translated as "farm" in T2, and the follower responded with an acknowledgment.

The pattern of results for negative evidence of grounding is more puzzling, especially for native English speakers. Mandarin speaking followers used fewer query words when they had two MT outputs rather than one, consistent with the idea that two translations helped reduce confusion about the meaning of the messages. However, English speaking followers used more queries with two translations. One possibility is that the English speakers were unable to identify when grounding work was required in the one translation case. Chinese bilingual speakers may have been more sensitive to potential translation issues.

For both positive and negative evidence of grounding, English speaking followers use similar amount of words when communicating via two MT outputs vs. English as a common language. This similarity echoes their self-report rating of the difficulty of understanding, and further suggests that ease of grounding is similar in these two conditions. We could not perform a similar comparison for Mandarin speakers because the experimental design required them to speak different languages in these two conditions.

Task Performance

The benefits of having a second translation were also revealed in task performance. As we hypothesized, dyads with two MT outputs performed significantly better than those with one MT output. In fact, their performance did not differ significantly from that of pairs using English as a common language. This suggests that for tasks that require grounding of many different referents (here, the landmarks on the map), the benefits from easy, successful grounding accumulate over the course of a conversation.

As we found no performance differences between using two translations and the use of English as a common language, one might wonder why organizations would chose to use MT at all. We think the justification is that it improves the collaborative experience of non-native speakers. Using a non-native language can be cognitively taxing [35], and non-native speakers brainstorm better using MT than English [41]. There is also evidence that MT improves performance at the group level [16]. Our study suggests that the primary downside of using MT, problematic communication that degrades performance, can be partly overcome by providing two different translations.

Workload

While we had expected that adding a second translation would increase cognitive workload, this was not the case. Self-reported workload using one or two translations was higher than workload using English as a common language.

[1] Although we combined the two translation services in our analyses, the overall pattern of results is similar when comparing each translation service individually with the two translation condition. The two-translation condition improved understanding over Google ($p. < 01$) or Bing ($p. = .10$) alone and also led to superior performance than Google alone or Bing alone (both $p. < .001$). The pattern of results for workload and our speech measures was also similar.

We suspect the similarity in workload between one and two translations can be attributed to subtle variations in workload for two different cognitive tasks: reading a message and figuring out its intended meaning. The second translation may have added to the reading time but reduced the time needed to interpret the message.

The lower workload associated with using English, which was true regardless of the participant's native language, may stem from a different cognitive task, that of message formulation. While it is difficult to produce messages in a non-native language, it is perhaps even more difficult to compose messages that one thinks will be accurately translated by the tool. Since all participants in the MT conditions received as well as produced MT translated messages, they might have been sensitive to the need for careful message construction. These are only speculations and future work will need to use reaction time methods or other more sensitive measures to understand how using MT and English affects cognitive workload.

DESIGN IMPLICATIONS

Apart from the straightforward implication that showing two outputs in MT systems can be helpful, our results suggest some more general ideas for how to make MT systems (and other systems that use error-prone agents) more useful.

Adding channels of communication. While two translations facilitate conversational grounding and task performance, they do not reduce cognitive workload over the one translation case. To make it easier for people to process messages, MT tools might incorporate other channels, for example the rich visual cues that benefit text-based brainstorming [40]. An MT interface might show images that illustrate some part of the original (untranslated) message along with the translations to help with message interpretation. Other channels less explicitly tied to specific messages might also benefit MT. For example, an MT (or other CMC) system might use context sensing features to show aspects of people's physical setting like location, weather, or time. This contextual information might prove useful in interpreting MT output.

Make the presence of computer agency more transparent. Showing two translations rather than one makes the MT system more transparent, exposing seams in the underlying infrastructure. Showing the fact that there are alternatives doesn't just provide a resource for repair; it also foregrounds the fact that translation is happening in the first place and that it is imperfect. Functionally, this has benefits such as encouraging people to attribute problems to the technology rather than to collaborators, in turn improving social outcomes [11]. Philosophically, supporting reflection [30] on the way systems impact our communication is also beneficial, paralleling discussions about how search and recommendation algorithms affect our experience of information [20] and how systems that track our behavior make both correct and incorrect inferences about us [23]. Making alternatives visible is a fundamental way to do this.

Use algorithmic "by-products" as resources to help people make judgments. Systems that do filtering, translation, recommendation, and other tasks for us make choices. Those choices often have data attached to them that could help people decide whether the choices are appropriate. For instance, in recommender systems, the data that supports any given recommended item can be used to help people reason about the quality of the recommendation [e.g., 18, 27, 31, 34]. Similar ideas could be applied to MT systems. When MT algorithms compute translations, they produce information such as confidence levels, word alignment choices, and alternative word translations. Most MT interfaces hide this information and show only a final output—but it might be beneficial to present such information. For example, the interface could visualize the accuracy confidence of translations by putting icons next to the messages telling users whether they should be aware of potential errors, or highlighting the parts of a message where algorithms are not able to find an accurate word alignment.

LIMITATIONS AND FUTURE DIRECTIONS

There were several limitations to this study. First, although our results implied possible effect of participants' English language ability on the grounding process, we were not able to collect any direct measures of language ability. Future studies may want to measure language ability (i.e., TOEFL scores) and include this factor in the data analysis.

Further, a possible side effect of showing two MT outputs that we did not test is that two translations might increase people's awareness of the presence of MT and the possibility of MT errors. Because previous work has shown that beliefs about the presence or absence of MT affect communication and collaboration [11]., more research is needed to separate the effects of having two translations on message understanding vs. attributions about the causes of misunderstandings.

Finally, previous work in related fields has shown other ways to facilitate MT-mediated communication, such as keyword highlighting [10] or adding semantically linked pictures [40]. We can't directly compare our approach relative to these other solutions because of differences in the tasks and procedures of the studies, but this would be a worthwhile goal for future research.

CONCLUSION

Machine translation can allow multilingual collaborators to interact via their own native languages, but MT can introduce translation errors that make communication and collaboration difficult. We examined whether providing two different translations for each message, rather than only one, could improve conversational grounding and task performance. Overall, our results suggest that when people

speak different native languages, showing two translations has many benefits and few costs.

ACKNOWLEGEMENTS

This research was funded in part by National Science Foundation grant #1318899 and an unrestricted gift from Google Inc. We thank Leslie Setlock and Huaishu Peng for their assistance and the anonymous reviewers for their valuable comments.

REFERENCES

1. Anderson, A.H. (2006). Achieving understanding in face-to-face and video-mediated multiparty interactions. *Discourse Processes, 41,* 251-287.
2. Carletta, J., Isard, A., Isard, S., Kowtko, J.C., Doherty-Sneddon, G., & Anderson, A.H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23,* 13-31.
3. Chang, P.C. (2009). *Improving Chinese-English Machine Translation through Better Source-Side Linguistic Processing.* Doctoral Dissertation. Stanford University.
4. Clark, H.H. (1996). *Using Language*. Published by Press of Cambridge, Cambridge.
5. Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*, *13,* 127-149.
6. Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22,* 1-39.
7. Diamant, E.I., Fussell, S.R., & Lo, F.L. Where did we turn wrong? Unpacking the effects of culture and technology on attributions of team performance. In *Proc. CSCW 2008*, 383-391.
8. Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied, 3,* 105–125.
9. Feely, A.J., & Harzing, A.W. (2003). Language management in multinational companies. *Cross Cultural Management, 10,* 37–52.
10. Gao, G., Wang, H.C., Cosley, D., & Fussell, S.R. (2013). Same translation but different experience: The effects of highlighting on machine-translated conversations. In *Proc. CHI 2013*, 449-458.
11. Gao, G., Xu, B., Cosley, D., & Fussell, S.R. (2014). How beliefs about the presence of machine translation impact multilingual collaborations. In *Proc. CSCW 2014*, 1549-1560.
12. Gao, G., Yamashita, N., Hautasaari, A.M.J., Echenique, A., & Fussell, S.R. (2014). Effects of public vs. private automated transcripts on multiparty communication between native and non-native English speakers. In *Proc. CHI 2014*, 843-852.
13. Gergle, D., Kraut, R.E., & Fussell, S.R. (2004). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology, 23,* 1-27.
14. Hart, S.G., & Staveland, L.E. (1998). Development of NASA-TLX: Results of empirical and theoretical research. *Advances in Psychology, 52,* 139-183.
15. Harzing, A. W., & Feely, A. J. (2008). The language barrier and its implications for HQ–subsidiary relationships. *Cross Cultural Management: An International Journal, 15,* 49–60.
16. Hautasaari, A. (2010). Machine translation effects on group interaction: An intercultural collaboration experiment. In *Proc. ICIC 2010*, 70-78.
17. Henderson, J.K. (2005). Language diversity in international management teams. *International Studies of Management and Organization, 35,* 66–82.
18. Herlocker, J.L., Konstan, J.A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proc. CSCW 2000*, 241-250.
19. Hutchins, W.J. (2005). Machine translation and human translation: In competition or in complementation? *International Journal of Translation, 13,* 5-20.
20. Introna, L.D., & Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matters. *The information society, 16,* 169-185.
21. Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116,* 26-37.
22. Ishata, T. (Eds.). (2011). *The Language Grid*. Published by Press of Springer, Heidelberg.
23. Khovanskaya, V., Baumer, E.P.S., Cosley, D., Voida, S., & Gay, G. (2013). Everybody knows what you're doing: a critical design approach to personal informatics. In *Proc. CHI 2013*, 3403-3412.
24. Li, N., & Rosson, M.B. (2014). Using annotations in online group chats. In *Proc. CHI 2014*, 863-866.
25. Lim, J., & Yang, Y.P. (1998). Exploring computer-based multilingual negotiation support for English-Chinese dyads: Can we negotiate in our native languages? *Behaviour and Information Technology, 27,* 139-151.
26. Locke, W. N., & Booth, A. D. (Eds.). (1995). *Machine Translation of Languages: Fourteen Essays*. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York.
27. McNee, S.M., Lam, S.K., Konstan, J.A., & Riedl, J. (2003). Interfaces for eliciting new user preferences in recommender systems. In *Proc. UM 2003*, 178-187.

28. Miyabe, M., & Yoshino, T. (2009). Accuracy evaluation of sentences translated to intermediate language in back translation. In *Proc. IUCS 2009*, 30-35.
29. Neeley, T. B. (2013). Language matters: Status loss and achieved status distinctions in global organizations. *Organization Science, 24,* 476-497.
30. Sengers, P., Boehner, K., David, S., & Kaye, J. (2005). Reflective design. In *Proc. AARHUS 2005*, 49-58.
31. Sharma, A., & Cosley, D. (2013). Do social explanations work?: Studying and modeling the effects of social explanations in recommender systems. In *Proc. WWW 2013*, 1133-1144.
32. Shigenobu, T. (2007). Evaluation and usability of back translation for intercultural communication. Usability and Internationalization. *Global and Local User Interface*. Springer, 259-265.
33. Slocum, J. A. (1985). Survey of machine translation: Its history, current status, and future prospects. *Computational Linguistics, 11,* 1-17.
34. Swearingen, K., & Sinha, R. (2002). Interaction design for recommender systems. *Designing Interactive Systems, 6,* 312-334.
35. Takano, Y. & Noda, A. (1995). Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing. *Lang. Learning, 45*, 657–681.
36. Tange, H., & Lauring, J. (2009). Language management and social interaction within the multilingual workplace. *Journal of Communication Management, 13,* 218-232.
37. Tenzer, H., Pudelko, M., & Harzing, A. W. (2013). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies, 45,* 508-535.
38. Veinott, E.S., Olson, J., Olsen, G.M., & Fu, X. (1999). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proc. CHI 1999*, 302-309.
39. Vilar, D., Xu, J., D'Haro, L.F., & Ney, H. (2006). Error analysis of statistical machine translation output. In *Proc. LREC 2006*, 697-702.
40. Wang, H.C., Cosley, D., & Fussell, S.R. (2010). Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proc. CSCW 2010*, 103-106.
41. Wang, H.C., Fussell, S. R., & Cosley, D. (2013). Machine translation vs. common language: Effects on idea exchange in cross-lingual groups. In *Proc. CSCW 2013*, 935-944.
42. Xu, B., Gao, G., Fussell, S.R., & Cosley, D. (2014). Improving machine translation by showing two outputs. In *Proc. CHI 2014*, 3743-3746.
43. Yamashita, N., Inaba, R., Kuzuoka, H., & Ishida, T. (2009). Difficulties in establishing common ground in multiparty groups using machine translation. In *Proc. CHI 2009*, 679-688.
44. Yamashita, N., & Ishida, T. (2006). Effects of machine translation on collaborative work. In *Proc. CSCW 2006*, 515-523.
45. Zhang, R.P., Pan, Y., & Yang, Y. (2006). A comparative case study of Google and Bing translation. In *Proc. ICERI 2012,* 3669-3673.