Machine Translation Effects on Group Interaction: An Intercultural Collaboration Experiment

Ari Hautasaari
Department of Social Informatics
Kyoto University
Sakyo-ku, Yoshida-Hommachi, 606-8501 Kyoto, Japan
+81-75-753-4820
arihau@ai.soc.i.kyoto-u.ac.jp

ABSTRACT
Even with the increasing use of machine translation to overcome language barriers it is still unclear how machine translation affects communication and interaction in intercultural groups. In this paper, we present the results of a laboratory experiment on intercultural distributed groups using machine translation-mediated chat as a communication tool. English-speaking participants from Finland and Japan worked with Japanese-speaking participants in a trading game scenario using machine translation-mediated chat as their main communication tool. Based on previous research we predicted that machine translation will help overcome the language barrier, but furthermore we predicted that machine translation would have a positive effect on social and relational communication as well as on overall group performance. In a controlled laboratory setting, machine translation proved to increase the amount of positive socioemotional messages and overall group performance in intercultural distributed groups with disparate language abilities.

Author Keywords
Machine translation, chat, group interaction, intercultural collaboration, laboratory experiment.

ACM Classification Keywords

General Terms
Experimentation, Human Factors, Languages

INTRODUCTION
With the increase of collaboration in online environments, the cultural background and the language ability of the collaborators becomes more variant. Collaboration and communication is done over great distances and often involves participants with very different linguistic backgrounds. More often than not, members of a multicultural working group have to communicate in their non-native language resulting in difficulties in communication and collaboration [9].

Multicultural organizations are engaging in intercultural collaboration in their daily routines. The multilingual and multicultural environment poses new challenges for information and knowledge sharing as well as on collaboration in distributed intercultural work groups as discussed in [13]. In recent years, chat-based collaboration tools have become popular both in real applications and in various research fields. At the time of writing, Google Wave is only available for limited preview by invitation. Google Wave, for example, is a communication and collaboration tool, which combines different types of content (text, image, video, wikis, etc...) in real-time collaboration in one online application.

Multilingual and multicultural collaboration is also being supported by various virtual team building environments to increase productivity, motivation, social interaction and collaboration skills. Existing applications include solutions for companies, such as the Finnish NoviCraft developed by TeamingStream, and support systems for open environments such as Second Life.

Machine Translation-Mediated Communication
Machine translation has been proposed as one technology to overcome cultural and language barriers in intercultural collaboration. Several multilingual communities and projects using machine translation as a supporting tool for communication already exist online and offline.

The Language Grid Project is a joint project between industry, government, universities and citizens. The

1 http://wave.google.com
2 http://wave.google.com/about.htm
3 http://www.teamingstream.com
4 http://secondlife.com
5 http://langrid.nict.go.jp
Language Grid Project aims to connect the existing language services in one infrastructure. The available language resources include machine translators, multilingual dictionaries, morphological analyzers and parallel texts. In the Language Grid, users are able to combine existing language resources to create new composite language services, such as a Japanese-English translator with a customized dictionary. Moreover, the Language Grid supports intercultural activities by providing tools for intercultural collaboration in multilingual communities. Currently these include document translation, multilingual chat, multilingual BBS and dictionary creation tools, among others [8].

Machine translation-mediated chat programs in virtual spaces have produced promising results in machine translation-mediated communication. In [14], a machine translation-mediated chat program was implemented to Second Life. The results of the study included that machine translation-mediated chat needs to follow the same design patterns as a normal chat, meaning that extra steps in posting a message in the chat program would be a hindrance for the users. Furthermore, if the participants are not properly informed about the system they tend to use more colloquial language often resulting in poor quality machine translation. However, if the users are informed about the system and its limitations they are able to communicate if they share a common motivation [14].

The evaluation of machine translation for communication use has been commonly done based on adequacy and fluency of single translated sentences. A problem with the evaluation of single sentences or utterances is that even when individual sentences are adequate and fluent the asymmetries caused by machine translation hinder the natural referring behavior of the conversational partners. Hence, there is a need for a new evaluation method for translation quality for machine translation-mediated communication [11].

Machine translation can be an effective communication tool especially in groups that share a common ground. In machine translation mediated communication in a group, it is not only important to establish common ground between the speaker and addressee, but also between the addressees. In order to effectively establish common ground between the addressees, they have to be able to monitor the interaction between the speaker and the addressee [12].

**HYPOTHESES**

In this paper, we observe the effects of machine translation in intercultural group interaction and use quantitative and qualitative data to analyze two hypotheses. Firstly, it is expected that due to a language barrier between two groups with disparate language abilities, messages sent through a chat program tend to be highly task-based with some negative socioemotional messages indicating frustration and a low amount of positive socioemotional messages.

H1: Machine translation support in chat-based interaction in intercultural groups with disparate language abilities will increase the amount of positive socioemotional messages.

Secondly, it is expected that in intercultural groups not sharing a common language, the lack of socioemotional communication will result in uneven performance within the groups. More precisely we predicted that participants would show less solidarity and collaborative effort towards other participants when socioemotional interaction is lacking.

H2: High ratio of socioemotional messages to task-based messages between the members of intercultural groups with disparate language abilities using chat to communicate has a positive effect on the overall group performance.

We conducted a laboratory experiment to test the two hypotheses. In this study, we collected chat logs, video logs and logs of game interactions as quantitative data, as well as pre- and post-experiment questionnaires. After each experiment session a short group interview was conducted and recorded with the participants.

**LABORATORY EXPERIMENT**

A laboratory experiment to test the hypotheses on the effects of machine translation in intercultural group interaction was conducted. The Shape Factory experiment [3][4] was adapted as a culturally non-bias experiment setting for the purposes of this study. The experiment participants were recruited from Kyoto University in Japan and the university of Oulu, Finland. The participants were divided into groups of three people according to their language ability. In the Japanese university, both English-speaking and Japanese-speaking groups were formed, whereas in the Finnish university the participants were proficient in Finnish and English. The experiment setting included 10 groups of three people divided to five experiment sessions and three different experiment conditions.

**The Shape Factory Experiment**

In order to study the effects of machine translation in intercultural distributed group work, a modified version of the Shape Factory experiment framework was used. In previous studies, the Shape Factory experiment has been used, for example, to study in-group and out-group effects in distributed teams [3] and the effect of collocation in distributed work [4]. In this paper, we adapted the Shape Factory experiment to a scenario where the participants use a chat program to communicate with a remote group not sharing similar language ability.

In the Shape Factory experiment, the goal of each participant was to collect a given set of shapes in each round of the experiment. The participants were able to produce a certain set of shapes with a predetermined price in each round. Consequently, since one participant could not produce all the shapes in each round, they had to buy
shapes from other participants in order to complete the given set. In a case where a participant could not complete the given set of shapes within the time limit a penalty was enforced. The penalty was given for every missing shape, but the cost of the penalty was not disclosed to the participants. In this experiment, the penalty was set to match the price of the most expensive shape sold in each round.

In each round each participant had one specialty shape that they could produce with the lowest cost for the round. In addition to the specialty shape, each participant could produce a medium priced shape and an expensive shape. Hence, all of the cheapest shapes could never be produced within the collocated groups. The set of shapes which one participant could produce changed in every round. An example of an experiment setting of shapes produced by different players in one round is illustrated in Picture 1.

![Picture 1: Example of an experiment setting. Players are indicated in color and their most expensive shapes are on the left and the cheapest on the right, respectively.](image)

<table>
<thead>
<tr>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Shapes" /></td>
<td><img src="image" alt="Shapes" /></td>
</tr>
</tbody>
</table>

The participants were divided in two groups of three people, one Japanese-speaking group and one English-speaking group. The participants were allowed to communicate with the other members in the collocated group by any means they felt appropriate, including a private chat function. Every participant was assigned to a color, which was used as the name of the participant both in the game interactions and chat (Blue, Green, Red, Orange, Yellow and Black). The remote group could be reached only with a chat program provided with a private chat window to each participant. The participants were allowed to choose freely in which language they would communicate with each other.

The participants were instructed on the game interaction before the experiment. The first round was set to 20 minutes in order for the participants to get familiar with the scenario, and subsequent rounds were set to 15 minutes in order to create time pressure on the transactions. Time pressure forces the participants to make a choice on who to give their time to and who to collaborate with.

After each round the participants were evaluated by cost efficiency. The total price paid for the bought shapes was deducted from the difference of the total cost of produced shapes and the total price of sold shapes. Hence, the performance of a participant was measured by how low the participants’ final score was. The winner was determined by the totaling the scores of individual rounds. An example illustration of the game interaction interface is shown in Picture 2.

**Machine Translation Supported Chat**

In the laboratory experiment, the participants in both groups were able to communicate with each other via a private chat system running on a web browser. Since the two groups were not collocated, the chat system was the only communication channel between the two groups. The system included a private chat window with complete message history between each participant, meaning that every participant had 5 private chat windows open simultaneously (Picture 3). The collocated participants were instructed to communicate freely with the media of their choice.

![Picture 3: The machine translation-mediated chat interface for Orange player.](image)

In the experiment conditions, the multilingual groups were given either a chat program with machine translation functionality or a normal chat with a possibility to write in their native language, but without machine translation support. The chat program was developed specifically for the purposes of this experiment while still keeping the design intuitive to experienced chat users. In the machine translation supported chat, both the original message and the translated message were displayed to the participants.
In order to support collaboration in the native language of the participants, machine translation functionality was implemented directly in the chat system. In this experiment, a Japanese – English machine translator (J-Server) provided by the Language Grid was used. The functionality of the chat was preserved between the two conditions so that participants could write a message and press the enter-key once in order to send the message. The original language of the message was identified and translated accordingly immediately after a participant pressed the enter-key.

A back-translation field was added in every private chat window in order to give monolingual users an option to evaluate the machine translation quality of their messages. Back-translation is essentially an automatic translation of a machine translated text back to the original language, a function which can be used to improve the accuracy of the translation [7]. In the experiment scenario, the back-translation field had two functions in the participants’ point of view; firstly, to give an opportunity to correct misunderstandings caused by misspellings or phrasing in the machine translated messages post hoc, and secondly, to give transparency to the translated messages hence increasing participants’ trust on the system. The reason for post hoc correction of sentences was to preserve the natural flow of a chat program, as noted also in [14]. Furthermore, since the English-speaking participants were not able to read Japanese (average response was 1.95 on a Likert scale of 1=Do not understand and 5=Native in the pre-experiment questionnaire about Japanese language proficiency) there was a need for a method to confirm that a message has been translated and sent accordingly.

**Experiment Conditions**

In this experiment, we conducted the aforementioned Shape Factory scenario in three conditions. In the first condition, Japanese-speaking groups were paired with English-speaking groups without machine translation support. In the second condition, Japanese-speaking groups were paired with English-speaking groups with machine translation support. In the third condition, we paired two English-speaking groups without machine translation support as a control group. In all conditions the participants were able to use the chat program in their native language, but only the second condition included the machine translation functionality.

**INTERACTION PROCESS ANALYSIS USING THE BALES’ CATEGORIES**

In order to categorize the content of the interaction in the experiment setting and analyze the effects of machine translation in collaborative tasks, an adaptation of the Bales’ interaction process analysis (IPA) system [1] was used. A previous modification of the IPA was done in [6] for the purposes of analyzing message content in an online discussion board. The modification in [6] was used as the basis for the adapted version used in this study.

The Bales’ system consists of 12 categories for small group interaction analysis. The categories are divided further in to four areas: Social-Emotional Area (Positive Reactions), Social-Emotional Area (Negative Reactions), Task Area (Attempted Answers), and Task Area (Questions) [1]. All of the categories have a corresponding opposite in another area, for example Category 4 (Gives Suggestion) and Category 9 (Asks For Suggestion). The adaptation of the Bales’ system for online synchronous chat interaction analysis with complete explanation of the categories can be found in Appendix A.

The socioemotional messages are used for social and relational maintenance between discussion participants. The negative socioemotional messages, on the other hand, indicate frustration and unwillingness to collaborate, whereas the amount of positive social messages can be seen as an indicator of a group producing high quality work. Task-based messages are related to answering and asking questions about procedures or content of the task [6].

In this study, the unit of analysis was one thought unit; a complete thought conveyed through the chat system, for example, “I would like to buy your triangle” or “Thank you”. One chat message could include multiple lines and sentences, and thus multiple units of analysis. In order to overcome the difficulties with colloquial style and certain specific characteristics of chat-based communication, some thresholds were set for the analysis process. In general, one unit of analysis was a complete sentence, but in cases where colloquial language, chat-specific expressions (i.e. emoticons) or non-complete sentences were used the unit was measured by line change or any punctuation mark.

As shown in [6], the Bales’ system adapts well for online discussion with minor adaptation and clarification to the categorization. In addition to the modifications in [6], spam and flooding were added to Category 11, meaning repetitive unmodified messages in the chat room indicating frustration. Excessive use of capital letters and punctuation marks (for example exclamation marks) was also added in this category. Since synchronous chat differs in the method of posting a message from an asynchronous media, such as a message board observed in [6], a threshold for separate messages had to be established. In this experiment, we used 4 seconds as a threshold to separate flooding, spamming and accidentally sent messages from clarifying messages. In addition, the threshold for the number of consequent identical messages identified as spam was set to three. For example [Do you have a]...[heart?] was handled as one message, but [Do you have a heart]...[?] was handled as separate messages if the delay between the messages exceeded 4 seconds. [Do you have a heart?] posted unmodified over three times in less than 4 second intervals was considered as spam and categorized accordingly.

**ANALYSIS**

In this paper, we concentrate on the content of the interaction and how machine translation as a supporting
tool affects the communication and collaboration between intercultural groups. Furthermore, we will discuss the correlation with the message content on overall group performance, specifically the effect of positive socioemotional messages in distributed intercultural groups in a laboratory setting. As mentioned in [5], the success of the evaluation of machine translation systems depends highly on the choice of application for the task in hand. In this study, we do not evaluate the quality of the machine translation per se, but the effect that machine translation has on the overall interaction in distributed intercultural groups with disparate language abilities.

Previous studies on online communication suggest that the proportion of positive socioemotional messages in small group interaction is dependent on the nature of the task. In recreational communication, the proportion of socioemotional messages compared to task oriented messages is high, whereas in instrumental communication the opposite is true, in both face-to-face and mediated communication [10]. In previous research, the positive social messages have been proposed to indicate cohesion, solidarity and better overall group performance [2][6]. In this study, our hypotheses were that machine translation increases the amount of positive socioemotional messages and coincidentally the overall group performance.

The experiment was conducted in five two-hour sessions. From the private chat logs between the participants, 2583 units of analysis were collected including the machine translated messages with the corresponding original messages. In addition to the chat logs, video logs of the intra-group interaction and post-experiment interviews as well as 38 questionnaire items from each participant were collected. The log data was analyzed together with the data from the Shape Factory game interaction logs.

**Content Analysis**

Bales’ interaction process analysis coding scheme was adapted and used to analyze the inter- and intra-group interaction through the chat program (Appendix A). In this paper, we will concentrate on the interaction conducted through the provided chat program, hence the intra-group face-to-face interaction is not in the scope of this study.

The chat logs were derived and imported after each experiment session. In the machine translation supported chat scenario, the original message was preserved along with the translated message (Japanese-English message pairs). The chat logs included a timestamp for each message pair in the form of [hh:mm:ss]. The timestamp was used in the categorization of chat-specific messages and behavior.

The complete game interaction logs were coded from the user interface and later checked for accuracy. The individual performance, as well as group performance, were calculated and compared against the interaction data. The pre- and post-experiment questionnaires were used to determine the participants’ experience with computers, language ability and subjective experiences on the experiment scenario and performance.

**Chat Interaction Examples**

The following section includes real examples derived from the chat logs gathered in this experiment. Machine translation results are indicated by (T) and shown under the corresponding original sentences.

Example 1: A completed transaction in the machine translation support condition. Red participant (Japanese) and Black participant (English).

<red> 何を持っていますか？
<red> (T) What do you have?
<black> 1 have a heart, a plus and an arrow
<black> (T) 私は心臓、プラス、および矢を持っている。
<red> それぞれいくらですか？
<red> (T) How much is it respectively?
<black> 100 for heart, 60 for plus and 30 for arrow
<black> (T) 心臓のための100、プラスのための60、および矢のための30
<red> わかりました。あなたは何が欲しいですか
<red> (T) I see. What do you want?
<black> Do you have a triangle?
<black> (T) 三角形を持っているか？
<red> あります。40です。
<red> (T) I have that. It's 40.
<black> I'd like to buy it
<black> (T) 私は、それを買いたい。
<red> わかりました。売ります。
<red> (T) I see. It's sold.
<black> Thank you
<black> (T) ありがとうございます

Example 1 includes an incorrect translation (“..for heart.”), where the utterance is translated to Japanese literally instead of conveying the intended meaning of the message. Regardless, the transaction was completed successfully.

Example 2: Positive socioemotional messages in machine translation support condition. The simplest form of positive socioemotional message, “Thank you very much”. Green participant (Japanese) and Orange participant (English).

<orange> Would you like to buy triangle for 30?
<orange> (T) 30に三角形を買いたいか？
<green> ありがとうございます。
<green> (T) Thank you very much. I'd like to buy it.

Example 3: Negative socioemotional messages. Example of spamming as indication of frustration. Green participant (English control group).

<green> Hello
<green> Hello
<green> 件の件についてお話ししましょう。
<green> (T) Let's talk about the issue.
<green> うるさい。
<green> (T) Noisy.
<green> なぜ話していますか？
<green> (T) Why are you talking?
<green>  Yorkshire.
<green> (T) Yorkshire.
RESULTS

The first hypothesis presented in this paper concerned the amount of positive socioemotional messages (e.g. “Thank you for the items”) as opposed to task-based messages (e.g. “I would like to buy a heart from you”) sent between the experiment participants through the chat-system. In the experiment, we observed the effect of machine translation on the distribution of messages in the Bales’ interaction analysis scheme in intercultural groups who do not share similar language abilities. In both intercultural groups (with machine translation and without machine translation support), the average response on English language proficiency was 2.45 for the Japanese-speaking participants, whereas the average response on Japanese language proficiency was 1.95 for the English-speaking participants on a Likert scale of 1=Do not understand and 5=Native.

Table 1 illustrates the means and standard deviations of message categories in intercultural groups without machine translation support (Japanese-English). The average number of messages was calculated by categorizing the messages from every round played in this condition. The six task-based message categories (Categories 4-9) are handled as one area, whereas positive socioemotional messages (Categories 1-3) and negative socioemotional messages (Categories 10-12) are handled separately. The inter-rater agreement in all experiment conditions was high (kappa=0.96).

Table 1: Means and standard deviations of average number of messages per category. Japanese-English groups without machine translation support.

<table>
<thead>
<tr>
<th>Category</th>
<th>Average # of Messages per Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>M 40.50, SD 21.57</td>
</tr>
<tr>
<td>Positive Socioemotional</td>
<td>M 8.66, SD 10.13</td>
</tr>
<tr>
<td>Negative Socioemotional</td>
<td>M 9.83, SD 7.87</td>
</tr>
</tbody>
</table>

Note: Averages are based on the number of messages produced in each category.

The participants in this condition sent a high amount of task-based messages compared to positive socioemotional messages. What is also notable is the amount of negative socioemotional messages that tended to increase towards the end of each experiment session. Negative socioemotional messages indicate disagreement, frustration and antagonism towards other participants. In this experiment, the negative messages were almost exclusively directed towards the remote group members. In the no machine translation support condition, passive rejection by not answering requests (Category 10) and messages indicating frustration (Category 11) increased proportionally towards the end of each experiment session.

Table 2 illustrates the means and standard deviations of message categories in intercultural groups with machine translation support (Japanese-English). In this condition, the participants were able to communicate in their preferred language, meaning that it was possible for a Japanese participants to communicate in English if they chose to do so and vice versa.

Table 2: Means and standard deviations of average number of messages per category. Japanese-English groups with machine translation support.

<table>
<thead>
<tr>
<th>Category</th>
<th>Average # of Messages per Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>M 52.17, SD 19.80</td>
</tr>
<tr>
<td>Positive Socioemotional</td>
<td>M 29.25, SD 13.66</td>
</tr>
<tr>
<td>Negative Socioemotional</td>
<td>M 2.75, SD 2.59</td>
</tr>
</tbody>
</table>

Note: Averages are based on the number of messages produced in each category.

In the groups with machine translation support, the amount of positive socioemotional messages increased to almost four times the amount per participant compared to the groups without machine translation support (t [22] = 4.19, p < .01). Besides the positive socioemotional messages the amount of task-based messages also increased slightly indicating higher amount of transactions, or transaction attempts, in the game scenario. The amount of negative socioemotional messages was almost non-existent compared to the group without machine translation support, consisting mostly of disagreements (Category 10) (t [22] = 2.96, p < .01).

Table 3: Means and standard deviations of average number of messages per category. English-speaking groups without machine translation support.

<table>
<thead>
<tr>
<th>Category</th>
<th>Average # of Messages per Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>M 82.83, SD 18.10</td>
</tr>
<tr>
<td>Positive Socioemotional</td>
<td>M 33.67, SD 19.51</td>
</tr>
<tr>
<td>Negative Socioemotional</td>
<td>M 27.67, SD 22.92</td>
</tr>
</tbody>
</table>

Note: Averages are based on the number of messages produced in each category.
Table 3 illustrates the means and standard deviations of message categories in a control group (English-English) without machine translation support. In the control group, the average response on English language proficiency was 4.2 on a Likert scale of 1=Do not understand and 5=Native.

As expected, the control group produced the highest amount of messages overall. The average amount of positive socioemotional messages sent per participant was roughly equal to the intercultural groups with machine translation support, and considerably higher than in the groups without machine translation support ($F [2, 27] = 9.51, p < .01$). The amount of task-based messages was also higher compared to the two other experiment conditions. The control group produced a high amount of negative socioemotional messages, however, as the standard deviation indicates the negative messages were not as equally distributed between the participants as in the other groups. Most of the negative socioemotional messages were sings of frustration (Category 11) resulting from anxiety caused by perceived bad performance or stress over the time limit. In fact, against initial prediction the time limit set to individual rounds proved not to be a major obstacle for groups with machine translation support, but to the groups with plain chat functionality.

**Group Performance**

The second hypothesis in this paper concerned the correlation between overall performance in intercultural groups and the amount of positive socioemotional messages. In the previous chapter, we showed that machine translation support did have a positive impact on the translation support of the Japanese-speaking and English-speaking groups are calculated separately along with the control group in which the remote English-speaking group was located in Finland.

In this experiment, machine translation group 1 (MT Group 1) performed the best overall. The average round scores within the distributed groups were relatively low, optimum being 120. Since in this experiment setting the participants could not complete the given task in each round without buying shapes from the remote group, the average scores of the distributed groups indicates a high amount of collaboration in intra- and inter-group.

Low amount of penalties is also an indication of high amount of inter-group collaboration. Since most of the shapes were available in the collocated group with a higher production cost, the penalties were in general given to participants who did not manage to gather the shapes from the remote group. To avoid the penalty, participants tended to buy the expensive shapes in the collocated group with a higher price at the end of the round if they failed to buy them from the remote group.

The intercultural groups without machine translation support (Chat Group 1 and 2) scored fairly low in the experiment. In the Chat Group 1, the difference between the English-speaking and Japanese-speaking average group score was also high indicating higher amount of collaboration in the English-speaking intra-group. Both Chat Groups received a notable amount of penalties, meaning unfilled orders of shapes to complete the task. As seen in Table 1, the average amount of task-based messages indicates that effort was made to complete the task (collect the given set of shapes), but with little success.

The Chat Groups sent the least amount of positive socioemotional messages (13.3% and 15.9% of all messages sent in one distributed group), whereas the MT Groups sent a high amount of positive socioemotional messages (32.8% and 38.3% of all messages sent in one distributed group). Consequently, the overall group performance in both Chat Groups was significantly worse compared to both MT Groups ($t [22] = 2.81, p = .01$). The MT Group 1 scored the best in the overall performance, but what is also notable is the positive impact of machine translation support on the performance of the Japanese-speaking groups ($t [10] = 4.34, p < .01$).

**CONCLUSION**

The analysis of the chat log data revealed that machine translation support in a chat system increases the amount of positive socioemotional messages in intercultural groups with disparate language abilities (H1). Although the laboratory experiment was done in a closed environment,
and furthermore, only a limited vocabulary was needed to successfully complete the tasks, the experiment shows clear improvement in overall group performance with machine translation support (H2).

Not only did the amount of positive socioemotional messages increase in groups using machine translation, but also the average amount of task-based messages increased while the amount of negative socioemotional messages decreased. Higher amounts of task-based messages, along with lower amount of penalties in the game scenario and lower amount of negative socioemotional messages, indicate that machine translation enhanced collaboration and overall performance in the intercultural distributed groups.

The evaluation method for machine translation-mediated communication was chosen in accordance to [5] and [11]. Hence, the direct evaluation of accuracy and fluency of machine translation was omitted from this study. Instead we asked the participants in a post-experiment interview whether they “…found communication difficult?”, “…found the machine translation as a hindrance?” and “…felt that the used technology affected the game performance?”. Overall, the participants answered the questions negatively, which indicates that content-based evaluation method was suitable for the purposes of this study.

**DISCUSSION**

The laboratory experiment scenario used in this study provided a surprisingly large amount of data in the form of chat logs. A closed environment, such as our experiment scenario, does not facilitate free discussion well due to strict time limits. Lengthening the round time would, however, omit the time pressure from the game interactions. Though it is not clear whether the round length affects the amount of informal chat, an experiment with longer rounds should be conducted.

In this study, we found that the quality of the machine translation (adequacy and fluency) was not a large factor in the experiment performance and communication. This is most likely due to the limited vocabulary used in the experiment and a shared common ground through the game interactions. We also found that the back-translation field and the chat functionality helped make the communication fluent and natural. Back-translation provided a way to confirm that the message was translated correctly, and the one-click interface was found intuitive to use. Furthermore, making the machine translation functionality transparent and instructing the participants to be aware of the system increased the success of communication, as also stated in [14].

**Limitations of the Experiment**

Even though we had some cultural diversity in the experiment participants, the range of languages, and subsequently participants with different cultural backgrounds, was limited. In this experiment, we looked into communication and collaboration between Japanese-speaking and English-speaking groups. The reason for choosing English as one language seems apparent since in a real life setting English is often used as a second language when no common language is available. One of the reasons for choosing Japanese-speakers for the second language group in this experiment was the quality and availability of current machine translators. The Language Grid offers multiple Japanese-English machine translators with an acceptable quality, but as noted in this study, the quality of a direct translation was not a major factor in communication in a controlled task-oriented environment.

In the future, we are planning to expand the languages used in the experiment setting. Naturally, the quality of the machine translation has an effect in choosing the language pairs, but the overall quality of English machine translators, especially to Latin-based languages, should be sufficient for small group interaction.

The most prominent reason for the limited amount of intercultural participants and supported languages was resources. Computer supported intercultural collaboration experiments can be organized over the Internet relatively easily, but organizing experiments with two culturally homogenous groups usually requires presence in two corresponding countries. Also, the participants in our experiment were either university students or university staff members. In the future experiments, it would be advisable to extend the demographic of the participants outside university students. An ideal experiment group would consist of a real life organization engaging in multilingual computer supported collaborative work in their daily routines.

**FUTURE WORK**

In the future work, we are planning to extend the experiments to include other languages and participants from different cultures. Also, even though the Shape Factory scenario used in this study was adapted keeping cultural bias in mind we need to evaluate whether modifications to the scenario are needed for experiments in more variant cultural settings.

In this experiment, we collected a large amount of data on communication, interaction and collaboration. In the future, we are planning to use the existing data along with data from complementing experiments to analyze social trust creation in intercultural groups. Furthermore data on intra-group interaction was not discussed widely in this paper. In the future work, intra-group face-to-face interaction will be included to the analysis.

We also plan to use the machine translation chat in experiments on other cultural aspects of collaboration. If machine translation can successfully help intercultural teams to overcome language barriers, the next step would be to overcome the cultural barriers as well.
ACKNOWLEDGMENTS
This work was supported by a Grant-in-Aid for Scientific Research (A) (21240014, 2009-2011) from Japan Society for the Promotion of Science (JSPS), and Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society.

REFERENCES
Appendix A: Modification of the Bales' Interaction Analysis Coding Scheme (Modification from [6] by author)

1. Shows solidarity, raises other’s status, gives help, reward. Examples: I think we did a great job!; I really like the example you used; thank you.
2. Shows tension release, jokes, laughs, shows satisfaction. Examples: Well, I guess I need to learn to spell!; lol; I’m sorry; apologies; explanation for not meeting an expectation.
3. Agrees, shows passive acceptance, understand, concurs, complies, confirms others opinions or analysis. Examples: I agree with what you have said, I know what you mean; mention of another person or the group in agreeing.
4. Gives suggestion or direction, implying autonomy for others; we use this category for procedural statements about the task or assignment at hand. Examples: Here is what we need to do next; I suggest we leave this question and come back; maybe we should do this in a chat room; I want to buy X from you if you can sell it to me.
5. Gives opinion, evaluation, analysis, expresses feelings, wishes. These statements are about feelings or opinions on the task. Example: my analysis was; I think; personal examples.
6. Gives orientation, information, repeats, clarifies, confirms. These are facts or supported information from a source (observable) or the text about the content of the task. Example: the text said X; the definition of this was X; expressing agreement with the book; citing examples from the book; I want to buy X.
7. Asks for orientation, information, repetition, confirmation. These questions are similar to those in Category 6 in that they are questions about facts and supported (observable) information. Examples: How does the book define X; didn’t the text say this; the definition, the statement, or the question was X.
8. Asks for opinion, evaluation, analysis, expression of feeling. This category is similar to 5, except that it encompasses asking for these things rather than offering them. These questions may be directed to individuals or the group as a whole. Examples: what do you think about; how do you feel; what did you think when reading; what is your take on this, what do you think?
9. Asks for suggestions, direction, possible ways of action. This category is similar to 4 in that we have determined these are questions about procedures, clarifying the task and asking how to accomplish the task. Examples: are we doing this right; should we do this in a chat room; when do we have to have this done; is it possible to work together?
10. Disagrees, shows passive rejection, formality, withholds help. These comments show negative social emotional reactions. Examples: I can’t agree to that idea; I don’t believe you are correct in that remark; I don’t have anything to say about that.
11. Shows tension, asks for help, withdraws out of field. Examples: I don’t think we can get this done in time; I don’t know how we are supposed to do this without more help; spam or flooding; excessive use of capital letters and punctuation marks; PLEASE TALK TO ME!!!!.
12. Shows antagonism, deflates another’s status, defends or asserts self. Examples: Since I’m a mother and you are not, I think I would know; I can’t believe anyone would make the comments you made; come on boy, be reasonable; my group is stupid.

Note: Categories 1 through 3 and 10 through 12 are about relational group work (helping to sustain or weaken relational ties within the group). Categories 4 and 9 are about procedures and how to go about completing the task. Categories 5 through 8 are content-related comments or questions. [6]

Note (by author): Added examples to the modification in [6] by author are written in italics.

Note (by author): Highly task oriented messages (attempted answers) were divided between categories 4 and 6 depending on the content. For example, [I want to buy a box from you if you can sell it to me] was considered to be in Category 4, and [I want to buy a box] in Category 6.