# Analyzing Misconceptions in Multilingual Computer-Mediated Communication

Naomi Yamashita
Social Communication Lab.
NTT Communication Science Labs.
2-4 Hikaridai Seika-cho
Soraku-gun Kyoto
Japan
naomi@cslab.kecl.ntt.co.jp

Toru Ishida
Department of Social Informatics
Kyoto University
Yoshida-Honmachi
Sakyo-ku Kyoto
Japan
ishida@i.kyoto-u.ac.jp

## ABSTRACT
Multilingual communities using machine translation to overcome language barriers are showing up with increasing frequency. However, when a large number of translation errors get mixed into conversation, it becomes difficult for users to fully understand each other. In this paper, we focus on misconceptions found in high volume in actual online conversations using machine translation. By comparing responses via machine translation and responses without machine translation, we extract two response patterns, which may be strongly related to the occurrence of misconceptions in machine translation-mediated communication. The two response patterns are that users tend to respond to short phrases of the original message and tend to trip on the wording of the original message when responding via machine translation.

## Categories and Subject Descriptors
H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work, Asynchronous interaction

## General Terms
Human Factors, Experimentation.

## Keywords
Multilingual Groups, Machine Translation, Computer-Mediated Communication, Misconception.

## INTRODUCTION
Multilingual communities using machine translation to overcome language barriers are showing up with increasing frequency. Such communities are centered in Eastern Asia, where their lingua franca is English, but few are actually proficient in the language. The number of such communities is expected to grow in the future.

However, pitfalls exist in communications using machine translations, especially when the translation quality is too low. One pitfall is the occurrence of misconceptions among participants. When translation quality is low (as in Eastern Asia due to grammatical construction dissimilarity), the load on participants to "guess" the mistranslated part of comments becomes high. Under such circumstances, participants may falsely assume that others are speaking and understanding on the basis of the same information and interpretation. In this paper, we refer to such a phenomenon as "misconception."

Such misconceptions can be very problematic, especially in group work. For example, in collaborative situations where discussion members do not realize the emergence of misconceptions, it is possible for the confounding of expectations to occur, having repercussions for subsequent conversations.



## MISCONCEPTIONS IN OUR CASE STUDY

### Case: Asia Broadband Project
The Asia Broadband Project is a project conducted in 2003 by Chinese and Japanese universities and research institutes with the support of the Japanese government.

Eighteen Japanese members and sixteen Chinese members joined the project. Their mission was to discuss and settle on an implementable tool to encourage intercultural collaboration within one month. During the project, all the discussions took place on a machine translation-embedded BBS, which automatically translates Chinese and Japanese messages and displays both original and translated messages. Since none of the members understood both Chinese and Japanese, all members posted and read the messages in their native languages.

### Massive Amount of Misconceptions
In the Asia-broadband Project, we've found a massive amount of said misconceptions in the Asia-broadband Project. Most misconceptions were detected in response messages posted from different country members.

Although the conversations were riddled with misconceptions, it seemed that the participants rarely noticed that they misconceived each other. Indeed, more than half of the participants answered (in our interview in English) that they could "often" understand the general outline of the translated messages.

The accumulation of misconceptions caused serious communication breakdowns. For example, Chinese and Japanese members differed in their perceptions of what they agreed on as

conclusions of the one-month discussion. Interestingly, they did not even know that there was a mismatch between conclusions reached between Chinese and Japanese members.

## Misconception Types

Based on our observation, we classify misconceptions into two types: *unsaid* and *said*.

Unsaid misconception occurs when a speaker falsely assumes that discussion members share mutual knowledge and proceed with much unsaid. Misconceptions of this type are usually found in elliptic discourse, where people "believe" that they share a lot of knowledge.

Said misconception occurs when a listener gets the meaning of a speaker's comments wrong. Misconception of this type is a sort of mis-meaning between discussion members, occurring typically from clearly stated comments. For example, while a speaker is talking about A, listeners may think that the speaker is talking about something else.

Although most misconceptions in usual conversations occur from the unsaid part, most misconceptions found in the Asia-broadband Project were said misconceptions. In the following, we focus on said misconceptions.

# RESPONSE PATTERNS

We analyze the response pattern of machine translation-mediated communication and investigate how such patterns generated confusion among members. In the following, we call response pairs *parent-child pairs*, where child indicates the response to the parent.

## Method

A basic assumption underlying this investigation is that people guess the meaning of others' messages based on words and guess how the messages are related based on lexical cohesion between messages, especially when discussing via low-quality machine translation. Misconceptions occur when such guesses are wrong. Thus, insights about how machine translation generated misconceptions may be gleaned using lexical cohesion analysis.

In this paper, we measured lexical cohesion between messages based on lexical items and synonyms they share. We refer to the lexical items and synonyms shared between messages as cohesive lexical items.

We first gathered message pairs in direct responses. Next, we divided the pairs into two groups based on whether parent and child are both posted from the same country or from different countries. We compared the two groups' response patterns using lexical cohesion.

## Responses Tripping on Others' Wording

We compared parent-child pairs posted from the same country and from different countries by the number of cohesive lexical items shared between parent-child pairs.

The ratio of parent-child pairs where cohesive lexical items are not shared at all is similar (10 to 15%) between those posted by members from the same country and those posted by members from different countries. From further detailed analysis on such parent-child pairs, we found that the content of the child message tend to be simple, such as greetings, agreement, short comments, etc.

Meanwhile, the ratio of parent-child pairs including more than five cohesive lexical items differ significantly between those posted by members from the same country (around 60%) and those posted by members from different countries (around 40%). Further t tests proved that parent-child pairs posted from the same country share significantly more cohesive lexical items than parent-child pairs posted from different countries ($F=16.078$, $p=0.000$).

From a further detailed analysis, we found many cases where response messages posted from different countries trip on the parent message's wordings. Such responses tended to be incoherent and inadequate as a response even if the parent-child messages shared some cohesive lexical items.

## Responses Focusing on Short Phrases

In general, the translation quality of a sentence decreases as its length increases. Thus, it must be difficult to fully understand translated messages when they contain long sentences. We expect that members respond especially to short phrases, since translation quality of short phrases is relatively high and thus understandable.

To investigate our hypothesis, we compared parent-child pairs posted from the same country and from different countries by the distribution of the parent message's length of sentences, where the sentence and the child message share cohesive lexical items.

From the comparison, we found out that when the length of a sentence (in a message) is long, responses to the message from other countries tend "not" to mention the long sentence. Further t tests proved that the average length of sentences where the sentence and the child message share cohesive lexical items is significantly different between the two groups; response messages from members in other countries are significantly more focused on short phrases of its parent message than the response messages of members from the same country ($F=4.816$, $p=0.029$).

The result supports our hypothesis. Members tend to respond only to short phrases, which are easy to understand.

# CONCLUSIONS

Reaching mutual understanding over lean media using low-quality machine translation is a difficult task. When members do not fully understand others' comments, people tend to speak and understand what is said on the basis of their own information and interpretation of the situation, falsely assuming that the other speaks and understands on the basis of that same information and interpretation.

Based on our experimental research, we found out that response messages via low-quality machine translations tend to trip on the parent message's wordings and tend to focus on short phrases of its parent message.

By searching such communication patterns, we may be able to automatically assess the likeliness that each dialogue includes misconceptions. By incorporating such a mechanism into a system, we may be able to alert discussion members with the occurrence of misconceptions.