Assessing the Impact of Real-Time Machine Translation on Requirements Meetings: A Replicated Experiment

Fabio Calefato, Filippo Lanubile
University of Bari
Dipartimento di Informatica
Bari, Italy
calefato,lanubile@di.uniba.it

Tayana Conte
Universidade Federal do Amazonas
Instituto de Computação
Manaus, Brazil
tayana@dcc.ufam.edu.br

Rafael Prikladnicki
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre, Brazil
rafael.prikladnicki@pucrs.br

ABSTRACT
Opportunities for global software development are limited in those countries with a lack of English-speaking professionals. Machine translation technology is today available in the form of cross-language web services and can be embedded into multiuser and multilingual chats without disrupting the conversation flow. However, we still lack a thorough understanding of how real-time machine translation may affect communication in global software teams.

In this paper, we present the replication of a controlled experiment that assesses the effect of real-time machine translation on multilingual teams while engaged in distributed requirements meetings. In particular, in this replication we specifically evaluate whether non-English speaking groups benefit from communicating in their own native languages when their English is not fluid enough for a fast-paced conversation.

Categories and Subject Descriptors
D.2.9 [**Management**]: Programming teams

H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing.

I.2.7 [**Natural Language Processing**]: Machine translation.

General Terms
Experimentation, Human Factors.

Keywords
Controlled experiment; global software engineering; machine translation; requirements meetings.

1. INTRODUCTION
Global Software Development (GSD) is characterized by the dispersion of stakeholders across different countries, continents and time zones. Requirements engineering is one of the most communication-intensive activities in software development and, thus, it suffers much from language difficulties in global software projects [10], [11], [20]. Language is indeed an important factor that largely accounts for the success of offshore IT work in countries with strong English language capabilities, such as Ireland, the Philippines, India, and Singapore [6], [15].

However, there are several other countries, considered followers in global competition, which are increasing their presence in the global IT market. Brazil is one real example of this situation [7]. Brazil's IT industry is large – A.T. Kearney consultancy estimates that the sector employs 1.7 million people, including programmers, systems analysts, and managers [17] – and it is growing by 6.5% a year on average since 2005 [2], although the vast majority of the IT companies are focused on domestic clients and do not export. For those who export, US companies are the main clients, accounting for over 80% of demand, followed by Latin America (especially Argentina, Chile, Colombia and Mexico), and Europe (especially Germany, Spain, France, England and Portugal). Nearly 100% of Brazil's IT export clients have time zone overlap with this country [10]. However, in order to take full advantage of the time zone overlap, Brazilian sites should create richer interactions with their foreign partners. This could avoid problems such as coordination breakdown, asynchronous and not so frequent communication, lack of interactive work, among other problems that lack of rich interaction may cause. And one key element for this is more effort on the English. Unfortunately, A.T. Kearney estimates that Brazil has only 10.2 million of English speakers, or 5.4% of the population. Chile, for example, has 34.7% of English speakers; India has 8.2% (which represents 90.6 million). Another study published by KPMG in 2009 indicated that one of the disadvantages of Latin American countries is the lack of English speaking professionals [18]. In this context, there are several initiatives going on, for example, in order to include English in the qualification of the IT professionals in Brazil [7]. However, this may be not enough and, to stay competitive in the global IT market these countries we will have to search for alternative solutions. For this reason, distributed project meetings, such as requirements workshops, can benefit from machine translation, as this technology is today available in the form of cross-language chat services and it might be used in countries, such as Brazil, where there are at the same time opportunities for global projects and the lack of English speaking professionals.

Machine translation (MT) is an established technology that uses software to translate text or speech from one natural language to another. The idea of using digital computers for translation of natural languages was proposed 50 years ago [14]. The technology available today – i.e., real-time, online conversation – is experiencing tremendous growth of interest, mostly because of

ESEM'12, Sept. 19–20, 2012, Lund, Sweden.

the Internet continuous expansion. The rise of social networking has also contributed to this growing interest, allowing users of social media to speak different languages to communicate with each other. Despite the recent progress of the technology, we still lack a thorough understanding of how real-time machine translation affects communication.

In our previous works, we first run a simulated study, which proved that state-of-the-art machine translation services, such as Google Translate, could be embedded into synchronous text-based chat with a negligible extra time [3]. However, the simulation could not say anything about completing complex group tasks while communicating with multiple native languages.

Then, we conducted a controlled experiment to investigate whether real-time machine translation could be successfully used instead of English in distributed multilingual requirements meetings [5]. We could observe that, despite far from 100% accuracy, real-time machine translation was not disruptive of the conversation flow and, therefore, accepted with favor by participants. However, since we involved only groups with high English proficiency, we concluded that stronger effects could be expected to emerge when language barriers are more critical.

Now, we have replicated the former study by means of a controlled experiment which involves participants who are not proficient in English, that is, they are not able to communicate in English as in their mother tongue. With respect to the initial study, other than changing the level of English proficiency, Portuguese-speaking participants came from the North of Brazil rather than the South Region of Brazil. From the initial experiment, we reused the research questions, the experimental plan, the variables and the instrumentation.

The remainder of this paper is structured as follows. Section 2 describes the experiment, including the design, the variables, the instrumentation and execution. Section 3 presents the results from data analysis. Section 4 discusses the results and the differences between the two replications. Threats to validity are described in Section 5. Finally, conclusions and future research activities are presented in Section 6.

2. THE EXPERIMENT

We are interested to further evaluate the effect of real-time machine translation on multilingual groups of individuals. Thus, the research questions we inherit from the previous study are the following:

RQ1: *Can machine translation services be used in distributed multilingual requirements meetings, instead of English?*

RQ2: *How does the adoption of machine translation affect group interaction in distributed multilingual requirements meetings, as compared to the use of English?*

Since a better command of language provides better opportunities of steering communication during meetings, one could reasonably argue that machine translation is more useful to those who are not proficient in English. Therefore, we add the following research question:

RQ3: *Do individuals with a low English proficiency level benefit more than individuals with a high level when using their native language, assisted by real-time translation?*

We have investigated these research questions by means of a replication of the original controlled experiment. The 16 participants involved in the replication were graduate and undergraduate students from Brazil and Italy. The former were from the Federal University of Amazonas in Manaus, while Italian students were from University of Bari. In particular, students interacted in groups of four people, two from Italy and two from Brazil, using two different communication modalities, that is, their respective native language (i.e., Italian or Portuguese), with the help of machine translation (*MT*), and English (*EN*), as a common, non-native *lingua franca* [22].

During the experiment, the multilingual groups were involved in a Planning Game activity, a requirements prioritization technique used in agile development. In particular, they had to complete two tasks. During the first task (*T1*), acting as customers, they separated a few vital requirements from the many elicited in a software development effort. Then, during the second task (*T2*), acting as developers, they completed a release plan. The task material, adapted from a previous work by Berander [1], was selected because the domain chosen for task execution is that of mobile phones, about which students typically have a rather equal knowledge gained through daily usage.

In order to assess whether machine translation is more beneficial to individuals with low English skills, we measured the English proficiency level for each study participant. We chose a placement test made publicly available online by Cambridge University[1], which includes 40 questions to be answered within 20 min. The test originally placed subjects into one of four distinct categories. In this replication, we selected participants at the *Low* level (scores 0-20), which will be then compared to the *High* level (scores 21-40) participants from the former experiment.

2.1 The Study Design

We followed a fractional factorial design [19] (see Table 1) in which each group participated in two meetings (Run 1 and 2), using a combination of the *communication mode* (MT and EN) and *task* (T1 and T2). Each multilingual group included 4 subjects, 2 speaking Italian and 2 speaking Portuguese as native language. As in the original experiment, each planning task in the replication was executed by two groups (i.e., 8 subjects), one group using MT (4 subjects) and one group using EN (also 4 subjects).

Table 1. Experimental plan

	Original experiment		Replicated experiment	
	MT	EN	MT	EN
Run 1	Gr1, Gr3 execute T1	Gr2, Gr4 execute T1	Gr6, Gr8 execute T1	Gr5, Gr7 execute T1
Run 2	Gr2, Gr4 execute T2	Gr1, Gr3 execute T2	Gr5, Gr7 execute T2	Gr6, Gr8 execute T2

In each run, each participant used different communication modes and different tasks. For example, a group that communicated

[1] www.cambridge.org/us/esl/venturesadulted/placement_test.html

through MT in the first run, used English language (EN) in the second run and vice versa. In addition, a group that executed task T1 in the first run executed task T2 in the second run.

The design allows an experimenter to do two comparisons, namely: in run 1, between the groups that executed task T1, and in run 2, between the groups that executed task T2. In addition, with this design, it is possible to analyze the influence of the communication mode both at the team and individual level.

2.2 Instrumentation, Training & Execution

Multilingual group meetings were run using eConference [4], a tool built on Eclipse RCP, the primary functionality of which is a closed group chat, augmented with agenda, meeting minutes editing, and typing awareness capabilities. In addition, we extended the tool developing an *ad hoc* plugin that enables the automatic translation of incoming messages, using the Google Translate APIs. In fact, whenever a new message is processed by eConference, the MT plugin invokes the MT web-service in order to show the translated messages, along with the original text, as shown in Figure 1.

Before each meeting, the group involved was trained to use the tool. First, a half-hour demo was given to students by one of the researchers. Then, a training session was set up, during which involved groups had to perform two training tasks, interacting first using their native language, exploiting the MT plugin, and then in English. As for the training tasks, we selected two riddles, described in English, which had to be completed within half an hour each.

During the training, two of the four students involved in each session were randomly selected to act as moderator or scribe. The extra duties of being a moderator include starting the meeting once every participant is online, keeping track of time limit, and so forth. The session scribe, instead, is enabled by the moderator to edit the tool whiteboard, a shared editor where all the group decisions and the final task solution were logged. We note that the groups of students were kept the same in the training sessions and in the actual experiment.

Each meeting required two hours in order to complete the experimental run. Two of the researchers, one in Brazil and one in Italy, were available to students during the runs, in order to provide technical help and prevent undesired interactions to occur outside of the tool, as pairs were collocated at each site.

During each run, groups were required to solve one of two tasks adapted from [1]. Both tasks were described in English. The first one (T1) was a requirements prioritization task to be completed within 30 minutes. Group's participants received a list of 16 features that described the desired functionalities of a mobile phone (e.g., alarm, calendar, MMS, notes, etc.). Then, they acted as a distributed group of customers who had to produce a prioritized list of requirements by dividing them into three distinct piles (i.e., less important, important, and very important). Further task constraints required that requirements within each pile were ranked by importance, and that no more than 13 requirements were assigned to one pile (85%).

The second task (T2) was about release planning and consisted of two consecutive steps, which had to be executed from a developer's perspective and completed within 60 minutes. In the first step, group participants had to distribute an overall amount of 1000 story points between the same 16 requirements from task T1, thus assigning the relative costs of implementing each of them. In the second step, the goal was to plan three releases of the product, based on the priorities, obtained from the outcome of T1, and the cost estimates, just assigned in the previous step. The following constraints were also given to participants. For the first release, they were allowed to assign 150-200 story points, whereas, for the second and third releases the ranges were 300-350 and 450-550, respectively.

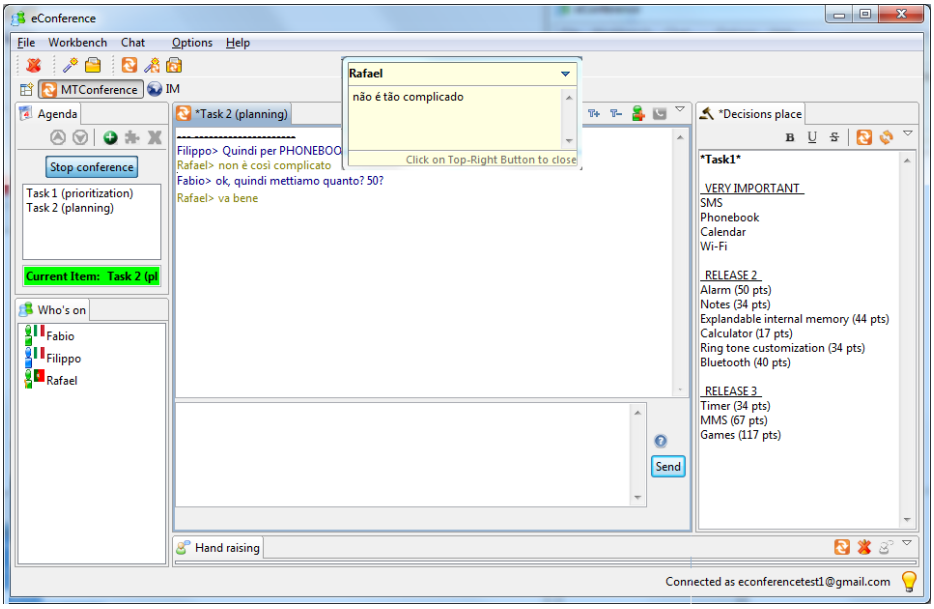

Figure 1. eConference with MT plugin

Finally, we note that, no matter what the language/task combination was, for each run the shared solutions were always edited in English.

2.3 Dependent Variables and Measures

The data sources considered in this study are the questionnaires, which were administered to the students upon the conclusion of the two tasks, and the meeting logs.

A large existing body of research on Group-Decision Support Systems (GDSS) identified domination, peer pressure, and social consensus among the problems faced in group communication. For example, according to DeSanctis & Gallupe [12], the health of group communication is observed through the equality of participation and the levels of satisfaction perceived with respect to process interaction and outcome. Such problems are expected to be even harder in multilingual groups, due to language barriers.

Therefore, for the two post-task questionnaires, inherited from the first study and written in English, we adopted a 4-point Likert scale (anchored with '4=strongly agree' and '1=strongly disagree' values), which was formulated with the aim of assessing the subjects' perception about the two constructs of i) *engagement and comfort with communication*, and ii) *satisfaction with task performance*. The questionnaires listed 16 closed questions, plus an open question, where subjects could freely report any thought or consideration about the whole experience, and a few "control" questions, in order to ensure that task execution was not hindered by the tool flaws or by unclear instructions and objectives. In addition, the post-T2 questionnaire also contained four extra questions that aimed at assessing the differences between the overall subjects' perception when using machine translation and English, at the end of both experimental runs.

From the chat logs collected at the end of the meetings, the *# of utterances* entered by group participants were counted to assess equality of participation. We also note that the chat logs were collected at both the Italian and the Brazilian site, since tasks executed using native languages produced two versions of the same discussion, one in Italian and one in Portuguese.

Finally, to gain more insight on the effects of machine translation, we looked at the very basic goal of communication, which is establishing a shared understanding. In fact, although machine translation helps people to cope with language barriers, it also poses hurdles to establishing mutual understanding due to translation inaccuracies and errors, which may cause both lack of mutual understanding (i.e., being aware that there is a problem that must be clarified) and misunderstandings (i.e., realizing that something that was initially considered understood correctly was actually wrong) [23]. In such situations, people become aware that there is a problem of a *lack of common ground*. Common ground is the knowledge that participants have in common when communicating and the awareness of it [8]. A common ground is dynamically established through grounding, an interactive process in which participants exchange evidence about what they do or do not understand over the course of a conversation. One the one hand, one could expect more *clarification requests* to emerge during MT meetings due to translations errors and inaccuracies. On the other hand, however, it could also be argued that low proficiency in a non-native language can be the cause of mistakes and inaccuracy, as well. Therefore, we need to investigate whether it is MT technology inaccuracy or low English proficiency to cause more clarifications requests by participants in a conversation. To quantify our construct of clarification requests, we performed a content analysis of the meeting logs.

3. RESULTS

In this section, we report the results from the analyses of the quantitative and qualitative data collected from the eight experimental runs. We first present quantitative analysis of the meeting logs. Then, we illustrate the findings from the quantitative analysis of the questionnaires. Finally, we report the results obtained from the content analysis performed on the logs. For the sake of clarity, where necessary, new results from this replication and findings from the original study are reported side by side.

3.1 Quantitative analysis of meeting logs

Table 2 provides some descriptive measures of the new meetings executed in the replication (i.e., with low proficiency groups Gr5-Gr8), along with those executed in the original experiment (i.e., with high proficiency groups Gr1-Gr4). To characterize them, we computed the time (in minutes) spent for executing the tasks, the overall number of utterances presented by participants, the frequency (expressed as utterance per minute – upm), and the average delay between two consecutive answers (in seconds).

Looking at the amounts of time spent for executing tasks, we note that results vary for run 1, whereas they are all somewhat comparable for run 2, for which all the groups took the whole time allowed (1 hour). In fact, the amounts time spent for run 1 range between 16 (Gr2) and 40 minutes (Gr1 and Gr7), who took about 10 extra minutes to complete the prioritization. As for Gr1 and Gr7, looking at the transcripts we realized that the delay was not related to the communication mode. Instead, the larger amount of time spent was due to the fact that group both group decided to adopt a time consuming approach. As such, every participant came up with a priority list, from which they eventually built a shared solution. The other groups, instead, adopted a more practical approach, that is, one participant proposed an initial priority list and the others suggested amendments until a shared solution was reached through discussion. With respect to Gr6 and Gr3, instead, we note that they took 35 and 67 minutes to execute run 1 and run 2, respectively. In both cases, the few extra minutes were granted to recover from a brief network disconnection that occurred at one site. Besides, Gr3 and Gr7 proved to be the most "active" groups overall, as they exhibited the highest frequency (6.33 and 6.90 upm; 6.44 and 6.75 upm, respectively) and the lowest average delay at typing utterances (10 and 8 sec.; 9 and 9 sec., respectively) over the two tasks. With respect to delays, the comparisons between the average delays in English meetings (mean 13.2 sec.) and MT meetings (mean 11.6 sec.) confirm that the subjects spent a little extra time in elaborating messages using the non-native English language.

Table 2. Descriptive measures for the eight meetings

Group		Communication mode	Time (min.)	# Utterances	Frequency (upm)	Average delay (sec.)
Gr1 (High)	Run 1	MT	40+	159	3.95	15
	Run 2	EN	61	322	5.28	11
Gr2 (High)	Run 1	EN	16	68	4.25	15
	Run 2	MT	59	346	5.86	10
Gr3(High)	Run 1	MT	30	190	6.33	10
	Run 2	EN	67*	462	6.90	8
Gr4 (High)	Run 1	EN	16	52	3.25	20
	Run 2	MT	54	169	3.13	14
Gr5(Low)	Run 1	EN	28	92	5.41	11
	Run 2	MT	59	358	6.17	10
Gr6(Low)	Run 1	MT	35*	140	4.38	14
	Run 2	EN	59	164	2.83	21
Gr7 (Low)	Run 1	EN	41+	264	6.44	9
	Run 2	MT	60	405	6.75	9
Gr8 (Low)	Run 1	MT	43+	240	5.58	11
	Run 2	EN	67+	354	5.28	11

* *Extra minutes granted to recover from network disconnection*
\+ *Exceeded the time limit for the task*

Table 3. A breakdown of participation level of subjects during task executions

Group (level)	# Utterances (%) – English				Δ Most - Least prolific member	# Utterances (%) -- Machine Translation				Δ Most - Least prolific member
	Member 1 (mod)	Member 2	Member 3	Member 4		Member 1 (mod)	Member 2	Member 3	Member 4	
Gr1 (High)	181 (56%)	61 (19%)	52 (16%)	28 (9%)	**47%**	64 (41%)	43 (27%)	26 (16)%	25 (16%)	**25% ↓**
Gr2 (High)	16 (24%)	28 (41%)	15 (22%)	9 (13%)	**28%**	71 (21)%	117 (34%)	91 (26%)	67 (19%)	**15% ↓**
Gr3 (High)	133 (29%)	106 (23%)	76 (16%)	147 (32%)	**16%**	60 (32%)	47 (25%)	39 (21%)	44 (23%)	**11% ↓**
Gr4 (High)	24 (43%)	12 (23%)	5 (10%)	11 (21%)	**33%**	65 (38%)	39 (23%)	24 (14%)	41 (24%)	**24% ↓**
Gr5 (Low)	29 (32%)	19 (21%)	24 (26%)	20 (22%)	**11%**	94 (26%)	130 (36%)	61 (17%)	73 (20%)	**19%**
Gr6 (Low)	33 (20%)	37 (23%)	53 (32%)	41 (25%)	**12%**	38 (27%)	38 (27%)	36 (26%)	28 (20%)	**7% ↓**
Gr7 (Low)	136 (52%)	46 (17%)	40 (15%)	42 (16%)	**37%**	212 (52%)	65 (16%)	56 (14%)	72 (18%)	**38%**
Gr8 (Low)	127 (36%)	79 (22%)	81 (23%)	67 (19%)	**17%**	71 (30%)	55 (23%)	62 (26%)	52 (22%)	**8% ↓**

In order to verify equality of participation during meetings (i.e., no domination by any group participants), we computed the number of utterances presented by each participant during a meeting to see how the use of machine translation affect the participation extent of subjects (see Table 3). We also computed the percentages because the release planning task executed during run 2 takes longer than the prioritization task of run 1, and so, regardless of the communication mode, any participant is supposed to have contributed more utterances during the former. We then computed the deltas between the percentage of utterances presented by the most and the least prolific subjects for each task execution. Comparing the two columns, we observe that: 1) there is an increase of participation (i.e., a smaller delta) of the least prolific subject, always at expense of the most prolific member during MT-enabled tasks, both for high and low proficiency groups; 2) the deltas are usually higher in high proficiency groups from the original experiment than in low proficiency groups from this replication; 3) the deltas from high proficiency groups always decrease when using MT, whereas, as for low proficiency groups, Gr7 delta remain unchanged after the communication mode switch, and Gr5 delta even increased by 8 percentage points when using MT.

Finally, we compared the percentages of utterances presented by group members with the lowest English proficiency skills during the EN and MT meetings (see Table 4). The new results with low proficiency groups confirm the same general tendency observed in our previous study with high proficiency groups, that is, the percentage of utterances presented by the least proficient subjects tend to increase when switching from English to their native language, with one exception (Gr7 and Gr3, respectively).

3.2 Questionnaires analysis

In this section we report the findings from our quantitative analysis of post-task questionnaires administered to low proficiency groups in the replicated experiment.

In order to measure the *satisfaction with task performance*, we defined a 4-item 4-point Likert scale to assess participants' perception of whether tasks were performed positively, with participants feeling actively involved in group communication when reaching shared decisions. In other words, we aimed to verify that information exchanged, and optionally translated, was properly processed when using both communication modes.

As a nonparametric alternative to the t-test for two paired samples, we performed a Wilcoxon signed rank test [9] on the responses to the four questions shown in Table 5. The test failed to reveal any difference in the levels of satisfaction with performance perceived by low proficiency subjects when using English rather than their native language.

Likewise, to measure the levels of *engagement and comfort with communication mode* perceived by low proficiency subjects, we used a 6-point Likert scale to assess discussion contentment. Again, we performed a Wilcoxon signed rank test, as shown in Table 6, which failed to reveal any difference between the use of English and MT in terms of being involved in an open and useful discussion with others.

Table 4. Gain in participation of the least proficient subject for each group when using native language with machine translation

Group (level)	Least proficient subject (nationality)	% of utterance	
		EN	MT
Gr1 (High)	Student #7 (Brazilian)	19%	27% ↑
Gr2 (High)	Student #4 (Brazilian)	22%	26% ↑
Gr3 (High)	Student #16 (Brazilian)	32%	23%
Gr4 (High)	Student #12 (Brazilian)	10%	14% ↑
Gr5 (Low)	Student #17 (Italian)	21%	36% ↑
Gr6 (Low)	Student #22 (Italian)	20%	27% ↑
Gr7 (Low)	Student #27 (Brazilian)	15%	14%
Gr8 (Low)	Student #32 (Brazilian)	23%	26% ↑

Table 5. Evaluation of satisfaction with performance for the four low proficiency groups Gr5-Gr8 (N=16)

	EN > MT	EN < MT	EN = MT	Wilcoxon Signed Rank Test
Q8. "*I actively participated in the discussion*"	4	1	11	Z = -1.41 p =.157
Q12. "*I had the sensation of wasting time*"	7	4	5	Z = -1.4 p =.163
Q13. "*It was easy to reach a common decision"*	3	4	9	Z = -1.28 p =.26
Q16. "*I had a positive global impression of the performance*"	3	1	12	Z = -1.0 p =.317

Table 6. Evaluation of engagement and comfort with communication mode for the four low proficiency groups Gr5-Gr8 (N=16)

	EN > MT	EN < MT	EN = MT	Wilcoxon Signed Rank Test
Q1. *"I had enough time to perform the activity*"	2	3	11	Z = -.707 p =.48
Q6. "*It was easy to communicate with others*"	4	5	7	Z = -.183 p =.855
Q7. "*I had adequate opportunity to participate in the discussion"*	4	3	9	Z = -.378 p =.705
Q9. "*I was encouraged to discuss contrasting solutions with others* "	5	2	9	Z = -1.265 p =.206
Q10. "*Other participants adequately answered my questions*"	6	2	8	Z = -1.613 p =.107
Q11. "*I felt involved in the discussion"*	0	1	15	Z = -1.0 p =.317

Finally, we asked four questions (Q17-Q20) to collect subjects' overall perceptions and preferences for each communication mode (i.e., "*Group activity has benefited from the suggested translations / chatting directly in English*", "*If I should choose a meeting environment, I would prefer a tool with the MT service / without the MT service*"). We again applied a Wilcoxon signed-rank test. The results, reported in Table 7, show that, overall, subjects perceived no particular benefit from using MT. Conversely, the test revealed a statistically significant difference at the 1% level (Z = -2.801, p=.005) only for low proficiency groups who showed a preference towards using MT-enabled communication rather than English.

3.3 Content analysis

Measuring the level of common ground that people achieve in group communication is generally a challenging task [8]. However, to determine if the adoption of machine translation affects group interaction in multilingual group meetings, we rather looked at evidence of lack of common ground. We operationalized the construct of lack of common ground in terms of clarification request. Receivers provide negative evidence during communication when messages are improperly or incompletely understood. Therefore, the higher the number of ill-defined messages presented, due to either MT inaccuracy or poor English proficiency, the more negative evidence presented by receivers and, consequently, the more requests for clarification.

To quantify our construct of clarification requests, we performed the content analysis of some logs collected from low proficiency group meetings. Content analysis, also called coding [21], is a mix of quantitative and qualitative analysis that transforms qualitative data (e.g., written text, as in our case) into quantitative data by applying a coding schema, which classifies content according to a finite set of possible thematic units (i.e., categories). We applied the same coding schema that was proposed as a result from our original study, but here we augmented the original set of nine categories with an extra one, called *Unknown*, to cope with those cases when poor English or an inaccurate translation made a message incomprehensible and, consequently, its categorization impossible. Two of the researchers performed the content analysis separately and then, intercoder agreement was measured by Cohen's Kappa to ensure the concordance level between the resulting categorizations. We opportunistically performed such analysis only on the logs from low proficiency groups Gr5 and Gr7, for which some subjects reported on comprehension difficulties in the questionnaires.

In addition, as one can observe in Table 3, these two are the only cases where equality of participation was not affected or even decreased when using MT.

Table 8 shows the breakdown of the content analysis performed. We note that unit percentages are reported, rather than occurrences, as a necessary normalization due to the large differences in the lengths of task discussions. Also, for the sake of space, we only report results for those thematic units that contribute to quantify the construct of clarification requests, namely *Check Misunderstanding*, *Check Provisional*, and *Unknown* categories. In particular, the Check Misunderstanding unit categorizes any utterance providing evidence that a previous message was not fully accepted (e.g., *"Not sure I get your question…"*, *"What?"*). The Check Provisional unit, instead, categorizes utterances that explicitly look for confirmation of acceptance through provisional, try-marked statements (e.g., *"So we decided for color screen, right?"*). Finally, the *Unknown* unit categorizes utterances that could not be coded by the raters, because the meaning was unclear, and, at the same time, misunderstood by other meeting participants.

Table 7. Evaluation of overall communication mode preference for both high and low proficiency groups (N=16)

	A vs. B	A > B	A < B	A = B	Wilcoxon Signed Rank Test
High group	***"Group activity benefited from using…"*** MT vs. English	7	4	5	Z = -.711 p =.477
	"Another time, I would rather communicate using…" MT vs. English	11	3	2	Z = -1.904 p =.057
Low groups	***"Group activity benefited from using…"*** MT vs. English	4	5	7	Z = -.061 p =.951
	"Another time, I would rather communicate using…" MT vs. English	11	1	4	**Z = -2.801 p =.005***

* Statistically significant results at the 0.05 level are indicated in bold

As compared to EN runs, the results show a higher number (8%) of Unknown (i.e., unclassifiable) utterances and checks for misunderstandings during MT runs. Although only partial, these results seem to suggest that the inaccuracy of state-of-the-art machine translation technology poses more hurdles to common ground than language barrier.

4. DISCUSSION

Table 9 compares the context variables and the results between the original study and the current replication. The original study involved only subjects with a high proficiency level of English, thus suggesting that the usefulness of MT would improve when used by individuals who are not able to communicate in English as in their mother tongue. Therefore, in the replicated experiment, we only involved subjects with a low proficiency level in English.

Table 8. The result of content analysis on Gr5 and Gr7 logs

	EN (Run 1)			MT (Run 2)		
	Check Mis.	Check Prov.	Unk.	Check Mis.	Check Prov.	Unk.
Gr5 (Low)	0%	2.2%	0%	2.9%	5.9%	4.3%
Gr7 (Low)	1.9%	3.8%	.9%	1%	1.2%	3.2%

Table 9. Comparison with the original experiment

		Former experiment	Current replication
Context	number of data points	4 teams, 16 subjects	4 teams, 16 subjects
	Subjects	(South) Brazilian students Italian students	(North) Brazilian students Italian students
	Proficiency level in English	High level	Low level
	number of tasks	2 tasks in 2 consecutive runs	2 tasks in 2 consecutive runs
Results	frequency of messages & delay between utterances	MT = EN	MT = EN
	equal participation	EN < MT	EN < MT
	checking misunderstanding and provisional	* NA	EN = MT

** Data not available*

In the following, we compare the findings from the two studies, with respect to the research questions presented earlier. In general, the new results confirm prvious findings.

More specifically, regarding the first research question RQ1 (*Can machine translation services be used in distributed multilingual requirements meetings, instead of English?*), Table 2 shows that in both the original study and in this replication the frequency of presented messages (measured by utterance per minute rate – upm) is substantially similar between EN and MT runs. This is also confirmed by the average delay between two consecutive utterances, a measure that is correlated to message frequency, since faster interaction means lower delay. However, we are not able to distinguish between the delay due to message comprehension and message production.

Results in Tables 3 and 4 are more interesting because they confirm that, no matter what their English proficiency level is, members of multilingual groups participate in more balanced discussions when using their native language with the help of MT, instead of English. In fact, Table 3 shows that the delta (i.e., the difference in participation) between the most prolific and the least prolific subjects tends to reduce in MT-enabled discussions. In addition, especially the least proficient subject of a multilingual group seemed to benefit from machine translation, as their percentage of contributed messages grows when language switched from English to native (see Table 4).

Overall, these findings from the two studies allow us to affirm that machine translation is not disruptive of the conversation flow, even during the execution of complex group tasks, such as distributed requirements meetings, and that it is accepted with favor independently of subjects' English proficiency level.

With respect to the research question RQ2 (*How does the adoption of machine translation affect group interaction in distributed multilingual requirements meetings, as compared to the use of English?*) one of the results from our original study was the definition of a coding schema that emerged from the inspection of meeting logs in the original experiment. In this replication, we opportunistically applied that coding schema to a couple of the logs of the low level groups. The results of the related content analysis are shown in Table 8. In one case (Gr5) we can observe a higher number of utterances coded as checks to avoid misunderstandings during MT meetings than in English meetings. However, the opposite happens in the other case (Gr7). These findings suggest the need to further our understanding by completing the content analysis of logs from both high and low proficiency groups. Instead, we can observe a higher number of utterances that could not be coded because the meaning was unclear, during the two runs with native language. Such finding suggests that inaccurate translations may impair the development of shared understanding more than low English skills. In addition, a percentage as high as the 4% of utterances that cannot be coded due to poor performance of the MT service raises questions on the feasibility of supporting multilingual groups with real-time translation in professional contexts for executing crucial tasks. More specifically, although such inaccuracies neither break the communication flow nor impair interaction to the extent that a task cannot be carried out, they force participants to fix them nonetheless. And, even if such a lack of common ground can be resolved by exchanging further utterances, this requires extra time, thus decreasing the efficiency of a meeting.

Finally, with respect to the research question RQ3 (*Do individuals with a low English proficiency level benefit more than individuals with a high level, when using their native language assisted by real-time translation?*), in terms of the levels of satisfaction and comfort perceived during the experimental runs, questionnaire analyses failed to reveal any difference (see Tables 5 and 6), which, on the one hand confirm findings from the original study with highly proficient subjects. On the other hand, however, these results (surprisingly?) suggest that, as of now, state-of-the-art MT technology is no more beneficial to individuals with low English proficiency than it is to people with high skills in a foreign language. The only statistical significant difference observed is that people with low English skills are more incline to use MT again in multilingual group interaction, despite some flaws of the current technology (see Table 7).

5. THREATS TO VALIDITY

One of the key issues in experimentation is evaluating the validity of results [22]. In this section we discuss the potential threats that are relevant for our study and how they are addressed.

5.1 Construct Validity

Construct validity concerns the degree of accuracy to which the variables defined in the study measure the constructs of interests. We identified a couple of such kind of threats.

We acknowledge the need to perform factor and scale reliability analyses on the responses to the questionnaires, in order to determine the validity of the constructs of *engagement and comfort with communication* and *satisfaction with task performance*. Instead, to ensure the validity of the *clarification requests* construct, two of the researchers independently applied the coding schema to chat logs. Then, inter-rater agreement was measured by Cohen's K index. The computed indexes are .88 and .91, meaning almost perfect agreement between the raters.

5.2 Internal Validity

Threats to internal validity influence the conclusions about a possible causal relationship between the treatment and the outcome of a study. The following rival explanations for the findings have been identified.

A learning effect occurs when subjects learn more about how to perform the required task, and are better the next time. The experimental design minimized this threat. We assigned the groups in such a way that, for each run, we are able to compare MT and EN on the same task (T1 in run 1, T2 in run 2) between different groups. Thus, for each comparison, the subjects have the same amount of accumulated experience.

An instrumentation effect occurs when differences in the results may be caused by differences in experimental material. Because in this study there are two different planning tasks, we cannot exclude that task complexity could have been a confounding factor, since subjects experience a communication mode with one task only. A selection effect occurs due to the natural variation in human subjects' performance. Random assignment of subjects to experimental conditions usually reduces this threat, but our experimental design is heavily influenced by the small amount of groups. We control this threat by design, restricting the level of groups to high and low proficiency (respectively, in the original study and its replication), and consequently assigning to groups any student whose proficiency do not alter the designed level.

5.3 External validity

External validity describes the study representativeness and the ability to generalize the results outside the scope of the study.

We identified the following threats to external validity. For any academic laboratory experiment the ability to generalize results to industry practice is restricted by the usage of students as study participants. Although the students may not be representative of the entire population of software professionals, it has been shown that the differences between students and real developers may not be as large as assumed by previous research [14]. Another issue with the representativeness of subjects is related to their familiarity with the use of synchronous, text-based communication. Computer science students are very accustomed with text-based interaction. Nevertheless, synchronous, text-based communication tools, such as chat and IM, are increasingly being adopted in the workplace, not only in the field of software development, to complement email [13].

5.4 Conclusion validity

Conclusion validity is concerned with the relationship between the treatment and the outcome.

We acknowledge that the small number of data points is not ideal from the statistical point of view. Small sample sizes, especially when the key experimental unit is at the team level, are a known problem difficult to overcome, especially for cross-country controlled experiments with participants interacting from different time zones.

6. CONCLUSIONS

The work presented here is part of an ongoing research, the purpose of which is understanding to what extent real-time machine translations can be beneficial for distributed, multilingual teams located in countries where professionals are not proficient in one common language. In particular, in this replication we specifically assessed whether non-English speaking groups benefit from communicating in their own native languages when their English is not fluid enough for a fast-paced conversation.

The results of this replication confirmed that real-time machine translation is not disruptive of the conversation flow, is accepted with favor, and grants a more balanced discussion. However, the findings also show that state-of-the-art MT technology is no more beneficial to individuals with low English proficiency than it is to people with high skills in a foreign language. Content analysis suggests that this might be due to machine translation inaccuracies, which slow down the development of a common ground.

As future work, we plan to (a) analyze the results in order to learn about the effects of human typos on machine translation accuracy; (b) execute further runs to obtain more data points and strengthen the conclusion validity; (c) replicate the experiment involving professionals.

ACKNOWLEDGMENTS

This research is partially funded by the Rio Grande do Sul State funding agency (FAPERGS), by the FTS-Brasil Project CNPq (483125/2010-5), and by the European Territorial Cooperation Operational Programme "Greece-Italy 2007-2013" under the project Intersocial. We would also like to thank all the students who took part in the experiment.

REFERENCES

[1] P. Berander, "Using Students as Subjects in Requirements Prioritization," *Int'l Symposium on Empirical Software Engineering (ISESE'04)*, pp. 167-176, 2004.

[2] Brazil IT-BPO Book 2008-2009, published by Brasscom, Brazilian Association of Information Technology and Communication Companies, São Paulo, SP, Brazil, 2010.

[3] F. Calefato, F. Lanubile, and P. Minervini, "Can Real-Time Machine Translation Overcome Language Barriers in Distributed Requirements Engineering?", *Proc. 5th Int'l Conference on Global Software Engineering (ICGSE'10)*, Princeton, NJ, USA, Aug. 23-26, pp. 257-264, 2010.

[4] F. Calefato and F. Lanubile, "Using Frameworks to Develop a Distributed Conferencing System: An Experience Report", *Software: Practice and Experience*, 2009, vol. 39, no. 15, pp. 1293–1311.

[5] F. Calefato, F. Lanubile, and R. Prikladnicki, "A Controlled Experiment on the Effects of Machine Translation in Multilingual Requirements Meetings", *Proc. 6th Int'l Conference on Global Software Engineering (ICGSE'11)*, Helsinki, Finland, August 15-18, 2011.

[6] E. Carmel, and R. Agarwal, "Tactical Approaches for Alleviating Distance in Global Software Development," IEEE Softw., vol. 18, no. 2, pp. 22-29, Mar. 2001.

[7] E. Carmel and R. Prikladnicki, "Does Time Zone Proximity Matter for Brazil? A Study of the Brazilian I.T. Industry." Technical Report, 2010, available at http://ssrn.com/abstract=1647305.

[8] H.H.Clark, and S.E. Brennan. *Grounding in Communication, in Perspectives on Socially Shared Cognition*, American Psychological Association, Wash. DC, 1991, pp. 127-149.

[9] W.J. Conover, *Practical Nonparametric Statistics*. Wiley, New York, 1980.

[10] D. Damian and D. Zowghi, "Requirements Engineering Challenges in Multi-Site Software Development Organizations", *Requirements Engineering Journal*, 8-3, 2003, pp. 149-160.

[11] D. Damian, "Stakeholders in Global Requirements Engineering: Lessons Learned from Practice", *IEEE Software*, 24-2, 2007, 21-27.

[12] G. DeSanctis and R. B. Gallupe, "A Foundation for the Study of Group Decision Support Systems", *Management Science*, 33(5): 589-609, May 1987.

[13] J.D. Herbsleb, D.L. Atkins, D.G. Boyer, M. Handel, and T.A.Finholt, "Introducing Instant Messaging and Chat into the Workplace." *Proc. Int'l Conference on Computer-Human Interaction (CHI '02)*, Minneapolis, MN, USA, 2002.

[14] M. Höst, B. Regnell, B. and C. Wohlin. "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment." *Empirical Software Engineering*, Vol. 5, No. 3, 2000, pp. 201-214.

[15] Y. Hsieh, "Culture and Shared Understanding in Distributed Requirements Engineering," *1st Int'l Conf. on Global Software Engineering (ICGSE'06)*, Florianopolis, Brazil, Oct. 2006.

[16] D. Jurafsky and J. H. Martin, *Speech and Language Processing* (2nd ed), Prentice Hall Series in Artificial Intelligence, Prentice Hall, 2008.

[17] A.T. Kearney. "Destination Latin America: A Nearshore Alternative, Technical Report, 2007.

[18] KPMG, Nearshore Attraction: Latin America Beckons as a Global Outsourcing Destination, Technical Report, 2009.

[19] D.C. Montgomery. *Design and Analysis of Experiments*. J. Wiley & Sons, New York, 1996.

[20] B. Nuseibeh, and S. Easterbrook, "Requirements engineering: a roadmap," *Proc. Int'l Conf. on the Future of Software Engineering* (ICSE '00), pp. 35-46, June 2000.

[21] S. Stemler, "An Overview of Content Analysis", *Practical Assessment, Research & Evaluation*, vol. 7, no. 17, 2001.

[22] C. Wohlin, P. Runesson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering, An Introduction*. Kluwer Academic Publishers, 2000.

[23] N. Yamashita, R. Inaba, H. Kuzuoka, and T. Ishida. "Difficulties in establishing common ground in multiparty groups using machine translation." *Proc. 27th Int'l Conf. on Human Factors in Computing Systems (CHI '09),* Boston, USA, April 4-9, 2009, pp, 679-688.