A Controlled Experiment on the Effects of Machine Translation in Multilingual Requirements Meetings

Fabio Calefato, Filippo Lanubile
Dipartimento di Informatica
Università degli Studi di Bari "A. Moro"
Bari, Italy
calefato,lanubile@di.uniba.it

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

Abstract—**Requirements engineering is a communication-intensive activity and thus it suffers much from language difficulties in global software projects. Remote requirements meetings can benefit from machine translation as this technology is today available in the form of cross-language chat services. In this paper, we present the design of a controlled experiment to investigate the effects of automatic machine translation services in requirements meetings. Experiment participants, using either Italian or Portuguese as native language, are asked to interact with a communication tool from a distance in order to prioritize and estimate requirements. First results show that real-time machine translation is not disruptive of the conversation flow and is accepted with favor by participants. However, concrete effects are expected to emerge when language barriers are critical.**

Keywords-machine translation; language barrier; requirements engineering; empirical study.

I. Introduction

Requirements engineering is a communication-intensive activity and thus it suffers much from language difficulties in global software projects [11], [12], [24]. 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 [8], [17].

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 [9]. 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 [20] – and it is growing by 6.5% a year on average since 2005 [4], 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). In this scenario, English language is required as a skill for every person working in the global market. 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 [21]. 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 [9]. 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.

In our previous work [6] we run a simulated study to evaluate the feasibility of adopting an automatic, cross-language translation to communication-intensive activities, such as distributed requirements engineering. Although our work proved that state-of-the-art machine translation services could be embedded into synchronous text-based chat with a negligible extra time, being only a simulation, the study did not allow us to hypothesize whether the quality of machine translation services would be good enough to allow participants to complete a complex group task while communicating with their own native language. To further our research, in this paper we investigate by means of an experiment how real-time machine translation can be effectively used during distributed requirements engineering meetings involving multilingual groups. Thus, we propose the following research questions for study:

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?*

The remainder of this paper is structured as follows. In Section 2 we briefly overview the state of the art in machine translation services. Section 3 presents the controlled experiment in detail, whereas early results are presented in

Section 4. Threats to validity are described in Section 5 and a general discussion follows in Section 6. Finally, conclusions and future research activities are presented in Section 7.

II. Machine translation background

Machine translation (MT) is an established technology, some 50 years in the making, which may be defined as the use of a computer to translate a text from one natural language, the source language, into another one, the target language [13]. The technology available today – i.e. real-time, online conversation – is experiencing tremendous growth of interest, on the heels of the Internet continuous expansion.

MT is difficult mainly because translation *per se* involves a huge amount of human knowledge that must be encoded in a machine-processable form. In addition, natural languages are highly ambiguous, as two languages seldom express the same content in the same way [1]. Although hybrid approaches also exist, MT systems can be broadly classified into two main categories, corpus-based and rule-based, according to the nature of the linguistic knowledge being used. The *rule-based* MT systems use knowledge in the form of rules, explicitly coded by human experts, which attempt to codify the translation process. Instead, *corpus-based* MT systems use large collections of parallel texts (i.e. pairs consisting of a text in a source language and its translation into a target language) as the source of knowledge from which the engine learns how to perform translations.

Compared to the rule-based approach, the corpus-based approach is particularly appealing to researchers because systems can be trained automatically, without any direct human intervention. Google Translate[1] is an example of corpus-based MT system that applies statistical learning techniques to build language and translation models from a large number of texts, both monolingual text in the target language and text consisting of examples of human translations between the source and the target languages. The Google Translate service can be used by third-party applications because it exposes a RESTful interface [26] that returns responses encoded as JSON[2] results. As of this writing, Google Translate supports the translation between any two pairs of over 50 languages, although not all at the same quality level. In our previous work [6], according to a set of human raters, Google Translate was found to produce better (i.e. more accurate) automatic translation than the rule-based Apertium[3] service.

Accurate computer translation is particularly appealing because it is quicker, more convenient, and less expensive than human translators are. An interesting research study was conducted by Yamashita *et al.* [30] [32] who investigated the effects of machine translation on mutual understanding. The study found that shared understanding is affected by the asymmetry of machine translation since the sender of a message does not know how well it has been translated to the target language. A limitation of this study is that the researchers employed picture description as the experimental tasks in one-to-one chat communication. Recently, the EU commission funded the MOLTO project (Multi-lingual Online Translation)[4] with the goal of producing accurate machine-translations of the official documents and save a billion euro currently spent per year to translate them in the 20+ official languages of the Union.

Aside from research prototypes or projects (e.g. for further reading, see [2], [15] [18], [25]) also commercial tools that offer cross-language chat services are available, such as IBM Lotus Translation Services for Sametime[5] and, lately, VoxOx[6], which provide cross-language translations for most of the existing instant messaging networks. Recently Google has even pushed MT goal further releasing a Google Translate app for Android [29], which integrates automatic translation with voice recognition for the English-Spanish pair.

III. The empirical study

We designed an experiment that aims at assessing the effects of employing an MT service (specifically Google Translate) in synchronous, text-based requirements meetings. The study participants, both graduate and undergraduate students, have to work in teams composed of four people, of whom two from the University of Bari, Italy, and two from PUCRS in Porto Alegre, Brazil (the terms students, subjects, and study participants are used interchangeably). Teams use two different communication modalities during the collaborative sessions. More specifically, the teams collaborate using both native languages (i.e. Italian or Portuguese), with the help of MT, and English, as a *lingua franca* [22] (i.e. common, non-native language).

During the experiment, the subjects have to complete two tasks during which first, as customers, they separate a few vital requirements from the many elicited in a software development effort, and then, as developers, they complete a release plan. The basis for the experiment is a requirements prioritization technique called Planning Game, an eXtreme Programming practice. The task material was adapted from a previous work by Berander [1] for the following reasons: i) it was publicly available; ii) the domain chosen for task execution was that of mobile phones, about which students typically have a rather equal knowledge gained through daily usage; iii) the original study, which compared the adoption of students and professionals as experimental subjects, found that subjects' commitment accounted for difference more than experience.

Because during a meeting a better command of language provides better opportunities of steering communication, one could reasonalby argue that MT is more useful to those who are not proficient in English (i.e. individuals who are not able to communicate in English as in their mother tongue). Thus, the English proficiency of the subjects to be involved in our study has to be evaluated. We chose a placement test, made publicly available online by Cambridge University[7], which includes 40 questions to be answered within 20 min. The test allows us to

[1] http://translate.google.com

[2] http://json.org

[3] www.apertium.org

[4] www.molto-project.eu

[5] www-01.ibm.com/software/lotus/sametime

[6] www.voxox.com

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

place subjects into one of four distinct categories, namely Level 1 (poor), Level 2 (basic), Level 3 (average), and Level 4 (advanced).

So far four experimental runs were executed and involved sixteen subjects. The results of the English proficiency test of the subjects involved are show in Table I. The majority scored at Level 3 (average) or 4 (advanced), whereas only two subjects proved to have basic or poor skills.

TABLE I. RESULTS OF THE ENGLISH PROFICIENCY TEST

Subject	Nationality	Score	Level	Group
1	Italian	39 / 40	4 (advanced)	Gr1
2	Italian	39 / 40	4 (advanced)	
3	Brazilian	39 / 40	4 (advanced)	
4	Brazilian	12 / 40	1 (poor)	
5	Italian	37 / 40	4 (advanced)	Gr2
6	Italian	35 / 40	4 (advanced)	
7	Brazilian	14 / 40	2 (basic)	
8	Brazilian	37 / 40	4 (advanced)	
9	Italian	26 / 40	3 (average)	Gr3
10	Italian	30 / 40	3 (average)	
11	Brazilian	27 / 40	3 (average)	
12	Brazilian	26 / 40	3 (average)	
13	Italian	31 / 40	3 (average)	Gr4
14	Italian	31 / 40	3 (average)	
15	Brazilian	32 / 40	4 (advanced)	
16	Brazilian	30 / 40	3 (average)	

A. *Design*

Table II shows the experimental plan, which corresponds to a 2^3 factorial design [23]. We consider three independent variables, each having two levels:

- *Communication mode:* levels *MT* and *EN;*
- *Task:* levels *T1* and *T2*;
- *English proficiency:* levels *average* and *advanced*.

Altough we have designed to run the experiment for eight groups, as of this writing we have executed four runs only for groups Gr1-Gr4 (shown in bold in the Table II).

The sixteen students were assigned to the four groups depending on their nationality, since we had the constraint of including two students from each country per group, and their English test results. We decided to consider a group to be at Level X if the majority of its members scored at Level X.

For instance, when three group members have advanced skills (Level 4) whereas one has basic skills (Level 2), the group is considered at Level 4. Since the majority of students scored at Level 3 (average) or 4 (advanced), in these four runs, the subjects at Level 1 and 2 from Brazil had to be in different groups and, thus, we assigned them to either Gr1 or Gr2. Therefore, for Gr1 and Gr2 to be considered advanced, we randomly assigned to them four of the six Italian students and two of the three Brazilian students who scored at Level 4. The remaining students were all at Level 3 except one, and so they were randombly assigned to either group Gr3 or group Gr4, which thus are both average groups (see Table II).

The groups for the remaining runs will be arranged according to the same logic and, therefore, the overall study will only take into account groups with average or advanced English skills.

TABLE II. THE 2^3 FACTORIAL DESIGN OF THE EXPERIMENT

Communication Mode	Task	English proficiency	Groups
MT	T1	advanced	**Gr1**, Gr5
EN	T1	advanced	**Gr2**, Gr6
MT	T2	advanced	**Gr2**, Gr6
EN	T2	advanced	**Gr1**, Gr5
MT	T1	average	**Gr3**, Gr7
EN	T1	average	**Gr4**, Gr8
MT	T2	average	**Gr4**, Gr8
EN	T2	average	**Gr3**, Gr7

B. *Preparation*

Before each experimental run can be executed, the students have to be trained for using the eConference tool [7], which is a text-based distributed meeting system. The primary functionality provided by the tool is a closed group chat, augmented with agenda, meeting minutes editing, and typing awareness capabilities. The tool is built on Eclipse RCP[8], a pure-plugin platform that allows for full extensibility. We developed a machine translation plugin[9] for eConference that allows selecting the language pair to employ for automatically translating incoming messages exploiting the Google Translate APIs in both one-to-one and group-chat sessions. When a new message is processed by eConference, the MT plugin invokes the web-service in order to show the translated messages along with the original text.

First, a half-hour demo is given to students by one of the researchers. Then, a training session is set up, using groups arranged as explained before. Then, the groups have to perform two training tasks, interacting first using their native language (i.e. Portuguese or Italian), exploiting the MT plugin of the tool), and then in English. As for the training tasks, we selected two riddles which have to be completed within half an hour each.

During the training, one student per group is randomly selected to act as moderator, whose extra duties include starting the meeting once every participant is online, keeping track of time limit, and saving chat logs. For the sake of simplicity, we decided that we would keep the same moderators during the actual experimental runs. Finally, during each meeting, one or more participants can act as the session scribes, i.e., they are enabled by the moderator to edit the tool's whiteboard, which is a shared editor where to log all the group decisions and the final task solution.

[8] www.eclipse.org

[9] http://code.google.com/p/econference-mt-plugin

C. *Execution*

Each of the four runs were executed over two days, in a two-hour session per day. Two of the researchers, one in Brazil and one in Italy, were available to students during the sessions, in order to provide technical help and prevent undesired interactions to occur outside of the tool.

During the meetings, the groups had to solve two tasks. The first one (T1) was a *requirements prioritization* task, which had to be executed from a customer's perspective and completed within 30 minutes. During T1 the two groups received a list of 16 features that described the desired functionalities of the mobile phone they were supposed to develop (e.g. alarm, calendar, MMS, notes, etc). The participants acted as a distributed group of customers who have to divide the set of requirements into three distinct piles, namely *Less important* (LI), *Important* (I), and *Very important* (VI). They were constrained to assign at most 13 requirements to one pile (i.e. 85%). Furthermore, the subjects were also instructed to rank the requirements within the piles in order to get a prioritized list as the final outcome of task T1.

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, the participants had to assign the relative cost of implementing each of the 16 requirements from the same list of the previous task. More specifically, they had to distribute an overall amount of 1000 story points (representing the whole implementation cost) between the 16 requirements available. In the following step, the goal was to plan three releases of the product, based on the priorities, obtained from T1, and the cost estimates, just assigned in the previous step. The following constraints were also given to the participants. For the first release they were allowed to assign 150-200 story points. Instead, for the second and third releases, they were allowed to assign 300-350 and 450-550 story points, respectively.

IV. Early Results

In this section we report the results from the analysis of the quantitative and qualitative data collected from the first four experiment runs. The data sources are the questionnaires, which were administered to the students upon the conclusion of each task, and the related chat logs.

For the two post-task questionnaires we adopted a 4-point Likert scale, anchored with '4=strongly agree,' and '1=strongly disagree' values. Both questionnaires listed 16 questions, formulated with the aim of assessing the subjects' perception about their i) *engagement level and comfort with communication*, and ii) *satisfaction with task execution*. In addition, a few ones were formulated as "control" questions in order to ensure that the tool itself did not suffer from flaws that hindered task execution, and that task description and goal were all clear.

A. *Quantitative analysis from meeting logs*

Table III provides some descriptive measures of the eight task meetings executed during the four experimental runs. 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).

The amounts of time spent for executing tasks are all comparable, except for Gr1 who took 40 minutes (10 over the limit of 30 minutes) to complete the first task (prioritization), while the other two groups took only 16 minutes, and Gr3, who did not complete the second task (planning). As for Gr1, 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 Gr1 decided that every participant had to come up with a priority list, from which they would eventually build a shared solution. This approach was more time consuming than that used by other groups, 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 Gr3, instead, we note that they took 67 minutes to execute task T2. The few extra minutes were granted to recover from a brief network disconnection that occurred to Italian students. The group, however, failed to complete the task as the students did not respect the story point range constraints for the three releases. Later asked if the network disconnection influenced the task execution, the two Italian students agreed that "*[they] wasted most of the of the time in the beginning, to agree on assigning story points to features*" and, hence, had almost no time to properly arrange the three releases. Besides, Gr3 proved to be the most "active" group over both tasks, as they exhibited the highest frequency (6.33 and 6.90 upm, respectively) and the lowest average delay at typing utterances (10 and 8 sec., respectively). With respect to delays, the comparisons between the average delays in English meetings (mean 13.5 sec.) and MT meetings (mean 12.5 sec.) confirm that the subjects spent a little extra time in elaborating messages using the non-native English language.

Finally, we compared how the change of communication mode affected the participation extent of subjects. In particular, we were interested in observing changes (if any) that occurred to group members with the lowest English proficiency skills. Therefore, we compared the percentages of utterances presented by each participant during the EN and MT meetings. We used the percentages rather than the number of utterances because task T2 (release planning) was longer and more complex than T1 (prioritization) and so, regardless of the communication mode, any participant contributed more utterances during the second task. From all the eight logs collected we found as expected that, regardless of the task/communication-mode combination, the participant selected as moderator is the first or second most active subject of the group. This results is explained by the coordination duties that moderators had to perform during tasks execution. Besides, our findings reported in Table IV show that, generally, the percentage of utterances presented by the least proficient subjects in the groups increased when they could interact using their native language, except for the Brazilian student 16 in Gr3, who conversely contributed more utterances in English rather than in Portuguese..

TABLE III. DURATION OF MEETINGS, OVERALL NUMBER OF UTTERANCES EXCHANGED, FREQUENCY AS UTTERANCE PER MINUTE, AND AVERAGE DELAY BETWEEN TWO CONSECUTIVE UTTERANCES

Group	Task	Communication mode	Time (min.)	# Utterances	Frequency (upm)	Average delay (sec.)
Gr1	T1 (prioritization)	MT	40	159	3.95	15
	T2 (cost estimation)	EN	61	322	5.28	11
Gr2	T1 (prioritization)	EN	16	68	4.25	15
	T2 (cost estimation)	MT	59	346	5.86	10
Gr3	T1 (prioritization)	MT	30	190	6.33	10
	T2 (cost estimation)	EN	67	462	6.90	8
Gr4	T1 (prioritization)	EN	16	52	3.25	20
	T2 (cost estimation)	MT	54	169	3.13	14

TABLE IV. GAIN IN PARTICIPATION OF SUBJECTS LEAST PROFICIENT IN ENGLISH WHEN USING NATIVE LANGUAGE WITH THE HELP OF MACHINE TRANSLATION

Group	Subject id/ nationality	Eng. proficiency level (score)	% of utterance	
			EN	MT
Gr1	Brazilian Student #7	Level 2 (14/40)	9%	16%
Gr2	Brazilian. Student #4	Level 2 (12/40)	13%	19%
Gr3	Brazilian Student #16	Level 3 (30/40)	32%	23%
Gr4	Brazilian. Student #12	Level 3 (26/40)	10%	14%

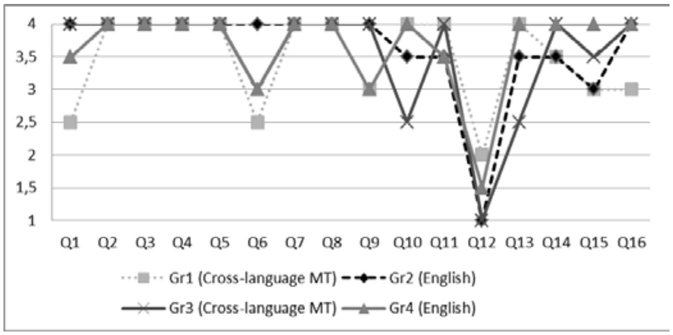

Figure 1. Post-T1 questionnaire responses (medians)

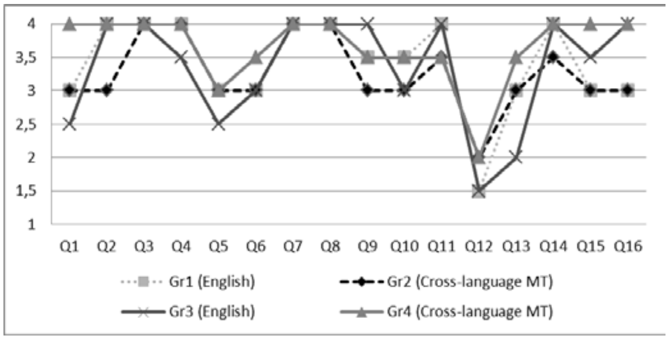

Figure 2. Post-T2 questionnaire responses (medians)

B. *Quantitative analysis from questionnaires*

In this section we report the findings from our quantitative analysis of post-task questionnaires. We only discuss significant differences observed in subjects' perception, distinguishing between the two tasks.

Figure 1 shows the medians of the responses to the questionnaire administered at the end of task T1 (prioritization). With respect to the *engagement level and comfort with communication mode* the study participants seem to have different opinions about the *"ease of communication with other participants"* (Q6). In fact, the English groups show moderate to strong agreement in their response (medians 3 and 4), whereas the cross-language MT groups have a mixed opinion (medians 2.5 and 3). Instead, as regards the level of *satisfaction with task execution*, the English teams felt to *"have enough time to perform the required activity"* (Q1, medians 3.5 and 4), whereas the cross-language MT groups show different opinions (medians 2.5 for Gr1 vs. 4 for Gr4). Finally, *"the global impression"* (Q16) of both the English and MT groups was positive (medians 3 and 4).

The responses to post-task T2 (release planning) questionnaire are shown in Figure 2. In this case, there is a general agreement between the cross-language groups and the English group, and no large differences are worth of mention, except for Q1, which suggests that English group felt more time pressure during T2 (medians 2.5 and 3) than MT groups (medians 3 and 4). Besides, from the medians of the responses to "control" questions Q2-Q5 and Q14-Q15, see the two previous figures, we are able to ensure, respectively, that the descriptions and goals of both T1 and T2 were clear to subjects, and that the tool itself did not suffer from flaws that hindered interaction during task execution.

C. *Qualitative analysis from questionnaires*

The post-T2 questionnaire contained an open question where subjects could freely report any thought or consideration about the whole experience. Although so far we have only collected the responses from 16 questionnaires, a few subjects provided useful insights. Italian subject 10 and Brazilian subject 11, both from Gr4, reported that *"[interaction over MT] was not as smooth as English-only interaction"*. In particular subject 10, later asked to elaborate on this, clarified that, during the meeting, the meaning of the comments were fully understood most of the times, despite of a few grammar mistakes or some wrong word choices. However, on some occasions, the automatic translation was below a threshold of tolerance, so that *"[they] had to ask the sender to rephrase the last comment, thus slowing things down*." For the same reason, Brazilian subject 8 questioned *"the applicability of MT as is in professional contexts"* where people might be less tolerant to mistranslations. Brazilian subject 16 suggested to use an automatic text correction service before translation as he felt that *"[the translation service] sometimes was tricked by my typos"*. Finally, Italian subjects 1 and 5 and Brazilian subject 3 (all at Level 4 of proficiency) reported that they could only see the usefulness of automatic MT *"when [one] is not so skilled in English."*

D. Qualitative analysis from meeting logs

In project meetings both lack of 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) can be detrimental. In such situations, people become aware that there is a problem of a *lack of common ground*, The common ground, or shared understanding, is the knowledge that participants have in common when communicating and the awareness of it [10]. A common ground is dynamically established through an interactive process, called grounding, through which participants exchange evidence about what they do or do not understand over the course of a conversation.

In order to collect evidence of the lack of common ground during meetings we look at the presence of *clarification dialogues*. Clarification dialogues, which are typically initiated by rephrasing statements in their own words, often in form of questions (e.g. "*Am I right if I interpret your statement as follows…?*", "*You mean that… , right?*"), provide evidence that a sentence presented by a speaker was not properly received by recipients [22]. Thus, we argue that the more and longer the clarification dialogues, the larger the lack of common ground. As a consequence, to quantify our construct of clarification dialogue, we started to perform the content analysis of the chat logs available, in order to count the number and the lengths of clarification such clarification dialogues.

Content analysis, also called coding [27], is a mix of quantitative and qualitative analysis that transforms qualitative data (e.g. written text, as in our case) into quantitative data (i.e. numbers) by applying a coding schema, which classifies content according to a finite set of possible thematic units (i.e. categories). A number of coding schemas have been proposed in the literature (e.g. [5] [13]), but none of them specifically covered the lack of common ground and, hence, we had to define a new one, shown in Table V. The coding schema lists the nine thematic units identified using the chat logs from the first three experimental runs as a training set.

As the next runs are executed, two of the researchers will independently apply the schema to the new chat logs and, then, inter-rater agreement (K index) will be computed to assess the concordance level between the resulting categorizations.

V. Discussion

The experiment presented in this paper is part of an ongoing study, the purpose of which is understanding to what extent real-time machine translations can be beneficial for distributed teams located in countries where professionals are not proficient in one common language. In this paper we conducted a follow-up study of our previous work by designing a controlled experiment with teams of students from Italy and Brazil who used this technology to participate in requirements meetings.

The experiment is in progress and thus we cannot make any strong statement. So far, in fact, four experimental runs have been executed. In general, first results suggest that concrete effects are expected to emerge when language barriers are critical.

More specifically, regarding the RQ1 (*Can machine translation services be used in distributed multilingual requirements meetings, instead of English?*), based on the data collected so far, we have found evidence that the use of MT is accepted with favor by participants and is not disruptive of the conversation flow, even during the execution of complex group tasks, such as distributed requirements meetings. Such finding is interesting because, as already shown by our previous study [6], state-of-the-art MT services are still far from 100% accuracy. Thus, the point is whether meeting participants always require exact word translations as long as it is still understood to a large extent. In other words, it is yet to be determined what an acceptable error rate is for automatic translation to be effective. We expect such rate to vary largely, depending on the criticality of the task to execute. In addition, data confirm that MT interaction is faster when it comes to contributing utterances, since native language is used, but overall it takes longer to complete the task, due to repairs (i.e. extra sentences) needed when mistranslations occur. Such findings are in line with results obtained by previous studies on MT (e.g. see [31], [32]).

Finally, with respect to RQ2 (*How does the adoption of machine translation affect group interaction in distributed multilingual requirements meetings, as compared to the use of English?*), we could not find any evidence of differences between MT and English interactions so far, although there are some clues (e.g. increase of participation of least proficient subjects) suggesting that differences might become evident with basic levels of English skills, but we still don't have concrete results at this time. In other words, it is yet to be determined the level of English proficiency below which the use of MT might become effective. This suggests that our research questions should be refined to also take into account English proficiency as a source of variation. More insights are expected to emerge from the content analysis of the logs.

A. *Practical Implications*

We believe that this work has important implications for the practice of requirements engineering in the context of Global Software Engineering (GSE). First, once machine translation services become stable and accurate, companies will be able to use it within teams whose members are not proficient in one common language. Second, if in the next rounds the experiment results indicated some evidence that machine translation services can be used by professionals with lower levels of English proficiency without prejudice of the project quality and productivity, companies would start to consider the possibility of recruiting people that have strong technical skills but still lack a higher level of proficiency in English (or the language spoken among the project team members).

However, IT professionals should be careful and still look for improvement in their language skills (mainly English). Machine translation services could work effectively for teams that have team members distributed across several countries, but face to face meetings will still need higher levels of English communication.

B. *Threats to Validity*

One of the key issues in experimentation is evaluating the validity of results [30]. In this section we discuss the potential threats that are relevant for our study and how they are addressed.

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 selection effect occurs due to the natural variation in human subjects' performance. Because we evaluate the interaction between participants using English as a *lingua franca*, the differences in their English proficiency might act as confounding factors. However, we control this threat by design, restricting the proficiency level of groups to advanced and average, and consequently assigning to them any student whose proficiency do not alter the group's designed level.

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 [16]. 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 [14].

TABLE V. THE PROPOSED CODING SCHEMA WITH THEMATIC UNITS (CATEGORIES).

Thematic unit (category)		**Description**
	QUESTION	A simple yes/no question (e.g. *"Web browser feature in the second release?", "Yeah"*) or a complex question (e.g. *"How do we arrange the first release? Complex features first?"*). It may also express the need for extra information or start a clarification dialogue
	ANSWER	A reply to a question that may take a few words (e.g. *yes, no, yeah, "correct, MMS"*) or more, depending on the complexity of the question. It may end a clarification dialogue.
CHECK	PROVISIONAL	Any utterance that explicitly looks for confirmation of acceptance through provisional, try-marked statements (e.g. *"So we decided for color screen, right?"*). It is normally followed by an AGREEMENT or an ANSWER.
	VERBATIM COPY	Any utterance that explicitly gives confirmation of acceptance by verbatim copying a previous utterances (e.g. *"Expandable memory is next"*, *"Ok, expandable memory next"*). It is normally followed by an AGREEMENT.
	MISUNDERSTANDING	Any utterance that provides evidence that a previously entered utterance was not accepted (e.g. *"I'm not sure I get the question"*, *"What?"*). It may initiate a request for clarification and is normally followed by a TASK or an ANSWER.
	ACKNOWLEDGMENT	Any utterance that explicitly demonstrates that a previously entered utterance has been understood and accepted (e.g. *ok*, *k*, *fine*), but not after a CHECK or QUESTION. It may end a clarification dialogue.
	TASK	Any task-related utterance, presented not in response to a question, which does not express acknowledgement or (dis)agreement (e.g. for providing clarification or extra information).
	AGREEMENT	Expresses agreement with a previously entered utterance, but not as an affirmative answer to a question, including smileys (e.g. *yes*, *yep*, *y*, *k*, *yeah*, *ok*, *right*, *I see*, *I agree*). It normally appears after a QUESTION, CHECK, or TASK utterance and may also end a clarification dialogue.
	DISAGREEMENT	Expresses disagreement with a previously entered utterance, but not as a negative answer to a question (e.g. *no*, *nope*, *n*). It may also initiate or continue a clarification dialogue.
	REPAIR	Any fragment entered to repair an error, typically in case of typos (e.g. *"It would be hard to surf the Internet without a color displays"*, *"...display"*) or clarifications necessary upon mistranslations.
	OTHER	Off-topic communication, not related to task, such as technical issues, preparation, activity coordination, and social messages. It may include smileys (e.g. *"Sorry, I'm late!"*, *"LOL!"*).

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 threats to construct validity. With respect to the two constructs of *engagement level and comfort with communication mode* and *satisfaction with task execution* defined for the questionnaire analysis, in our follow-up we will overcome this threat by executing principal component and scale reliability analysis to assess the extent to which a set of questions measures a single latent variable. As for the construct of *clarification dialogue*, we acknowledge the need for completing the content analysis. In fact, we will assess the effectiveness of the proposed coding by having two of the researchers apply the schema to the new chat logs and, as a consequence, the concordance level between the resulting categorizations can be evaluated computing the inter-rater agreement (K index).

VI. Conclusions & Future Work

In this paper we investigated the use of automatic, cross-language translation to communication-intensive activities, such as distributed requirements engineering and distributed project management. More specifically, we presented the design of a controlled experiment with students in Brazil and Italy in order to evaluate (1) if machine translation services can be used in distributed multilingual requirements meetings, instead of English; and (2) how the adoption of machine translation affects group interaction in distributed multilingual requirements meetings, as compared to the use of English.

As of this writing, four experimental runs have been executed. Since the experiment is still in progress we cannot make any strong statement. However, first results show that real-time machine translation is not disruptive of the conversation flow and is accepted with favor by participants. Besides, it seems that concrete effects are expected to emerge when language barriers are critical.

As future work, we plan to (a) use the obtained results to refine the design of the experiment by defining hypotheses that take into account the subjects' English proficiency levels; (b) replicate the experiments involving volunteers from industry, both in Brazil and Italy, thus making our experimental closer to a real distributed project environment. Finally, we will make the experiment material available for replications with subjects from other countries.

Acknowledgment

The first two authors are partially funded by IBM Faculty Award. The third author is funded by the PDTI program, financed by Dell Computers of Brazil Ltd. (Law 8.248/91) and Ci&T. The authors thank all the students who participated in the experiment and Teresa Mallardo for the help provided.

References

[1] D. Arnold, "Why translation is difficult for computers", In *Computers and Translation: A translator's guide*. Benjamins Translation Library, 2003.

[2] Bangalore, S., Murdock, V., and Riccardi, G. "Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system." *Proc. 19th Int'l Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, Aug. 24 – Sep. 1 2002, Volume 1, doi:10.3115/1072228.1072362.

[3] P. Berander, "Using Students as Subjects in Requirements Prioritization," *Int'l Symposium on Empirical Software Engineering (ISESE'04)*, pp. 167-176, 2004 , doi: 10.1109/ISESE.2004.1334904.

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

[5] J. Carletta, S. Isard, G. Doherty-Sneddon, A. Isard, J.C. Kowtko, and A.H. Anderson, "The Reliability of a Dialogue Structure Coding Scheme", *Computational Linguistics*, vol. 23, no. 1, pp. 13-31, Mar. 1997.

[6] 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, doi:10.1109/ICGSE.2010.37.

[7] 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, doi: 10.1002/spe.937.

[8] 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, doi:10.1109/52.914734.

[9] 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.

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

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

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

[13] N. Hara, C.J. Bonk, and C. Angeli, "Content Analysis of Online Discussion in an Applied Educational Psychology", *Instructional Science*, vol. 28, no. 2, pp. 115-152, 2000.

[14] 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.

[15] C. Hogan and R. Frederking, "WebDIPLOMAT: a Web-based interactive machine translation system." *Proc. 18th Int'l Conference on Computational Linguistics - Volume 2*, Saarbrücken, Germany, Jul. 31 – Aug. 04, 2000, pp. 1041-1045, doi:10.3115/992730.992801.

[16] 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.

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

[18] S. Jones and G. Parton. "Collaboration Across the Multinational Battlespace in Support of High-stakes Decision Making - Instant Messaging with Automated Language Translation", Technical report, The Mitre Corporation, 2008.

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

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

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

[22] B. Lutz, "Linguistic Challenges in Global Software Development: Lessons Learned in an International SW Development Division", *Proc. 4th Int'l Conf. Global Software Engineering (ICGSE'09),* Limerick, Ireland, Jul 13-16, 2009.

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

[24] 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, doi:10.1145/336512.336523.

[25] W. Odgen. “A Task-Based Evaluation Method for Embedded Machine Translation in Instant Messaging Systems,” in *Advanced Decision Architectures For The Warfighter: Foundations and Technology* (P. Mcdermott And L. Allender eds.), chapter 19, pp. 341-357, Aug. 2009.

[26] C. Pautasso, O. Zimmermann, F. Leymann, “RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision”, *Proc. 17th Int’l Conf. on World Wide Web (WWW ‘08)*, pp.805-814, 2008. doi=10.1145/1367497.1367606

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

[28] L.D. Paulson, “Translation technology tries to hurdle the language barrier,” *IEEE Computer*, vol. 34, no. 9, 2001, pp. 12-15.

[29] A.Verma, “A new look for Google Translate for Android”, The Official Google translate Blog, January 12, 2011, http://googletranslate.blogspot.com/2011/01/new-look-for-google-translate-for.html

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

[31] N. Yamashita and T. Ishida. “Effects of machine translation on collaborative work.” *Proc. 20th Int’l Conference on Computer Supported Cooperative Work (CSCW '06)*, Banff, Alberta, Canada, November 04-08, 2006, pp. 515-524, doi:10.1145/1180875.1180955.

[32] 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, doi:10.1145/1518701.1518807.