Empir Software Eng
DOI 10.1007/s10664-015-9372-x

Assessing the impact of real-time machine translation on multilingual meetings in global software projects

Fabio Calefato · Filippo Lanubile · Tayana Conte · Rafael Prikladnicki

Abstract Communication in global software development is hindered by language differences in countries with a lack of English speaking professionals. Machine translation is a technology that uses software to translate from one natural language to another. The progress of machine translation systems has been steady in the last decade. As for now, machine translation technology is particularly appealing because it might be used, in the form of cross-language chat services, in countries that are entering into global software projects. However, despite the recent progress of the technology, we still lack a thorough understanding of how real-time machine translation affects communication. In this paper, we present a set of empirical studies with the goal of assessing to what extent real-time machine translation can be used in distributed, multilingual requirements meetings instead of English. Results suggest that, despite far from 100 % accurate, real-time machine translation is not disruptive of the conversation flow and, therefore, is accepted with favor by participants. However, stronger effects can be expected to emerge when language barriers are more critical. Our findings add to the evidence about the recent advances of machine translation technology and provide some guidance to global software engineering practitioners in regarding the losses and gains of using English as a *lingua franca* in multilingual group communication, as in the case of computer-mediated requirements meetings.

Keywords Global software development · Machine translation · Distributed meetings · Computer-mediated communication · Controlled experiment

Communicated by: Nachiappan Nagappan

F. Calefato (✉) · F. Lanubile
Dipartimento di Informatica, University of Bari, Bari, Italy
e-mail: fabio.calefato@uniba.it

F. Lanubile
e-mail: filippo.lanubile@uniba.it

T. Conte
Instituto de Computação, Universidade Federal do Amazonas, Manaus, Brazil
e-mail: tayana@icomp.ufam.edu.br

R. Prikladnicki
Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil
e-mail: rafael.prikladnicki@pucrs.br

Published online: 26 February 2015

1 Introduction

Language is a critical factor that largely accounts for the success of global software projects in countries with strong English language capabilities, such as Ireland, the Philippines, India, and Singapore (Carmel and Agarwal 2001; Hsieh 2006; Shah et al. 2012). However, there are several other countries increasing their presence in the global IT market, but English-speaking professionals are limited in number. As an example, Brazil is a country where significant global software development operations have been located recently (Prikladnicki and Carmel 2013; Kearney 2007; Brazil IT-BPO Book 2013). However, the country lacks a bigger body of professionals who are able to communicate in English with confidence (KPMG 2009).

Machine translation (MT) is an established technology, some 60 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 (Jurafsky and Martin 2008). Machine translation 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 (Arnold 2003). Over the years, research has cyclically investigated the progress of state-of-the-art machine translation technology per se, in order to assess whether it has finally evolved to live up to its original and ultimate promise (Paulson 2001). Some researchers believe that machine translation technology available today is not reliable enough to be used "as is", since the best machine translation systems still make mistakes and they always will, except for translating simple texts or grasping the general meaning of complex ones (Raybaud et al. 2011). Therefore, post-editing of automatic translations has become a popular practice in the translation industry in order to assist human translators (Aziz et al. 2012). Nevertheless, machine translation is experiencing a tremendous growth of interest, on the heels of the Internet continuous expansion. As for now, machine translation technology is particularly appealing because it might be used, in the form of cross-language chat services, in countries where there are at the same time opportunities for global projects and the lack of English speaking professionals. However, despite the recent progress of the technology, we still lack a thorough understanding of how real-time machine translation affects communication.

Our goal is to identify to what extent real-time machine translation can be used in distributed, multilingual requirements meetings instead of English. In this paper, we present an empirical investigation that integrates former studies (Calefato et al. 2010, 2011, 2012b) and doubles the experimental runs of the controlled experiment (from 16 to 32) and the experimental size (from 32 to 64 subjects, with two levels of English proficiency). Thanks to this extension, we are in a better position to assess whether real-time machine translation can be used instead of English in distributed multilingual meetings. We focused on requirements meetings because requirements engineering is one of the most communication-intensive activities in software engineering and then it is specially challenged in global software projects (Damian and Zowghi 2003; Damian 2007).

The remainder of this paper is structured as follows. Section 2 describes our research goal and questions. Section 3 summarizes the results of the simulation study. Section 4 describes the experiment, including the design, the variables, the instrumentation and execution. Section 5 discusses the results of the quantitative and qualitative analysis. Threats to validity are described in Section 6. Finally, conclusions and future research activities are presented in Section 7.

2 Research Questions

Overall, our research goal is to further the understanding of the effect of real-time machine translation on multilingual groups collaboration in the context of global software projects. In contrast with previous research in the field of machine translation, our empirical investigation is more challenging because distributed requirements meetings are a complex, realistic and communication-intensive task.

We start by looking at the technical feasibility, that is, we first assess whether state-of-the-art machine translation services can be embedded into synchronous text-based chat without introducing errors or a delay to an extent that breaks the communication flow. Therefore, we define the first research question as follows:

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

Once the technical feasibility of using machine translation is verified, we can look at the effects on communication and group interaction. In particular, we want to know whether machine translation is accepted, and identify which differences exist in communication (i.e., interaction style, reaching understanding) when individuals can interact using their mother language as compared to English. Therefore, the second research question is:

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

Finally, since a better command of language provides better opportunities for steering communication during meetings, one could reasonably argue that machine translation is more useful to those who are not proficient in English. Therefore, we also define 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?*

3 Simulation

In order to answer RQ1, we organized a simulation, the goal of which is to evaluate the feasibility of adopting a machine translation service in a synchronous, multilingual text-based chat. In particular, the simulation compared the performance (i.e., the effectiveness and efficiency) of two machine translation services, Apertium[1] and Google Translate,[2] which represent two completely different approaches to machine translation. In fact, machine translation 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* machine translation systems use knowledge in the form of rules explicitly coded by human experts, which attempt to codify the translation process. Such systems heavily depend on linguistic knowledge, such as bilingual dictionaries (Arnold 2003). The most notable advantage of the rule-based machine translation approach is the ability to encode specific linguistic knowledge that automatic systems can process (e.g., morphological and bilingual dictionaries, lexical and structural transfer rules). Nevertheless, this approach has the

[1] http://www.apertium.org

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

drawback of being costly, as a considerable human effort is required in order to develop the necessary linguistic resources. Apertium is an open source, rule-based machine translation platform, supporting over 30 languages as of this writing. Once installed, Apertium can be also remotely accessed through a REST or an XML-RPC API, respectively for web and desktop applications, without further installations.

Corpus-based machine translation systems, conversely, 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 it is cheaper as systems are trained automatically, without any direct human intervention. The downside of this approach is that it requires huge amounts of training data, which may not be available for all languages and domains. Google Translate is an example of corpus-based machine translation 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 source and target language pairs. As of this writing, Google Translate supports the translation between any two pairs of 80 languages, although not all at the same quality level. Since both machine translation paradigms have different strengths and shortcomings, recently hybrid approaches have also emerged (Burchardt et al. 2013). Being a public service, Google Translate cannot be installed on a corporate server. Yet, it can be used by third-party applications because it exposes a RESTful interface that returns responses encoded as JSON[3] results.

Back in 2011, when we selected one service for each class of machine translators, we considered only those alternatives that provided a freely available API and included support for Italian. Moses,[4] Google Translate, and Bing Translator[5] were the three freely available services selected for representing the class of corpus-based statistical translators. The first one was discarded because a corpus of texts translated into Italian need to be provided for training. Then, Google Translate was preferred over Bing Translator after an internal evaluation of quality. As per the class of rule-based systems, instead, the other candidate besides Apertium was OpenLogos,[6] which supported Italian but provided no API.

As per the evaluation of performance, while the effectiveness of a machine translation service relates to the *quality* of the translated output (i.e., the fluency and fidelity of the translation), the efficiency relates to the amount of *time* necessary to translate the original input text (i.e., speed). Efficiency is fundamental in our scenario because if the use of the machine translation service involves a large amount of additional time, then it would break the real-time feature of a chat and consequently hamper the synchronicity of communication.

3.1 Evaluation of Translation Quality

The automatic translation of a text is a process affected by grammatical errors, mistranslations, and not translated words (Ogden et al. 2009). Evaluating the quality of a translation is an extremely subjective task and disagreements about evaluation methodology are rampant (Altman 1991; Mitkov 2003; Wisniewski et al. 2012). Nevertheless, evaluation is essential. In this study, we entailed four human raters to evaluate accurately each translation in terms of *adequacy* (Jurafsky and Martin 2008). Specifically, in our simulation the raters assessed the

[3] http://json.org

[4] http://www.statmt.org/moses

[5] http://www.bing.com/translator

[6] http://logos-os.dfki.de

adequacy of translations assigning scores to output sentences produced by the two machine translation services, judging whether each translation contains the information that existed in the original sentence. In order to evaluate the adequacy of translations properly, the raters also had to take into account the context, that is, all the sentences exchanged before the utterance at hand.

The scoring scheme adopted is a 4-point Likert scale (see Table 1), anchored with values *4= completely inadequate* and *1=completely adequate*. This scale was adapted from the intelligibility scale proposed in (Arnold et al. 1994) and considered appropriate to our goal because: (a) it is not too fine grained, i.e., it does not consist of too many values; (b) it can be easily applied as descriptions are well defined, i.e., it can be uniformly interpreted by evaluators; (c) and there is no middle value, i.e., it helps to avoid central tendency bias in ratings by forcing raters to judge the output as either adequate or not (Garland 1991; Johns 2005).

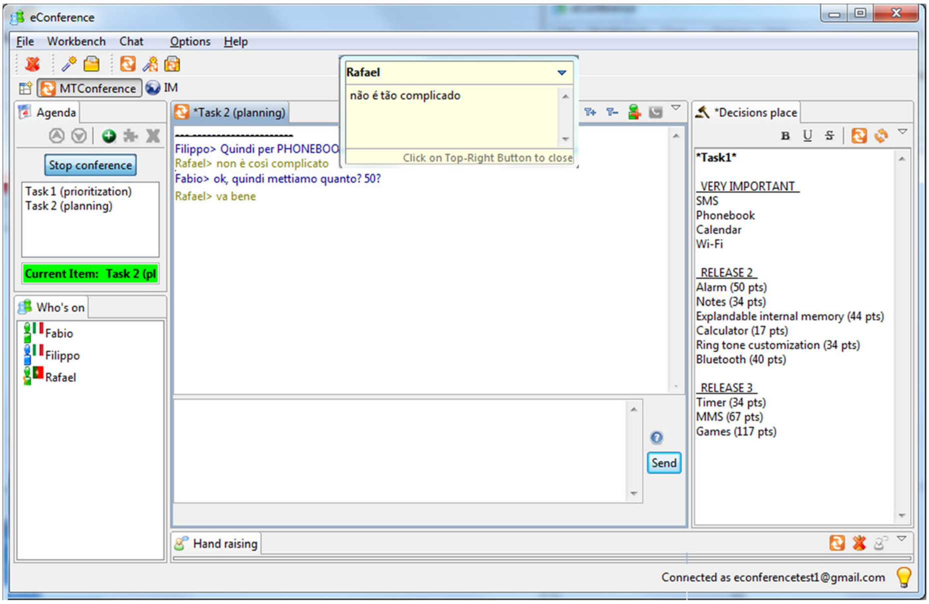
Before the official scoring session was held, the raters participated in a training session in which they become acquainted with the scale. The raters were all master students completing their thesis project in our laboratory at the University of Bari, and were selected among those who proved to have a good knowledge of English.

3.2 Instrumentation & Execution

eConference (Calefato and Lanubile 2007) is a closed group chat, augmented with agenda, meeting minutes editing, and typing awareness capabilities. In addition, we developed an ad hoc plugin, named eConferenceMT, which enables the automatic translation of text messages. The plugin allows the selection of a machine translation service (i.e., either Google Translate or Apertium) and a language pair for automatically translating incoming messages during one-to-one and group chat sessions. When eConference receives a new message, the plugin invokes the configured machine translation service, using the proper web-service interfaces, in order to show the translated messages along with the original text. Figure 1 shows a

Table 1 Adequacy scale (adapted from Arnold et al. 1994)

Value	Description
1	*Completely adequate* The translation clearly reflects the information contained in the original sentence. It is perfectly clear, intelligible, grammatically correct, and reads like ordinary text.
2	*Fairly adequate* The translation generally reflects the information contained in the original sentence, despite some inaccuracies or infelicities of the translation. It is generally clear and intelligible and one can (almost) immediately understand what it means.
2	*Fairly adequate* The translation generally reflects the information contained in the original sentence, despite some inaccuracies or infelicities of the translation. It is generally clear and intelligible and one can (almost) immediately understand what it means.
3	*Poorly adequate* The translation poorly reflects the information contained in the original sentence. It contains grammatical errors and/or poor word choices. The general idea of the translation is intelligible only after considerable study.
4	*Completely inadequate* The translation is unintelligible and it is not possible to obtain the information contained in the original sentence. Studying the meaning of the translation is hopeless and, even allowing for 3context, one feels that guessing would be too unreliable.



Rafael

não é tão complicado

Click on Top-Right Button to close

Filippo> Quindi per PHONEBOO

Rafael> non è così complicato

Fabio> ok, quindi mettiamo quanto? 50?

Rafael> va bene

Task1

VERY IMPORTANT
SMS
Phonebook
Calendar
Wi-Fi

RELEASE 2
Alarm (50 pts)
Notes (34 pts)
Explandable internal memory (44 pts)
Calculator (17 pts)
Ring tone customization (34 pts)
Bluetooth (40 pts)

RELEASE 3
Timer (34 pts)
MMS (67 pts)
Games (117 pts)

Connected as econferencetest1@gmail.com

Fig. 1 The eConference system augmented with the machine translation plugin

screenshot of eConference, with the machine translation plugin installed, as an example of multilingual group chat. The original sentence in Portuguese from Rafael ("*Não é tão complicado*"[7]) appears in the tooltip on mouse over, whereas its translation to Italian ("*Non è così compilcato*") shows up in the message board.

In order to run the simulation, as a first step, we modified our machine translation plugin in order to process the chat logs from recorded meetings. The plugin was modified in order to spawn several threads, one for each participant identified in the chat logs, and sent out the messages as loaded from the chat log. Each thread also received any message sent and then invoked the translation service. Because all the messages logged by our tool are timestamped, we were able to send them with the same timing and order as in the real workshops, that is, we recreated a realistic condition similar to the one that would have happened if the actual communication had relied upon our machine translation system. Besides, we also put each translation service under the same stress condition in which messages sent at the same time would have caused the translation service to be invoked concurrently by each participant in the workshop.

The text corpus used to run the simulation is composed of chat logs, written in English, and collected from five requirements workshops run during an experiment on the effects of text-based communication in distributed requirements engineering (Calefato and Lanubile 2007).

We used one workshop log (CL1) to train the raters, whereas the remaining four (CL2-CL5) were employed as the test set during the simulation. Overall, the test set accounted for over 2000 utterances to be translated by both machine translation services. Participants in each workshop ranged from five to eight undergraduate students attending a requirements engineering course at the University of Victoria, Canada. During a workshop the participants, either acting as a client or as a developer, had first to elicit the requirements specification of a

[7] In English: "*It's not so complicated*"

web application (first session); then, they had to negotiate and reach closure on the previously collected requirements (second session). Table 2 contains an excerpt of the chat logs, showing the messages exchanged between two clients and one developer.

The raters completed the evaluation of the whole corpus in two weeks. For each of the four chat logs to be evaluated, the raters received a spreadsheet containing the original body of sentences and the two translations. The sheets containing the translations produced by Apertium and Google Translate were added in random order and renamed TR-A and TR-B. Furthermore, because Apertium marks unknown words using the symbols *, #, and @, we searched for and removed these markers from its translation results. Such a setup allowed us to avoid any potential bias of order and prevent the raters from identifying the service.

The simulation was executed on a box running Debian GNU/Linux, with two 2GHz Dual-Core AMD Opteron CPUs and 4 GB of memory. Finally, in order to compare the performance of Apertium and Google Translate, the simulation was run twice on the same text corpus and on the same machine, once for each machine translation service.

3.3 Results

Our analysis focused on evaluating both the effectiveness and the efficiency in order to assess, respectively, the goodness of translations in terms of adequacy and the extra amount of time taken to translate over 2000 sentences from the original language (English) to the target language (Italian).

3.3.1 Translation Quality Results

The four coders performed the rating separately. We measured the inter-rater agreement by computing the Fleiss' kappa index for multiple raters (Fleiss 1981). Kappa index is a statistical measure applied for assessing the reliability of agreement between multiple raters (i.e., more than two) involved in assigning categorical ratings to a number of items. In particular, for the Apertium service, the Fleiss' kappa index shows a fair agreement level (k=0.37) (Altman 1991). Instead, for Google Translate, the kappa index measured shows a moderate agreement level between the raters (k=0.47) (Altman 1991).

In order to identify a difference between the quality of translation, as perceived by the raters, produced by the two machine translation services, we first calculated how many sentences were evaluated as adequate (i.e., belonging to categories 1 and 2) and inadequate (i.e., belonging to categories 3 and 4). Fig. 2 shows that, for Google Translate, over a half of the whole test suite (2053 sentences) was judged adequate (63.3 %). Conversely, for Apertium over the 62.2 % of the translated sentences was judged inadequate. In addition, we found that the mean and median ratings for Google Translate were, respectively, 2.17 and 2.0. Instead, for Apertium the mean and median ratings were 2.8 and 3.5, respectively.

Table 2 An excerpt from the chat logs

Participant	Message
Client 1	we don't necessarily need the conversations to be stored in a DB…
Client 2	We also need application sharing. IE - letting someone else access a single window on my computer.
Client 1	and yeah, we do need application sharing
Dev 1	Ok
Dev 1	we have questions about that so just wanted an overview

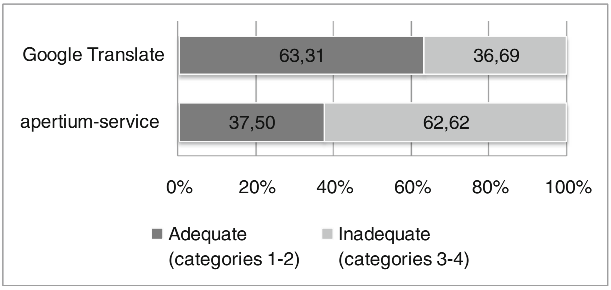


Fig. 2 Percentage of adequate vs. inadequate ratings

Afterwards, we performed a paired *t*-test for two related samples. We summed the ratings from each rater for each translated utterances, thus obtaining $N=2053$ summed scores for each machine translation service. The summed scores obtained ranged between 4 (best) and 16 (worst). The paired *t*-test result, shown in Table 3, revealed a statistically significant difference at the 1 % level ($t=-27.06$, $p=.000$) in favor of Google Translate, which thus was found to produce more accurate translations than Apertium. As a non-parametric alternative to the *t*-test for two related samples, we also performed the Wilcoxon test, which confirmed the same results of the former statistic.

3.3.2 Effect of Typos on Translation Quality

One can easily argue that the presence of both syntactical and grammatical errors in source sentences can affect the output of machine translation. Typos, in particular, can easily occur when typing fast. For example, in the over 2000 utterances from the chat logs taken into account in this simulation, an automatic spell checker (Mudge 2009) identified as much as 379 utterances containing at least one spelling error, and 682 with at least one grammatical error (spelling errors included). Therefore, we analyzed how well the machine translation services are able to cope with the presence of errors to understand whether a machine translation-augmented chat might benefit from using an integrated and automatic spell checker that highlights errors in a typed text.

For each of the two machine translation services, we first divided the set of the translated sentences into two subsets, one obtained from the translation of original utterances containing at least one error and one from those containing none. Then, for every utterance in each subset, we computed an aggregated rating – from 4 (best) to 16 (worst) – by summing the scores from all the four raters. Results are shown in Figs. 3 and 4 for Google Translate and Apertium, respectively. In the case of Google Translate, the average aggregate score for utterances *with errors* was 9.23 (SD=3.78), whereas for those *without errors* it was 8.38 (SD=4.28). In the

Table 3 Results from the paired *t*-test

	Mean	Std. Dv.	N	Diff.	Std. Dv. Diff.	t	df	*p*
Apertium	11.19	4.06	2053	−2.58	4.23	**−27.06**	2052	**.000**[a]
Google translate	8.66	4.13						

[a] Statistically significant results at the 0.01 level are indicated in bold

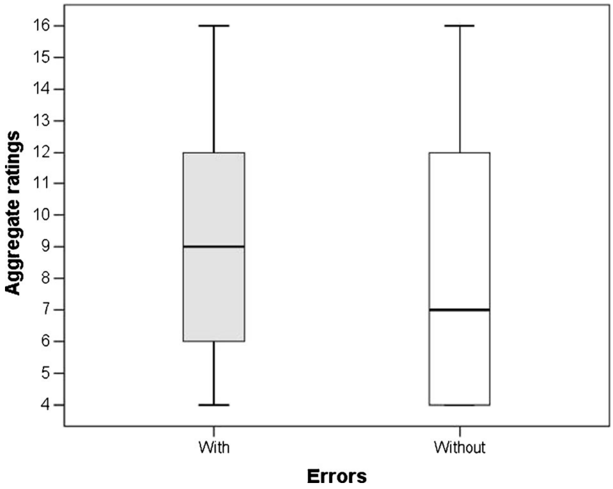


Fig. 3 Box-and-Whiskers plot of ratings for utterances translated by Google Translate from original sentences with and without errors (the lower, the better)

case of Apertium, instead, the average aggregate score for utterances *with errors* was 12.14 (SD=3.39), whereas for those *without errors* it was 10.71 (SD=4.28). Finally, we performed an independent *t*-test on the two unpaired subsets to assess whether the effect of errors on translation was significant (see Table 4). Both in the case of Google Translate (t=4.42, *p*=.000) and Apertium (t=7.59, *p*=.000), we found a significant difference at the 1 % level in the aggregate scores for utterances *with errors* and *without errors*. In addition, as a non-parametric alternative to the *t*-test for two independent samples, we also performed the Mann–Whitney *U* test, which confirmed the same results of the former statistic.

However, given the large size of the sample and the small differences in the mean values obtained, we complemented the result of the *t*-test by computing the effect size through

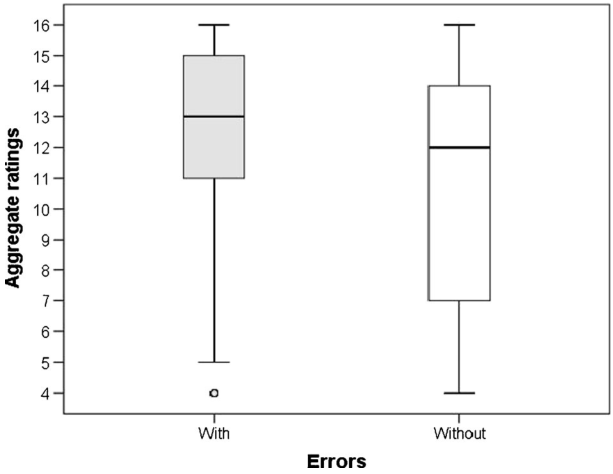


Fig. 4 Box-and-Whiskers plot of ratings for utterances translated by Apertium from original sentences with and without errors (the lower, the better)

Table 4 Results from the unpaired *t*-test to assess the effect of errors on translation

	Errors	N	Mean	Std. deviation	t	df	*p*
Apertium	With	682	12.14	3.39	4.42	2051	*p*=.000
	Without	1371	10.71	4.28			
Google translate	With	682	9.23	3.77	7.59	2051	*p*=.000
	Without	1371	8.38	4.27			

Cohen's d (Cohen 1992). The effect size measure captures the magnitude of mean differences in two groups. In the case of Google, the effect size computed is d=0.21. In the case of Apertium, instead, the effect size is d=0.37. According to the scale suggested by Cohen, both values indicate a small effect size. Therefore, we conclude that errors in original sentences did not have an impact on translation quality.

3.3.3 Time Performance Results

We collected and compared the response times of the two services when completing a growing number of concurrent translations. In particular, the data points for this time performance analysis were obtained as an average measure of 256 repeated translation requests.

Figure 5 shows that Apertium response times are lower than those of Google Translate are. In fact, in the worst case, Apertium took on the average less than 30 ms to complete the repeated translation requests, whereas Google's service took more than twice the amount of time (over 70 ms). Conversely, the graph of response times shows that Google Translate performance does not depend on the length of the sentences, as in the case of Apertium.

Figure 6 plots the response times of the two services when completing concurrent translation requests from an increasing number (from 1 to 8) of clients. The data points were again collected as an average measure of 256 repeated requests for translating the longest sentences available in the whole data set (362 characters). The graph shows that Apertium performances

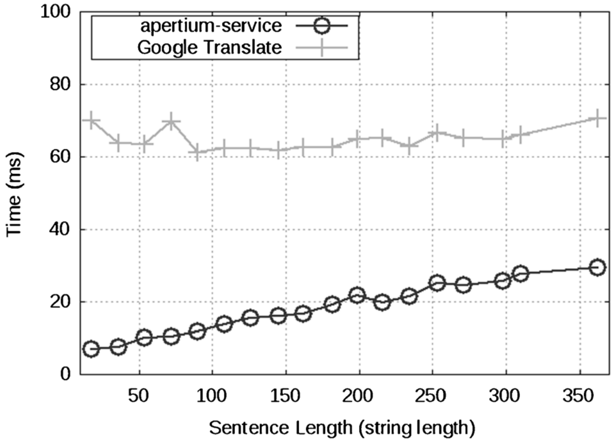


Fig. 5 A comparison between the amounts of time (in ms) taken by Google Translate and Apertium services to translate sentences of growing lengths

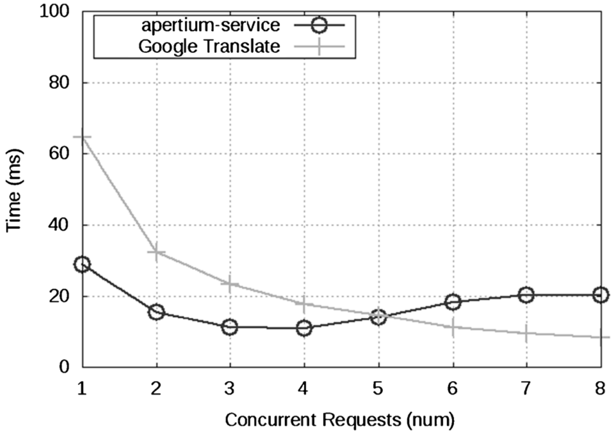


Fig. 6 A comparison between the amounts of time (in ms) taken by Google Translate and Apertium to complete 1024 translation tasks requested from an increasing number of concurrent clients

are better when the numbers of concurrent requests are low (less than 4), whereas Google Translate is better able to cope with a growing number of concurrent clients.

4 The Controlled Experiment

The simulation described in the previous section provides us with data to assess the technical feasibility of embedding using machine translation into synchronous text-based chat without disrupting real-time interaction (RQ1). However, the simulation results do not provide insights about whether complex group tasks can be carried out while participants communicate in their own native language with the help of machine translation. Therefore, to gain further data about RQ1 and also be able to answer RQ2 and RQ3, we designed a controlled experiment with the following characteristics:

(1) multilingual groups who interact to complete a knowledge- and communication-intensive task;
(2) subjects with different levels of English proficiency;
(3) a machine translation service for two-way translations.

The controlled experiment involved 64 participants, both graduate and undergraduate students from Brazil and Italy. The Brazilian students were from either the Federal University of Amazonas in Manaus or PUCRS in Porto Alegre, whereas the Italian students were from University of Bari. None of the students who volunteered to take part in the experiment knew about the experiment before or beyond what was shared during the training and execution sessions (see next section).

The participants interacted in groups of four people, two from Italy and two from Brazil, using two different communication modalities, that is, their respective native language, Italian or Portuguese, with the help of machine translation (*MT*), and English (*EN*), as a non-native *lingua franca* (Lutz 2009). 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*), they completed a release plan acting as developers. The task material, adapted from a previous work by Berander (2004), 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,[8] which includes 40 questions to be answered within 20 min. The test originally placed subjects into one of four distinct categories. For this experiment, instead, we collapsed the four categories in just two, placing the participants at either the *Low* level (scores 0–20) or the *High* level (scores 21–40).

4.1 The Study Design

As shown in Table 5, we followed a fractional factorial design (Montgomery 1996) in which each group participated in two meetings (Runs 1 and 2), using a combination of the *communication mode* (MT and EN) and *task* (T1 and T2). Each of the overall 16 multilingual groups included 4 subjects, 2 speaking Italian and 2 speaking Portuguese as their native language.

In the two runs, all participants had the chance to perform both tasks using either communication mode. For example, a group that communicated through machine translation (MT) to execute T1 in the first run, instead, used the English language (EN) to perform T2 in the second run and vice versa. This design allows an experimenter to do two comparisons, that is, 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 at both team- and individual-level.

4.2 Instrumentation, Training & Execution

Multilingual group meetings were run using eConferenceMT. Before each meeting, the groups involved were trained to use the tool. First, a half-hour demo was given to participants by one of the researchers. Then, a training session was set up, during which the groups had to perform two training tasks, interacting first using their native language, exploiting the machine translation 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 each training session, two of the four participants were randomly selected to act as moderator or scribe. The extra duties of being a moderator included starting the meeting once every participant is online, keeping track of time limit, and so forth. The session scribe, instead, edited the content of the whiteboard, a shared editor where all the group decisions and the task solution were logged.

As per the experiment group meetings, we kept the same groups of participants arranged for the training sessions. Each experimental meeting required the two groups to complete both runs in two hours. Two of the researchers, one in Brazil and one in Italy, were available to participants during experimental meetings, in order to provide technical help and prevent undesired interactions to occur outside of the tool, as same language participants were collocated at each site. During each run, the two groups were required to solve the two tasks one after the other. The tasks were described in English on printings handed to the participants just before the start.

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

Table 5 Experimental plan

	MT	EN
Run 1	Gr1, Gr3, Gr6, Gr8, Gr9, Gr11, Gr13, Gr15 *execute T1*	Gr2, Gr4, Gr5, Gr7 Gr10, Gr12, Gr14, Gr16 *execute T1*
Run 2	Gr2, Gr4, Gr5, Gr7, Gr10, Gr12, Gr14, Gr16 *execute T2*	Gr1, Gr3, Gr6, Gr8, Gr9, Gr11, Gr13, Gr15 *execute T2*

The first one (T1) was a requirements prioritization task to be completed within 30 min. The 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 the features into three distinct piles (i.e., less important, important, and very important). Further task constraints required that the features within each pile were ranked by importance, and that no more than 13 features 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 min. In the first step, the participants had to distribute an overall amount of 1000 story points between the same 16 features 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.

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

4.3 Dependent Variables and Measures

Two are the data sources considered in this study: a questionnaire, which was administered to the participants upon the conclusion of each task, and the meeting logs, automatically collected by the conferencing tool.

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 and Gallupe (1987), 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 (written in English), we adopted a 4-point Likert scale anchored with '4=strongly agree' and '1=strongly disagree' values. The scale 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 questionnaire listed 16 closed questions, plus an open question, where subjects could freely report any thought or consideration about the whole experience, and a few other "control" questions, included only to ensure that task execution was not hindered by the tool flaws or by unclear instructions and objectives (as such, they are neither reported nor discussed in the rest of the paper). 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, upon the execution of both experimental runs.

Instead, from the chat logs collected at the end of the meetings, the *# of utterances* entered by group participants were counted to assess the 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) (Yamashita et al. 2009). In such situations, people become aware that there is a problem of *lack of common ground*. Common ground is the knowledge that participants have in common when communicating and the awareness of it (Clark and Brennan 1991). 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 could expect more *clarification requests* to emerge during machine translation-enabled meetings due to translations errors and inaccuracies. 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 machine translation technology inaccuracy rather than low English proficiency to generate more clarifications requests by participants in a conversation. To quantify our construct of clarification requests, we performed a content analysis of the meeting logs.

4.4 Results

In this section, we report the results from the analysis of data collected from the 32 experimental runs (16 groups, two runs each). We first present the 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 some of the logs.

4.4.1 Quantitative Analysis of Meeting Logs

Table 6 provides some descriptive measures of the meetings executed in the experiment, grouped by proficiency. 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 tend to vary more for run 1 than for run 2, during which most of the groups took the whole time allowed (1 h). In fact, the amounts time spent for run 1 range between 16 min (Gr2 and Gr4, both High) and 40 min (Gr1 – High and Gr7 – Low), 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. More specifically, every participant came up with their priority list, from which they eventually built a shared solution. The other groups, instead, adopted a more practical approach: first, one participant

Table 6 Descriptive measures for the eight meetings

Group		Communication mode	Time (min.)	# Utterances	Frequency (upm)	Average delay (sec.)
Gr1 (High)	Run 1	MT	40[b]	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[a]	462	6.90	8
Gr4 (High)	Run 1	EN	16	52	3.25	20
	Run 2	MT	54	169	3.13	14
Gr9 (High)	Run 1	MT	23	132	5.91	10
	Run 2	EN	64[b]	322	5.07	12
Gr10 (High)	Run 1	EN	24	276	11.73	9
	Run 2	MT	62[b]	378	6.14	10
Gr12 (High)	Run 1	EN	35[b]	241	6.86	9
	Run 2	MT	43	252	5.94	10
Gr13 (High)	Run 1	MT	31[b]	277	9.13	7
	Run 2	EN	58	464	7.96	8
Gr5(Low)	Run 1	EN	28	92	5.41	11
	Run 2	MT	59	358	6.17	10
Gr6(Low)	Run 1	MT	35[a]	140	4.38	14
	Run 2	EN	59	164	2.83	21
Gr7 (Low)	Run 1	EN	40[b]	264	6.44	9
	Run 2	MT	60	405	6.75	9
Gr8 (Low)	Run 1	MT	43[b]	240	5.58	11
	Run 2	EN	67[b]	354	5.28	11
Gr11 (Low)	Run 1	MT	35[b]	213	6.17	10
	Run 2	EN	50	307	6.14	10
Gr 14 (Low)	Run 1	EN	30	208	6.92	9
	Run 2	MT	57	474	8.4	8
Gr15 (Low)	Run 1	MT	29	142	5	12
	Run 2	EN	41	177	4.36	14
Gr16 (Low)	Run 1	EN	27	135	5.11	12
	Run 2	MT	46	280	6.18	11

[a] Extra minutes granted to recover from network disconnection

[b] Exceeded the time limit for the task (30 min. for T1, 60 min. for T2)

proposed an initial priority list and, then, 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 min 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, Gr10 (High), Gr13 (High), and Gr14 (Low) proved to be the most "active" groups overall, as they exhibited the highest frequency (upm between 6.92 and 11.73) and the lowest average delay at typing utterances (delay between 7 and 10 s.) over the two tasks. Finally, with respect to the communication mode, the comparison between the upm and average delays in English meetings (respectively, 5.86 and 11.8 s.) and machine translation-enabled meetings (respectively, 5.93 and 10.7 s.) confirms that the subjects spent a little extra time in elaborating and/or composing messages using the non-native language. This difference appears more evident in English meetings involving participants with low skills, in the case of which the average delay is 12.13 s. Instead, in the other three cases the delays are almost identical, ranging between 10.25 and 10.75 s.

In order to verify the equality of participation during meetings (i.e., no domination by any group member), we computed the number of utterances presented by each participant during a meeting to see how the use of machine translation affected the participation extent of the subjects (see Table 7). More specifically, we calculated the percentages because the release planning task executed during run 2 takes longer than the prioritization task of run 1 and, therefore, any participant is expected to have contributed more utterances during the former, regardless of the communication mode. We also 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) in general (Gr1-8, Gr13, Gr15-16), during machine translation-enabled tasks there is an increase of participation (i.e., a smaller delta) of the least prolific subject, typically at the expense of the most prolific one, regardless of the English proficiency level; 2) the deltas decrease when using machine translation in 6 of the 8 high proficiency groups, whereas, as for low proficiency groups, the decrease of deltas occurs in only 4 of the 8 groups.

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 8). These results reveal that the percentage of utterances presented by the least proficient subjects tend to increase when switching from English to native language (8 cases, ~50 %); we also observed that the percentage remained the same in 6 cases (~37 %) and it decreased in two cases only (Gr3 – High and Gr15 – Low, ~13 %).

4.4.2 Questionnaires Analysis

In this section, we report the findings from the quantitative analysis of post-task questionnaires. In order to measure the *satisfaction with task performance*, we included in the questionnaire five closed questions to assess participants' perception of whether the tasks were performed positively, with participants feeling actively involved in group communication when reaching shared decisions. The questions (i.e., "*The task was easy to perform*", "*I actively participated in the discussion*", "*I had the sensation of wasting time*", "*It was easy to reach a common decision*", "*I had a positive global impression of the performance*") aimed to verify that information exchanged, and optionally translated, was properly processed when using both communication modes (see Table 9). As a nonparametric alternative to the *t*-test for two paired samples, we performed a Wilcoxon signed rank test (Conover 1980) on the responses to the five questions. The test failed to reveal any difference in the levels of satisfaction with performance perceived by both high and low proficiency subjects when using English rather than their native language.

Likewise, to measure the levels of *engagement and comfort with communication mode* perceived, we designed six questions (i.e., "*I had enough time to perform the activity*", "*It was easy to communicate with others*", "*I had adequate opportunity to participate in the discussion*", "*I was encouraged to discuss contrasting solutions with others*", "*Other participants adequately answered my questions*", "*I felt involved in the discussion*") to assess discussion contentment (see Table 10). Again, we performed a Wilcoxon signed rank test,

Table 7 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 % ↓
Gr9 (High)	54 (17 %)	105 (33 %)	57 (18 %)	106 (33 %)	15 %	28 (21 %)	50 (38 %)	19 (14 %)	35 (27 %)	24 % ↑
Gr10 (High)	71 (26 %)	98 (36 %)	68 (25 %)	39 (14 %)	22 %	105 (28 %)	135 (36 %)	92 (24 %)	46 (12 %)	24 % ≈
Gr12 (High)	65 (27 %)	66 (27 %)	57 (24 %)	53 (22 %)	5 %	65 (24 %)	61 (24 %)	58 (23 %)	69 (27 %)	4 % ↓
Gr13 (High)	230 (50 %)	83 (18 %)	117 (25 %)	34 (7 %)	43 %	130 (47 %)	47 (17 %)	72 (26 %)	28 (10 %)	37 % ↓
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 % ↓
Gr11 (Low)	109 (36 %)	59 (19 %)	71 (23 %)	68 (22 %)	17 %	82 (38 %)	31 (15 %)	50 (23 %)	50 (23 %)	23 % ↑
Gr14	64 (31 %)	50 (24 %)	58 (28 %)	36 (17 %)	14 %	134 (28 %)	129 (27 %)	157 (33 %)	54 (11 %)	22 % ↑

Table 7 (continued)

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	
(Low)										
Gr15 (Low)	62 (35 %)	30 (17 %)	49 (28 %)	36 (20 %)	18 %	66 (46 %)	28 (20 %)	23 (16 %)	25 (18 %)	30 % ↑
Gr16 (Low)	66 (49 %)	20 (15 %)	25 (19 %)	24 (18 %)	34 %	141 (50 %)	36 (13 %)	57 (20 %)	46 (16 %)	37 % ≈

Table 8 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)	Participant #7 (Brazilian)	19 %	27 % ↑
Gr2 (High)	Participant #4 (Brazilian)	22 %	26 % ↑
Gr3 (High)	Participant #16 (Brazilian)	32 %	23 % ↓
Gr4 (High)	Participant #12 (Brazilian)	10 %	14 % ↑
Gr9 (High)	Participant #33 (Italian)	33 %	38 % ↑
Gr10 (High)	Participant #37 (Italian)	36 %	36 % ≈
Gr12 (High)	Participant #48 (Brazilian)	24 %	23 % ≈
Gr13 (High)	Participant #50 (Italian)	25 %	26 % ≈
Gr5 (Low)	Participant #17 (Italian)	21 %	36 % ↑
Gr6 (Low)	Participant #22 (Italian)	20 %	27 % ↑
Gr7 (Low)	Participant #27 (Brazilian)	15 %	14 % ≈
Gr8 (Low)	Participant #32 (Brazilian)	23 %	26 % ↑
Gr11 (Low)	Participant #44 (Brazilian)	23 %	23 % ≈
Gr14 (Low)	Participant #49 (Italian)	28 %	33 % ↑
Gr15 (Low)	Participant #57 (Italian)	28 %	16 % ↓
Gr16 (Low)	Participant #63 (Brazilian)	19 %	20 % ≈

which also failed to reveal any difference between the use of English and machine translation in terms of being involved in an open and useful discussion with others.

Furthermore, since the groups experienced both communication mode (MT and EN) performing two different tasks, we also grouped all subjects' responses by task, thus obtaining two separate data sets, one related to T1 and one related to T2. Then, we executed again the Wilcoxon test on these two datasets. The result of this repeated test proved that, overall, the differences in tasks did not account for differences in our evaluation. Anyway, we found a significant difference in only two cases. More specifically, as per the *satisfaction with task performance*, the signed rank test revealed a statistically significant difference at the 1 % level only for Q13 in the case of low proficiency groups ($Z=-2.667$, $p=.008$), that is, the participants had a better perception of a positive performance in T1 when using English. As

Table 9 Evaluation of satisfaction with performance for high and low proficiency groups

	High proficiency groups (N=32)				Low proficiency groups (N=32)			
	EN>MT	EN<MT	EN=MT	Wilcoxon Signed Rank Test	EN>MT	EN<MT	EN=MT	Wilcoxon Signed Rank Test
Q5. "*The task was easy to perform*"	9	11	12	Z=−0.968 p=.33	12	7	13	Z=−1.127 p=.26
Q8. "*I actively participated in the discussion*"	3	3	26	Z=−0.333 p=.74	5	8	19	Z=−0.832 p=.4
Q12. "*I had the sensation of wasting time*"	12	11	9	Z=−1.92 p=.85	13	10	9	Z=−1.166 p=.24
Q13. "*It was easy to reach a common decision*"	13	11	9	Z=−0.646 p=.52	14	8	10	Z=−1.285 p=.2
Q16. "*I had a positive global impression of the performance*"	6	8	18	Z=−0.632 p=.527	6	9	17	Z=−0.943 p=.35

Table 10 Evaluation of engagement and comfort with communication mode for both high and low proficiency groups

	High proficiency groups (N=32)				Low proficiency groups (N=32)			
	EN>MT	EN<MT	EN=MT	Wilcoxon Signed Rank Test	EN>MT	EN<MT	EN=MT	Wilcoxon Signed Rank Test
Q1. “*I had enough time to perform the activity*”	5	15	12	Z=−1.974 p=.48	6	8	16	Z=−0.852 p=.39
Q6. “*It was easy to communicate with others*”	15	7	10	Z=−1.99 p=.047	7	14	11	Z=−0.821 p=.41
Q7. “*I had adequate opportunity to participate in the discussion*”	8	4	20	Z=−1.291 p=.198	8	15	9	Z=−1.406 p=.16
Q9. “*I was encouraged to discuss contrasting solutions with others*”	13	7	12	Z=−0.597 p=.55	8	7	17	Z=−0.268 p=.79
Q10. “*Other participants adequately answered my questions*”	11	6	5	Z=−1.166 p=.24	12	12	8	Z=0.0 p=1.0
Q11. “*I felt involved in the discussion*”	6	9	20	Z=−0.247 p=.80	5	9	18	Z=−1.213 p=.22

Table 11 Evaluation of overall communication mode preference for both high and low proficiency groups (N=32)

	A vs. B	A>B	A<B	A=B	Wilcoxon Signed Rank Test
High groups	"*Group activity benefited from using…*" MT vs. EN	11	13	8	Z=−0.934 p=.35
	"Another time, I would rather communicate using…" MT vs. EN	17	9	6	Z=−1.113 p=.27
Low groups	"*Group activity benefited from using…*" MT vs. English	9	11	12	Z=−0.512 p=.61
	"Another time, I would rather communicate using…" MT vs. EN	18	4	10	**Z=−2.679 p=.007**[a]

[a] Statistically significant results at the 0.01 level are indicated in bold

per the *engagement and comfort with communication mode*, instead, the test revealed a statistically significant difference at the 1 % level only for Q6 in the case of high proficiency groups (Z=−2.653, p=.008), indicating that in T1 the participants with better English skills found easier to interact when using English instead of machine translation.

We also analyzed the answerer to the four questions (Q17-Q20) meant 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* machine translation *service / without the* machine translation *service*"). We again applied a Wilcoxon signed-rank test. Overall, the results reported in Table 11 show that the subjects perceived no particular benefit from using machine translation. Conversely, the test revealed a statistically significant difference at the 1 % level (Z=−2.679, p=.007) only for low proficiency groups who showed a preference towards using machine translation-enabled communication rather than English.

Finally, we reviewed the answers to the open question included in the post-T2 questionnaire. Here subjects were free to report any thought or consideration about the whole experience. Although mostly gave suggestions to possible tool extensions (like using an automatic text correction service), a few interesting answered were collected from members of two groups. Specifically, Italian subject 20 and Brazilian subject 2, from Gr5 and Gr7 respectively, reported that "*[interaction over* machine translation*] was not as smooth as English-only interaction.*" In particular, subject 18, later asked to elaborate on this, clarified that, during the meeting, she could fully understand the meaning of the comments most of the times, despite of a few grammar mistakes or some wrong word choices. However, on some

Table 12 The result of content analysis on Gr5 and Gr7 logs

	EN (Run 1)			MT (Run 2)		
	Check misunderstanding	Check provisional	Unknown	Check misunderstanding	Check provisional	Unknown
Gr5 (Low)	0 %	2.2 %	0 %	2.9 %	5.9 %	4.3 %
Gr7 (Low)	1.9 %	3.8 %	0.9 %	1 %	1.2 %	3.2 %

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.*"

4.4.3 Content Analysis

To determine if the adoption of machine translation affects group interaction in multilingual group meetings, we had to reverse the perspective and look for evidences of *lack of common ground* (Clark and Brennan 1991). 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 machine translation 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 on some of the logs collected from low proficiency group meetings. Content analysis, also called coding (Stemler 2001), 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, also known as categories. We specifically developed an ad hoc coding schema for this study, composed of ten categories (see Table 13). Two of the researchers performed the content analysis separately. We opportunistically performed such analysis only on the logs from low proficiency groups Gr5 and Gr7, for which some subjects reported about comprehension difficulties in the questionnaires. To ensure the concordance level between the resulting categorization, we applied Cohen's kappa (Cohen 1960), which is the statistical measure to assess the reliability of agreement between two raters. The computed indexes are k=0.88 and k=0.91, meaning almost perfect agreement between the raters (Landis and Koch 1977).

Table 12 shows the breakdown of the content analysis performed. We note that thematic unit percentages are reported, rather than occurrences, as a necessary normalization due to the large differences in the lengths of task discussions. In addition, 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 were impossible to code because their meaning was incomprehensible due to either poor English (i.e., poorly written source) or inaccurate translations (i.e., inaccurately translated output).

As compared to EN runs, the results show a higher number of Unknown (i.e., unclassifiable) utterances in machine translation runs. Although partial, these results suggest that the inaccuracy of state-of-the-art machine translation technology poses more hurdles to common ground than language barrier.

5 Discussion

The main contribution of this paper is the empirical evidence from a couple of studies (one simulation and one controlled experiment) that furthered our understanding of the effect of real-time machine translation on multilingual group collaboration in global software projects.

Previous studies on machine translation (e.g., Yamashita and Ishida 2006; Yamashita et al. 2009; Gao et al. 2013) have employed ad-hoc experimental tasks, such as puzzle solving, often in one-to-one chat sessions. On the contrary, in our studies we used as experimental tasks requirements prioritization and planning, two technical and complex activities that capture all the subtleties and challenges of real-world group communication in global software engineering. In particular, our objective evaluation focused on (i) assessing real-time machine translation services in terms of translation quality and time performance, as well as (ii) evaluating the impact of machine translation technology on group interaction.

Our findings, discussed in the remainder of this section, add to the evidence about the recent advances of machine translation technology and provide some guidance to global software engineering practitioners in regarding the losses and gains of using English as a *lingua franca* in multilingual group communication, as in the case of computer-mediated requirements meetings.

5.1 RQ1 – Machine Translation is a Viable Alternative to Using English as a Lingua Franca

In order to answer the first research question, we needed to assess whether state-of-the-art machine translation technology available today can be used instead of English to perform distributed, multilingual requirements meetings. Therefore, as a first step we have run a simulation study in which we compared two real-time machine translation services, that is, Google Translate and Apertium. In particular, we compared their performance in terms of both effectiveness, i.e., adequacy of translation produced, and efficiency, i.e., the amount of time taken to perform translations.

As far as translation quality is concerned, we employed four raters to assess the adequacy of the output produced by the two services. The use of raters to evaluate translations according to some scale is the standard approach in the field (Arnold et al. 1994), although recently research has started to focus on the identification of predictive quality measures that would allow MT systems to automatically assess their output (e.g., see Wisniewski et al. 2012). The raters first assigned each translation to categories 1–4 (1=completely adequate, 4=completely inadequate). Then, we evaluated the inter-rater agreement by computing multiple Kappa index, which was measured 0.36 for Apertium and 0.46 for Google Translate. One plausible explanation of the fair to moderate Kappa values measured is the employment of non-bilingual raters for the evaluation of translations quality. In fact, we recruited four Italian master students who are not native English speakers, although knowledgeable in software engineering and thus fully able to understand the context of conversations. Hence, possible disparities in raters' English skills may account for the moderate agreement levels achieved.

Overall, we found Google Translate to provide significantly more adequate translations than Apertium. In fact, on the average, the translation quality for Google Translate is 2.17 (median 2.0), with over the 63 % of translations falling into category 1 or 2, that is, judged completely or partially adequate by the raters. Conversely, for Apertium the average translation quality is 2.8 (median 3.5), with most of the translation produced (~63 %) falling in category 3 or 4, that is, judged partially or completely inadequate by the raters. Besides, we analyzed how the presence of errors (e.g., both spelling and grammatical) affects the adequacy of the resulting translation. For both translation services, we found errors in the original sentences to have no significant impact on translation quality.

We identified a couple of reasons why Apertium achieved lower adequacy ratings than Google. The first reason is the low quality of the translation rules defined in the English-Italian pair, which is still in an early stage of development. We tried to address this issue using one of the five chat logs available – the one not used by raters for the evaluation – to find and add the missing linguistic knowledge to the EN-IT pair[9] in the Apertium platform by updating morphological and bilingual dictionaries and translation rules. Nevertheless, adequacy ratings were not only lower than those produced by Google Translate, but also worse than those obtained using Apertium with the full-fledged English-Spanish pair, which we evaluated informally. The second one is that Google service was better able to cope with the colloquialism typical of text-based chats (e.g., short and slang forms, such as "*hes*" instead of "*he's*", "*dont*" instead of "*don't*"), which conversely Apertium proved not to manage as well. As per Google Translate performance, the result of our simulation (~63 % of adequate translations) is in line with other services. For example, Rautio and Koponen (2013) report the performance of MOLTO[10] machine translation service during a simulation, the setting of which is very similar to ours. They found that, on average, 61.8 % of the sentences were judged adequate by human raters. In addition, our test set comprises sentences that use computer science jargon, thus making translation even harder. Finally, despite achieving better adequacy results than Apertium, Google Translates still suffers from at least a couple of drawbacks. The first one is due to the statistical approach used, which prevents Google Translate from being improved, as in the case of rule-based systems, to which specific domain knowledge can be added in form of new dictionaries and translation rules. The second drawback is a limitation that raises privacy concerns. When using Google Translate, one cannot install the machine translation service on a company's private server, meaning that private data must be sent to Google servers for being translated.

With respect to time performance, we found good time responses for both services, which also proved to scale up well as the length of sentences and the number of clients – i.e., concurrent requests – increase (see Figs. 5 and 6, respectively). However, the time performance of Apertium (less than 30 ms in the worst case) is better than Google (around 70 ms in the worst case). This difference in speed is observed probably because Google Translate service is publicly available and only reachable through Internet (i.e., other people might be using it at the same time of our tests). As such, the load of requests to the well-known Google Translate service was reasonably much higher than that served by our installation of the Apertium, which was run instead as a private service and accessed through our university LAN, thus making negligible the queuing and propagation delays. Still, albeit a hosted service, Google Translated response times are still adequate. Besides, we also found Google Translate performance to be rather stable, whereas the response times of Apertium showed a tendency to increases with the length of the input sentence.

The findings about efficiency are no less important than those about adequacy are. In fact, a hypothetical, 100 % accurate machine translation system could not be used in real time communication system if it took several seconds to complete translations because the extra delays would break the communication flow. Instead, we found that, the delays introduced by the automatic translation with both machine translation systems (ranging between ~5-70 ms) is negligible and does not disrupt real time interaction.

Furthermore, in the controlled experiment we compared the interaction of subjects who carried out a couple of complex tasks communicating in either English or their native languages, assisted by machine translation. We reported in Table 6 some measures computed to describe the meetings executed in the experiment. Such measures show that the frequency of message (measured by utterance per minute rate – upm) in the case of English meetings (mean 5.86) was almost identical to that of machine translation-enabled meetings (mean 5.93 s.). This finding is also confirmed by the average delay between two consecutive utterances (i.e., 11.8 s.

[9] SVN revision 19832

[10] www.molto-project.eu

for English and 0.7 s. for machine translation), a measure that is correlated to message frequency, since faster interaction means lower delay. In other words, low proficiency participants spent a little extra time when using the non-native language. Such difference seems to be mostly due to those meetings involving low proficient subjects and English communication mode, in the case of which the measured average delay is higher, namely 12.13 s. Unfortunately, we are unable to distinguish whether such extra time was spent by subjects deliberating on the meaning of incoming messages, reviewing their own messages before sending, or, most likely, a performing both activities.

From a practical standpoint, these results indicate that state-of-the-art machine translation technology is not disruptive of the conversation flow, even during the execution of complex group tasks, such as distributed requirements engineering. As such, machine translation technology available today is a viable alternative to using English as a lingua franca.

5.2 RQ2 – Machine Translation Ensures a More Balanced Discussion While Decreasing Efficiency

The main purpose of running the controlled experiment was to answer the second research question concerning how the adoption of machine translation affect group interaction in distributed, multilingual requirements meetings, as compared to the use of English. Therefore, we designed a controlled experiment during which 16 multilingual groups of subjects collaborated using a machine translation service for two-way translations in order to complete a Planning Game activity. Then, we analyzed the data collected, looking at how the use of machine translation affects group interaction in terms of (i) equality of participation and (ii) reaching an understanding.

In order to assess the equality of participation during the experimental runs, we checked whether any member dominated group discussions. Like group/social pressure, domination is just one of the dysfunctional aspects that are intrinsic to group communication in general and even harder in requirements group approaches, which do require the contributions of every party involved to be successful (Macaulay 1996; Gottesdiener 2002; Calefato et al. 2012b). Thus, we first computed the percentage of utterances presented per participant. Then, we calculated the delta between the most and least prolific subjects. From our findings, we could observe that a more balanced discussion tend to occur when native language with machine translation technology was used. In other words, we observed a decrease in the deltas in 10 out of 16 groups (~63 %). Besides, the increase of participation (i.e., smaller delta) of the least proficient subject in a group generally occurred at the expense of the most prolific participant. Finally, we also observed that in 50 % of the cases, the least proficient subjects increased the number of utterances contributed after switching from English to native language with machine translation (we further discuss the implication of English proficiency level in the next section). This finding is a very relevant contribution of our work because it shows that the use of machine translation technology, although still not 100 % accurate, is already capable of limiting domination, a well-known challenge in achieving efficient group communication.

The second evaluation of the effects of machine translation on group interaction concerned the detection of differences in reaching an understanding as compared to English. More specifically, we looked for evidence of misunderstandings or, lack of common ground, which was operationalized in terms of *clarification requests.* Specifically, our goal was to understand whether such clarification requests spawned more because of translations inaccuracies rather than poor English or vice versa. Therefore, we opportunistically performed content analysis on logs from those groups whose participants reported difficulties in understanding each other. Content analysis identified about 3–4 % of utterances from machine translation-enabled runs that could not be coded due to inadequate translations. Although partial, the data analyzed suggests that the inaccuracies of state-of-the-art machine translation technology may impair the development of shared understanding more than low English skills. In addition, a percentage as high as the 4 % of unclassifiable utterances 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. Besides, 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.

To the best of our knowledge, an acceptable error rate for effective automatic translation is yet to be determined. Besides, we would expect such rate to vary, depending on the domain and the criticality of the task to execute. However, our findings suggest that, at least for group approaches in the requirements engineering domain, an error rate below 5 % is adequate to ensure task completion (efficacy) at the expense of some interaction speed (efficiency).

5.3 RQ3 – No Evidence that Machine Translation is more Beneficial to Individuals with low English Proficiency

In our empirical study, we also wanted to assess whether individuals with a low English proficiency level would benefit from using their native language more than those with a high English proficiency level (RQ3). Therefore, in order to answer this question, we conceptualized the extent to which subjects appreciated the experimental tasks in terms of the *engagement and comfort with communication* and *satisfaction with task performance*.

We analyzed the data collected from the post-task questionnaires (see Section 4.4.2), first grouping them by respondents' English proficiency level. We expected that subjects with lower proficiency in English would significantly appreciate machine translation-enabled meetings more than meetings in English because a better command of a language provides better opportunities for proper communication. However, our findings show that English communication skills did not affect subjects' appreciation of experimental tasks in terms of either *engagement and comfort with communication* or *satisfaction with task performance*.

Furthermore, during the experiment, participants carried out two different experimental tasks. In fact, T1 was a requirements prioritization task, whereas T2 was about release planning. Hence, since we cannot exclude that such differences might have influenced subjects' perception of task execution, we reanalyzed the data collected from the questionnaires, this time grouping them also by task. With respect to *satisfaction with task performance*, we found that low proficiency participants perceived a better performance in T1 when using English. This result is surprising, also considering that lowly proficient participants were found to be significantly more prone to use machine translation for future multilingual group tasks than those with high English skills (see Table 10). One possible explanation of this finding might be related to the different complexity of the two experimental tasks. Compared to release planning (T2), requirement prioritization (T1) is a less complex task in the sense that it was less time-consuming, by design, and less ambiguous, since it concerned expressing personal preferences about phone features from a customer's perspective (Lehtola et al. 2004). Besides, T2 involves larger amounts of negotiation and decision-making than T1 to meet all the constraints in terms of story points distribution (1000 points over 16 features) and releases planning (the amount of story-points per release within the given range) (Cohn 2005). Therefore, in the case of a simpler task, it appears that even participants with lower English skills perceive the tradeoff between inaccurate translations and the comfort of communicating in the native language not to be beneficial. In fact, looking at the logs from the requirements prioritization (T1) runs, we noticed that all the lowly proficient groups except one (Gr7) followed the same approach: first one participants came up with his or her prioritized list of features, then the others suggested emendations. Therefore, following this approach, they had no need to formulate complex sentences in English and, consequently, no particular need for machine translation. Finally, with respect to *engagement and comfort with communication*, our analysis revealed that participants with better English skills found easier to interact during T1 when using English instead of machine translation. Such result is not surprising as the previous one because machine translation technology is still far from perfect. As such, one could expect that participants fluent in English would feel frustrated while repairing inaccurate translations with extra utterances. In other words, this result confirms that, due to the flaws of current machine translation technology, for individuals with good communication skill in a foreign language the costs of recovering from inaccurate translations outweigh the benefits of having better command in their native language. Besides, together, these two findings suggest that the less complex the task, the lower the perceived usefulness of machine translation.

In conclusion, our results suggest that, overall, state-of-the-art machine translation technology is accepted with favor but it is also perceived to be no more beneficial to individuals with low English proficiency than it is to people with high skills in a foreign language, especially for the execution of less complex tasks.

6 Threats to Validity

One of the key issues in experimentation is evaluating the validity of results (Wohlin et al. 2000). In this section, we discuss the potential threats that are relevant for our findings and how we addressed them.

6.1 Construct Validity

Construct validity concerns the degree of accuracy to which the variables defined in the study measure the constructs of interests. We identified the following threats in our study.

The constructs of *satisfaction with task performance* and *engagement and comfort with communication* have been adapted from a previous study on media effects in requirements meetings (Calefato et al. 2012a). The measures of the two constructs are based on self-reported questionnaire items as opposed to objectively measured ones. As such, they might have been influenced by subjective perception of subjects during tasks execution. We can only acknowledge this threat since the outcomes of the experimental tasks are not unique and cannot be evaluated against an optimal solution – i.e., all groups in the experiment reaching an agreement and delivering a release plan. However, because they are intrinsically subjective, there are no objective measures to capture satisfaction, engagement and comfort levels. Furthermore, our questionnaire analysis show that our results are reliable. In fact, to ensure the validity of the constructs, we verified that questions actually loaded on the two factors as intended by performing a principal component analysis with varimax rotation. Then, we also performed scale reliability analysis to determine further the construct validity by assessing the extent to which a set of questions measures a single latent variable. For this purpose, we computed the Cronbach's alpha coefficient, which represents the most-widely used index of internal consistency in social science (Cronbach 1951). The alpha indexes for the scale in the post-T1 and the post-T2 questionnaires were 0.68 and 0.75, respectively. The values are, respectively, very close to and even above the threshold of 0.70 suggested to affirm scale reliability (Nunnally 1978).

With regard to the construct of the lack of common ground, because the measure of *clarification requests* was obtained from the content analysis of the meeting logs, we mitigate the threat to construct validity by using two independent coders and measuring the intercoder agreement by Cohen's kappa index (Cohen 1960). The computed indexes are k=0.88 and k=0.91, meaning almost perfect agreement between the raters.

6.2 Internal Validity

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

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 machine translation and English 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 the study there are two different experimental tasks, we cannot exclude that task complexity could have been a confounding factor, since subjects experience a communication mode (i.e., EN or MT) 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. Yet, we further controlled it by design, restricting the level of groups to high and low proficiency and assigning participants to groups accordingly.

6.3 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 in our study.

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 (Höst et al. 2000). 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 nowadays commonly adopted in the workplace, not only in the field of software development, to complement email (Herbsleb et al. 2002).

Furthermore, also the requirements definition tasks used in this experiment may not be representative of industrial practice. However, unlike many other experiments in the field of machine translation, we did not use generic, puzzle-like tasks. Instead, although simulated, we designed our requirements prioritization and planning tasks to be as close as possible to their real-world counterparts, so that a high cognitive load and a realistic effort were required for accomplishing them.

6.4 Conclusion Validity

Conclusion validity is concerned with the relationship between the treatment and the outcome. We identified the following threats to conclusion validity in our study.

When the experimental unit is at the team level, small sample sizes are a known problem that is difficult to overcome, especially for cross-country controlled experiments with participants interacting from different time zones, as in our case. As such, we can only acknowledge that the small sample of experimental groups (16) represents a useful, yet less than ideal circumstance in which we furthered our understanding in the field of machine translation applied to requirement engineering, and replications of our study in settings with more resources available are encouraged. Nevertheless, to mitigate this threat, where necessary we used nonparametric tests in the statistical analysis because they do not rely on any assumed distribution of the underlying data and can be valid for even a small sample size. Besides, this threat does not apply to the other statistical analyses conducted when the experimental unit is at utterance level or subject level. In fact, we run our simulation on a text corpus consisting of over 2000 utterances and the controlled experiment involved 64 subjects.

7 Conclusions

We reported the findings from the empirical investigation of real-time machine translation as an help for distributed multilingual meetings in global software projects, with a focus on requirements meetings as an example of communication-intensive activities in software engineering.

Our findings indicate that state-of-the-art machine translation technology is already a viable solution for multilingual group communication since it is not disruptive of the conversation flow, it does not prevent group to complete complex tasks, and it even grants discussions that are more balanced. Yet, machine translation technology currently available is still far from 100 % accurate and, as such, its adoption comes with costs. In fact, translations inaccuracies needs to be repaired by rephrasing the original content, thus causing a decrease in efficiency. Finally, our findings challenge the expectation that machine translation would be more beneficial to individuals with lower English skills. In fact, our study showed that individuals with lower English skills perceive group performance to be poorer when using machine translation, probably because of these inaccuracy-and-repairs diversions from the regular communication flow. Nonetheless, the same individuals are significantly more prone to use machine translation for future multilingual group tasks than those with high English skills.

A common limitation of previous studies is the employment of experimental tasks like picture description or idea exchange, often in one to one chat. Settings like these are likely to miss out some the facets and subtleties occurring in complex software engineering tasks. The use of comunication-intensive, real world tasks is one of the key contribution of our work to the field machine translation.

Although we focused on text-based machine translation technology only, future research might use this work as a basis for investigating the use of real-time speech translation in similar settings. In fact, we would expect a bigger impact when speech is involved because, when hearing, participants have less time to deliberate on the supposed meaning of foreign words and sentences. We are aware that speech recognition is not perfect either. This suggests that future work should carefully consider the propagation of errors from inaccurate transcriptions through resulting translations and how their combined effect would harm comprehension and task performance.

Acknowledgments This research is partially funded by the Rio Grande do Sul State funding agency (FAPERGS) under projects 11/2022-3 and 002061-2551/13, and by CNPq (483125/2010-5, and 309000/2012-2). We would like to thank all the participants who took part in the experiment and Pasquale Minervini for the support he provided with the management of the Apertium service.

Appendix: Experimental Data

Table 13 Categories of the coding schema (thematic units) in content analysis

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!”).
Unknown		Machine translation is poor and the general meaning of the sentence so hard to get that it is not possible to identify the appropriate category.

References

Altman DG (1991) Practical statistics for medical research. Chapman and Hall, London

Arnold D (2003) Why translation is difficult for computers. In Computers and translation: a translator's guide. Benjamins Translation Library

Arnold D, Balkan L, Meijer S, Humphreys RL, Sadler L (1994) Machine translation: an introductory guide. NCC Blackwell, London

Aziz W, Sousa SCM, Specia L (2012) PET: a tool for post-editing and assessing machine translation. Proc 8th Int'l Conf Lang Resour Eval (LREC'12):3982–3987

Berander P (2004) Using students as subjects in requirements prioritization. Int'l Symp Empir Softw Eng (ISESE'04):167–176. doi:10.1109/ISESE.2004.34

Brazil IT-BPO Book 2010–2011 (2013) Published by Brasscom, Brazilian Association of Information Technology and Communication Companies, São Paulo, Brazil

Burchardt A, Tscherwinka C, Eleftherios A, Uszkoreit H (2013) Machine translation at work. Computational Linguistics, Studies in Computational Intelligence, Springer, 458:241–261. doi:10.1007/978-3-642-34399-5_13

Calefato F, Lanubile F (2007) Using frameworks to develop a distributed conferencing system: an experience report. Softw Pract Exper 39(15):1293–1311. doi:10.1002/spe.937

Calefato F, Damian D, Lanubile F (2007) An empirical investigation on text-based communication in distributed requirements engineering. Proc 2nd Int'l Conf Glob Softw Eng (ICGSE'07):27–30. doi:10.1109/ICGSE.2007.9

Calefato F, Lanubile F, Minervini P (2010) Can real-time machine translation overcome language barriers in distributed requirements engineering? Proc 5th Int'l Conf Glob Softw Eng (ICGSE'10):257–264. doi:10.1109/ICGSE.2010.37

Calefato F, Lanubile F, Prikladnicki R (2011) A controlled experiment on the effects of machine translation in multilingual requirements meetings. Proc 6th Int'l Conf Glob Softw Eng (ICGSE'11):94–102. doi:10.1109/ICGSE.2011.14

Calefato F, Damian D, Lanubile F (2012a) Computer-mediated communication to support distributed requirements elicitations and negotiations tasks. Empir Softw Eng J 17(6):640–674. doi:10.1007/s10664-011-9179-3

Calefato F, Lanubile F, Conte T, Prikladnicki R (2012b) Assessing the impact of real-time machine translation on requirements meetings: a replicated experiment. Proc of the 6th Int'l Symp Empir Softw Eng and Meas (ESEM'12):251–260. doi:10.1145/2372251.2372299

Carmel E, Agarwal R (2001) Tactical approaches for alleviating distance in global software development. IEEE Softw 18(2):22–29. doi:10.1109/52.914734

Clark HH, Brennan SE (1991) Grounding in communication, in perspectives on socially shared cognition. American Psychological Association, Washington DC, pp 127–149

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. doi:10.1177/001316446002000104

Cohen J (1992) A power primer. Psychol Bull 112(1):155–159. doi:10.1037/0033-2909.112.1.155

Cohn M (2005) Agile estimating and planning. Prentice Hall

Conover WJ (1980) Practical nonparametric statistics. Wiley, New York

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334

Damian D (2007) Stakeholders in global requirements engineering: lessons learned from practice. IEEE Softw 24(2):21–27. doi:10.1109/MS.2007.55

Damian D, Zowghi D (2003) Requirements engineering challenges in multi-site software development organizations. Requir Eng J 8(3):149–160. doi:10.1007/s00766-003-0173-1

DeSanctis D, Gallupe RB (1987) A foundation for the study of group decision support systems. Manag Sci 33(5): 589–609. doi:10.1287/mnsc.33.5.589

Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York, pp 38–46

Gao G, Wang H-C, Cosley D, Fussell SR (2013) Same translation but different experience: the effects of highlighting on machine-translated conversations. In Proc SIGCHI Conf Hum Factors Comput Syst (CHI '13):449–458. doi:10.1145/2470654.2470719

Garland R (1991) The mid-point on a rating scale: is it desirable? Mark Bull 2:66–70

Gottesdiener E (2002) Requirements by collaboration: workshops for defining needs. Addison-Wesley Longman Publishing Co., Inc.

Herbsleb JD, Atkins DL, Boyer DG, Handel M, Finholt TA (2002) Introducing Instant messaging and chat into the workplace. Proc Int'l Conf Comput Hum Interact (CHI '02):171–178. doi: 10.1145/503376.503408

Höst M, Regnell B, Wohlin C (2000) Using students as subjects - a comparative study of students and professionals in lead-time impact assessment. Empir Softw Eng 5(3):201–214. doi:10.1023/A:1026586415054

Hsieh Y (2006) Culture and shared understanding in distributed requirements engineering. 1st Int’l Conf Glob Softw Eng (ICGSE’06):101–108. doi:10.1109/ICGSE.2006.261221

Johns R (2005) One size doesn’t fit all: selecting response scales for attitude items. J Elections Public Opin Parties 15(2):237–264. doi:10.1080/13689880500178849

Jurafsky D, Martin JH (2008) Speech and language processing 2nd ed. Prentice Hall Series in Artificial Intelligence, Prentice Hall

Kearney AT (2007) Destination Latin America: a nearshore alternative, technical report

KPMG (2009) Nearshore attraction: Latin America Beckons as a global outsourcing destination. Technical Report

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174. doi:10.2307/2529310

Lehtola L, Kauppinen M, Kujala S (2004) Requirements prioritization challenges in practice. LNCS: Product Focused Software Process Improvement 3009:497–508. doi:10.1007/978-3-540-24659-6_36

Lutz B (2009) Linguistic challenges in global software development: lessons learned in an International SW Development Division. Proc 4th Int’l Conf Glob Softw Eng (ICGSE’09):249–253. doi:10.1109/ICGSE.2009.33

Macaulay LA (1996) Requirements engineering, Springer-Verlag Telos

Mitkov R (2003) The Oxford handbook of computational linguistics. Oxford Handbooks in Linguistics, Oxford University Press

Montgomery DC (1996) Design and analysis of experiments. Wiley, New York

Mudge RS (2009) After the deadline – language checking technology. Automattic. http://open.afterthedeadline.com

Nunnally JC (1978) Psychometric theory (2nd ed.). McGraw-Hill, New York

Ogden W, Zacharski Z, An S, Ishikawa Y (2009) User choice as an evaluation metric for web translation in cross language instant messaging applications. Proc Mach Transl Summit VII

Paulson LD (2001) Translation technology tries to hurdle the language barrier. Computer 34(9):12–15. doi:10.1109/MC.2001.947080

Prikladnicki R, Carmel E (2013) Is time zone proximity an advantage for software development? The case of the Brazilian I.T. industry. Int’l Conf Softw Eng (ICSE’13):973–981. doi:10.1109/ICSE.2013.6606647

Rautio, J, Koponen, M (2013) “MOLTO evaluation and assessment report.” Technical Report

Raybaud S, Langlois D, Smaïli K (2011) ‘This sentence is wrong’. Detecting errors in machine-translated sentences. Mach Transl 25(11):1–35. doi:10.1007/s10590-011-9094-9

Shah YH, Raza M, UlHaq S (2012) Communication issues in GSD. Int’l J Adv Sci Technol 40:69–76

Stemler S (2001) An overview of content analysis. Practical assessment, research & evaluation, 7(17)

Wisniewski G, Kumar Singh A, Yvon F (2012) Quality estimation for machine translation: some lessons learned. Mach Transl 27(3–4):213–238. doi:10.1007/s10590-013-9141-9

Wohlin C, Runesson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2000) Experimentation in software engineering, an introduction. Kluwer Academic Publishers

Yamashita N, Ishida T (2006) Effects of machine translation on collaborative work. Proc 20th Int’l Conf Comput Supported Coop Work (CSCW ‘06):515–524. doi:10.1145/1180875.1180955

Yamashita N, Inaba R, Kuzuoka H, Ishida T (2009) Difficulties in establishing common ground in multiparty groups using machine translation. Proc 27th Int’l Conf Hum Factors Comput Syst (CHI ‘09):679–688. doi:10.1145/1518701.1518807