ISSN 2277-3061
Volume 15 Number 10
International Journal of Computer and Technology

An Analysis of Comment-Revision Thresholds in Bilingual Electronic Meetings

Bart Garner, Milam Aiken
School of Business Administration, University of Mississippi, University, MS 38677
bgarner@bus.olemiss.edu
maiken@bus.olemiss.edu

ABSTRACT

Prior studies have shown that providing participants in bilingual or multilingual, electronic meetings with the capability of revising comments can increase the accuracy of translations to other languages. This is often done via a round-trip translation (RTT) in which the source text is translated to another language, translated back again, and compared with the original. If the similarity falls below a certain threshold, e.g. 50%, the originator may wish to revise the comment before final submission. However, minor changes might not be needed, and it is not clear where the threshold lies between acceptable and non-acceptable text. In this study, we seek to determine at what point accuracy can be improved by comment revision. Results show that the threshold did not affect the target-language comprehension, but higher thresholds substantially increased the cognitive burden for users in the form of alerts and comment revisions.

Indexing terms/Keywords

Electronic meetings, Group support systems, Machine translation, Bilingual groups

Academic Discipline And Sub-Disciplines

Computer Science (machine translation)

TYPE (METHOD/APPROACH)

Lab Experiment

INTRODUCTION

The United States Bureau of Labor Statistics suggests demand for human interpreters in the United States will grow by 29% between 2014 and 2024 due to increased globalization and immigration. The result is more people are increasingly using free, online, translation services for communication because human interpreters are expensive or unavailable for immediate service. However, automatic machine translation continues to suffer from poor accuracy in many cases and must be used with caution. For example, one study [16] reports on the use of *Google Translate* to provide medical information to patients because no human interpreter was available for the language required, and a later analysis revealed that the communication was only about 57% accurate. This could be critical when life-or-death decisions are being made.

On the other hand, translation accuracy in an informal bilingual or multilingual electronic meeting might not be as important. Further, as more languages are added to a meeting, it becomes more difficult to obtain human interpreters conversant in the foreign tongues. Finally, as all group members are typing simultaneously, it becomes nearly impossible for a human, or group of humans, to provide concurrent translation [7]. Thus, automated translation becomes necessary.

Although accuracy can be poor with Web-based translation services, improvements can sometimes be made by the user. For example, if something is not understood in a conversation, people often try to express their thought using different words. In an electronic meeting, typing errors, acronyms, slang, poor grammar, and idioms can adversely affect the accuracy [4, 14, 15]. If the originator of a comment could have some indication of the translation accuracy before it is submitted to the group, he or she could make revisions. Round-trip-translation (RTT) is one technique that could provide this evaluation capability.

Using RTT, text is translated to another language in a forward translation (FT), and then the results of this FT are translated back into the original language in a backward translation (BT). Combined, they form an RTT. If there are differences between the backward translation and the original comment, the content of the message might be misunderstood, as several studies have found significant, positive correlations between FT and BT accuracies (e.g., [2, 5, 6, 10, 19, 20]).

If group members could be alerted to the fact that their comments might not be translated accurately, they might revise the original text to make it more understandable, e.g., if the similarity between the original and back translation falls below a certain threshold. However, it is not clear at what limit revision benefits the conversation. This paper describes an experiment with 14 groups in bilingual meetings with automatic translation using thresholds ranging from 0 to 100.

LITERATURE REVIEW

Early multilingual, electronic meeting systems translated among different languages, but it was not easy to determine if all group members understood the text that was exchanged [3]. For example, one study [23] found that 1 out of 20 translated comments in an electronic meeting included misconceptions, while the comment originators did not know they were misunderstood.

In an attempt to correct these translation errors, *Amikai's AmiChat* introduced the capability for group members to indicate whether or not a comment was understood [12]. By clicking an onscreen button, users could send a message to the originator of the text that a particular passage was not clear. However, some users might not want to make the effort or might be too shy to admit miscomprehension.

Instead of relying on the reader to alert the originator of possible muddled text, another electronic meeting system called *AnnoChat* was developed using RTT for error detection [24]. The back translation was presented to the originator for perusal, and the author could determine if changes might need to be made before final submission. Using this system, Japanese users were able to improve translations to English with RTT, but they were not able to improve translations from Japanese to Chinese and Korean [13]. When more than two languages are used in a meeting, changing a comment could increase the accuracy of a translation into one language, but it might decrease the translation accuracy to others.

In another study [9], 22 students used automatic translation in a bilingual (English-German) meeting that provided an RTT on every comment, while another group of 18 students used the system without automatic comment evaluation. With the first group, an arbitrary threshold of 80% was selected for notification. That is, if the reverse translation words were less than 80% similar, the originator of the comment was alerted and given a chance to revise the text before final submission. In this way, group members would not need to review a back translation for every comment written, thus saving them time and perhaps increasing satisfaction. Results showed that evaluation and a chance for comment revision improved comprehension of the translated text from English to German from 83.12% to 87.69%, but the improvement was not statistically significant. However, a further analysis revealed that if the threshold had been changed to 50%, there would have been significant improvement. That is, modifying only the most garbled translations yielded meaningful differences in comprehension.

EXPERIMENTAL STUDY

Purpose

The purpose of this study was to further explore RTT thresholds to determine at what point revisions improve translation accuracy.

Subjects and Task Description

A total of 126 business students from a large public university in the southern United States participated in 14 meetings with sizes ranging from 4 to 11 to discuss (in English) for about 10 minutes the parking problem on campus, a topic that has been used frequently in electronic meeting research (e.g., [21]) and a time determined to be sufficient for a full discussion [22]. The students had knowledge of the topic and were motivated to find a solution, but had no ultimate control over the outcome.

Seven similarity thresholds were used (0, 20, 40, 50, 60, 80, and 100) with 0 indicating no back translation was shown and no revision was possible because no percentage would be less than 0%. On the other hand, groups using a threshold of 100 saw every comment's back translation, as the similarity score never rose above this. Similarly, groups using the limit of 50 saw back translations only if the similarity score was 50% or below.

All of the students spoke English fluently and few knew any other language. During the meetings, the group facilitator added pre-written German comments that were automatically translated to English, thus, simulating a bilingual session. A German speaker evaluated the students' comment translations to German after the experiment. Group members answered a short survey after each meeting assessing how useful and easy to use the system was and whether or not they could detect comments translated from German.

Meeting Software

In some earlier studies of bilingual electronic meetings, reverse translations were shown for all comments, possibly annoying group members and slowing down the meeting. In the latest study, one arbitrary threshold was selected to discriminate among the translations so that only the worst translations were captured, alleviating some of the annoyance. However, results showed that this threshold was probably too high. In addition, the software provided details on the RTT on a separate screen, requiring the meeting participant to switch between pages, further hindering the process.

We used a multilingual electronic meeting system linked with *Google Translate*, an online service that currently supports 103 languages [18]. Instead of showing the RTT on a separate screen, this system shows the back translation at the bottom on the current screen, only if it falls below a certain threshold set by the meeting facilitator, as shown in Figure 1. The meeting participant simply types a comment in the textbox at the top of the screen using the selected language (English), and then clicks on a button to submit it. In this case, the system translates the comment into a target foreign language (German) and then translates this back to English, showing it along with the similarity percentage if it falls below the predetermined limit. If the author decides that the comment is acceptable, despite the low score, he or she can click "Submit as is" and the text is available to the group (in English and German). Otherwise, the user clicks "Revise comment," and the focus is placed back in the textbox

In Figure 1, the text was translated to German as "Jeder sollte seinen eigenen Parkplatz" or "Everyone should his own parking spot." In German, as in English, verbs are sometimes omitted during informal conversations. In this case, "haben" or "have" was omitted. As a result, the back translation also omitted the "have." The poor back translation is grammatically correct but confusing. However, the forward translation can still be understood. Thus, a BT might be perfect, while the FT is misunderstood, and another BT might be inaccurate, while the FT is still comprehensible.

Results

Table 1 shows a summary of the experimental results. Students believed the meeting system was useful and easy to use, and there was no significant difference among the group thresholds in terms of these two variables ($F = 0.23$, $p = 0.96$ and $F = 0.75$, $p = 0.61$, respectively). However, there was a significant difference in the number of comments written per person in the same amount of time ($F = 3.49$, $p < 0.01$). Those in the 0% threshold groups wrote only 1.5 comments per person in comparison with the overall average of 3.95, even though they were not making revisions.

Students generally were not able to recognize those comments translated from German, indicating that perhaps the translations were grammatically correct and comprehensible. Self-assessed comprehension scores of the discussions ranged from 96.1% to 99.5%, significantly above the 72.45% threshold required by many American graduate schools for admission [8] ($t = 55.7$, $p < 0.01$). There was also no significant difference among the group thresholds in terms of German recognition ($F = 0.86$, $p = 0.53$) and comprehension ($F = 0.87$, $p = 0.52$).

There was a significant correlation between comprehension and perceptions of ease of use ($R = 0.265$, $p = 0.01$) and usefulness ($R = 0.30$, $p < 0.01$), and those who found it easy to use also found it useful ($R = 0.44$, $p < 0.01$). None of the other variables showed significant correlations.

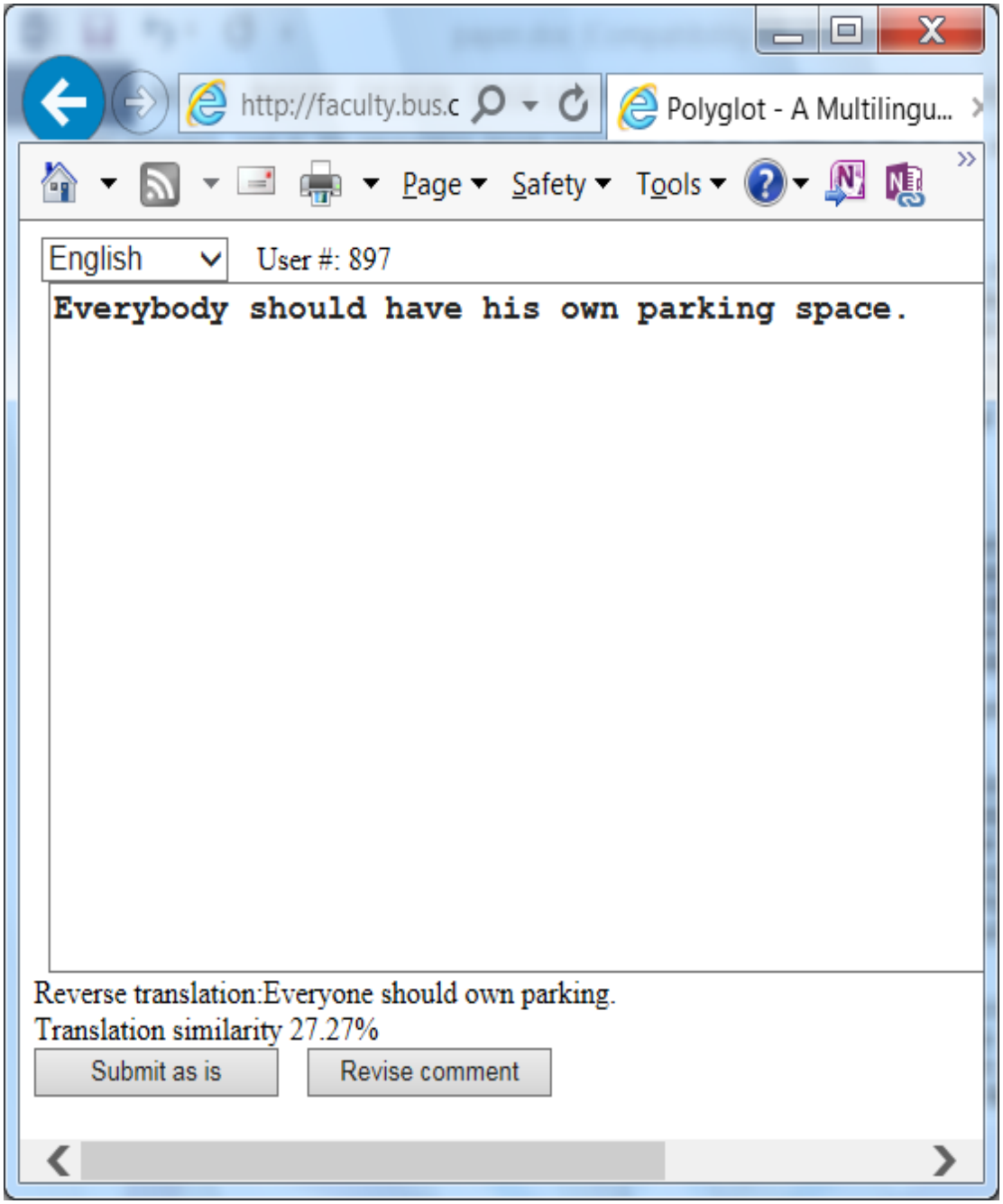

Figure 1. Round-trip translation comment error detection

Table 1. Overall summary of variables

Threshold	% Understood	German recognized	Easy-to-Use	Useful	Comments / Person
Overall					
Mean	98.56	2.24*	6.06*	5.60*	3.95
Std Dev	4.31	1.66	1.04	1.25	2.84
0%					
Mean	99.5	2.90*	5.90*	5.50*	*1.5*
Std Dev	1.5	2.02	1.3	1.36	0.67
20%					
Mean	99.33	2.33*	6.00*	5.40*	3.2
Std Dev	2.49	1.58	0.97	1.45	2.1
40%					
Mean	99.41	2.59*	6.24*	5.82*	3.88
Std Dev	2.35	1.72	1	0.98	2.47
50%					
Mean	98.75	2.13*	5.63*	5.50*	4
Std Dev	3.31	1.45	1.11	0.87	1.58
60%					
Mean	98.33	1.67*	6.39*	5.72*	6
Std Dev	5	1.37	0.76	1.15	3.99
80%					
Mean	97.25	2.38*	6.00*	5.38*	4.25
Std Dev	6.53	1.87	0.71	1.58	1.98
100%					
Mean	96.11	1.78*	5.78*	5.67*	3.67
Std Dev	6.57	1.23	1.31	1.25	1.83

* Significantly different from the neutral measure of 4 on the Likert 1 (disagree) – 7 (agree) scale at α = 0.01

Text analysis

Table 2 shows the numbers of comments written by members in each threshold group, the similarity percentage between the original text and the back translation, the numbers of alerts (messages with BTs falling below the threshold), the numbers of revisions, and the percentage of the translations to German understood by the reviewer. The RTT similarity scores were fairly consistent among the groups. As expected, the alert numbers rose steadily with the threshold, but the numbers of revisions did not consistently rise. Those in the 80% threshold groups revised 12.6% of the alerted comments, but those in the 60% threshold groups revised 39.0% of the alerted comments. The reviewer understood 84.8% of the text, significantly above the 72.45% minimum required ($t = 1.8$, $p = 0.04$), but significantly below the 98.6% comprehension of the English-speaking group members ($t = -2.2$, $p = 0.02$).

Several examples from the transcripts illustrate problems that occurred when English was translated to German:

- "When there are only few slot left, its only up to your luck whether you will get the slot or not." → "Wenn es gibt nur wenige Slot links, es ist nur bis zu Ihrem Glück, ob Sie den Schlitz erhalten wird oder nicht." The word "left" as in "remaining" was translated to "links" as in direction. However, the grammatical error "its" did not adversely affect the translation.
- "Knock down the football stadium and use that space to build additional parking. Roll Tide." → "Klopfen Sie das Fußballstadion nach unten und nutzen diesen Raum zusätzliche Parkplätze zu bauen. Roll Tide." "Roll Tide" is a rallying cry of a football team in the United States. Although most students in the southern part of the USA probably understood what was meant, a German speaker from another country would probably not recognize it.
- "I've got 99 problems but a parking spot aint one." → "Ich habe 99 Probleme bekam aber einen Parkplatz aint ein." The American vernacular word "ain't" was misspelled, and *Google Translate* simply repeated the word. If the contraction had been spelled correctly, however, the translation would have been perfect.
- "Freshman can't bring their cars on campus." → "Freshman können ihre Autos auf dem Campus zu bringen." The word "Freshman" was translated correctly in some comments, but not here. Also, the sentence translation gives the direct opposite meaning of what was intended.
- "invest in a pair of heeleys" → "investieren in ein Paar heeleys" This was a good translation, but the German reviewer did not realize that "heeleys" are shoes.

The following examples of round-trip translations provide further details of what occurred in the meetings:

- "We need bigger spots." → "Wir brauchen größere Flecken." → "We need more spots." 75% similarity In this case, "more" does not provide the same meaning, and "Flecken" means "dirty spots." The student decided it was good enough, however, and submitted it as is.

- "having a 10 am is practically having a 8am because of parking." → "a 10.00 mit praktisch, weil der Park ein 8.00 mit." → "a 10:00 with convenient because of a Park with 8:00." (37.5%) The German FT is as incomprehensible as the BT, but the student decided to submit it as is.
- "everyone get a bike" → "jeder bekommen ein Fahrrad" → "each get a bicycle" (75%) The student revised the comment to "get a bike" with a 100% similarity score less than a minute later. The German comprehension was also improved.

- "Build another parking garage" → "Bauen Sie ein anderes Parkhaus" → "Build another park" (50%) The student submitted it as is, although clearly not similar to the original. However, the German translation was comprehensible.

Groups using a threshold RTT similarity score of 50% generated comments with the lowest German comprehension (72.1%), but the average comprehension for the groups below increased to an average of 89.1% while the average for the groups above the threshold also rose to 84.7%, on average. Thus, it is not clear where the threshold should be set based solely upon comprehension. However, groups at the 50% threshold and below revised only 1.7% of all comments, on average, while those above revised 10.9%. Just transitioning from the 50% to the 60% threshold increased the burden by over three times (3.7% of all comments to 12.3%). Therefore, a setting of 50% might be optimal based upon comprehension and participant effort, in agreement with what an earlier study [9] predicted.

Table 2. Comment analysis

Threshold	Comments	Overall Similarity %	Alerted % / similarity %	Revised % / similarity %	% understood in German (Mean / Std Dev)
Overall	224	72.80%	31.0% / 51.5%	5.6% / 49.0%	84.8% / 34.4%
0	21	63.80%	0 / NA	0 / NA	81.0% / 39.3%
20	28	79.80%	0 / NA	0 / NA	*93.0% / 25.5%*
40	102	78.50%	15.7% / 28.4%	2.9% / 34.6%	93.4% / 24.8%
50	54	71.70%	16.7% / 40.2%	3.7% / 40.0%	72.1% / 44.8%
60	130	74.40%	31.5% / 46.6%	12.3% / 50.7%	80.6% / 39.5%
80	60	77.20%	53.3% / 58.9%	6.7% / 54.0%	81.7% / 38.7%
100	30	64.00%	100 / 83.6%	13.8% / 65.6%	91.8% / 27.4%

CONCLUSION

Summary

In this study, group members exchanged comments in English with translations to German via a bilingual electronic meeting system. Seven RTT similarity thresholds were used in attempt to find the optimal setting for maximum accuracy, but results were not clear. German comprehension above and below the 50% threshold were both high, but the extra burden of comment revision increased with the threshold. Thus, we conclude 50% might be optimal.

Limitations

The first limitation is that the bilingual meeting was only simulated with German comments entered by the facilitator and translations reviewed afterward. Knowing this, perhaps the participants were not sufficiently motivated to submit understandable comments, especially in the case of the 50% threshold groups. However, it is difficult to recruit sufficient foreign-language-speaking subjects for experiments.

A second limitation is that only English and German were used. Other language combinations, for example, Georgian to Swahili, are likely to produce far less translation accuracy [1].

Third, only one reviewer was used to evaluate the translations from English to German. Different reviewers might have comprehended more or less of the comments.

Future Research

Because of the inconsistent results with comprehension below and above the 50% threshold, further study on an optimal setting is necessary. In addition, an evaluation of the experimental subjects' time and effort during comment revision could yield further insight into providing the most productive multilingual meeting.

In addition, prior studies have focused on RTT with just two languages, but it is much more difficult when many languages are involved in a meeting [17]. For example, as noted in the [13] study, increasing the accuracy on one translation might decrease the accuracy of another. Nevertheless, studies should investigate how translation accuracy can be improved in these multilingual meetings as well.

Finally, the text analysis showed that some users did not revise their comments even when the RTT was poor. This suggests possible poor understanding of the task at hand or a lack of user engagement. Additional work should be done to determine if improved training or increased user engagement improves FT accuracies.

REFERENCES

[1] Aiken, M. and Balan, S. (2011). An analysis of Google Translate accuracy. Translation Journal, 16(2) April.

[2] Aiken, M. and Hazarika, B. (2012). Evaluation and revision of translation-mediated communication in a multilingual electronic meeting. International Journal of Business and Systems Research. 6(4), 379-394.

[3] Aiken, M., Martin, J., Paolillo, J., and Shirani, A. (1994). A group decision support system for multilingual groups. Information & management, 26(3), 155-161.

[4] Aiken, M., Park, M., and Garner, B. (2012). Translation of relevant and irrelevant multilingual group support system comments. International Journal of Intercultural Information Management, 3(1), 2012, 45-58.

[5] Aiken, M. and Park, M. (2010). The efficacy of round-trip translation for MT evaluation. Translation Journal. 14(1).

[6] Aiken, M. and Park, M. (2009). Enhancing bilingual electronic group meeting comprehension with round-trip translations. International Journal of Systems and Change Management. 4(2), 103-116.

[7] Aiken, M., Park, M., and Lindblom, T. (2011). A comparison of oral and electronic bilingual meetings. Proceedings of the 42nd Annual Meeting of the Decision Sciences Institute, Boston, MA, November 19-22, 951-956.

[8] Aiken, M., Park, M., Lindblom, T, and Wee, J. (2011) Multilingual group support system comprehension sufficiency. International Journal of Information and Operations Management Education, 4(2), 146-162.

[9] Aiken, M., Posey, J., and Reithel, B. (2015). Comment evaluation and revision in a bilingual meeting. International Journal of Management & Information Technology, 10(7), 2311-2317.

[10] Chan, S. (2006). A Dictionary of Translation Technology. Hong Kong: Chinese University Press.

[11] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221-233.

[12] Flournoy, R. and Callison-Burch, C. (2000). Reconciling user expectations and translation technology to create a useful real-world application. Proceedings of the 22nd International Conference on Translating and the Computer. 16–17 November, London, UK.

[13] Ogura, K., Hayashi, Y., Nomura, S. and Ishida, T. (2004). User adaptation in MT-mediated communication. The First International Joint Conference on Natural Language Processing (IJCNLP-04). 596–601.

[14] O'Hagan, M. and Ashworth, D. (2002). Translation-mediated communication in a digital world – Facing the challenges of globalization and localization. Multilingual Matters: Toronto, Ontario, Canada.

[15] Park, M., Aiken, M., Lindblom, K., and Vanjani, M. (2010). Spelling and grammatical errors in electronic meetings. Issues in Information Systems, 11(1), 384-391.

[16] Patil, S., & Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. BMJ, 349, g7392.

[17] Pepper, W., Aiken, M., and Garner, B. (2011). Usefulness and usability of a multilingual electronic meeting system. Global Journal of Computer Science and Technology (GJCST), 11(10), 35-40.

[18] Posey, J. and Aiken, M. (2014). Large-scale, distributed, multilingual, electronic meetings: A pilot study of usability and comprehension. International Journal of Computers & Technology, 14(3), 5578-5585.

[19] Shigenobu, T. (2007). Evaluation and usability of back translation for intercultural communication. Usability and Internationalization: Global and Local User Interfaces. Berlin: Springer, pp.259–265.

[20] Shigenobu, T., Yoshino, T., Nadamoto, A. and Ishida, T. (2007). Accuracy and usability of back translation. International Workshop on Intercultural Collaboration (IWIC2007), 477–482.

[21] Valacich, J., Jung, J., and Looney, C. (2006). The effects of individual cognitive ability and idea stimulation on idea-generation performance. Group Dynamics: Theory, Research, and Practice. 10(1), 1-15.

[22] Wong, Z. and Aiken, M. (2006). The effects of time on computer-mediated communication in group meetings: An exploratory study using an evaluation task. International Journal of Information Systems and Change Management, 1(2), 138-158.

[23] Yamashita, N. and Ishida, T. (2006a). Automatic prediction of misconceptions in multilingual computer-mediated communication. Proceedings of the 11th International Conference on Intelligent User Interfaces. Sydney, Australia, 62–69.

[24] Yamashita, N. and Ishida, T. (2006b). Effects of machine translation on collaborative work. Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work. Banff, Alberta, Canada, 515–524.

Authors' biographies

Dr. Bart Garner is a Clinical Assistant Professor of Management Information Systems in the School of Business Administration at the University of Mississippi. His research interests include multilingual meeting systems and business ethics.

Dr. Milam Aiken is a Professor and Chair of Management Information Systems in the School of Business Administration at the University of Mississippi. His research interests include machine translation and multilingual meeting systems.