# Towards an Objective Test for Meeting Browsers: The BET4TQB Pilot Experiment

Andrei Popescu-Belis[1], Philippe Baudrion[2], Mike Flynn[1], and Pierre Wellner[1]

[1] IDIAP Research Institute,
Centre du Parc, Av. des Prés-Beudin 20,
Case postale 592,
CH-1920 Martigny, Switzerland
{andrei.popescu-belis,mike.flynn,pierre.wellner}@idiap.ch
[2] University of Geneva, School of Translation and Interpretation,
40, bd. du Pont d'Arve,
CH-1211 Geneva 4, Switzerland
philippe.baudrion@adm.unige.ch

**Abstract.** This paper outlines first the BET method for task-based evaluation of meeting browsers. 'Observations of interest' in meetings are empirically determined by neutral observers and then processed and ordered by evaluators. The evaluation of the TQB annotation-driven meeting browser using the BET is then described. A series of subjects attempted to answer as many meeting-related questions as possible in a fixed amount of time, and their performance was measured in terms of precision and speed. The results indicate that the TQB interface is easy to understand with little prior learning and that its annotation-based search functionality is highly relevant, in particular keyword search over the meeting transcript. Two knowledge-poorer browsers appear to offer lower precision but higher speed. The BET task-based evaluation method thus appears to be a coherent measure of browser quality.

**Keywords:** Multimedia meeting browsers, task-based evaluation, human-computer interaction, human factors.

## 1 Introduction

As more and more meetings are being recorded and stored, the demand for applications which access this data to find relevant information increases as well. The goal of this paper is to outline the BET evaluation method for meeting search and browsing interfaces, and to argue that this method captures significant aspects of meeting browser quality, based on the analysis of first-time usage of several meeting browsers. The BET evaluation of the TQB interface aims first at finding the most useful features of the meeting browser, i.e. the ones that appear to be used in correlation with the highest BET scores. In addition, the experiment aims also at comparing these with those obtained for other browsers, and to assess the validity of the BET method itself.

A. Popescu-Belis, S. Renals, and H. Bourlard (Eds.): MLMI 2007, LNCS 4892, pp. 108–119, 2007.

The BET will be briefly explained and discussed in Section 2, followed by a comparison with other approaches in Section 3. The features of the TQB annotation-based meeting browser will be described in Section 4. The details of the main evaluation experiment reported here appear in Section 5, while results are discussed in Section 6. These results are compared to a similar experiment with two knowledge-poorer browsers in Section 7. Perspectives for further analyses appear in Section 8.

## 2 The Browser Evaluation Test (BET)

### 2.1 Designing a BET Evaluation

The BET is an extensive framework containing guidelines and tools that allow evaluators to construct a browser-independent evaluation task, and then to test the performances of a given browser on that task [1]. Each evaluation task is meeting-specific and consists of a set of *observations of interest* determined by a pool of *observers* who have watched closely the meeting recording, and have noted the most salient facts and events that occurred in the meeting (observers are not meeting participants). The observations are sampled, and possibly edited, to produce a final list for each meeting. The actual testing of a browser requires *subjects* to answer as many binary-choice *test questions* as possible in a fixed amount of time, by using the meeting browser to access the meeting. The binary-choice test questions are pairs of true/false statements constructed by the observers from their observations of interest, as explained below.

Using the BET requires therefore a one-time investment in collecting and possibly annotating the corpora, collecting and preparing the observations, and possibly running benchmark tests with baseline browsers such as media players. Subsequent browser tests take advantage of this one-time effort to run tests and to produce comparable scores. While the details of the testing protocol can vary according to the evaluators' goals, we believe that the list of observations will remain a valuable resource associated to these meetings, which should be extended to other meetings in the future.

From the very first BET experiments [1], two important parameters characterized the subjects' performance. The first one is *precision* (or accuracy), i.e. the proportion of correctly answered questions among all true/false statements that were seen, a number between 0 and 1. The second one, called *speed*, is the average number of questions that were processed per minute. These scores parallel somewhat the precision and recall scores used in information retrieval. None of them is sufficient alone to capture the overall quality of a meeting browser, as trivial strategies can maximize them independently, but not jointly. However, while it is certainly possible to compute the average of precision and speed, a more nuanced integrative score, which factors out the different strategies of the subjects (maximizing either precision or speed), must yet be found.

### 2.2 Collecting the Observations

BET questions are derived from observations of interest produced by a set of observers using dedicated interfaces. Observers can see the full recordings of every media source—audio, video and slides—for each meeting they work on. There is no time limit, but observers are asked to produce a minimal amount of observations, for instance 50 observations for a 50-minute meeting.

Each observer is instructed to produce observations about facts or events *that the meeting participants appeared to consider interesting*. The instruction is kept generic on purpose, in order not to influence observers towards a particular type of observation. Such a definition is compatible with many types of observations, even though it is possible to argue that some facts which do not seem important to participants could be important to an external observer, depending on their interest, which is a relative notion. We argue however that by selecting various subsets from the lists of observations produced using the BET, one can accommodate a wide range of evaluation objectives.

Observers create first a list of observations, which are automatically time-stamped by the BET observer interface with the media time. Observers are also asked to estimate the "locality" of each observation, i.e. whether it applies around the current media time or throughout the meeting. Observations should also be difficult to guess without access to the recording, and must be stated in a simple and concise manner. After they have completed their list, observers are asked to rate the importance of observations (on a five-point scale) and to create a false version of each of them. The result for each observation is a complementary pair of statements, one true and one false, both of which will be later presented to subjects during testing.

### 2.3 Validation, Editing, Grouping and Ordering

Once collected, observation pairs (a true and a false statement) are discussed by the BET experimenters and by browser designers. At this stage, some of the observations can be rejected for a number of reasons, which are carefully explicited to ensure that they are browser-neutral, and do not select observations that are better suited to a particular kind of browsing technique. These reasons are: (1) statements that are true at one moment but false at another moment of the meeting; (2) statements that are considered incomprehensible to native English speakers because of serious grammatical or typographical errors, or unclear formulation; (3) statements that are too easily guessable; (4) true and false statements that aren't parallel enough, or are not mutually exclusive; (5) statements based on "censored" material, i.e. on segments which participants had asked to be left out of the recording. Rejection of observations requires consensus among different experimenters, working on potentially very different browser designs. In addition to rejection, only very limited editing of statements is also allowed (for any of the above reasons) in order to avoid rejecting too many observations.

In many cases, different observers make similar observations—a proof of inter-observer agreement which is exemplified below (Section 2.4). Therefore,

observations must be manually grouped by the experimenters, so that subjects are tested using a single representative from the group, in order to avoid redundancy. The representative observation of the group is manually selected by the experimenters based on the following criteria, which (again) avoid favouring one type of browser over another. The selected pair of true/false statements must (1) meet the validity criteria stated above; (2) be concise and crisply expressed; (3) express, if possible, only one factual point; (4) share the same keywords as the whole group; and (5) be difficult to guess.

The edited representative observations are finally ordered, first by size of group, because this represents the number of times the observation was made by independent observers, then (for groups of equal size) by median importance adjusted per observer, then by mean adjusted importance, and finally by media time. The ordering can be changed to suit the evaluators' goals, though in some cases the answer to one question could reveal the answer to following ones.

### 2.4 Resulting Test Material

Three meetings from the AMI Corpus [2] were selected for the observation collection procedure: IB4010, IS1008c, and ISSCO-Meeting_024. The meetings are in English, and involve four participants, native or non-native English speakers. In the first meeting, the managers of a movie club select the next movie to show; in the second one, a team discusses the design of a remote control; in the third one, a team discusses project management issues. Although the first two meetings are in reality enacted by researchers or students, the movie club meeting (IB4010) appears to be more natural than the remote control meeting (IS1008c), probably due to the familiarity of the participants with the topic. For each of these three meetings, BET observations were collected, edited and ordered, this resource being now publicly available at `http://mmm.idiap.ch`. In the evaluations below, the order based on importance was kept constant.

For these meetings, respectively 222, 133 and 217 raw observations were collected, from respectively 9, 6 and 6 observers, resulting in respectively 129, 58 and 158 final pairs of true/false observations. As initial observations are grouped according to their similarity, as explained above, the average size of the groups (1.72, 2.29 and 1.37 observations per group) provides a measure of inter-observer agreement. While these values are not very high with respect to the number of observers, it is more eloquent to consider only the agreement for the observations that were answered by at least half of the subjects in the experiments on TQB (i.e. 16 for IB4010 and 8 for IS1008c). As these were ranked by importance, the average number of observers having made these observations was around 5 for both meetings, i.e. 55% and 83% of the observers agreed upon them.

As an example, the first two pairs of true/false observations for IB4010 were: "The group decided to show The Big Lebowski" vs. "The group decided to show Saving Private Ryan", and "Date of next meeting confirmed as May 3rd" vs "Date of next meeting confirmed as May 5th". For the IS1008c meeting the first pair is: "According to the manufacturers, the casing has to be made out of wood" vs. "According to the manufacturers, the casing has to be made out of rubber".

## 3 Comparison to Other Approaches

The evaluation of interactive software, especially of multi-modal dialogue systems, is still an open problem [3,4,5]. A possible approach in the case of meeting browsers is based on the ISO/IEC standards for software evaluation, especially for task-based evaluation or for evaluation in use [6,7]. The three main aspects of quality that are evaluated in such approaches are often summarized as *effectiveness* (the extent to which the system helps the user to successfully accomplish a task), *efficiency* (the speed with which the task is accomplished) and *user-satisfaction* (measured using questionnaires). Depending on the nature of the system that is evaluated, these three broad quality characteristics can be substantially particularized and/or extended [8]. A well-known approach to dialogue system evaluation, PARADISE [9], predicts user satisfaction from task completion success and from a number of computable parameters related to dialogue cost. However, depending on the specificity of the modules of a dialogue system, each of them can also be evaluated separately using black-box methods [10].

The goal of the BET is to provide an evaluation framework that sets as few *a priori* constraints as possible on the task of meeting browsing and on the functionalities of a meeting browser. The BET differs from classic usability testing as the details of the task are not predetermined by designers or evaluators. The process of collecting observations determines only in an indirect manner what the users of a meeting browser would primarily look for in a meeting (information type and content), and therefore the BET can be applied independently of the specifications of a meeting browser. This approach tempers undue influence of each observer's own special interests, and avoids the introduction of experimenter bias regarding the relative importance of particular types of meeting events, e.g. related to a particular modality that a given browser might focus on. The precision and speed of the subjects using a specific browser to answer questions based on BET observations respectively reflect the effectiveness and efficiency of the meeting browser, if the subjects' abilities are factored out across a large pool of subjects. Finally, in the setting described here, the BET measures browser quality at first-time usage, and not occasional or long-term usage, which would require extensive training of the subjects.

## 4 The Transcript-Based Query and Browsing Interface

In the study reported here, the TQB interface [11] was submitted to the BET via a web browser. TQB was designed to provide access to the transcript of meetings and to their annotations, as well as to the meeting documents, as these language-related modalities are considered to be the main information vector for human interaction in meetings. Along with the transcript, the following annotations are stored in a database to which the TQB interface gives access: segmentation of individual channels into utterances, labelling of utterances with dialogue act tags (e.g. statement, question, command, or politeness mark), segmentation of

the meeting into thematic episodes, labelling of episodes with salient keywords, and document-speech alignment using explicit references to documents [2,12]. For the BET evaluation, manual transcripts and annotations from the AMI Corpus are used, in order to focus the test on the quality of the interface and not on the quality of automatic annotation.

TQB allows users to search, within a given meeting, for the particular utterances that satisfy a set of constraints on the transcript and the annotations (Figure 1). TQB displays in one frame the annotation dimensions that are searchable for the selected meeting, with a menu of possible values for each of them, except for the transcript, which can be searched as free text. The results of a query, i.e. the utterances that match all the constraints, are displayed in another frame. These utterances can be used as a starting point to browse the meeting, by clicking on one of the retrieved utterances, which makes the transcript frame scroll automatically to its position. The enriched transcript and the meeting documents constitute the two principal frames occupying the center of the TQB interface. Browsing through these frames is enhanced by the possibility to listen to the recording of each utterance, and to display documents that are explicitly referred to at a given point of the conversation [11].

To increase the informativeness of the BET evaluation, the users' interactions with TQB are closely monitored. The logging mechanism has two components: the first one logs all the queries to the database with their timestamps, while the second one logs the actions performed by the user in the other frames. These include the state and position of the audio player and of the scrollbars (sampled every 30 seconds), and the user's mouse clicks.

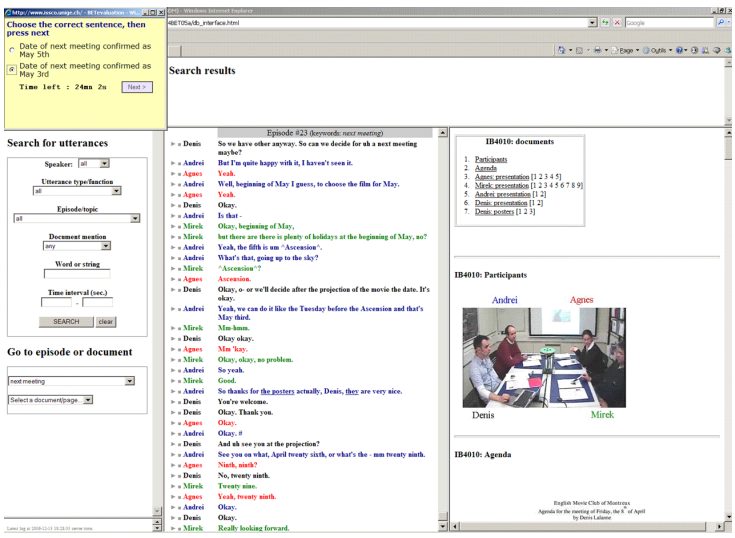


**Fig. 1.** View of the Transcript-based Query and Browsing Interface (TQB) with BET true/false observations displayed in the upper-left corner

## 5 BET Setup for TQB Evaluation

Two of the three meetings for which BET observations exist were used for the evaluation of TQB, namely IB4010 and IS1008c. The evaluation proceeds as follows. The subjects register with their email as a unique identifier and state their proficiency in English. The subjects then read the instructions for the experiment on their computer screen, explaining first the BET guidelines, and then the principles of the TQB interface, using a snapshot and 4-5 paragraphs of text. The subjects did not have the opportunity to work with TQB before being tested on the first meeting, so this was their very first occasion to explore the functions of TQB.

The BET master interface displays one by one the pairs of true/false statements corresponding to observations, following the order described above. Using TQB to browse the meeting, each subject must determine the true statement and (implicitly) the false one. When the choice is validated, the BET interface automatically displays the following pair of statements, and so on until the time allowed for the meeting is over. After a short break, the subject proceeds to the second meeting. The duration allowed for each meeting was half the duration of the meeting: 24'40" for IB4010, and 12'53" for IS1008c; the timing was managed by the BET master interface.

TQB was tested with 28 subjects, students at the University of Geneva, mainly from the School of Translation and Interpreting. Results from 4 other students were discarded for not completing the two meetings. The average proficiency on a 4-point scale (from 'beginner' to 'native') was 2.6, median value being 3 ('advanced'). Half of the subjects started with IB4010 and continued with IS1008c, and the other half did the reverse order, thus allowing for differentiated results depending on whether a meeting was seen first or second within the trial. Performance was measured using precision and speed, as defined above.

## 6 BET Results for TQB

### 6.1 Overall Scores and Variations

The overall precision, averaged for 28 subjects on two meetings, is 0.84 with a $\pm 0.05$ confidence interval at 95% level (confidence intervals will be regularly used below). The overall average speed is $0.63 \pm 0.09$ questions per minute. These values do not vary significantly (less that 1%) when they are computed for the two subgroups of 14 subjects who saw the meetings in a different order (IB/IS or IS/IB): only the confidence intervals increase, by 20 to 50%.

The average speed and precision vary more markedly across the two meetings, though however these differences are not significant at the 95% confidence level: speed and precision are $0.67 \pm 0.10$ and respectively $0.85 \pm 0.05$ for IB4010, both higher than the respective values for IS1008c, $0.57 \pm 0.13$ and $0.79 \pm 0.10$. If the statistical significance was higher, one could conclude that IB4010 is easier than IS1008c from the BET perspective.

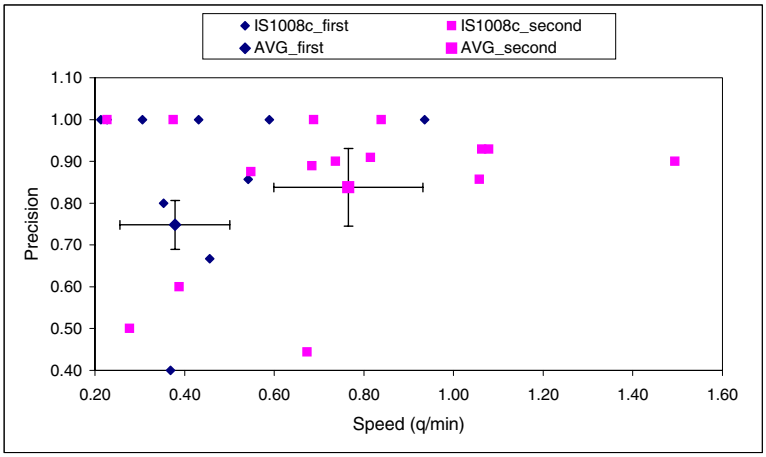


**Fig. 2.** IS1008c: scores of each subject and average score with 95% confidence intervals when the meeting is seen first (diamonds ◇) and when it is seen second (squares □). Speed appears to be significantly higher in the second case, but not precision.

The performance of each group (IB/IS vs. IS/IB) on the IS1008c meeting is shown in Figure 2: diamonds (◇) correspond to subjects seeing IS1008c as their first meeting, while squares (□) represent subjects seeing IS1008c as their second meeting, after training on IB4010. The average values of precision and speed are higher when the meeting is seen second, certainly because subjects were able to get training with the TQB interface. The 95% confidence intervals are however strictly disjoint only for speed ($0.38 \pm 0.12$ vs. $0.77 \pm 0.17$ questions per minute) but not for precision ($0.75 \pm 0.17$ vs. $0.84 \pm 0.09$). Similarly, for IB4010, speed is significantly higher (exactly with 94% confidence) when the meeting is seen second than when it is seen first, while precision increases less significantly.

These results point to an important property of the TQB interface that is highlighted by the BET evaluation, namely its learnability. A single previous trial appears to be sufficient to improve scores, indicating that TQB is an easily learnable interface. These results seem to hold independently of the individual performances of the subjects (which might maximize either precision or speed), but an assessment using more meetings would enable us to extend the study of the learning curve beyond the first two ones.

### 6.2 Use of TQB Features by BET Subjects

The analysis of TQB features used during the experiments shows that queries to the transcript and annotation database are quite extensively used to browse meetings. Subjects submit on average $2.50 \pm 0.54$ queries for each BET question, with no significant differences between the two groups (IB/IS: $2.40 \pm 0.92$ vs. IS/IB: $2.59 \pm 0.57$). There is however a significant difference between the two meetings: the remote control one elicits twice as many queries per BET question

as the movie club one (IS1008c: $4.16 \pm 1.53$ vs. IB4010: $2.12 \pm 0.44$), a fact that could be related to the differences in meeting difficulty alluded to above.

When subjects use TQB queries, they click on average in 35% of the cases on one or more utterances returned by the query, to visualize them in context within the transcript window—this figure provides thus a measure of the relevance of query results. Again, while the average is basically the same for the two groups, there is a difference between meetings: 39% for IB4010 vs. 27% for IS1008c. So, although more queries per BET question are used for IS1008c, clicking on the results is less frequent for this meeting, both facts being consistent with a higher perceived difficulty. Viewing the utterances within the meeting transcript appears to be sufficient to answer BET questions, as listening to the related audio is very infrequent, only about twice per meeting.

Another measure of the importance of TQB queries is the increasing correlation, from the first to the second meeting of a trial, of the number of queries and the precision of the answers. Pearson correlation (across subjects) between the precision scores and the average number of queries launched for each BET question is 0.49 overall (IS/IB group: 0.70 and IB/IS group: 0.37). For each meeting, the correlation increases after learning: for IB4010 it goes from 0.33 to 0.76, and for IS1008c from $-0.39$ to 0.28. Quite naturally, speed is however negatively correlated with the number of queries per BET question, overall at $-0.32$. Put simply, these figures show that using queries helps subjects to increase their precision, while at the same time slowing them down slightly.

Statistics over all the 550 queries produced by the 28 subjects indicate that most of the queries are keyword related: 43% look only for a specified word (or character string), while an additional 31% look for a specific word uttered by a particular speaker, and 7% for a word within a given topic episode. Some other constraints or combinations of constraints are used in 1–3% of the queries each: word(s) + dialogue-act, word(s) + person + dialogue-act, topic, person, word(s) + topic + person, etc. The fact that subjects use the query functionality mainly to do keyword search over the transcripts probably reflects the influence of popular Web search engines, and suggests that annotations other than transcript could better be used for automated meeting processing (e.g. for summarization) rather than directly for search by human users.

### 6.3 Question-Specific Scores

Moving further into the analysis of subjects' answers, it is possible to compute the above statistics separately for the correct answers, and for the wrong ones, and to compare the results. For instance, the average number of queries per BET question, computed only for the questions to which a subject answered correctly, is $2.41 \pm 0.58$, while the same average over the wrong answers is $2.01 \pm 0.53$. The difference is more visible for the more difficult meeting, IS1008c ($3.38 \pm 1.10$ queries per correct answer vs. $2.37 \pm 1.43$ for wrong answers) than for IB4010 ($2.09 \pm 0.47$ vs. $1.88 \pm 0.74$).

The detailed analysis of scores indicates that these vary considerably with each question. As the order of the questions was kept constant, there are less

and less available answers per question as one moves forward through the list. For instance, all 28 subjects answered the first eight questions for IB4010, but only the very first question for IS1008c was answered by all subjects. When IS1008c was seen first, two subjects spent all their time on the first question only, and only 8 subjects (out of 14) managed to answer the first five questions; when IS1008c was seen second, 13 subjects managed to answer the first five questions. As an example, average values for precision and speed are shown in Figure 3 for the first six questions of IS1008c: while scores generally increase after learning, there is considerable variation across questions, e.g. improvement of precision is not the same for all questions, while speed sometimes even degrades in the second round (for the 4th and 6th questions).

These results indicate that performances should be analyzed separately for each question, as their nature requires different competencies and browser functionalities. To take an example, it appears that the most clicked utterance among all those retrieved through TQB queries is the following one, from IB4010: "Uh Goodfellas, I didn't see it". This utterance was clicked 18 times, out of which 16 were in relation to the fifth BET question for IB4010: "No one had seen Goodfellas" vs. "Everyone had seen Goodfellas". Quite obviously, in this case, finding this utterance provides implicitly the answer to the BET question through an immediate inference, as the second statement cannot be true.

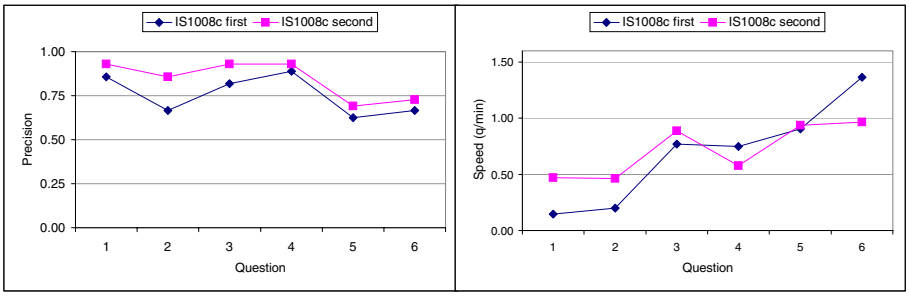


**Fig. 3.** IS1008c: precision and speed for the first six BET questions, when the meeting is seen first (diamonds ◇) and when it is seen second (squares □). Performances generally increase in the second case, but there is considerable variation across questions.

## 7 BET Results for Other Browsers

In another series of experiments [13], conducted by the IDIAP Research Institute and the University of Sheffield, four meeting browsers or "conditions" were tested with the BET, in a slightly different setting than the one described above. Usable data was obtained from 39 subjects: each subject performed a calibration task (answering questions using a very simple browser), and one of the following browsers: base (15 subjects), speedup (12 subjects), and overlap (12 subjects). Unlike TQB, none of these meeting browsers relied on manual annotation of the

data or on human transcripts. The ISSCO-Meeting_024 was used for calibration, and the other two meetings (IB4010 and IS1008c) were used alternatively in the different conditions.

The calibration condition presented a large slide view, 5 video views, the audio, a timeline, and slide thumbnails. The base condition played audio and included a timeline, scrollable speaker segmentations, a scrollable slide tray, and headshots with no live video. The speedup condition was exactly like the base condition except that it allowed accelerated playback with a user-controlled speed between 1.5 and 3 times normal speed. The overlap condition duplicated the speedup condition by offering simultaneously the first half of meeting on the left audio channel of the subject's headphone, and the second half of the meeting on the right channel, requiring the subjects to focus on one channel at the time.

Raw performance scores for both meetings were as follows for the three conditions (see [13, Section 3] for more details). For the base condition, average precision and speed were respectively 0.77 and 1.2 questions per minute; for the speedup condition, 0.83 and 0.9 questions per minute; and for the overlap condition, 0.74 and 1.0 questions per minute. The average precision is generally below the values obtained by TQB ($0.84 \pm 0.05$ for TQB), while speed is always higher ($0.63 \pm 0.09$ for TQB). These results are quite surprising, as TQB provides access to the transcript, which should considerably improve its information extraction capabilities. In addition, although TQB subjects were not native English speakers unlike those of the other two browsers, data from TQB shows that proficiency is in fact better correlated with precision (at 0.65 level) and much less with speed, therefore the proficiency factor might not explain the difference in precision. Other factors must thus be found, by analyzing experimental logs, to account for these differences.

## 8 Perspectives

The results of the Browser Evaluation Test method presented here show that the BET captures a number of properties related to browser quality, which match our *a priori* intuitions and therefore contribute to validate the BET evaluation method itself. These results must be further confirmed through future analyses, in particular question-specific ones, and possibly through experiments with more browsers and subjects. Future analyses could better model, for instance, the notion of 'strategy', i.e. a subject's bias towards maximizing either precision or speed, in order to construct a more global performance score, as an "average" of precision and speed.

The BET method offers a generic, task-based solution to the problem of evaluating very different meeting browsers, as it sets few constraints on their functionalities. The set of BET observations created for three meetings will constitute a valuable resource for future evaluations, along with the scores obtained in the experiments presented here, which will provide an initial baseline to which future interfaces can be compared.

**Acknowledgments.** This work has been supported by the Swiss National Science Foundation, through the IM2 National Center of Competence in Research, and by the European IST Programme, through the AMIDA Integrated Project FP6-0033812. The authors would like to thank Gerwin Van Doorn, who implemented the speedup and overlap browsers at IDIAP.

## References

1. Wellner, P., Flynn, M., Tucker, S., Whittaker, S.: A meeting browser evaluation test. In: CHI 2005 (Conference on Human Factors in Computing Systems), Portland, OR, pp. 2021–2024 (2005)
2. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)
3. Gibbon, D., Mertins, I., Moore, R.K. (eds.): Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation. Kluwer Academic Publishers, Dordrecht (2000)
4. Devillers, L., Maynard, H., Paroubek, P.: Méthodologies d'évaluation des systèmes de dialogue parlé: réflexions et expériences autour de la compréhension. Traitement Automatique des Langues 43(2), 155–184 (2002)
5. Dybkjær, L., Bernsen, N.O., Minker, W.: Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43(1-2), 33–54 (2004)
6. ISO/IEC: ISO/IEC TR 9126-4:2004 (E) – Software Engineering – Product Quality – Part 3: Quality in Use Metrics. International Organization for Standardization / International Electrotechnical Commission, Geneva (2004)
7. Bevan, N.: International standards for HCI and usability. International Journal of Human-Computer Studies 55, 533–552 (2001)
8. Möller, S.: A new ITU-T recommendation on the evaluation of telephone-based spoken dialogue systems. In: LREC 2004 (4th International Conference on Language Resources and Evaluation), Lisbon, pp. 1607–1610 (2004)
9. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: ACL/EACL (35th Annual Meeting of the Association for Computational Linguistics), Madrid (1997) pp. 271–280 (1997)
10. Traum, D.R., Robinson, S., Stephan, J.: Evaluation of multi-party virtual reality dialogue interaction. In: LREC 2004 (4th International Conference on Language Resources and Evaluation), Lisbon, pp. 1699–1702 (2004)
11. Popescu-Belis, A., Georgescul, M.: TQB: Accessing multimodal data using a transcript-based query and browsing interface. In: LREC 2006 (5th International Conference on Language Resources and Evaluation), Genova, pp. 1560–1565 (2006)
12. Popescu-Belis, A., Clark, A., Georgescul, M., Zufferey, S., Lalanne, D.: Shallow dialogue processing using machine learning algorithms (or not). In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 277–290. Springer, Heidelberg (2005)
13. Cremers, A., Post, W., Elling, E., van Dijk, B., van der Wal, B., Carletta, J., Flynn, M., Wellner, P., Tucker, S.: Meeting browser evaluation report. Technical report, AMI Project Deliverable D6.4 (December 2006)