IDIAP RESEARCH REPORT

The AMI Meeting Corpus: A Pre-Announcement[1]

Jean Carletta Simone Ashby
Sebastien Bourban Mike Flynn
Mael Guillemot Thomas Hain
Jaroslav Kadlec Vasilis Karaiskos
Wessel Kraaij Melissa Kronenthal
Guillaume Lathoud Mike Lincoln
Agnes Lisowska Iain McCowan
Wilfried Post Dennis Reidsma
Pierre Wellner

IDIAP–RR 05-82

July, 2005

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-82

The AMI Meeting Corpus: A Pre-Announcement[1]

Jean Carletta Simone Ashby Sebastien Bourban Mike Flynn
Mael Guillemot Thomas Hain Jaroslav Kadlec Vasilis Karaiskos
Wessel Kraaij Melissa Kronenthal Guillaume Lathoud Mike Lincoln
Agnes Lisowska Iain McCowan Wilfried Post Dennis Reidsma
Pierre Wellner

July, 2005

Abstract. The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. It is being created in the context of a project that is developing meeting browsing technology and will eventually be released publicly. Some of the meetings it contains are naturally occurring, and some are elicited, particularly using a scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The corpus is being recorded using a wide range of devices including close-talking and far-field microphones, individual and room-view video cameras, projection, a whiteboard, and individual pens, all of which produce output signals that are synchronized with each other. It is also being hand-annotated for many different phenomena, including orthographic transcription, discourse properties such as named entities and dialogue acts, summaries, emotions, and some head and hand gestures. We describe the data set, including the rationale behind using elicited material, and explain how the material is being recorded, transcribed and annotated.

1 Introduction

AMI is a large, multi-site and multi-disciplinary project with the aim of developing meeting browsing technologies that improve work group effectiveness. As part of the development process, the project is collecting a corpus of 100 hours of meetings using instrumentation that yields high quality, synchronized multi-modal recording, with, for technical reasons, a focus on groups of four people. All meetings are in English, but a large proportion of the speakers are non-native English speakers, providing a higher degree of variability in speech patterns than in many corpora. We expect the corpus to become an invaluable resource to a range of research communities, since it should be of interest to those working on speech, language, gesture, information retrieval, and tracking, as well as being useful for organizational psychologists interested in how groups of individuals work together as a team. We describe the data set and explain how the material is being recorded, transcribed and annotated.

2 The shape of the corpus

Any study of naturally-occurring behaviour such as meetings immediately encounters a well-known methodological problem: if one simply observes behaviour "in the wild", one's results will be difficult to generalize, since not enough will be known what is causing the individual (or individuals) to produce the behaviour. [1] identifies seven kinds of factors that affect how work groups behave, ranging from the means they have at their disposal, such as whether they have a way of communicating outside meetings, to aspects of organizational culture and what pressures the external environment places on the group. The type of task the group is trying to perform, and the particular roles and skills the group members bring to it, play a large part in determining what the group does; for instance, if the group members have different roles or skills that bear on the task in different ways, that can naturally increase the importance for some contributions, and it can also be a deciding factor in whether the group actually needs to communicate at all or can leave one person to do all of the work. Vary any of these factors and the data will change in character, but using observational techniques, it is difficult to get enough of a group history to tease out these effects. One response to this dilemma is not to make completely natural observations, but to standardize the data as much as possible by eliciting it in a controlled manner for which as many as possible of the factors are known. Experimental control allows the researcher to find effects with much greater clarity and confidence than in observational work. This approach, well-established in psychology and familiar from some existing corpora (e.g., [2]), comes with its own danger: results obtained in the laboratory will not necessarily occur outside it, since people may simply behave differently when performing an artificial task than they do in their daily lives.

Our response to this methodological difficulty is to collect our data set in parts. The first consists of elicited material using a design task in which the factors that [1] describe are all fixed as far as they can be. Since it constitutes the bulk of the data, the details of how it was elicited are important, and so we describe it below. The second consists of other, less controlled elicitations for different tasks. For instance, in one set of five meetings, forming one coherent set, which draws personnel from an existing work group to plan where to place people, equipment, and furniture in a fictionalized move to a new site that simplifies a real situation the group faces. These again provide more control than in natural data, but give us a first step towards thinking about how one combines data from disparate sources. The third contains naturally occurring meetings in a variety of types, the purpose of which is to help us validate our findings from the elicitation and determine how well they generalize by seeing how badly variation in the factors affects our models. The goal in this part of the collection was not to constrain the type of meeting in any way apart from keeping the recording manageable, but to allow the factors to vary freely. Taking histories that would allow us to classify the groups by factor would be a formidable task, and so the recorded data is included "as is", without supplementary materials.

3 The meeting elicitation scenario

In our meeting elicitation scenario [3], the participants play the roles of employees in an electronics company that decides to develop a new type of television remote control because the ones found in the market are not user friendly, as well as being unattractive and old-fashioned. The participants are told they are joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype of the new remote control. We chose design teams for this study for several reasons. First, they have functional meetings with clear goals, so making it easier to measure effectiveness and efficiency. Second, design is highly relevant for society, since it is a common task in many industrial companies and has clear economic value. Finally, for all teams, meetings are not isolated events but just one part of the overall work cycle, but in design teams, the participants rely more heavily on information from previous meetings than in other types of teams, and so they produce richer possibilities for the browsing technology we are developing.

3.1 Participants and roles

Within this context, each participant in the elicitation is given a different role to play. The *project manager* (PM) coordinates the project and is responsible overall. His job is to guarantee that the project is carried out within time and budget limits. He runs the meetings, produces and distributes minutes, and produces a report at the end of the trial. The *marketing expert* (ME) is responsible for determining user requirements, watching market trends, and evaluating the prototype. The *user interface designer* (UI) is responsible for the technical functions the remote control provides and the user interface. Finally, the *industrial designer* (ID) is responsible for designing how the remote control works including the componentry. The user interface designer and industrial designer jointly have responsibility for the look-and-feel of the design.

For this elicitation, we use participants who are neither professionally trained for design work nor experienced in their role. It is well-known that expert designers behave differently from novices. However, using professional designers for our collection would present both economic and logistical difficulties. Moreover, since participants will be affected by their past experience, all those playing the same role should have the same starting point if we are to produce replicable behaviour. To enable the participants to carry out their work while lacking knowledge and experience, they are given training for their roles at the beginning of the task, and are each assigned a (simulated) personal coach who gives sufficient hints by e-mail on how to do their job. Our past experience with elicitations for similar non-trivial team tasks, such as for crisis management teams, suggests that this approach will yield results that generalize well to real groups. We intend to validate the approach for this data collection both by the comparisons to other data already described and by having parts of the data assessed by design professionals.

3.2 The structure of the elicited data

[4] distinguishes the following four phases in the design process:

- *Project kick-off*, consisting of building a project team and getting acquainted with both each other and the task.
- *Functional design*, in which the team sets the user requirements, the technical functionality, and the working design.
- *Conceptual design*, in which the team determines the conceptual specification for the components, properties, and materials to be used in the apparatus, as well as the user interface.
- *Detailed design*, which finalizes the look-and-feel and user interface, and during which the result is evaluated.

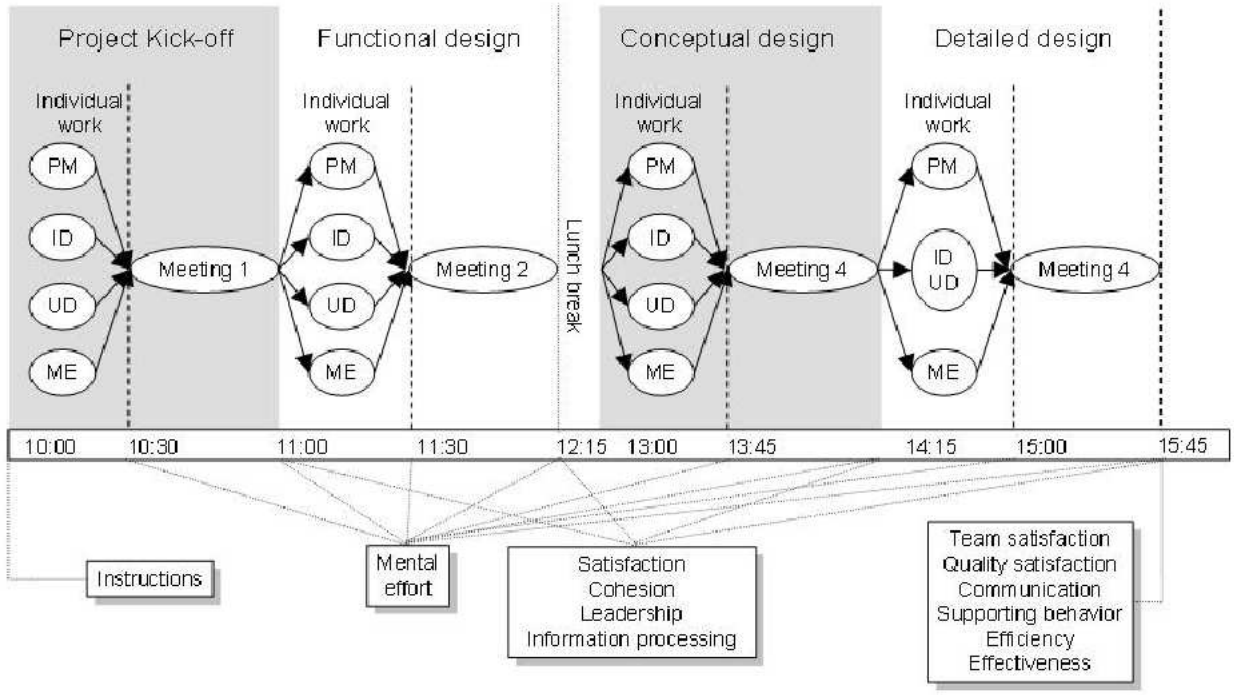


Figure 1: The meeting paradigm: time schedule with activities of participants on top and the variables measured below. PM: Project Manager; ID: industrial designer; UI: user interface designer; ME: marketing expert.

We use these phases to structure our elicitation, with one meeting per design phase. In real groups, meetings occur in a cycle where each meeting is typically followed by production and distribution of minutes, the execution of actions that have been agreed on, and the preparation of the next meeting. Our groups are the same, except that for practical reasons, each design project was carried out in one day rather than over the usual more extended period, and we included questionnaires that will allow us to measure process and outcomes throughout the day. In future data collections we intend to collect further data in which the groups have access to meeting browsing technology, and these measures will allow us to evaluate how the technology affects what they do and their overall effectiveness and efficiency. An overview of the group activities and the measurements used is presented in fig. 1.

3.3 The working environment

Our collection simulates an office environment in which the participants share a meeting room and have their own private offices and laptops that allow them to send e-mail to each other, which we collect; a web browser with access to a simulated web containing pages useful for the task; and PowerPoint for information presentation. During the trials, individual participants receive simulated e-mail from other individuals in the wider organization, such as the account manager or their head of department, that are intended to affect the course of the task. These emails are the same for every group.

4 Data capture: Instrumented meeting rooms

The data is being captured in three different instrumented meeting rooms that have been built at different project sites. The rooms are broadly similar but differ in overall shape and construction

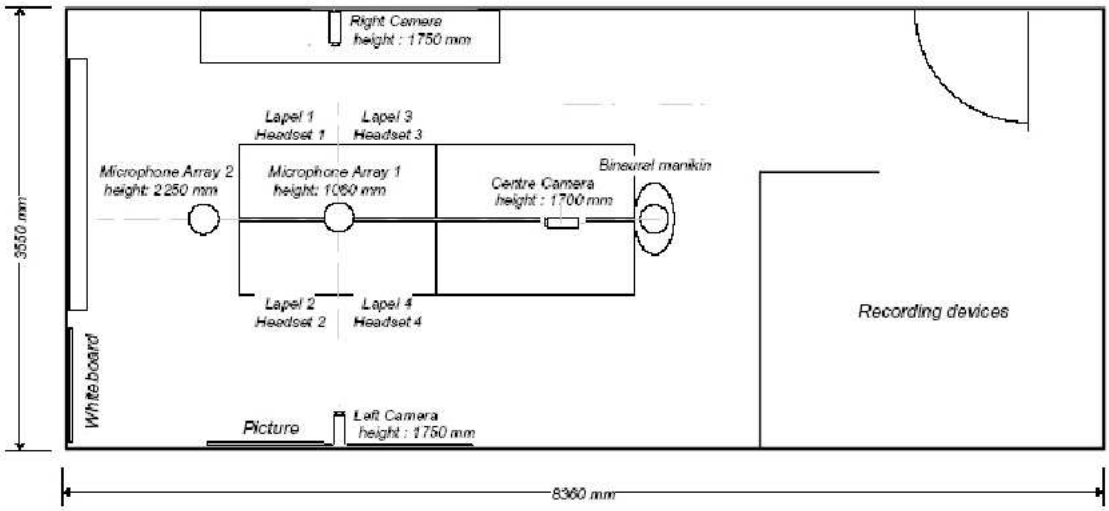


Figure 2: Overhead Schematic View of the IDIAP Instrumented Meeting Room.

and therefore in their acoustic properties, as well as in some recording details, such as microphone and camera placement and the presence of extra instrumentation. All signals are synchronized by generating a central timecode which is used to replace the timecodes produced locally on each recording device; this ensures, for instance, that videos same frames at exactly the same time and that we can find those times on the audio. An example layout, taken from the IDIAP room, is shown in figure 2.

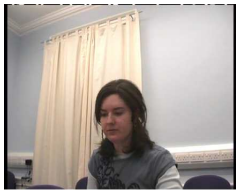
4.1 Audio

The rooms are set up to record both close-talking and far-field audio. All microphone channels go through separate pre-amplification and analogue to digital conversion before being captured on a PC using Cakewalk Sonar recording software. For close-talking audio, we use omni-directional lapel microphones and headset condenser microphones. Both of these are radio-based so that the participants can move freely. For far-field audio, we use arrays of four or eight miniature omni-directional electret microphones. The individual microphones in the arrays are equivalent to the lapel microphones, but wired. All of the rooms have a circular array mounted on the table in the middle of the participants, plus one other array that is mounted on either the table or the ceiling and is circular in two of the rooms and linear in the other. One room also contains a binaural manikin providing two further audio channels.

4.2 Video

The rooms include capture of both videos that show individuals in detail and ones that show what happens in the room more generally. There is one close-up camera for each of four participants, plus for each room, either two or three room view cameras. The room view cameras can be either mounted to capture the entire room, with locations in corners or on the ceiling, or to capture one side of the meeting table. All cameras are static, with the close-up cameras trained on the participants' usual seating positions. In two of the rooms, output was recorded on Mini-DV tape and then transferred to computer, but in the other, output was recorded directly. Figure 3 shows sample output from cameras in the Edinburgh room.

Figure 3: Camera views in the Edinburgh room.


Closeup


Corner


Overhead

4.3 Auxiliary Data Sources

In addition to audio and video capture, the rooms are instrumented to allow capture of what is presented during meetings, both any slides projected using a beamer and what is written on an electronic whiteboard. Beamer output is recorded as a timestamped series of static images, and whiteboard activity as timestamped x-y co-ordinates of the pen during pen strokes. In addition, individual note-taking uses Logitech I/O digital pens, where the output is similar to what the whiteboard produces. The latter is the one exception for our general approach to synchronization; the recording uses time-codes produced locally on the pen, requiring us to synchronize with the central timecode after the fact as best we can. We intend to subject all of these data sources to further processing in order to extract a more meaningful, character-based data representation automatically [5, 6].

5 Orthographic Transcription

Our first and most crucial annotation is orthographic transcription of the recorded speech.

5.1 The transcription process

Transcribers work to a written manual, the features of which are described in the next section. We use several steps in the transcription process in order to ensure the quality of the results.

First pass. First pass transcribers are expected to achieve a balance between speed and accuracy. They start not with the raw audio signals but with a blank transcription that uses a simple energy-based technique to segment silence from speech for each person in the meeting, a technique originally developed and tested in [7]. Transcribers only listen to and transcribe the areas identified as speech by the auto-segmentation, using special marks for transcription of which they are unsure or that is unintelligible. They adjust segment boundaries where the given ones clearly begin too late or end too early, but without care to be accurate at this stage.

Second pass. In this step the checker reviews all segments, both speech and silence. The first-pass transcription is verified, any missed speech is transcribed, segment boundaries are carefully reviewed and adjusted to better fit the speech, and any uncertainties (items in parentheses) are resolved. If a sequence remains unintelligible, it is marked permanently as such.

Some meetings also receive a third pass from a transcription manager as a quality control step. Each transcription is then validated using a script that checks for spelling errors against the evolving AMI dictionary, uninterpretable symbols, and problems with the data format before being marked as 'finished'.

It is important to manage any large transcription effort carefully in order to avoid inconsistencies in the set of transcriptions, as well as to keep the work flowing smoothly. We have found Wikis invaluable in this regard. We use them to allocate work to individual transcribers, record their progress, discuss and resolve difficulties with interpreting the manual or with the audio files, and create official spellings for words that are not already in the dictionary used for spell checking. The transcriptions themselves are held in a CVS repository with symbolic tags representing their status, to which the transcribers have access via a simple web form.

Figure 4: Transcription Sample

```
(ID) That's our number one prototype.
(PM)                                /@ like a little lightning in it.
(ID) Um do you wanna present the potato,
(ID) or shall I present the Martian?
(UI)    /Okay, um -
(PM)          /The little lightning bolt in it, very cute.
(UI)                            /What -
(UI) We call that one the rhombus, uh the rhombus.
(ME)        /I could -
(PM)                                    /The v- the rhombus rhombus?
(ID)                                                        /That's
(ID) the rhombus, yep.
(UI) Um this one is known as the potato, uh it's
(UI) it's a $ how can I present it? It's an ergonomic shape,
(ID)  /$
(ME)    /$
(UI) so it it fits in your hand nicely. Um,
(UI) it's designed to be used either in your left hand or or
(UI) in your right hand.
```

5.2 Features of AMI transcriptions

Speech is transcribed verbatim using British spellings, without correcting grammatical errors, e.g. 'I seen him', 'me and him have done this'. Additionally, certain common 'nonstandard' forms signifying linguistic reduction are employed, such as 'gonna' and 'kinda'. Normal capitalization on proper nouns and at the beginning and end of sentences is used, along with simplified standard English punctuation, including commas, hyphens, full stops and question marks. Other types of punctuation are used for specific purposes. Neologisms are flagged with an asterisk, e.g. 'bumblebeeish*'. Where mispronunciations are simply due to interference from the speaker's mother tongue, and therefore could be considered how one would expect a speaker of that language to pronounce the English word involved, they are ignored. Other mispronunciations are flagged with an asterisk as for neologisms, with the word transcribed using its correct spelling, not a spelling representing how it was pronounced. Discontinuity and disfluency, at the word or the utterance level, are indicated with a hyphen, e.g. 'I think basi- '; 'I just meant—I mean . . . '. Particular care is also taken with punctuation at the end of a speech segment, where it indicates either that the turn continues (comma or no punctuation) or does not (full stop, question mark or hyphen). Qualitative and non-speech markers are kept to a minimum. Simple symbols are used to denote laughing '$', coughing '%' and other vocal noises '#', while other types of nonverbal noises are not indicated in the transcription. Whispered or emphasized speech, for example, are not tagged in any special way. A special category of noises, including onomatopoetic and other highly meaningful sounds, are indicated with a meta-noise tag within square brackets, e.g. '[sound imitating beep]'.

Sample transcription given in a human-readable format is shown in figure 4. The transcribers used Channel Trans (http://www.icsi.berkeley.edu/Speech/mr/channeltrans.html), which adapts Transcriber (http://www.etca.fr/CTA/gip/Projets/Transcriber/) for multiple speakers. Transcribers worked from headset audio except in a few instances where the lapel audio was of higher quality.

6 Forced Alignment

Automatically generated word and phoneme level timings of the transcripts are provided. Firstly this allowed more effective annotation of higher level information, secondly the time-segmentation is provided with the corpus for further processing. As the process for obtaining the time-segmentation has several implications on future processing we include a brief description of the steps involved. The timings were generated using acoustic models of an automatic speech recognition system [8]. The system was specifically developed for the transcription of the AMI meetings using all input channels and is based on the Hidden Markov Model Toolkit (HTK, http://htk.eng.cam.ac.uk). The time level information itself was obtained in a multi-step process:

Preprocessing of transcripts. Normalisation of transcripts to retain only events that are describable by phonemes. Text normalisation to fit the following dictionary creation.

Generation of a pronunciation dictionary. For the alignment a pronunciation for each word is required. This is either a fully automatic or a semi-automatic process. Dictionaries are based on the UNISYN dictionary [9], pronunciations for words not in that dictionary were created using pronunciation prediction (for more details on this process see [8]). In the case of semi-automatic processing, the suggested pronunciation is manually checked.

Viterbi Alignment. The acoustic recordings from the independent headset microphones are encoded and processed using the Viterbi algorithm, and the text and dictionaries created in the previous steps. Utterance time boundaries are used from the previous segmentation. Two passes of alignment are necessary to ensure a fixed silence collar for each utterance.

The acoustic models used in this process are trained on data from conversational telephone speech recordings (CTS) and more than 100 hours of close-talking microphone recordings from meetings, including the AMI corpus.

Post-processing. The output of the alignment stage includes silence within words. This is corrected.

The output of the above process is an exact time and duration for each pronounceable word in the corpus according to close talking microphones. Furthermore phoneme level output is provided, again with exact timing. In each case times and durations are multiples of 10 milliseconds. Due to the automatic processing errors in the times are inevitable. Word level times should be broadly correct, however problems arise in the vicinity of overlapped speech (i.e. multiple speakers talking at the same time) and non-speech sounds (like door-closing etc). Furthermore problems can be expected where it was impossible to derive pronunciation for human generated sounds.

Phoneme level transcripts and timings should be used with caution. Meeting speech is conversational and spontaneous, hence similar in nature to CTS data. Greenberg et al. [10] have shown that there are considerable differences between human and automatic phone labelling techniques. Since the cost of manual labelling is prohibitive for corpora of this size one has to be aware of the properties of automatic methods as used here: Firstly, canonical pronunciations from dictionaries are used to represent arbitrary acoustic realisations of words. Secondly acoustic models for alignments make use of phoneme context. This and general model building strategies imply that phone boundaries can be inaccurate for frequently occurring phone sequences.

7 Annotation

In addition to orthographic transcription, the data set is being annotated for a wide range of properties:

- Named entities, focusing on references to people, artefacts, times, and numbers;
- Dialogue acts, using an act typology tailored for group decision-making and including some limited types of relations between acts;
- Topic segmentation that allows a shallow hierarchical decomposition into subtopics and includes labels describing the topic of the segment;

- A segmentation of the meetings by the current group activity in terms of what they are doing to meet the task in which they are engaged;
- Extractive summaries that show which dialogue acts support material in either the project manager's report summarizing the remote control scenario meetings or in third party textual summaries;
- Emotion in the style of FeelTrace [11] rated against different dimensions to reflect the range that occurs in the meeting;
- Head and hand gestures, in the case of hands focusing on those used for deixis;
- Location of the individual in the room and posture whilst seated;
- for some data, where on the video frames to find participant faces and hands; and
- for some data, at which other people or artefacts the participants are looking.

These annotations are being managed by a process similar to that used by the transcribers. For each one, reliability, or how well different annotators agree on how to apply the schemes, is being assessed.

Creating annotations that can be used together for such a wide range of phenomena requires careful thought about data formats, especially since the annotations combine temporal properties with quite complex structural ones, such as trees and referential links, and since they may contain alternate readings for the same phenomenon created by different coders. We use the NITE XML Toolkit for this purpose [12]. Many of the annotations are being created natively in NXT's data storage format using GUIs based on NXT libraries — figure 5 shows one such tool — and others require up-translation, which in most cases is simple to perform. One advantage for our choice of storage format is that it makes the data amenable to integrated analysis using an existing query language.

8 Release

Although at the time of submission, the data set has not yet been released, we intend to allow public access to it via http://mmm.idiap.ch, with a mirror site to be established at Brno University of Technology. The existing Media File Server found there allows users to browse available recorded sessions, download and upload data by HTTP or FTP in a variety of formats, and play media (through RTSP streaming servers and players), as well as providing web hosting and streaming servers for the Ferret meeting browser [13].

References

[1] McGrath, J.E., and A. Hollingshead. Interacting with technology: Ideas, evidence, issues and an agenda. In *Sage Publications, Thousand Oaks*, 1994.

[2] A.H. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The hcrc map task corpus. In *Language and Speech 34*, 1991.

[3] W.M. Post, A.H. Cremers, and O.B. Henkemans. A research environment for meeting behavior. In *In Nijholt, A., Nishida, T., Fruchter, R., Rosenberg, D., eds.: Social Intelligence Design, University of Twente, Enschede, the Netherlands*, 2004.

[4] G. Pahl and W. Beitz. Engineering design: a systematic approach. In *Springer, London*, 1996.

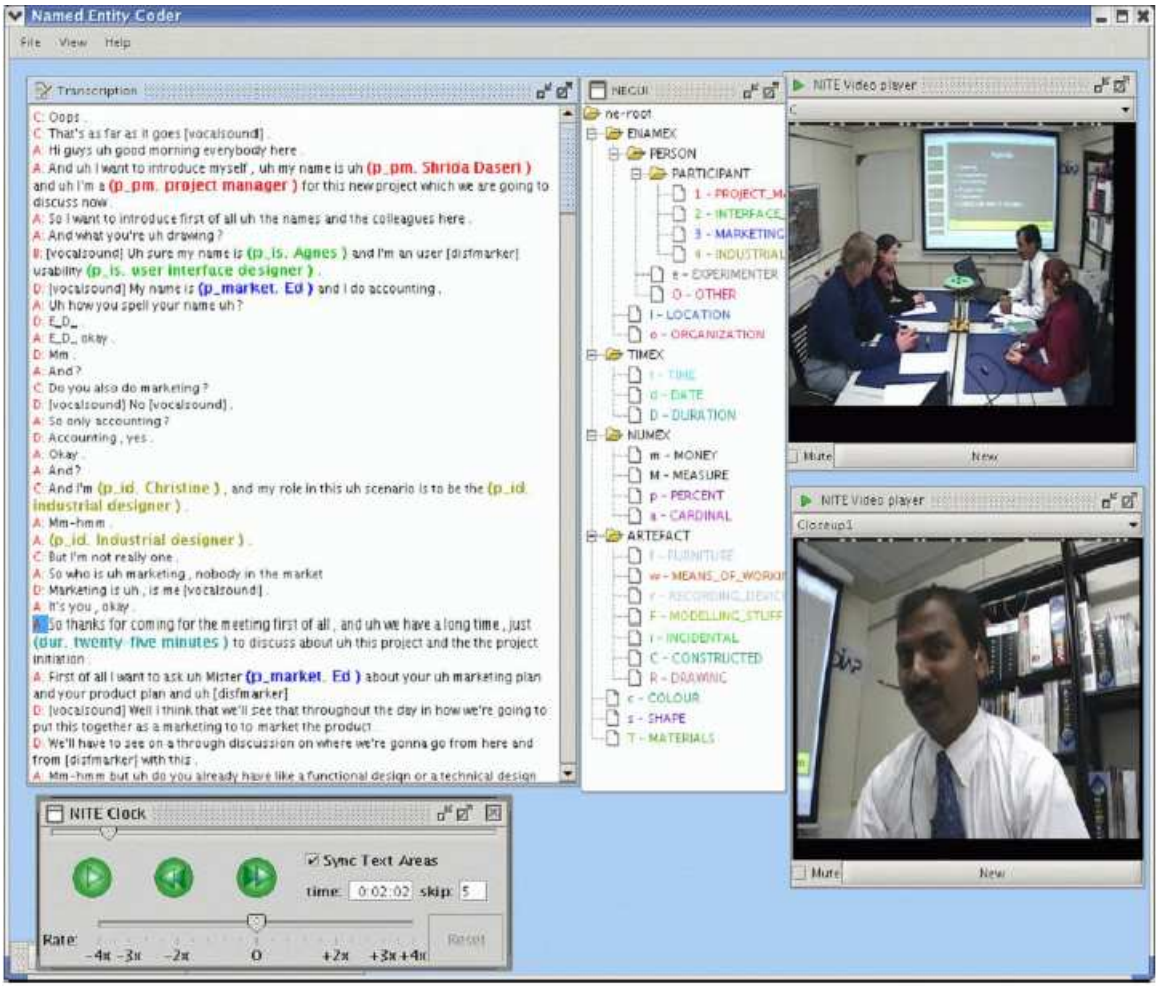


Figure 5: Screenshot of the named entity annotation tool.

[5] D. Chen, J.M. Odobez, and H. Bourlard. Text detection and recognition in images and video frames. In *Pattern Recognition 37*, 2004.

[6] M. Liwicki and H. Bunke. Handwriting recognition of whiteboard notes. In *In Marcelli, A., ed.: 12th Conference of the International Graphonomics Society, Salerno*, 2005.

[7] G. Lathoud, I.A. McCowan, and J.M. Odobez. Unsupervised location-based segmentation of multi-party speech. In *ICASSP-NIST Meeting Recognition Workshop*, 2004.

[8] T. Hain, J. Dines, G. Garau, D. Moore, M. Karafiat, V. Wan, R. Oerdelman, and S. Renals. Transcription of conference room meetings: an investigation. In *InterSpeech 2005*, 2005.

[9] S. Fitt. Documentation and user guide to unisyn lexicon and post-lexical rules. In *Technical report, Centre for Speech Technology Research, University of Edinburgh*, 2000.

[10] S. Greenberg. Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. In *ESCA Workshop on modelling pronunciation variation for automatic speech recognitio*, 1998.

[11] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. An instrument for recording perceived emotion in real time. In *Douglas-Cowie, E., Cowie, R., Schrder, M., eds.: ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000.

[12] J. Carletta, S. Evert, U. Heid, J. Kilgour, D. Reidsma, and J. Robertson. The nite xml toolkit. In *(submitted)*, 2005.

[13] P. Wellner, M. Flynn, and M. Guillemot. Browsing recorded meetings with ferret. In *Bengio, S., Bourlard, H., eds.: Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers. Lecture Notes in Computer Science 3361. Springer- Verlag, Berlin*, 2005.