LREC 2004, Workshop on Compiling and
Processing Spoken Language Corpora,
Lisbon, Portugal, May 2004

# The ICSI Meeting Corpus: Close-talking and Far-field, Multi-channel Transcriptions for Speech and Language Researchers

**Jane A. Edwards**
International Computer Science Institute, and
Institute of Cognitive Studies, UC Berkeley
edwards@ICSI.Berkeley.EDU

**Abstract**

The recently-completed ICSI Meeting Corpus is available through the LDC. It consists of audio and transcripts of 75 research meetings, ranging in size from 3 to 10 people, with an average of 6 people. The meetings were recorded by means of both close-talking (headset or lapel) microphones and far-field (table-top) microphones. The close-talking microphones enable separation of each person's audible activities from those of every other participant. The far-field microphones provide a view of the meeting as a whole. The transcripts preserve words and other communicative phenomena, displayed in musical score format, time-synchronized to the digitized audio recordings. The corpus is intended as a resource for both speech researchers and language researchers. This paper describes the methods used to prepare the corpus, some interesting challenges and solutions, and the benefits of using both close-talking and far-field microphones.

## 1. Introduction

The "ICSI Meeting Corpus" was recently completed and is now available through the Linguistics Data Consortium (LDC). It consists of audio recordings and transcripts of 75 naturally-occurring research meetings. The goal was to produce a high-quality resource for use by both speech researchers and language researchers.

All meeting participants were recorded both by close-talking microphones (usually a headset), and by far-field microphones (arranged along the tabletop). The close-talking microphones enable separation of each person's audible activities from those of the other participants. The far-field microphones provide a view of the meeting as a whole. The resulting audio recordings fill 9 DVD's.

The meetings were "natural" (not contrived): they would have occurred regardless of whether or not they were recorded. The meetings recorded were, in large part, regular weekly meetings of ICSI research groups (5 main groups), and usually lasted about an hour.

They ranged in size from 3 to 10 participants (with average size of 6). This is much larger than most multi-party interactions recorded in other corpora. The corpus contains 5 main types of meetings and 53 unique speakers.

The meetings differed in the degree to which they followed an agenda, and the degree to which power was centralized or distributed evenly among the participants. The participants knew each other well for the most part, and cared about the matters under discussion. This led to some degree of overlapping speech, from very little to very much, depending on the group.

Standard procedures were observed in terms of Human Subjects requirements for informed consent. The Consent Form asked participants for permission to use their data in the corpus, and let them know they would have access to the transcripts and audio prior to public release of the data, and that things would be excised from their speech in meetings if they requested such excisions.

This paper describes some of the methods used in preparing the ICSI Meeting Corpus. (For information on other aspects of the corpus, please see Morgan, et al., 2001 and 2003; Janin, et al., 2003). ICSI's is the first meetings corpus to be released with audio and transcripts for public use. It is important to mention in passing that corpora of meetings are also being prepared by CMU and NIST, among others. This is an active interest area and certain to become more so in the future.

## 2. The main goals of transcription for the ICSI Meeting Corpus

Even if ongoing international efforts toward increasing standardization of data encoding methods at least in their broad outlines (e.g., TEI, EAGLES, CES, MATE), it is still the case that projects are necessarily unique in certain ways, dictated by their specific goals, and intended audience.

This corpus was designed for use by two distinct research communities: speech recognition researchers on the one hand and language researchers (linguistics and discourse researchers) on the other hand.

The main goal was to produce a word-level transcript of each speaker's channel, time synchronized to the digitized audio recording. Non-word events were also captured, and comments were added concerning aspects which might be relevant to either speech recognition (e.g., voice quality or non-canonical pronunciation), or discourse research (e.g., situational comments). The transcription conventions were chosen to be as theory-neutral and as minimalistic as possible. Among other things, there was no attempt made to capture "short" vs. "long" pauses. Pauses could be noted if they stood out to a transcriber, but perceived pause length, which is found in virtually all discourse corpora, was beyond the scope of this project. Prominence was treated in a similar manner, as was intonation. Several annotation efforts are already underway, but the most immediate goal was to provide the most accurate basic transcript possible, to be embellished later as needed.

The use of individual microphones for everyone at the meeting was invaluable in disentangling the many overlaps which occurred during the meeting. In addition, it made it possible to capture such things as whispered comments, very quiet laughs, and the sudden inbreaths which occur prior to attempting to gain the floor -- most of which would be impossible with less sensitive and/or shared microphones.

In some cases, every inhale and exhale could be heard on a particular channel. But such breathing patterns were not preserved in the transcript due to being predictable and uninformative. In contrast breathing patterns which were potentially communicative were preserved (e.g., the sudden outbreath of frustration, a sudden inbreath of surprise, or a yawn). Although English is full of words such as “sigh,” “wheeze”, and “gasp”, these were usually not appropriate, because they seemed either overly negative or inappropriately dramatic. Instead, more neutral descriptions were used.

## 3. Multi-channel audio and visual representation

If an interaction has only a couple of participants, many transcription methods are viable (as discussed in Edwards, 2002). In the ICSI Meeting Corpus, however, overlaps could include as many as 10 people (if everyone laughed at a joke) and 3- or 4-way overlaps were not uncommon. In such cases, musical score notation has obvious natural advantages over other transcription methods (e.g., Edwards, 1992; Ehlich, 1993), since it enables simultaneous or partially overlapping events to be displayed one above the other, with reference to a common time line.

The musical score notation for this corpus needed furthermore to be time-sychronized with the audio recording for each speaker. This was accomplished by use of a computer interface called "Channeltrans" (www.icsi.berkeley.edu/Speech/mr/channeltrans.html). It is an extension of the "Transcriber" interface (Barras, Geoffrois, Wu, and Liberman, 2000). Both are available free of charge.

Both Transcriber and Channeltras preserve events and the time bins in which they occurred, and both of them do so in XML format. Channeltrans differs from Transcriber in that it preserves the channel number in addition to the time and event. That is, unlike Transcriber, which has only one display ribbon for speech, Channeltrans has as many display ribbons as there are participants. In addition, Channeltrans allows the time bins on each ribbon to be totally independent of those on all other ribbons. Both properties -- multiple ribbons and independent time segmentation -- were essential for the Meeting Corpus data, due to the large number of participants and great amount of overlapping speech in these meetings.

The basic strategy used in transcribing the data was to view each display ribbon as capturing the actions of a particular meeting participant, heard over the close-talking microphone which he or she was wearing (i.e., the dominant speaker on that channel). In cases of crosstalk, other speakers might be heard on the same channel, but only the events produced by the dominant speaker were transcribed on that speaker's ribbon. That is, even if an utterance could be heard on several channels, it was transcribed only on one channel, i.e., the channel corresponding to the person who spoke that utterance.

## 4. Some time-saving strategies in first-pass transcripts

The basic task of transcribing the data involved identifying the boundaries of an event (e.g., utterance, noise, happenstance) and transcribing the nature of the event itself.

Because the meetings often had so many participants, it was impractical to accomplish the time bin segmentation in a strictly manual way (i.e., having transcribers do all the segmentation into time bins). For an hour meeting with ten participants, for example, it would have required ten hours to listen to each channel exhaustively to find the time bins which required encoding. Viewing the energy waveform for activity might seem an effective solution, but in fact, it was not very reliable.

A highly effective approach turned out to be to apply a speech-nonspeech detector to the audio recordings to generate a preliminary segmentation into time bins to be adjusted later by human transcribers. (For details see Pfau, Ellis & Stolcke, 2001).

Undergraduate transcribers were encouraged to correct, adjust, or add new segmentations as needed. The time bins were intended simply as units of a manageable size with clean breaks on either side (i.e., no truncated words). Utterances might be contained in a single time bin or they might extend across several time bins. The time bins were not to intended as definable discourse units or prosodic units but simply as manageable units, which could be made more precise later if needed.

The project also used professional transcription agencies for some first-past transcripts. The presegmented versions were processed in such a way that all of the time bins which the presegmenter identified as containing events were strung together in a linear fashion, and recorded onto a cassette together with sequence numbers to prevent duplication or omissions of segments. The professional transcribers then transcribed each time bin chunk, together with its sequence number, and the resulting chunk-wise transcript was re-assembled at ICSI and double checked by the student transcribers.

## 5. Checking the transcripts for word-level accuracy

After a transcript was completed, it was submitted to a spell-checker, and then reviewed in its entirety while the checker listened to the audio recording. After this was completed, the process was repeated by one of two senior researchers.

Even though the data had by this time been seen by at least two and often three pairs of carefully trained eyes, these "read-throughs" by the senior researchers led to a number of corrections at the word and utterance level. This reflects two aspects of the meetings: the highly technical nature of the discussions, and the fact that many meeting participants were non-native speakers of English. The senior researchers have technical backgrounds which gave them an advantage over both the linguistically-trained student checkers and the professional transcribers. This experience is no doubt familiar to anyone who has ever prepared a transcript, and is a clear reminder of the extent to which a linguistic message is underdetermined by its acoustics and of the importance of context, intonation, pragmatic conventions and world knowledge in filling in the gaps.

In about twenty cases, there were words or phrases which seemed acoustically very clear but remained incomprehensible even to senior researchers. In these cases, the actual speakers were asked to listen and demystify them. Here are two examples:

> (1) .. now that of course we have sort of started to lick blood with this, {QUAL editor's note, speaker explained that "lick blood" is a German idiom meaning "having started something, and wanting more of it"}
>
> (2) From Michael Strube, I've heard very good stuff about the chunk parser that is done by FORWISS, {QUAL editor's note, speaker-verified names}

Once checking was completed by a senior researcher, the transcripts were made available for correction by the participants themselves. Very few errors were detected in this way.

## 6. Benefits of using both Close-talking and Far-Field Microphones

On the surface it would seem that close-talking microphones alone would be sufficient in that they provide a sensitive record of each person’s speech. However, there are several situations in which the far-field microphones are invaluable.

### a) *Compensating for some glitches on the close-talking recording*

When a word was unclear or even the speaker’s identity who spoke it, the far-field microphones often provided a sufficiently different “ear” on the situation to be able to clarify it.

### b) *Tracking discourse when some participants are out of the room*

In some meetings, a participant left the room while still wearing the microphone (e.g., to photocopy something for the meeting, or to arrange something with the administrative staff). If a transcript had been based only on the close-talking microphone, the outcome of these intermingled conversations would have been confusing or even bizarre. The situation became immediately clear when listening to the far-field microphone.

The next two are somewhat more subtle and will require more sophisticated approaches to use the far-field data, but are possible in principle, and not far from what is being done already.

### c) *Distinguishing self-oriented subvocalizations from shared communicative behavior*

Where should the researcher draw the line between that which is probably audible only on the close-talking microphone and that which was probably heard by others? This is almost a Heisenberg uncertainty problem. With poorer quality recordings in the past, the researcher could be sure that if he or she heard something, the other people in the interaction no doubt heard it too. But the Meeting Corpus includes some degree of "over-precision", that is, vocalizations which speakers may make for their own purposes, with no intention that they be part of the meeting as a whole. For example, there are cases in which a speaker says a word or two to him- or herself at low volume. And there is even a case in which a particular speaker said "uh-huh" an unexpectedly large number of times, but at such a low volume that it was inaudible to other participants at the meeting. If a backchannel occurs in a meeting and no one else hears it, is it still communicative?

The high quality far-field microphones provide the raw data which can in principle be used to estimate what the others may have heard.

### d) *Determining the best mix of channels to represent the meeting as a whole*

This is an extension of the previous problem. During the calibration of equipment at the beginning of the meeting, the technician often boosted the recording volume of close-talking microphones for "soft-talkers" relative to people who normally speak more loudly. When these channels are simply combined, the person with the soft voice will be louder than he or she was in the actual meeting relative to the other participants.

In contrast, the high quality far-field microphones record multiple participants without adjustments to the loudness of the participants individually, and when combined, can give a better approximation of the relative loudness of speakers at the meeting. Some work would be needed to match up the levels of the different tabletop microphones, but this is possible in principle. Without the far-field microphones, the problem would be data-limited and therefore unsolvable.

## 7. Conclusions

This paper has discussed the general structure of the Meeting Corpus, and some of the procedures it developed in meeting the challenges of transcribing 75 actual (rather than contrived) meetings in musical score format, time-synchronized to digitized audio recordings.

It also briefly described a minimalist approach to transcription which was chosen to serve the needs of two very different research communities: speech and language research.

Finally, it discussed the benefits of having both close-talking and far-field microphones, mentioning several types of problems which would be insurmountable without the use of both types of microphones.

## 8. Acknowledgments

Data gathering and preparation of the ICSI Meeting Corpus was made possible by the following sponsors: DARPA (through Communicator and a “ROAR” seedling), IBM, and the Swiss National Science Foundation (through NCCR’s IM2 project). In addition, I wish to thank Nelson Morgan, Barbara Peskin, and the LREC reviewers for their helpful comments on an earlier draft of this manuscript.

## 9. References

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. (2000). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication* special issue on Speech Annotation and Corpus Tools, Vol. **33**, No 1-2.

Edwards, Jane A. (1992). Design Principles for the Transcription of Spoken Discourse. In J. Svartvik (Ed.) *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, August 4-8, 1991* (pp. 129--147). NY: Mouton de Gruyter.

Edwards, Jane A. (2002) The Transcription of Discourse. In D. Tannen, D. Schiffrin, and H. Hamilton (eds). *The Handbook of Discourse Analysis*. NY: Blackwell (pp. 321-348).

Ehlich, K. (1993) HIAT: A Transcription System for Discourse Data. In J. A. Edwards & M. D. Lampert (Eds.) *Talking data: Transcription and coding in discourse research.* (pp. 123-148). Lawrence Erlbaum Associates, Inc; Hillsdale, NJ.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. & Wooters, C. (2003). The ICSI Meeting Corpus. *ICASSP-2003*, Hong Kong, April 2003. <ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icassp03-janin.pdf>.

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., and Stolcke, A. (2001) The Meeting Project at ICSI. *Human Language Technologies Conference*, San Diego, March 2001

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters C. Meetings About Meetings: Research at ICSI on Speech in Multiparty Conversations. ICASSP-2003, Hong Kong, April 2003. <ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icassp03meetings.pdf>

Pfau, T. Ellis, D. P. W. & Stolcke, A. (2001), Multispeaker Speech Activity Detection for the ICSI Meeting Recorder. *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy.