

Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus *

Jean Carletta
University of Edinburgh

December 11, 2006

Abstract.

The AMI Meeting Corpus contains 100 hours of meetings captured using many synchronized recording devices, and is designed to support work in speech and video processing, language engineering, corpus linguistics, and organizational psychology. It has been transcribed orthographically, with annotated subsets for everything from named entities, dialogue acts, and summaries to simple gaze and head movement. In this written version of an LREC conference keynote address, I describe the data and how it was created. If this is "killer" data, that presupposes a platform that it will "sell"; in this case, that is the NITE XML Toolkit, which allows a distributed set of users to create, store, browse, and search annotations for the same base data that are both time-aligned against signal and related to each other structurally.

Keywords: annotated corpora, meetings, discourse annotation

1. Introduction

The AMI Meeting Corpus has recently been released as a public resource, including signals, transcription, and a range of linguistic and behavioural annotations. Creating the corpus was an ambitious undertaking — perhaps more ambitious than the AMI Consortium itself realizes. It brings together techniques for eliciting behaviour that is natural but not wild, methods for synchronizing a wide range of signals, new technology for unifying annotations in a common framework, and even new ideas about licensing, in order to create material that should be a lasting resource for not one but several research communities. In this paper to accompany an LREC keynote, I describe the corpus, highlighting what is unique about the data set and the processes used to make it, and make some personal predictions about how it could change the course of the fields it is meant to serve.

* This paper is an extended version of a Keynote Address presented at the Language Resources & Evaluation Conference, Genoa, May 2006.

2. Background

AMI is a 15-member multi-disciplinary consortium dedicated to the research and development of technology that will help groups hold better meetings. The consortium's focus so far has been on developing meeting browsers that improve work group effectiveness by giving better access to a group's past meetings, although increasingly in future, we will be considering how related technologies can help group members joining a meeting late or having to "attend" from a different location.

Each of the consortium's constituent disciplines sees this problem through a different lens. The user requirements specialists think in terms of prototyping technologies on some set of meetings and asking users what they think. The signal processors are concentrating on indexing meetings for properties that an end user interface might need and determining what features can be used to derive them. The language technologists think that since end users should be focused on the semantic content of the meeting, technologies like dialogue act recognition and summarization will be crucial, with full semantic analysis and the discernment of argument structure an eventual goal. They naturally see good speech recognition as the most essential basis for any set of features, preferably from far-field microphones so that the users are unencumbered. Meanwhile, the organizational psychologists are acutely aware that introducing any new technologies will change what users do, and therefore are keen that we be able to test whether they really do improve group effectiveness when used in the way groups would actually use them.

Each of these ways of looking at the problem requires us to collect meeting data, but each puts different constraints on the collection. What counts as high quality data for signal processing — meeting recordings where there are no drop-outs, the cameras are fully calibrated, and the subjects even walk about a bit — may contain no interesting content whatsoever. On the other hand, by clubbing together we can justify a reasonably hefty data collection, as long as it is carefully planned to allow different research communities to benefit from the same basic data. Moreover, there is a real benefit to bringing together researchers from different disciplines to consider the same data, since it keeps any of them from missing the point because their field has become too narrowly focused. During the course of our first project, the consortium went from considering data collection a necessary evil to thinking that the data itself would be one of the major goods we produce.



Figure 1. Still image taken from one team’s detailed design phase.

3. Corpus Design

There are of course many different kinds of meetings carried out for different purposes (McGrath, 1984), all of which potentially benefit from different kinds of technological support. We have chosen to focus on design team meetings in which everyone has unique expertise that bears on the problem. Partly this is because these teams have particular problems when not all the members can be together, looking at the same materials, handling the same artefacts, and engaging in free-flowing, unencumbered discussion, but also because design teams often want to revisit earlier decisions to find out why they were made. The constraints imposed by our organizational psychologists immediately raise a problem: it may be easy to ask participants whether they like a technology and whether they think it helped, but how do we make it possible to tell whether it actually improves work group effectiveness?

Our answer to this question is experimental control — that is, we record meetings in which paid subjects play one of four roles (project manager, marketing, industrial designer, and user interface designer) in a fictitious team that is designing a new remote control. After initial training for the role, each team takes a design from start to prototype in a series of four meetings corresponding to the four phases of the design process. Apart from the fact that the four meetings are compressed into one day in order to facilitate collection, the groups we observe simulate real design teams closely. They produce presentations and minutes for their meetings and have access to a complete working environment that includes a meeting room, separate offices, business software, and email from other team members as well as from the wider organization. This gives us data that is realistic, but also amenable to measurement by comparing what the teams produce to the specification they are given. Figure 1 shows a still image from one of the close-up cameras during a detailed design phase.

Although this sort of control is rare for corpora used by language engineers, it is not without precedent; in the HCRC Map Task Corpus (Anderson et al., 1991), for instance, how well a pair does the task is measurable in terms of how different the routes are on their maps when they finish. With the control required to obtain measurable outcomes, of course, comes the risk of focusing too narrowly on idiosyncracies from the way the data was collected. In addition, even though some language engineering techniques currently require material that comes from a limited domain in order to make ontology building more manageable, we need to break past this requirement if our work is to be compelling. To mitigate these risks and provide open domain material that will encourage the required advance, 35 hours of the corpus is made up of real meetings which progressively push out first from the remote control design scenario into other types of new design teams, non-design teams, and finally a few other types of meetings entirely. Most of these meetings are “real” (that is, they would have occurred whether or not we had been recording) but a few are controlled, more loosely and in different domains than the bulk of the data. All meetings are in English, but we deliberately included a high proportion of non-native speakers because this is realistic in a European context and because of the challenges this creates for speech recognition.

4. Recording

The AMI Meeting Corpus is recorded in such a way as to provide many different signals for the same underlying events: close-talking and far-field audios, individual and room-view videos, plus the output of the data projector, electronic whiteboard, and individual digital pens. The meetings were actually recorded in three different rooms with different acoustic properties and layouts in order to give us some variation, but the rooms are broadly similar. Audio capture is highly redundant, with wireless omni-directional lapel and headset condensor microphones to capture individual speakers as well as two circular eight-microphone arrays for far-field capture. Video capture is via six cameras, four showing individual views and two showing room views that differ depending on which of the three room being used. Except for the digital pen output, all signals are synchronized using global timestamping.

5. Transcription

Speech corpora are most useful if they come with transcription. The AMI Meeting Corpus includes high quality, manually produced orthographic transcription for each individual speaker, including word-level timings that have derived by using a speech recognizer in forced alignment mode. Spelling is British, indicates what the speaker said without correcting “errors”, and uses reduced lexical forms such as “gonna”. Markup includes indications of mispronunciations and neologisms, some nonverbal noises such as laughter, and punctuation indicating turn structure. The transcribers worked from a set of written guidelines in a system that includes two separate passes to ensure high quality whilst giving early approximate results. In the first pass, transcribers were allowed to mark their work as uncertain and were provided with “empty” segments that had been identified automatically by applying a simple energy-based technique to segment silence from speech for each meeting participant. Second pass transcribers then listened thoroughly, resolving problems with the existing transcription and ensuring that any speech outside the identified segments was included. As a final step, a validation script was run to find unrecognized spellings and transcription codes.

6. Annotations

As well as orthographic transcription, large portions of the AMI Meeting Corpus have been annotated for a range of phenomenon that include both linguistic properties and behaviours visible on the video record. Simply agreeing what kinds of annotations to create and defining formal coding schemes was a major undertaking. In order to ensure fitness for purpose, our schemes were developed by working groups that included people with experience managing such efforts, along with the researchers planning to use the annotation to develop browser components and theorists who understand in the particular phenomenon being labeled. Where possible, the groups used or extended existing schemes in order to make it easier to pool data across corpora, but not at the expense of a good fit to the data or to the needs of our application developers. The annotations cover the following phenomena:

- *Named entities* using an adaptation of NIST’s “1999 Named Entity Recognition Task Definition” manual (Chinchor et al., 1999). For the remote control design data, the adaptation excludes some irrelevant categories and includes extra ones that relate to the task,

such as references to colours, shapes, materials, and artefacts in the room.

- *Dialogue acts* representing speaker intentions, where the acts cover all words in the transcription. The set of categories used is designed to separate social acts and quasi-acts (such as backchannels and unsuccessful attempts to take the floor) from ones that move the team towards task completion and from “reflexive” acts (West, 1996) in which the team discusses how to discuss their work. It also divides information exchange from comments and discussion of actions the team might take. As well as categorizing the dialogue acts, the annotation includes information about acts that were addressed to individuals instead of the entire group and some rudimentary argument relations between pairs of acts.
- *Topic segmentation* that hierarchically decomposes the transcription into topics and subtopics at a depth of up to three levels. Each topic segment contains a label; for the remote control design teams, the labels are chosen from a closed set, and for all of the data, the annotators used standard labels for things like meeting openings and closings, discussing the agenda, and purely social conversations.
- *Abstractive summaries*, consisting of free text giving a general abstract of the meeting plus specific explanations of the decisions made in it, any problems or issues encountered, and any actions set. Each of these headings was addressed in up to 200 words.
- *Extractive summaries*, in which the annotators identify the parts of the meeting that support the contents of an abstractive summary. This annotation is a many-to-many linking between individual dialogue acts from the transcription and individual sentences in the abstractive summary; roughly 10% of the dialogues are linked to at least one sentence.
- *Limited head gestures* that show communicative intentions, such as nodding.
- *Limited hand gestures* used in turn-taking and deixis, such as pointing at people or objects in the room.
- *Movement around the room*, showing, for instance, whether the meeting participants are sitting in their seats or standing at the electronic whiteboard.

- *Face, mouth and hand location* on some video frames, such as is needed for developing tracking software.
- *Coarse gaze annotation* indicating, for instance, whether a participant was looking at the whiteboard or another participant.

These annotations have at least two purposes: they help the analyst to explore and understand the data set, and they can be used to develop software components that annotate new material automatically using machine learning.

7. NXT: the technology behind a killer corpus

Wikipedia (Wikipedia contributors, 2006) defines a killer application as “a computer program that is so useful or desirable that it proves the value of some underlying technology, such as a gaming console, operating system, or piece of computer hardware.” The AMI Meeting Corpus is certainly intended to be desirable, not just for the signals and transcription — although these themselves have been enough to attract the speech community — or for individual annotations that could be used for individual tasks, such as named entity recognition, but also for the range of annotations available. Although some previous corpora, such as Switchboard (Godfrey et al., 1992), have accrued quite a few different kinds of annotation over the years, it is difficult to find and obtain them, much less figure out how they relate to each other beyond the simple facts of their start and end times. If the AMI Meeting Corpus is a killer corpus, then the NITE XML Toolkit (Carletta et al., 2003) is the technology whose value it proves, since it uniquely provides the infrastructure required for data of this type. Our discourse annotations were created using NXT, with transcription and time-stamped labellings imported from ChannelTrans (International Computer Science Institute, nd) and EventEditor (Sumeç, nd), respectively.

There are three special properties that NXT has which were essential to the success of our corpus annotation. The first is that it represents the structural relationships between annotations explicitly, including the ability to search using these relationships. Outside NXT, annotations are often simple time-stamped labellings of the signals. In the AMI corpus’ topic segmentation, segments are spans that inherit their timings from the words underneath. This leaves no ambiguity about what words the annotator considered to be in or out, unlike for the simpler representation. It also makes clear that decisions about the

boundaries were based solely on the speech and not, say, on communicative head gestures. Although the difference may sound trivial, a time-based representation makes it harder to replicate previous work and understand what data is fit for what purpose. Also, in the AMI corpus, extractive summaries do not just pull out segments of a meeting by time, but point to the dialogue act (and from there, the words) to be extracted, as well as any sentences in the meeting abstract that relate to the extracted segment. This arrangement makes the most sense of the data, but only NXT can support it.

NXT's second advantage is that it stores data as XML with "stand-off" out-of-file links between elements, which allowed us to divide our corpus annotations among multiple files. In the NXT representation of a corpus, one "metadata" file serves both as a formal description of the annotations available, including how they relate to each other, and as a catalogue explaining where to find signals and annotations on disk. Annotations of the same type that together span the length of a dialogue or meeting are kept together in one file. Where the annotations naturally form a tree, they can be placed in one file for which the XML structure matches the natural structure, making it easier to inspect the stored data. Without this sort of file management, we would not have been able to create such a large and heavily annotated corpus over what was a relatively short time, because we would have had difficulty integrating annotations created at different sites. In addition, since each individual file is simple, it is relatively easy to translate annotations from other formats and to process them for use by external tools such as machine learners and part-of-speech taggers.

Finally, NXT contains some configurable GUIs for common annotation tasks — figure 2 shows a screenshot of the named entity annotation tool as used for this corpus — and a set of libraries that makes writing tailored annotation and browsing tools easier. The libraries handle basic display properties such as providing a default rendering for transcription, synchronizing timed annotations against multiple signals as they are played, and highlighting search results, leaving the developer free to concentrate on the rest of the user interface. In any large scale data creation, tools are important. Even quite small quirks, like an extra mouse click, will make costs escalate and quality fall. People tend to stick to the tools they know even if the tool does not really fit the task, which limits the usefulness of what they produce. Using NXT's libraries allowed us to develop tools quickly for the annotations our researchers actually wanted, not the ones existing tools could produce. Without NXT, for instance, we would not have hierarchical decomposition for our topic segments or links between our extractive and abstractive summaries.

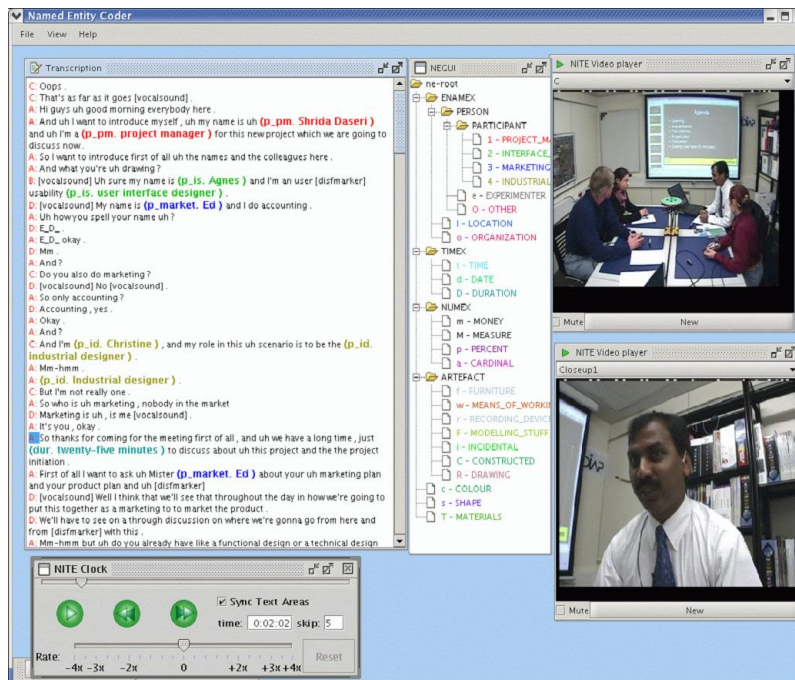


Figure 2. Screenshot of NXT’s named entity annotation tool, configured for use on the AMI Meeting Corpus.

8. Public release

It takes real effort to collect, transcribe, and annotate language corpora, and people who have undertaken this work naturally feel protective of their data. In the past, some corpora have been embargoed for “family” use only until they were quite old, and even ones that have been released can have quite restrictive licensing. This makes it difficult to compare different approaches to building the same technology.

Our philosophy is that because our data has value, we must get it out to the community as quickly as possible. For this reason, we have chosen a “ShareAlike” license from Creative Commons (Creative Commons, nd). This form of licensing is intended to create an environment in which people freely share what they have created. The corpus license allows users to copy, distribute, and display the data for non-commercial purposes as long as the AMI Consortium is credited. However, if the user wishes to distribute anything derived using the corpus, that can only be done under the same license as the original data. Although Creative Commons licensing is relatively new for data sets, it is similar to the GNU General Public License (Free Software

Foundation, nd), which is already in common use for research software, including NXT. The license does not bar us from distributing the data under other terms as well, but it does allow us to give the data away to the widest group possible without fear of being exploited.

Our main way of distributing the corpus is through the website <http://corpus.amiproject.org>. At the website, anyone can look at the signals for one meeting and read extensive documentation. After registration, users can browse meetings online using SMIL presentations, download their chosen data, and participate in a discussion forum. Registration is simple and free. Everything that has been released is on the website, apart from the full-size videos. These are too large for download, but the website gives a contact for ordering firewire drives that contain them, priced at the cost of production. In addition to the website, we have also produced a “taster” DVD that includes everything for a single meeting - signals, transcripts, and every available annotation, including samples of some types that have not yet formed part of the public release. The DVD can currently be ordered for free from the website.

Under the ShareAlike licensing, we expect researchers will produce new annotations that they wish to share. Although nothing in the licensing compels us to help distribute these, or others to distribute them through us, it would be useful if all the annotations could be found in the same place and in the same format. We intend to set up an annotation bank, starting with the current annotations, that will accept contributions. Since our annotations relate to each other structurally, we see NXT as the reference storage format for this corpus when used by the wider community, just as it is within the consortium. As far as we know, this has never been attempted on this scale before, so in the early days, we will be considering how best to make this happen.

9. Discussion

The AMI Meeting Corpus has required substantial investment. Although we produced it primarily to suit the AMI Consortium’s needs, it is already having an impact on the broader research community through its use in external evaluations such as NIST’s series on “rich transcription” (National Institute of Standards and Technology, 2006). The earliest users outside the project are interested solely in the signals and transcription, since these were made available first. As more people begin to use the data for more tasks, the corpus has the potential to change the nature of language engineering by challenging our current conceptions of what the field is about.

Statistical language processing is inevitably driven by the data that is available, and it is common to hear the complaint that too much work is done on too few corpora simply because they are there, without worrying about whether the results will be useful for anything. If nothing else, the AMI Corpus has a role to play as additional material. For some tasks, the data is of course very challenging, and we need to make allowances for that when we measure results. It is easier to work on diminishing returns for well-established data sets, but, one would hope, less rewarding, both for the researchers involved and for the community as a whole. However, the corpus is more than new fodder for old techniques. The range of annotations available is unique, and researchers will inevitably derive features from them that they have never considered using before, just to see if they will help for whatever they are doing. Meanwhile, the sheer range of people working on the corpus should create a stimulating environment. Common data will give common ground to researchers from different fields, helping us to see new problems and avoid old traps. We particularly hope that the involvement of groups outside the language and speech community will make it easier to move towards the system evaluations that we all know are necessary for real progress.

For these reasons, we expect the AMI Meeting Corpus to become an invaluable resource to the broader research community. The AMI Consortium will continue working together in the newly-funded AMIDA project, and therefore intends both to maintain the corpus and to take an interest in its growth. We are happy to have it used, and hope that it will attract researchers with other approaches to our own problems, but also be taken in new and unforeseen directions.

Acknowledgements

I thank the large number of researchers involved in the creation of the NITE XML Toolkit, both during the NITE project and afterwards, and in the collection, transcription, and annotation of the AMI Meeting Corpus, without whom these more personal reflections would not be possible. This work was funded by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811).

References

- Anderson, A. H., 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: 1991, 'The HCRC Map Task Corpus'. *Language and Speech* **34**(4), 351–366.
- Carletta, J., S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann: 2003, 'The NITE XML Toolkit: flexible annotation for multi-modal language data'. *Behavior Research Methods, Instruments, and Computers* **35**(3), 353–363.
- Chinchor, N., E. Brown, L. Ferro, and P. Robinson: 1999, '1999 Named Entity Recognition Task Definition Version 1.4'. Online at http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1.4.pdf; accessed 6 Dec 06.
- Creative Commons: n.d., 'Creative Commons'. Online at <http://creativecommons.org/>; accessed 11 Dec 06.
- Free Software Foundation: n.d., 'GNU General Public License'. Online at <http://www.gnu.org/copyleft/gpl.html>; accessed 11 Dec 06.
- Godfrey, J. J., E. C. Holliman, and J. McDaniel: 1992, 'SWITCHBOARD: Telephone speech corpus for research and development'. In: *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.* pp. 517–520.
- International Computer Science Institute: n.d., 'Extensions to Transcriber for Meeting Recorder Transcription'. Online at <http://www.icsi.berkeley.edu/Speech/mr/channeltrans.html>; accessed 11 Dec 06.
- McGrath, J.: 1984, *Groups: Interaction and Performance*. Englewood Cliffs, NJ, USA: Prentice-Hall.
- National Institute of Standards and Technology: 2006, 'Rich Transcription 2006 Spring Meeting Recognition Evaluation'. Online at <http://www.nist.gov/speech/tests/rt/rt2006/spring/index.html>; accessed 11 Dec 06.
- Sumec, S.: n.d., 'Event Editor'. Online at <http://www.fit.vutbr.cz/research/grants/m4/editor/index.htm.cs.iso-8859-2>; accessed 11 Dec 06.
- West, M.: 1996, 'Reflexivity and work group effectiveness: A conceptual integration'. In: M. West (ed.): *The Handbook of Work Group Psychology*. John Wiley, pp. 555–579.
- Wikipedia contributors: 2006, 'Killer application — Wikipedia, The Free Encyclopedia'. Online at http://en.wikipedia.org/w/index.php?title=Killer_application&oldid=88980227; accessed 21 Nov 06.