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

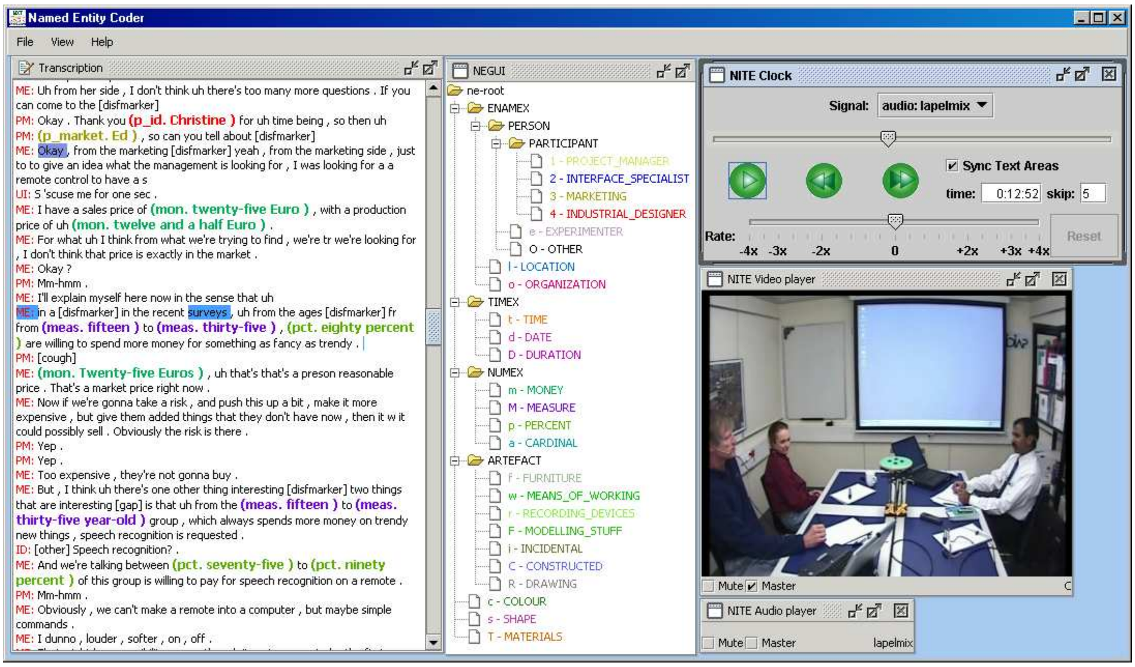
Jean Carletta

University of Edinburgh
2 Buccleuch Place
J.Carletta@ed.ac.uk

Abstract

Creating the AMI Meeting Corpus was an ambitious endeavour, probably more ambitious than the people who first thought of it realize even now. It contains 100 hours of meetings captured using a whole host of 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 keynote speech, I describe the data and talk about what it was like to manage the process of creating it, distributed over six sites. If this is "killer" data, that presupposes a platform that it will "sell"; in this case, that is the NITE XML Toolkit, which is the only viable way to create, store, and browse annotations for the same base data, at least where creation is on this large a scale or there are structural relationships among the annotations. I'll end with details of the data's imminent public release, but on the way I'll take in what happens when you mix engineers and psychologists, sing the praises of Wikis, and argue that this data has really been in the making since the 1950's.

Three rooms, 100 hours of meeting data.



ME: Uh from her side , I don't think uh there's too many more questions . If you can come to the [disfmarker]
PM: Okay . Thank you (p_id. Christine) for uh time being , so then uh
PM: (p_market. Ed) , so can you tell about [disfmarker]
ME: Okay , from the marketing [disfmarker] yeah , from the marketing side , just to to give an idea what the management is looking for , I was looking for a a remote control to have a s
UI: S 'scuse me for one sec .
ME: I have a sales price of (mon. twenty-five Euro) , with a production price of uh (mon. twelve and a half Euro) .
ME: For what uh I think from what we're trying to find , we're tr we're looking for , I don't think that price is exactly in the market .
ME: Okay ?
PM: Mm-hmm .
ME: I'll explain myself here now in the sense that uh
ME: in a [disfmarker] in the recent surveys , uh from the ages [disfmarker] fr from (meas. fifteen) to (meas. thirty-five) , (pct. eighty percent) are willing to spend more money for something as fancy as trendy .
PM: [cough]
ME: (mon. Twenty-five Euros) , uh that's that's a preson reasonable price . That's a market price right now .
ME: Now if we're gonna take a risk , and push this up a bit , make it more expensive , but give them added things that they don't have now , then it w it could possibly sell . Obviously the risk is there .
PM: Yep .
PM: Yep .
ME: Too expensive , they're not gonna buy .
ME: But , I think uh there's one other thing interesting [disfmarker] two things that are interesting [gap] is that uh from the (meas. fifteen) to (meas. thirty-five year-old) group , which always spends more money on trendy new things , speech recognition is requested .
ID: [other] Speech recognition? .
ME: And we're talking between (pct. seventy-five) to (pct. ninety percent) of this group is willing to pay for speech recognition on a remote .
PM: Mm-hmm .
ME: Obviously , we can't make a remote into a computer , but maybe simple commands .
ME: I dunno , louder , softer , on , off .

Signal: audio: lapelmix
☑ Sync Text Areas
time: 0:12:52 skip: 5
□ Mute ☑ Master
□ Mute □ Master lapelmix

NXT's named entity coder — one of the many tools used to annotate and view AMI data.