Sharing and Standards

Brian MacWhinney
Psychology -- Carnegie Mellon

http://talkbank.org

Funding from NSF - TalkBank, Diver, SCOTUS
Funding from NIH - CHILDES, Phon, AphasiaBank
Why video?

• Behavior emerges from the nonlinear intersection of forces on 7 time scales in the current moment.
• Experiments only control a single time dimension.
• Different social objects have different patterns of mesh to time frames.
• Video captures interactions in the moment and it can capture most objects.
8 standard research methods

- Microanalysis (CA, linguistic, ethology)
- Microgenetic analysis
- Group and treatment comparisons
  - Same consultant, different communities …
- Error analysis
- Diffusion analysis - Hall museum, Whiten chimps
- Longitudinal studies
- Large sample analysis
- Dynamic modeling
TalkBank Standards

• Standard transcription format that merges 10 common formats (CA, ISL, SALT, DT, AG, MT, Columns, SyncWriter, Phon, Praat)
• XML Schema definition for format translation
• XML verification and roundtrip
• Suite of analytic tools, transcription tools
• Linkage to media and tools for linkage
• Codec standardization
• Streaming media server, locally deployable
• Metadata: OLAC, OAI, ISBN
Technology is there

• Standards are available. Programs are available. They are stable and tested.
• Streaming is solved. Soon, Grid will be available -- Bennett’s SuperLab
• Collaborative commentary, event linkage are available.
  • WebDiver - Stanford
  • ProjectPad - Northwestern
  • TalkBankViewer, CLAN WebData - CMU
Data Sharing

- The CHILDES model.
- Data sharing not crucial for established researchers. It is crucial for the field.
- Google could not have happened without open data.
- Raw data sets are infinitely rich. No one can be scooped.
- Tenured faculty have a responsibility to share data, within IRB guidelines.
- Federal agencies have a responsibility to promote data sharing.
Transcripts linked to media

```plaintext
@Begin
@Transcriber: Tim Koschmann Last revision 012000 Johannes Wagner
@Participants: Be Betty, No Norman, Co Coach, Mar Maria, May, Jen Jenny
Lill, ? unidentified Person, Ps Pauses
@Dependent: yes
@Filename: MyTheory.ca Moviefile MyTheory.mov
@Time: 6 minutes
@Contents: fragment of tutor-group discussion
@Comment: numbering is by TCUs and pauses, not lines

Be: See what it said in here (. ) in - my theory (hhh) •
(0.4)
?
[khu-] (.hhh)
Be: [about this amnesio (. ) dysnomic aphasia, •
(0.3)
Be: u:hm (it) says the cause of lesion is usually deep in
(temporal lobe just like Kathy was saying presumably
interrupting connections of sensory speech areas with the
```

TalkBank Groups, Areas, Topics

- Child Language (CHILDES, PhonBank)
- Conversation Analysis (MOVIN, CA)
- SLA (SLAWeb, FLLOC, LIDES)
- Legal (Supreme Court)
- Aphasia (AphasiaBank)
- Classroom Discourse
- Linguistic Exploration
- Gesture (FORM, Elan, SuperLab)
- Sociolinguistics (SLX)
The CHILDES Model

• Data sharing is simply assumed.
• Over 2000 published articles based on CHILDES data
• Groups in individual languages (Chinese, Japanese, Dutch …)
• New tools: MOR, GRASP, PhonBank, Browser
• Careful treatment of Human Subjects
• Model for NSF CyberInfrastructure
Classroom Database

- TIMMS - six countries
- PBL - Koschmann, LeBaron
- Gravity - TERC
- Science Museum
  - Atmospheric light diffusion - Rahm
  - Electricity generation - Crowley
- Dresden - SLA English, French, Czech
- Grimshaw - Oral Defense
- Greeno - Garden Plot, numerical series
Classroom Database

- Numerical displays - Sfard, McClain, Cobb
- Lehrer - Carmen Curtis and quilt patterns
- Lectures -- MacWhinney gesture analysis
- Roth -- map lecture
- Stevens -- professional dialogs
- WorkGroup -- MacWhinney, CMU groups
- Moskovitch -- bilingual classroom, math
- Horowitz -- reading exercises
Tutorial Interactions

- Circle - Physics
- Frederiksen - statistics
- Graesser - statistics
- DISPEL - collaborative problem solving
TalkBank Lite

- No data-sharing
  - Medical College of University of Southern Denmark
  - University of Antwerp
  - University of Pittsburgh
- No transcript -- naked video
  - Lehrer year long study
  - Informedia Nursing Homes
  - McCune video
Example of TalkBank “Secondary” Analysis

- James Greeno, Brian MacWhinney, and Carla van der Sande
- Learning as the construction of mental models that explain device representations.
- Humans represent (explain) devices through
  - Perspectival embodiment
  - Spatial imagery
Gabriel’s Model
Dad’s Model
Gravity and Pprims
Garden Plots

Sally’s backyard is 40 feet wide by 72 feet long, and it is structured so that a central rectangle of grass is surrounded by an even border of flowers. The area of the border is \( \frac{5}{12} \) of the area of the whole garden. If Sally wants to be sure to walk her dog 1/4 mile, how many laps will she have to take around the grass?
Implications

- Learner models are fragmentary.
- Working devices must have full linkage.
- These linked systems can be annotated by embedded causal/perspectival links.
- Teachers can facilitate linkage formation by retracing perspectival links with focus on:
  - Proceduralized subcomponents
  - Missing links between levels
Cherry Picking?

- We are constructing complete propositional tree analysis.
- Coding must be reliable.
- Model must apply across all TalkBank datasets.
- Model will be elaborated in detail for Physics and Chemistry in NSF Pittsburgh Science of Learning Center
Evidential Database

- Claims pointing to evidence through media.
- Online collaborative commentary.
- Additional database components for teacher training, parent training, and educational standards. Also with collaborative commentary.
- Supporting pointers to PDFs, Picts of
  - Articles, Legal precedents, evidence
- Publishable as special issues
  - JLS, JOC, Discourse Processes, Cog&Inst
Conclusions

• Infrastructure is ready.
• We need more data, more awareness.
• Data sharing will provide resources for students, teachers, parents.
• Data sharing is not crucial for individual researchers. It is crucial for the emergence of coherence in the field.