A Transcript-Video Database for Collaborative Commentary
in the Learning Sciences

Brian MacWhinney, CMU, macw@mac.com

Researchers in the Learning Sciences have acquired great sophistication in the use of video recordings as ways of understanding instructional interactions. Conference and lecture presentations on instructional processes are now typically grounded on the analysis of authentic video material. For example, at the 2002 meeting of the American Educational Research Association (AERA), there were 44 scientific panels and symposia that relied on analysis of classroom video. Video is also used in teacher training programs (Derry, in press; R. D. Pea, 1999) and materials illustrating proposed nationwide educational standards (Daro, Hampton, & Reznick, 2004). The field also enjoys a great range of high-quality tools for the analysis of video interactions. Systems such as N-Vivo (www.qrsinternational.com), DIVER, TransAna (www.transana.org), ATLAS.ti (www.atlasti.com), Elan (www.mpi.nl/tools/elan.html), MacShapa (Sanderson & Fisher, 1994), CLAN (childes.psy.cmu.edu), VideoNoter/C-Video (Roschelle, Pea, & Trigg, 1990), Ethnograph (www.qualisresearch.com), Anvil (www.dfki.de/~kipp), Orion (Baecker, Fono, & Wolf, 2006; Goldman-Segall & Reicken, 1989), ePresence (Baecker, Fono, & Wolf, 2006), Informedia (Wactlar, Christel, Gong, & Hauptmann, 1999), and VideoPaper (Beardsley, Cogan-Drew, & Olivero, 2006) are allowing researchers to produce large quantities of well-analyzed video interactions.

Despite the high quality of video analysis methodology, the large quantity of data being produced, and the centrality of video to the scientific study of learning and instruction, there has not yet been a community-wide acceptance of the importance of a shared database of instructional interactions. There has been extensive discussion of the formation of collaboratories for the study of instructional interactions (Baecker, Fono,
Wolf, 2006; Edelson, Pea, & Gomez, 1996). However, without a general method for sharing data across projects, collaboratories are limited to datasets collected from single projects (Abowd, Harvel, & Brotherton, 2000). However, many of the most interesting questions in learning and instruction involve comparison between alternative teaching frameworks and situations. This type of diversity in the database can best be achieved by having data from many different laboratories and groups channeled into a uniform, but distributed database.

Development of a shared database is a crucial next step in the maturation of the Learning Sciences as a scientific discipline. In Genetics, projects such as the Human Genome Project (www.ornl.gov/hgmis), GenMapp (www.genmapp.org), or Protein Map (Aisenman & Berman, 2000) are now storing all published genetic sequences in forms that are open to analysis and data-mining through the web. In fact, gene sequences are not accepted for publication until they have been entered in these systems. In Paleontology, museums worldwide preserve fossils whose specific physical structure, radiological dating, and stratificational location are crucial to our reconstruction of the history of life and the earth. Electronic records and scans based on this evidence are now being made available electronically (www.ucmp.berkeley.edu/pdn/) for deeper analysis and data-mining. Internet databases are now fundamental to progress in Astronomy (van Buren, Curtis, Nichols, & Brundage, 1995), Physics (Caspar et al., 1998), Economics, Medicine, History, Political Science, Experimental Psychology, Linguistics, and other sciences.

To address this need, the TalkBank Project has begun an effort to construct a shared database for the Learning Sciences. TalkBank (http://talkbank.org) is an international collaborative effort that has been building a web-accessible database for spoken language
interactions. All of the video and audio media in TalkBank are fully transcribed and each transcribed utterance is linked directly to the corresponding segment of the media. The media and transcripts can be downloaded from the web. Users can also open a browser window, scroll through transcripts, play back the corresponding audio or video, and insert commentary regarding their analyses. The current TalkBank database has large collections of data in the areas of child language (CHILDES), aphasia (AphasiaBank), second language learning (SLABank), bilingualism (LIDES), formal meetings, and spontaneous conversational interactions (CABank and MOVIN).

A shared database for the Learning Sciences will have some interesting features unique to this area. It will be important to develop a taxonomy of educationally-relevant activities, events, and interaction types that can serve as metadata for coding and retrieval. It will also be important to supplement video records with additional ethnographic materials such as diaries, notebooks, drawings, and class records. However, the most powerful feature of a shared database in the Learning Sciences will certainly be its availability to collaborative commentary. The idea of scientific collaboratories has been developed and discussed elsewhere in this volume. With the context of collaboratories, projects such as Orion (Goldman, 2006), DIVER (R. Pea, in press), and WebCast (Baecker, Fono, & Wolf, 2006) have shown how a group of educational researchers can work together to analyze interactions and evaluate competing interpretations. However, for the process of collaborative commentary to work as a general model for the learning sciences, it must be linked to a commitment to the process of data-sharing. What is unique about the TalkBank Project is not its emphasis on collaborative commentary, but rather its emphasis on data sharing. However, the greatest
value for scientific progress arises when data sharing is joined with collaborative commentary.

**Data Sharing**

The goal of TalkBank is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of human and animal communication. The concept emerged from two ongoing initiatives that had already proven important to their respective user communities. The first is the Linguistic Data Consortium (http://ldc.upenn.edu) that has published some 288 large corpora over the past decade. The second is the CHILDES system (http://childes.psy.cmu.edu) that has constructed a database of 150 corpora of parent-child interactions in 20 languages. The data-sharing model for the new TalkBank project is based on the model from the CHILDES project (MacWhinney, 2000).

Having reviewed best practice in 12 very different research areas studying communicative interaction, Talkbank has identified these 7 shared needs:

1. guidelines for ethical sharing of data,
2. metadata and infrastructure for identifying available data,
3. common, well-specified formats for text, audio and video,
4. tools for time aligned transcription and annotation,
5. a common interchange format for annotations,
6. network based infrastructure to support efficient (real time) collaboration, and
7. dissemination of shared data, tools, standards and best practices to the research community.
Data on Instructional Interactions

Materials in our initial collection of video studies of instructional interactions include:

1. Studies of problem-based learning in medical school education from Tim Koschmann and Curtis LeBaron (Koschmann, 1999).
3. Recording of students engaged in group problem-solving in algebra from James Greeno and Carla Van Sande.
4. Materials from the TIMMS study (Stigler, Gallimore, & Hiebert, 2000) comparing math and science instruction in Australia, Czechoslovakia, Hong Kong, Japan, and the Netherlands.
5. College lectures on research methods in psychology from Brian MacWhinney.
6. A lecture on map reading to a 6th grade class contributed by Wolf-Michael Roth.
7. A discussion of a unit on camels in a 5th grade class contributed by Rosalind Horowitz.
8. A comparison of classroom, business, and meeting contexts contributed by Reed Stevens.
9. Teachers’ discussion of the basis of gravity from Beth Warren.
10. Science museum visit materials from Irene Rahm and Kevin Crowley.
11. Dyadic tutorial sessions on the f-ratio in the analysis of variance from Carl Frederiksen.
13. Dyadic tutorial sessions on how to play a video game from Nikolinka Collier.


In addition, the related CHILDES (Child Language Data Exchange System) database (http://childes.psy.cmu.edu) contains several major video studies of more informal learning in the home in English, German, Japanese, Spanish, Cantonese, and Thai.

In an early attempt to promote collaborative commentary, TalkBank promoted the creation of a CD-ROM for a special issue in the Journal of the Learning Sciences (Sfard & McClain, 2002). This CD-ROM contains articles commenting on two lessons on graphs in a 7th grade classroom. The PDF files for these articles contain links that replay the relevant video. In addition, there is a demonstration transcript that serves as a compendium of commentary on particular analyses.

We also hope to provide streaming video access to three large longitudinal classroom corpora. One corpus, from Carolyn Maher and her associates at Rutgers, contains 3000 hours recorded over a span of 12 years tracking the math learning of a group of 15 students. Another, from Rich Lehrer and Carmen Curtis, records a year’s worth of integrative geometry lessons from a 3rd grade classroom. The third, from Juliet Langman, compares alternative formats for bilingual classrooms.

**Related Corpora**

The TalkBank database contains a wide range of materials outside of this particular initial corpus for learning contexts. Among the major datasets in these other areas are:

1. CallFriend phone conversations in English, Spanish, and Japanese, provided by the LDC and transcribed by TalkBank.
2. The Santa Barbara Corpus of Spoken American English (SBCSAE), including conversations between college friends, lectures, and meetings.


4. European political television programs and other components of the MOVIN database from Johannes Wagner.

5. Informal interview materials from a special issue of the Journal of Communication.

6. A complete collection of all the oral arguments of the last 30 years of the Supreme Court of the United States (SCOTUS). This enormous dataset is currently being formatted into the TalkBank Schema and linked to the digitized media.

7. Over twelve corpora from second language learners and bilinguals.

8. A new database of 12 corpora from aphasic speakers.

**Ethics for Data-Sharing**

Public sharing of data over the web brings with it a variety of challenges regarding participant rights and professional ethics. These issues have been an ongoing topic of discussion within the TalkBank communities. The current result of this process is a set of ethical and practical guidelines adopted for all TalkBank data sets, described at http://talkbank.org/share and available for use beyond the TalkBank project. The centerpiece of this approach is the idea that participants can opt to provide releases for the use of their data at any one of eight different levels. The lowest level of protection allows for full web access to transcripts and video with no attempt at anonymization.
Higher levels of protection add anonymization in transcripts and media, password protection, and finally no access but only archiving for the future. The choice of an appropriate level for a given dataset is decided first by the Human Subjects review process at each institution and then by the participants themselves. In addition, TalkBank discourages any use of the data that is critical of the performance or motives of individuals recorded in the interactions. Groups that require further privacy and respect considerations include indigenous groups, speakers of endangered languages, clinical subjects, subjects in psychiatric treatment, and classroom teachers.

**Infrastructure for Annotation**

Researchers working with video records from learning contexts need support for a wide range of transcription, editing, and analysis functions. The TalkBank project has supported the development of a variety of annotation tools, including the AG ToolKit (agtk.sourceforge.net) (Maeda, Bird, Ma, & Lee, 2002), TransAna (www.transana.org), and Transcriber (www.ldc.upenn.edu/mirror/Transcriber/). However, the annotation tool that is most fully compatible with other TalkBank programs and which provides best integration with the database is CLAN (childes.psy.cmu.edu/clan). CLAN allows the transcriber to link directly to audio and video material, rewind the material, use CA transcription, and analyze files using a variety of search programs. By inserting metadata fields in CLAN, the contents of all TalkBank files can be published over the web in conformity with the standards of the OLAC (Open Language Archives Community) project (www.linguistlist.org/olac).
Interchange Formats for Data Annotation

The greatest challenge facing TalkBank has been the need to bring hundreds of corpora created in diverse ways into conformity with a common standard. The first step in this direction involved the specification of a proper XML Schema for the CHAT transcription system. The system involves three major steps. In the first step, the ANTLR parser generator creates a parse tree that is converted to a JAXB tree that is then serialized into XML. JAXB is Sun’s data binding framework that generates Java code for specialized DOM construction, validation, and serialization. In the second step, XSLT outputs CHAT. In the third step, a modified version of Unix DIFF compares the original CHAT with the converted CHAT and reports differences for correction. Once a corpus passes through this process with no errors, it is included in TalkBank. CHAT versions are zipped so that users can download complete datasets and the XML versions are shipped to the server (http://xml.talkbank.org) to support online transcript and media browsing.

Browsable Transcripts

Once the transcripts are in the TalkBank XML format, they can easily be rendered as HTML pages. The Java WebStart program called TBViewer then allows users to view transcripts over the web and play the underlying audio or video by clicking on utterances. After clicking on the first utterance, the user can decide to either stop playback or to allow the viewer to play back the whole transcript by highlighting each utterance as it is played. The transcript can be scrolled and paged to allow for full control of playback over the web. The playback application is installed by clicking on a link at the TalkBank page that then runs Java WebStart.
The playback facility relies on the use of hinted video and audio from Apple’s QuickTime Streaming Server. Figure 1 illustrates a typical page in the TBViewer. For audio-only corpora, the QuickTime window displays the utterance and plays the sound without any video.

![Figure 1: The TBViewer window with transcript and video](image)

**Collaborative Commentary**

We can define collaborative commentary as the process by which a research community engages in the interpretive annotation of electronic records. The goal of this process is the evaluation of competing theoretical claims. To achieve clear connections
between data and theory, commentators need to link their comments and other evidentiary materials to specific segments of either transcripts or electronic media. Now that researchers have access to browsable corpora representing various learning contexts, we can begin to think about how to implement the process of collaborative commentary.

To illustrate the goals of this process, let me use some of my own explorations through the database as the example. I was interested in exploring evidence for the neo-Vygotskian claim (Nelson, 1998) that word meanings are shaped through communicative interactions. While browsing through online media at the CHILDES childes.psy.cmu.edu site, I located several instances of videos of mother-child book reading in the McMillan and Rollins corpora. In these interactions, mothers help children turn the pages and name the animals or objects in the pictures. In some cases, children call the pictures by the wrong name. I wanted to see whether mothers would use these errors as opportunities to provide corrective positive feedback (MacWhinney, 2005b). For example, if the child calls a bear a “doggie,” the mother should respond, “no, that’s a bear, not a doggie.” Then, the child might engage in self-correction by saying “a bear.”

To initiate the process of collaborative commentary, I then write up a short summary of my analysis. In reality, I have not yet conducted this step, although an initial attempt can be found in MacWhinney (2005a). I would then want to make this analysis available in three ways. First, I would like to post my claim to some discipline-based commentary space on the web. Second, I would want to make sure that others who view the relevant segments from the Julie and Rollins corpora could see that I have provided detailed interpretive commentary regarding certain specific segments. Finally, I hope to receive feedback from other researchers regarding my interpretations and arguments.
To further illustrate how this process can work, let us consider a second example from the area of mathematical learning. In this fictive example, Harriet Keck is a developmental psychologist specializing in children’s concepts of number. She and her colleague Robin Clark are both interested in understanding how children solve problems such as $3 + 4 = ?$. Keck believes that children solve the problem in an internal mental model and then read out the solution to their fingers. Clark believes that children use their fingers to form external representations of the addends and then count their fingers visually. Keck’s model predicts that children will count directly across the fingers, whereas Clark’s model suggests that children will begin with placing one addend on each hand separately.

To explore this issue, Dr. Keck uses metadata search tools to find video cases in TalkBank format involving “four-year-old children AND counting”. Exploring these videos using the TalkBank Viewer, she finds that 70% of their gestures support her theory, whereas only 30% are in line with Clark’s account.

Over the next several weeks, Keck and her colleagues use CLAN to link each case of finger counting to comments that also point to a brief report summarizing her conclusions. Not surprisingly, Clark disagrees with Keck’s conclusions and responds by reinterpreting the same video cases that Keck has just analyzed. His analysis points to several counter examples that do not fit Keck’s theory. He also argues for including trials that have no overt finger counting in the denominator. Keck, in turn, responds to Clark’s criticism by asserting the gestures he has coded are inadvertent hand-movements and revises her paper to anticipate his objection. Keck and students submit their revised paper, including the video data, to the on-line edition of Cognitive Development. One of
the reviewers has a question on whether the authors have properly categorized a set of gestures from one of the videos. Keck responds with a close analysis of the gesture in question using a fine-grained analysis of the actual hand-movements. The reviewer is convinced by her response and the paper is published with links to the video data and analysis.

Although this scenario may seem a bit futuristic, it is not very different conceptually from forms of collaborative commentary we have already produced. One example is a special issue of the Journal of the Learning Sciences (Sfard & McClain, 2002) that focuses on learning about graphs and numerical distributions in a 7th grade classroom. The difference is that in this new framework, analyses will be directly linked to the data, rather than hidden within PDFs. Moreover, in this new framework, analyses will be directly accessible from browsers.

**Infrastructure for Collaborative Commentary**

The TBViewer is now being configured to support this process. In the browser window, pencil icons next to utterances represent the commentary field. When the user clicks on this field, a separate commentary window opens up. This commentary window allows the researcher to create the following elements:

1. A brief summary of the claim or analysis relevant to the current utterance or utterance sequence.
2. Typing of the claims and analysis into specific categories.
3. Explanations of the evidentiary role of the texts and media being referenced.
4. Links to other texts or claims that are relevant to the current claim.
5. Links to external web content, including material (HTML, PDF, Word) that presents the proposed analysis more fully.

6. Embedded HTML code.

Once this material has been entered into the commentary field, it can be redisplayed through the TBViewer facility. The QuickTime window echoes the comments that have been entered in the commentary database as streaming playback progresses. If a given segment has more attached commentary than can be displayed in the reserved segment of the QuickTime window, the window will have a final line listing the number and size of the comments that cannot be displayed.

**Naked Media**

Because TalkBank transcripts are subject to ongoing modifications, reference to line numbers is not stable. A more reliable method links commentary to time points in the media. The idea of linking commentary to media is also in accord with the theoretical emphasis in the Annotation Graph framework of Bird & Liberman (2001). Thinking of a database in this way also opens up a more general possibility for multimedia databases that we will refer to as “naked media.” Consider the case of a large database of classroom video data contributed to TalkBank by Rich Lehrer from the geometry lessons of Carmen Curtis. This database consists of 200 hours of classroom video with no accompanying transcripts. It would take perhaps a full year to transcribe all of these sessions. On the other hand, the video can be prepared for streaming web access in about a month. Once the naked video is posted on the web, it can be target for collaborative commentary. In cases of this type, collaborative commentary can operate effectively even without accompanying transcripts.
Conclusion

The TalkBank project has provided openly accessible databases for the study of spoken language interactions. We are now implementing support for collaborative commentary targeted to these databases. Construction of these new methods will open up many exciting new lines of investigation for each of the several disciplines studying human communication.

Acknowledgments

The discussion of the process of collaborative commentary and the example of the dialog between Keck and Clark was developed in collaboration with Roy Pea. Prabhu Raghunathan has programmed the TBViewer. Leonid Spektor has programmed CLAN. Franklin Chen has programmed XML support for TalkBank. Kelley Sacco has developed and maintained the TalkBank database.

From 1999 to 2005, TalkBank received support from NSF Grant BCS-9978056 to Carnegie Mellon University. TalkBank senior investigators included Steven Bird, Ken Dauber, Jerry Goldman, Mark Liberman, Brian MacWhinney, Roy Pea, and Howard Wactlar.
References


