TalkBank: Building an Open Unified Multimodal Database of Communicative Interaction

Brian MacWhinney

Carnegie Mellon University, Psychology
Pittsburgh, PA USA 15213
macw@cmu.edu

Steven Bird

University of Melbourne, Computer Science
Victoria 3010, Australia
sb@cs.mu.oz.au

Christopher Cieri

University of Pennsylvania, LinguisticData Consortium
Philadelphia, PA USA 15213
ccieri@ldc.upenn.edu

Craig Martell

Naval PostgraduateSchool, Computer Science
Monterey, CA 93943
cmartell@nps.edu

Abstract

The goal of the TalkBank project (http://talkbank.org) is to support data-sharing and direct, community-wide access to naturalistic recordings and transcripts of human and animal communication. Toward this end, we have constructed a web accessible database of transcripts linked to audio and video media within fields such as conversation analysis, classroom discourse, animal communication, gesture, meetings, second language acquisition, first language acquisition, bilingualism, tutoring, and legal oral argumentation. We discuss how we have taken discrepant databases from dozens of individual projects and merged them together into a well-structured uniform database in which transcripts can be opened online through browsers, allowing direct multimedia playback. To achieve translation across corpora, we have defined a general XML schema. The validity of this schema is checked by bidirectional conversion from alternative input formats to XML and back. The resultant transcripts are then linked to hinted media and XSLT is used to format web readable browsable multimedia transcripts playable through SMIL. A parallel pathway is used to support collaborative commentary and publication of PDF linked to media through special issues of journals in the relevant fields.

Data-Sharing

TalkBank (http://talkbank.org) is an interdisciplinary research project funded by the National Science Foundation. The goal of the project 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 transcripts of parent-child interactions in 20 languages. The data-sharing model for TalkBank is based on the model from the CHILDES project (MacWhinney, 2000). Having reviewed best practice in 12 very different research areas, Talkbank has identified these seven 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,
7. education of researchers to the existence of shared data, tools, standards and best practices.

Seven Interest Areas

TalkBank has constructed databases in 12 different areas. We will only discuss the seven areas that have the largest data collections.

Classroom Discourse

For years, educators have been relying heavily on video data to study classroom discourse. 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. However, none of these data are publicly sharable, largely because of technical limitations and the failure to link digitized video to transcripts. To demonstrate the use of TalkBank methods for this area, we organized a meeting on classroom discourse in 1999 and then constructed an initial database of materials from story-telling, math education, dyadic tutoring sessions, college lectures, and bilingual classrooms. In addition, we published a special issue in the Journal of the Learning Sciences that uses TalkBank methods to analyze a lesson on graphs in a 7th grade classroom. Within this area, we are now working to provide streaming video access to three large longitudinal classroom corpora. One corpus contains 3000 hours recorded over a span of 12 years tracking the math learning of a group of 15 students. Another records a year's worth of integrative geometry lessons from a 3rd grade classroom. The third compares alternative formats for bilingual classrooms.

Conversation Analysis

The TalkBank database for adult conversational interaction includes transcripts in CA notation for a subset of the CallFriend phone conversations available through the LDC, the Santa Barbara Corpus of Spoken American English (SBCSAE), recorded phone calls from the Nixon Whitehouse, European political television programs, informal interview materials from a special issue of the Journal of Communication, and a variety of classic materials from the field of Conversation Analysis (CA). Materials in this collection were originally transcribed in a variety of alternative forms of CA format. We have been working to standardize this format and to program convertors between CA and the general TalkBank Schema.

Meetings

Within this area, we have two videotapes of dissertation defense reviews and a series of multi-party university work-group meetings, initially transcribed in ISL format. The single largest audio collection in TalkBank includes the oral arguments of the last 30 years of the Supreme Court of the United States (SCOTUS). This dataset is currently being formatted into the TalkBank Schema and linked to the digitized media.

Gesture

There is an increasing interest in corpus-based research on sign language and gesture. With appropriate tools, researchers can create corpora containing grammatical information, discourse structure, facial expression, along with gesture. The resulting corpora can be used to test hypotheses concerning the relationship of the paralinguistic aspects of communication to speech and to meaning. To begin to address this need for a multi-modal corpus, the LDC developed FORM, a non-semantic, geometrically-based annotation scheme which allows an annotator to capture the kinematic information in a gesture just from videos of speakers. FORM stores this gestural information using Annotation Graphs (AG), allowing for easy integration of gesture information with other types of communication information. The FORM work so far has produced 30 minutes of FORM-annotated videos of Brian MacWhinney teaching at CMU. Five other gestural corpora are also available.

Bilingualism and Second Language

Data in this collection include the large European Science Foundation (ESF) study of guest workers in seven countries, five code-switching corpora, Chinese immigrants in Hungary, early second language teaching in Dresden, and the Southampton FLLOC database http://www.flloc.soton.ac.uk of English learners of French.

First Language Acquisition

All of the data in the CHILDES database are now formatted in accord with the TalkBank XML Schema. This database includes over 100 corpora from 30 languages. The corpora in English, Japanese, and Spanish have also been tagged for part of speech.

Sociolinguistics

A major branch of sociolinguistic research has used quantitative methods to analyze variation in vernacular speech observed empirically for the past forty years. Despite the gradual evolution of this field's methodology, it has not yet fully exploited recent technological advances. TalkBank has contributed data, tools, standards and examples of best practice to enrich corpus creation, analysis and sharing among sociolinguists. These contributions flow directly from TalkBank sponsored research on Annotation Graphs and associated tools and from TalkBank's support of the LDC project on Data and Annotations for Sociolinguistics (DASL). Specific outcomes of this collaboration include: the SLx Corpus of Classic Sociolinguistic interviews from a wide variety of regional and social dialects and speaking styles. SLx is a survey of more than 150 sociolinguistic variables giving examples of each from the interviews in which they occur. It was constructed using an Annotation Graph (AG) compliant tool based upon the TableTrans tool described below. The tool allows for browsing of the corpus and for creating further annotations. This system was introduced at a workshop on robust sociolinguistic methods at NWAVE 2003, the primary conference for this community.

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 spoken language data need support for a wide range of transcription, editing, and analysis functions. The TalkBank project has built several tools to support these functions plus infrastructure to support further development of annotation tools for new disciplines.

One set of tools has been developed within the AG framework (http://agtk.sourceforge.net) that characterizes annotations in terms of acyclic directed graphs (Bird & Liberman, 2001). The arcs on these graphs are linked to media through start and stop times. This form of transcription is ideal for datasets in which annotations are fully linked to media. Using the Annotation Graph ToolKit (AGTK) we have developed four useful sample applications: MultiTrans, TableTrans, TreeTrans, and InterTrans (Bird et al., 2002). With TableTrans (Figure 1), we have annotated a corpus of vervet monkey vocalizations (80 open-reel tapes) for variables such as call type, caller and recipient. TableTrans has also been productively used with sociolinguistic and classroom discourse data.

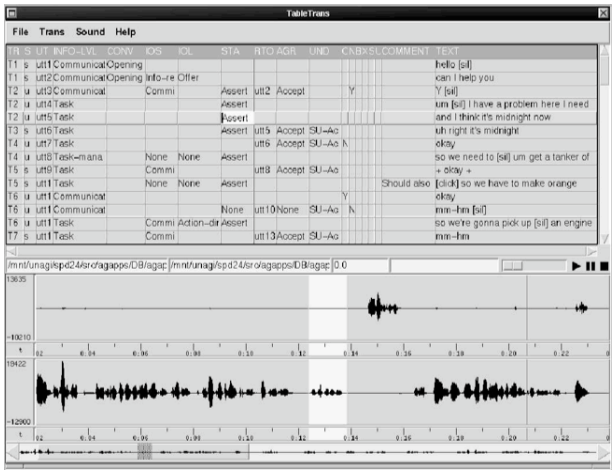

Figure 1: A screen from TableTrans

A second set of of tools uses the CHAT format that was originally designed for the transcription of child language corpora. This format provides an extensive set of codes for transcribing all manner of conversational structures, along with a variety of morphosyntactic and discourse coding systems. Files in the CHAT format can be created using the CLAN editor and analyzed by a set of CLAN search and concordance commands. The editor also supports transcription in the popular Conversation Analysis (CA) format.

TableTrans allows the user to create metadata specific to a file by entering information into dialog boxes. In CLAN, metadata is entered directly into @ID fields in the transcript. The harvesting and publication of metadata is done in conformity Open Language Archives Community (OLAC: http://www.linguistlist.org/olac) standards.

TalkBank has supported the development of several methods for linking transcripts to media, including the TransAna (http://www2.wcer.wisc.edu/Transana/) for video transcription and Transcriber for audio transcription (http://www.etca.fr/CTA/gip/Projets/Transcriber). In addition, the CLAN program supports direct linkage of CHAT files to either audio or video through Transcriber-style linkage, simulation of foot pedal control, numerical editing of time points, or highlighting of segments from a waveform.

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 JavaCC parser generator implements the lexer and parser for CHAT, generating a SAX event stream which optionally passes through schema validation before being serialized into an XML file. Errors are sent to files that can be browsed from CLAN for easy correction. We are currently working on an alternative to this method that uses the ANTLR parser generator to create 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. Getting the CHILDES database through this complete process required a full year of work correcting deviations from the schema; we are now about halfway through the process of converting all of the other TalkBank materials, including CA transcripts, to XML. The tools we have constructed during this process will be reusable as we continue to reformat additional corpora.

Browsable Transcripts

Once the transcripts are in the TalkBank XML format, they can be rendered as HTML pages. When opening a transcript through the browser, the user can either perform a canned analysis or view the complete file. Currently, we have only implemented a simple frequency count to illustrate the analysis function. To view the complete file, the user can select HTML (default), text, or XML. The display of HTML, which is done on the fly through XSLT within Cocoon, takes from one to five seconds depending on file size and connection speed. Once the file is displayed, users can play the audio or video media for a given utterance by clicking on the speaker ID hyperlink. Alternatively, the user can replay the entire file by clicking on the SMIL presentation link at the top of the file. SMIL playback relies on Apple

QuickText and configuring QuickTime for SMIL. The media for SMIL playback are a set of four QT hinted alternative movies compressed by Cleaner 6. The media server is a separate machine dedicated to running Apple's QuickTime Streaming Server to avoid overload and port conflicts. Figure 2 illustrates a typical page for browsing and Figure 3 shows the streaming QuickTime video playback with the transcript echoed below. For audio-only corpora, the QuickTime window displays the utterance and plays the sound without any video. While QuickTime is replaying, users can search and scroll the HTML to keep up with the media window. Eventually the two windows will be linked using Macromedia Flash so that the main window scrolls and highlights as the playback advances.

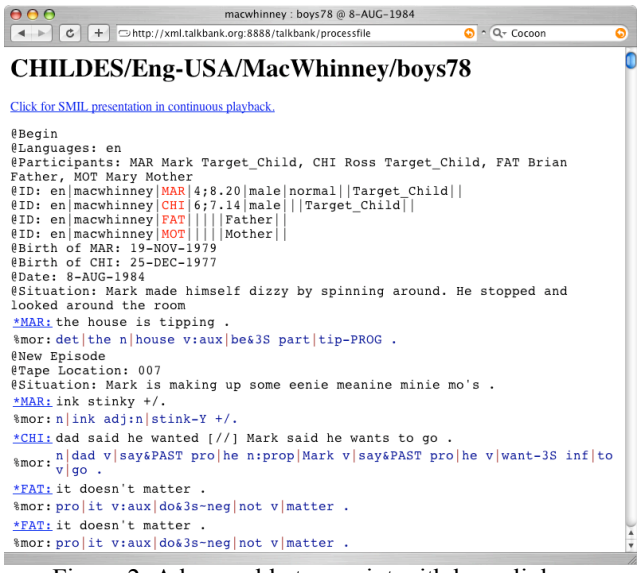

Figure 2: A browsable transcript with hyperlinks

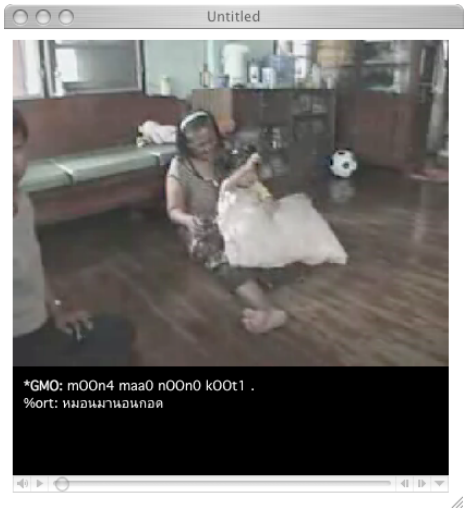

Figure 3: A SMIL frame in a Thai interaction

Collaborative Commentary

Users of the Internet may be familiar with a simple form of online commentary called blogging Currently, we are working with the Oyez project at Northwestern University www.oyez.org to incorporate their ProjectPad software to provide this function for TalkBank. The three areas in which collaborative commentary is currently being explored are classroom discourse, oral arguments at the Supreme Court, and tape recordings from the Nixon Whitehouse. This issue is discussed in greater detail in a panel presentation on Collaborative Commentary elsewhere in these proceedings.

Special Journal Issues

Another approach to the fostering of collaborative commentary relies more on publication through journal articles. Working with small groups of educational researchers, we have produced three special issues of journals that are linked to video materials. A special issue of Discourse Processes examines a problem-based learning (PBL) interaction between medical students. The attached CD-ROM included a QuickTime video movie and a printed transcript. Later, in a special issue of the Journal of the Learning Sciences, we were able to include both the video and transcript on the CD-ROM, along with a complete set of PDF files for the articles in the printed version of the issue. In addition, we used Adobe Acrobat to insert hyperlinks to video segments that corresponded to particular references from the articles. Readers of the special issue could load these PDF files, click on the hyperlinks, and directly replay the related video. A third special issue of the Journal of Communication used a similar PDF-hyperlink method.

One difficulty with this approach is that it requires distribution of the finalized transcript and video before authors begin composing their commentary. If this is not done carefully, references to the transcript and media will be inconsistent. Another problem is that, after the publisher has completed the final PDFs, press deadlines allow only a very small amount of time to insert hyperlinks in the PDF files. In the framework we plan to implement next, we will seek to minimize these two problems by encouraging authors to make direct links from commentary to materials on the web. Within their articles, they can link directly to start and stop times on media on the web.

Conclusion

TalkBank has now constructed an openly accessible database for the study of spoken language interactions. The implementation of additional methods for on-line browsing, analysis, and commentary will open up many new lines of investigation and thought for each of the several disciplines studying human communication.

References

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication, 33*, 23-60.

Bird, S., Maeda, K., Ma, X., Haejoong, L., Randall, B., & Zayat, S. (2002). TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools build on the Annotation Graph Toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 1-7). Paris: European Language Resources Association.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.