Talk Bank: A Multimodal Database of Communicative Interaction

1. Overview

The ongoing growth in computer power and connectivity has led to dramatic changes in the methodology of science and engineering. By stimulating fundamental theoretical discoveries in the analysis of semistructured data, we can to extend these methodological advances to the social and behavioral sciences. Specifically, we propose the construction of a major new tool for the social sciences, called TalkBank. The goal of TalkBank is the creation of a distributed, web-based data archiving system for transcribed video and audio data on communicative interactions. We will develop an XML-based annotation framework called Codon to serve as the formal specification for data in TalkBank. Tools will be created for the entry of new and existing data into the Codon format; transcriptions will be linked to speech and video; and there will be extensive support for collaborative commentary from competing perspectives.

The TalkBank project will establish a framework that will facilitate the development of a distributed system of allied databases based on a common set of computational tools. Instead of attempting to impose a single uniform standard for coding and annotation, we will promote annotational pluralism within the framework of the abstraction layer provided by Codon. This representation will use labeled acyclic digraphs to support translation between the various annotation systems required for specific sub-disciplines. There will be no attempt to promote any single annotation scheme over others. Instead, by promoting comparison and translation between schemes, we will allow individual users to select the custom annotation scheme most appropriate for their purposes. Codon will also facilitate the direct comparison of complementary and competing analyses of a given dataset. TalkBank will benefit four types of research enterprises:

1. **Cross-corpora comparisons**. For those interested in quantitative analyses of large corpora, TalkBank will provide direct access to enormous amounts of real-life data, subject to strict controls designed to protect confidentiality.
2. **Folios**. Other researchers wish to focus on qualitative analyses involving the collection of a carefully sampled folio or casebook of evidence regarding specific fine-grained interactional patterns. TalkBank programs will facilitate the construction of these folios.
3. **Single corpus studies**. For those interested in analyzing their own datasets rather than the larger database, TalkBank will provide a rich set of open-source tools for transcription, alignment, coding, and analysis of audio and video data. In some cases, confidentiality concerns will force researchers to use this mode of analysis.
4. **Collaborative commentary**. For researchers interested in contrasting theoretical frameworks, Codon will provide support for entering competing systems of annotations and analytic profiles either locally or over the Internet.

The creation of this distributed database with its related analysis tools will free researchers from some of the unnecessarily tedious aspects of data analysis and will stimulate fundamental improvements in the study of communicative interactions.

This proposal outlines the shape of TalkBank and the computational tools that will support its construction. The initiative unites ongoing efforts from the Linguistic Data Consortium (LDC) at Penn, the Penn Database Group, the Informedia Project at CMU, and the CHILDES Project at CMU. The initiative also establishes an ongoing interaction between computer scientists, linguists, psychologists, sociologists, political scientists, criminologists, educators, ethologists, cinematographers, psychiatrists, and anthropologists. Seven specific activities are proposed:

1. Needs assessment through workshops and workgroups.
2. Establishment of a framework for data formatting and coding called “Codon”.
3. The construction of demonstration TalkBank data sets.
4. Development of methods for confidentiality protection.
5. The development of tools for creating Codon data sets.
6. The development of tools for aligning and analyzing Codon data sets.
7. Dissemination of the programs, data, and results.

2. Background

Communicative interactions include face-to-face encounters, conversations across phone lines and video connections, as well as dialogs between humans and computers. Whatever the specific format, each communicative interaction produces a complex pattern of linguistic, motoric, and autonomic behavior. By studying behavioral patterns, social scientists have learned about underlying cognitive, linguistic, physical, and social competencies and how they develop in various social and cultural contexts [1-3].

Legacy technology. Most researchers studying communicative interactions are still relying on the videotape and audiotape technology of the 1970s. This technology uses VITC (either SMPTE or EDU) time-code generators to insert codes that support alignment of the video with the transcript. For audiotapes, tape counters are used to mark time points. Although these codes provide reasonably accurate alignment, access to segments of video or audio is dependent on the tedious process of rewinding of the tape. This process of rewinding creates a serious barrier between researchers and the data. Consider the example of a researcher, such as Adolph [4], who studies the ways a child learns to crawl up a steep incline. When the child tries to crawl or walk up an incline that is too steep, she may begin to fall. Adolph's theory makes a crucial distinction between careful falling and careless falling. The assignment of particular behaviors to one of these categories is based on examination in videotapes of a set of movement properties, including arm flailing, head turning, body posture, and verbalization. As Adolph progresses with her analyses, she often finds that additional indicators need to be added to assign behaviors to categories. However, access to the full video database involves rewinding hours of tape to access and reevaluate each episode during which the child begins to fall. This process is facilitated by Adolph's use of VITC time markers and coding within the MacShapa program [5], as well as by the use of high-end playback units that use time markers to access segments of the videotape. But, even with these tools, the access to data and annotations is so slow and indirect that even the original investigator avoids more than one or two passes through the data. For audiotapes, researchers rely on foot pedals to rewind the tape, so that small stretches of speech can be repeated for transcription. The process of replaying audio segments is so difficult that the time needed to transcribe an hour of spontaneous interactional dialog is usually about 25 times the length of the original segment. This legacy technology is extremely fragile, cumbersome, and unreliable.

New opportunities. Fortunately, there are new alternatives to this older approach. In terms of hardware, researchers now have access to large hard disks, removable storage, writable CD-ROM, DVD, and powerful processors. These advances make it possible to replace the older technology with a system based on completely digital analysis. This system provides instant retrieval of annotations and direct access to data from annotations. Moreover, new software tools can support the sharing and analysis of digital data across the Internet. We now have well-developed systems with recognized digital video file formats (QuickTime, MPEG, or AVI), middleware architectures (CORBA, DCOM), and programs for encoding and decoding (Active Movie, Fusion, Strata Video, Windows Media Player). Browsers such as Netscape and Internet Explorer are tightly linked to plug-ins that can play back various audio and video formats. Programming tools such as Java, Tcl/Tk, and Python, as well as the specification of Unicode and XML, facilitate the design of powerful, cross-platform retrieval engines.

A diversity of approaches. Although we can now build the tools we need to solve this problem; one important ingredient is still missing. Paradoxically, the chief roadblock to improvements in the study of human communication is a lack of communication between researchers. Responding to the new technological opportunity, dozens of projects have popped up, each attempting to solve part of the same basic set of interrelated problems. The result has been a Babel of formats, standards, and programs. A non-exhaustive list of formats with URLs includes Alembic[6], Annotator[7], Archivage [8], CA [9], CAVA [10], CES[11], CHILDES

[12], COALA [13], Computerized Profiling [14], CSLU[15], DAISY[16], DAMSL[17], Delta[18], Digital Lava [19], Discourse Transcription [20], DRI, Emu [21], Festival [22], GATE [23], HIAT [24], Hyperlex [25], Informedia [26], ISIP[27], LDC [28], LIPP[29], MacSHAPA [30], MATE [31], MediaTagger [32], ODF[33], Partitur [34], Praat [35], SABLE[36], SALT [37, 38], SDIS[39], Segmenter [40], SGREP[41], SignStream [42], ShoeBox[43], SNACK[44], SoundWriter [45], Speech Analyzer[46], Standoff [47], SUSANNE[48], SyncWRITER [49], TEI[50], Tipster [51], TreeBank[52], Transcriber [53], TransTool [54], VoiceWalker [55], and UTF[56]. If this list of formats were extended to include general-purpose database schemes, such as Folio, Excel, or Nud*ist, it would grow to several times this size. Communication between these projects has been minimal; and developers have often built systems and formulated standards without seeking input from the user communities involved.

This proliferation of formats and approaches can be viewed as a positive sign of intellectual ferment. The fact that so many people have devoted so much energy to fielding new entries into this bazaar of data formats indicates how important the computational study of communicative interaction has become. However, for many researchers, this multiplicity of approaches has produced headaches and confusion, rather than productive scientific advances. If this ferment is to lead to meaningful advances, we need to channel its creative energy into the articulation of a solid, useful system. We need a way to bring these many approaches together, without imposing some form of premature closure that would crush experimentation and innovation.

These diverse approaches share a common interest in utilizing the ability of computers to link audio and video records to annotated transcripts. Before the advent of digitized audio and video, it was not possible to achieve a direct linkage between annotations and data. Because this linkage is now possible, we are confronting a remarkable new opportunity. This ability to link annotations to data opens up the possibility for extensive data sharing, interdisciplinary linkages, and direct theoretical comparisons across the social sciences. To grasp this opportunity, we need to provide tools that are sensitive to disciplinary concerns. At the same time, researchers need to transcend disciplinary boundaries to achieve access to a shared library of data on communicative interactions. Developing this system will produce a qualitative improvement in the ways that social scientists make use of transcribed data from social interactions. If we fail to seize this moment of opportunity, it may become impossible to overcome the commitments made by individual communities to mutually unintelligible systems based on traditional disciplinary boundaries. We are clearly at a choice point.

The CHILDES and LDC Projects. The LDC and CHILDES projects have learned that community acceptance of programs and coding schemes is directly linked to the availability of shared data. In the case of the LDC, the widespread use of databases such as TIMIT, Switchboard, MapTask, CSR, CallHome, and CallFriend have fundamentally altered the culture of the international speech technology community. In the case of CHILDES, all of the major datasets on language acquisition have been placed into a common format in a shared database. As a result, new work in child language is done within a clearly established framework. The decision to link new standards to the development of a common database was crucial in the development of CHILDES and greatly supported the role it has played in the field of language acquisition. Without a concrete focus on real data, the formulation of a set of transcription standards and computer programs is an arid exercise.

The CHILDES Project has also learned about the importance of providing multiple annotation formats. In 1997, we worked with Noldus Information Technologies to develop a link between CHAT transcription and time-stamped behavioral events, as coded in The Observer [33]. More recently, we have completed a full computational implementation of the system of Conversation Analysis developed by Sachs, Schegloff, Jefferson [9] and others. Implementation of the CA mode was completed in February of 1999 and is now in use at several universities in Europe. The lesson that we learned during this process has important implications for the TalkBank project. First, we learned that TalkBank must provide tools that allow each research community to use its familiar analytic language. Second, we learned that markedly different research communities (CA

and child language) can share a strong common interest in developing shared access to digitized communicative interactions. Third, we have learned that, through the proper development of computational tools, we can advance the empirical and theoretical dialog between sub-disciplines within the social sciences.

Taking this as a basic lesson, we intend to follow a similar path in the development of TalkBank. Specifically, we propose the establishment of an annotation abstraction layer called "Codon". Codon will be designed to maximize our ability to support all of the various specific annotation schemes favored by sub-disciplines. Once these particular schemes are formalized, we will develop Codon representations for each particular annotation scheme. In addition, we will use Codon as an interlingua for translations between alternative annotation schemes. We will also construct a suite of search and analysis programs that use the Codon format. These tools will facilitate automated alignment of transcripts to audio and video, intelligent query and retrieval from the database, and Internet access to data and transcripts. We will also provide a series of import and export filters to facilitate the use of existing tools.

The personnel of the CHILDES, LDC, and the database group at Penn have highly complementary expertises. The CHILDES project has had close, ongoing contact with the complex features of analyzing face-to-face interactions, using video and audio linked to transcripts. The LDC project has developed a profound control over issues in audio processing and the use of very large shared databases in promoting scientific advance through the common task method. The database group at Penn has developed tools and languages for data integration, has taken a leading role in the development of query languages for semistructured data, and has recently been involved in the development of XML-QL. Together, these three groups have all the basic technological and analytic tools needed to build TalkBank.

Starting at very different initial points, these three groups have arrived independently at a common purpose and philosophy. The diversity of our various strengths will work in our favor. It is especially timely to be merging our efforts now. If we had attempted to work together on this problem three years ago, the effort would have been premature. If we wait another three years, the moment of opportunity will have passed and the research community will have divided into unproductive fractionation. No single group can solve this problem alone. Instead, we plan to rely on each other's strengths to address this fundamental new opportunity.

3. Research Design and Methods

The seven specific activities being proposed in the TalkBank Project are:

1. Needs assessment through workshops and workgroups.
2. Establishment of standards for data formatting and coding called "Codon".
3. Construction of demonstration TalkBank data sets.
4. Development of methods for confidentiality protection.
5. Development of tools for creating Codon data sets.
6. Development of tools for aligning and analyzing Codon data sets.
7. Dissemination of the programs, data, and results.

Project 1: Needs assessment

Before embarking on an effort of this scope, it is crucial to solicit input from investigators in the many fields impacted by the project. We will achieve this through a series of workshops. In these workshops, we will include representatives of all research fields engaged in empirical investigations founded on human communication. In the next paragraphs, we identify a group of over 50 researchers who have formally agreed to participate in the TalkBank project. Section 11 at

the end of this proposal lists these same initial participants and gives scanned copies of their letters of agreement to participate. This initial participant list is meant to be representative and illustrative, rather than exclusive, both in terms of researchers named and research areas discussed. As soon as the project begins, we will begin to expand this list to include several hundred researchers. Our goal in presenting this initial list of formal participants is to demonstrate the fact that many of the most prominent researchers in the study of communicative interaction are interested in TalkBank.

These researchers come from a wide variety of disciplines, including computer and information science, linguistics, psychology, sociology, political science, education, ethology, philosophy, psychiatry, and anthropology. However, the research interests of these communities typically cut across these traditional disciplinary boundaries. For example, the study of problem-based learning in classroom interactions is of interest to researchers in education, psychology, sociology, medicine, and computer science. Similarly, the study of language disabilities is of interest to linguists, psychologists, pediatricians, neurologists, psychologists, and workers in speech and hearing. In order to characterize these various communities, we have identified 17 areas that will be included in TalkBank. This list is clearly incomplete. For example, we believe that TalkBank will also be relevant to workers in areas such as criminology, cinematography, oral history, and marketing. However, we have not yet pursued contacts with researchers in these additional fields.

1. **Math and Science Learning**. Researchers in educational psychology have a long history of relying on videotape to study classroom interactions. For example, James Stigler (Psychology, UCLA) has collected an important database of videotapes comparing Japanese, German, Czech, and American instruction in mathematics at the High School level. This work uses a commercial program (Digital Lava) that has interesting overlaps with the Informedia and CHILDES tools. However, because Digital Lava is no longer under active development, Stigler and his collaborators are very interested in shifting their work to TalkBank. On the grade school level, Catherine Snow (Education, Harvard) and Lauren Resnick (Education, LRDC) have been at the forefront of the movement to establish a set of national eductional standards for math, science, and literacy. The process of formulating these standards has relied heavily on videotapes of specific instructional patterns. These videotapes play two roles in their work. First, they are used to assess children's expression of knowledge in situations outside of formal testing. Second, they help teachers understand the dynamics of activities such as problem-based learning for math and science. Closely related to the study of classroom discourse, is the study of tutorial discourse. Here, Kurt vanLehn (CS, Pitt) is directing a NSF KDI study of collaborative learning named CIRCLE. This project examines the learning through computer tutors, skilled tutors, peer tutors, and collaborative learning. The CIRCLE group has just now begun to compile a database of video recordings of tutorial sessions. If the databases being constructed by Stigler, vanLehn, Snow, Reznick, and their associates could make use of the Codon format, their value for both these groups and the wider educational research community would be greatly increased.

2. **Conversation analysis**. Conversation Analysis (CA) is a methodological and intellectual tradition stimulated by the ethnographic work of Harold Garfinkel and formulated by Harvey Sachs, Gail Jefferson, Emanuel Schegloff, and others. Recently, workers in this field have begun to publish fragments of their transcripts over the Internet. However, this effort has not yet benefited from the alignment, networking, and database technology to be used in TalkBank. The CHILDES Project has begun the process of integrating with this community. Working with Johannes Wagner (Odense), Brian MacWhinney has developed support for CA transcription within CHILDES. Wagner plans to use this tool as the basis for a growing database of CA interactions studied by researchers in Northern Europe. Representative of other active groups in this area include Charles Goodwin, (Applied Linguistics and TESOL, UCLA), Gene Lerner (Sociology, UCSB), and John Heritage (Sociology, UCLA).

3. **Text and discourse**. Closely related to Conversation Analysis is the field of Text and Discourse. Here, researchers such as Wallace Chafe (Linguistics, UCSB) and Herbert Clark (Psychology, Stanford) have focused on understanding the cognitions underlying complex social interactions. Focusing more on written discourse, researchers such as Tim Koschmann (Medical Education, Southern Illinois) and Arthur Graesser (Psychology, Memphis) have emphasized structured systems for text comprehension and verbal problem solving. This second type of research has strong implications for the study of math and science learning, since it provides a formal analysis of the way in which instruction leads to changes in specific cognitive structures. Both of these lines of research have developed highly articulated, analytic frameworks that will challenge and enrich the development of Codon and TalkBank.

4. **Second language learning**. Annotated video plays two important roles in the field of second language learning. On the one hand, naturalistic studies of second language learners can help us understand the learning process. The work of Ryuichi Uemura (Fukuoka Institute of Technology) represents this line of work. Uemura has collected a large database of videotaped and transcribed interactions of English speakers learning Japanese and Japanese speakers learning English. Similarly, Manfred Pienemann (Linguistics, Paderborn) has collected a database from learners of Japanese, French, and German, using the COALA program. These databases are intended for use by researchers and teachers, as they attempt to better understand the process of language learning. The second use of video in second language learning is for the support of instructional technology. By watching authentic interactions between native speakers, learners can develop skills on the lexical, phonological, grammatical, and interactional levels simultaneously. There are now dozens of sets of video-based materials for second language learners. However, distribution of these materials remains a major problem. Two researchers working in this framework who will be involved in the TalkBank project are Roger Anderson (Applied Linguistics and TESOL, UCLA), who has developed a CD-ROM of materials for learning Quechua and or Hal Schiffman (South Asian Studies, Penn), who has developed a CD-ROM of materials for learning Tamil.

5. **Corpus linguistics**. Although a great deal of corpus linguistics focuses on written documents, there is also several important corpora of spoken language. These include the British National Corpus, the London-Lund Corpus, the Australian National Database of Spoken Language, the Corpus of Spoken American English, and the materials in the Gallery of the Spoken Word. Eventually, the TalkBank project will be able to involve dozens of researchers in this tradition. However, during our planning phase, we are including Lou Burnard (Linguistics, Oxford), Geoffrey Sampson (Linguistics, Sussex), and Michael Seadle (Computer Scientist, Gallery of the Spoken Word), as representatives of this larger community.

6. **Speech production, aphasia, language disorders, and disfluency**. The facilities provided by TalkBank are also relevant to the areas that focus on segmental phonology, fluency, and intonational patterns. One area that can particularly benefit from access from data coded on this level is the study of language disorders. The establishment of norms for articulatory and auditory competencies across social groups and clinical populations should eventually be grounded on a database of actual spoken productions and target sounds for conprehension. Examples of projects that could benefit from the elaboration of such a database include Nan Bernstein-Ratner's (Speech and Hearing, Maryland) [57] work on stuttering, Frank Wijnen's (Linguistics, Utrecht) [58] studies of developmental speech disfluencies, and Julia Evans (Speech and Hearing, Madison) [59] studies of the interactional bases of Specific Language Impairment. Much of this type of work currently uses programs from the CHILDES Project. However, Kim Oller (Speech and Hearing, Maine) has developed a more powerful analytic approach called LIPP that can provide a model for further developments of analytic frameworks in this area.

7. **First language acquisition**. Over 1000 published studies of first language acquisition have relied on the use of the CHILDES database (Brian MacWhinney, Psychology, CMU). This work extends across the areas of phonology, morphology, syntax, lexicon, narrative, literacy, and discourse. Although CHILDES has been a great success in its current format, workers in this field are becoming increasingly aware of the need for a facility to link transcripts to audio and video. By providing this facility, TalkBank will open up new avenues for child language research.

8. **Gesture**. Researchers such as David McNeill (Psychology, University of Chicago) have developed sophisticated schemes for coding the relations between language and gesture. McNeill has shown how gesture and language can provide non-overlapping views of thought and learning processes. McNeill has constructed a videotape database of film descriptions from 12 languages that would be a very useful addition to TalkBank. Working in a very different framework, Justine Cassell (Media Laboratories, MIT) has developed programs that generate psychosocially appropriate gestures, movements, and intonations for computerized animations. Her work can help guide the development of systems for annotating and analyzing gesture in naturalistic interactions.

9. **Signed Language**. The NSF-sponsored SignStream project led by Carol Neidle (Linguistics, Boston University) has formulated programs for coding videotaped data of signed language [42]. The BU group has joined with the LDC and other Penn researchers to create a NSF-funded national resource for creating, archiving and distributing sign language and gestural data. The development of Codon as an interlingua between annotation schemes will allow us to include SignStream data in the distributed TalkBank database. Phyllis and Sherman Wilcox (Linguistics, New Mexico) have made creative use of video to illustrate the emergence of aspect marking in ASL. We will also include researchers examining the effects of cochlear implants on young deaf children [60, 61].

10. **Psychiatry, conflict resolution**. Psychiatrists such as Mardi Horowitz [62] (Psychiatry, USF) have been leaders in the exploration of transcript analysis and annotation. Because of privacy concerns, it is impossible to have open access to videotapes of clinical interviews. However, the application of the technology being developed here could provide a major boost to studies of clinical interactions. Moreover, data could be shared over the Internet with password protection for academic users who have signed releases. A related use of annotated multimodal data occurs in work on conflict resolution. For example, Preston Covey of the Center for Applied Ethics at CMU uses annotations of filmed interactions to display the operation of specific levels of conflict escalation and resolution. Similarly, Laurie Weingart (Graduate School of Industrial Administration, CMU) uses video to study negotiation processes in business transactions.

11. **Behavioral analyses**. Within both social and developmental psychology, research is often grounded on the detailed coding of behaviors from videotape. For example, Grazyna Kochanska (Psychology, Iowa) studies the development of conscience [63] by coding specific child behaviors as “fearful” or “fearless” and specific adult disciplinary behaviors as either “gently controlling”, “strongly controlling” or “non-directive”. Marc Bornstein (NICHD) has built a library of 3000 hours of videotaped records of children and their parents at ages 5, 12, 18, 25, and 36 months in 10 cultures. In each culture, there is a culturally meaningful group comparison. For example, in Israel the comparison is between Haifa and the Kibbutz. In Argentina, the comparison is between Buenos Aires and Native American groups. Bornstein codes these interactions to study the effects of parenting styles on cognitive and emotional development.

12. **Animal behavior**. Videotapes of animals and humans in experimental situations are often coded using tools such as The Observer [33] from Noldus Information Technology. The

Observer has video editing and playback functionality similar to those found in Digital Lava and Informedia, but with additional analysis facilities and support for streaming video across the Internet. Excellent examples of audio and video analysis for elephants, birds, and whales can be found in the work of Christopher Clark's Bioacoustics group at Cornell. Clark has developed a computer program, called Canary, which is now the standard for the study of bird vocalizations. Work on primate calls is illustrated in the research of Robert Seyfarth (Psychology, Penn). The formal issues in coding audio or video records of animal behavior are identical to those that arise for coding human interaction, although of course the content may be quite different.

13. **Anthropology**. Since the beginning of the century, ethnographers have pioneered the use of film documentaries to record the lives of non-Western peoples [64]. Much of this documentary material is still available and includes excellent video footage. Researchers such as David Zeitlyn (Anthropology, Kent) continue this tradition of audio and video recording in the field. However, anthropologists have also begun to utilize new technology. For example, Napoleon Chagnon (Anthropology, UCSB) has made his films of the Yanomamo available over the Internet, along with extensive linked commentary. Similarly, Brenda Farnell (Anthropology, Illinois) has produced a CD that documents the performances of traditional Assiniboine storytelling with signs. In addition to the original narratives, the CD includes the complete Labanotation texts of the Sign Talk gestures and a phonemic transcription of the texts with English translations. Other anthropologists have been working on developing the ideas of Sapir and Whorf regarding links between language, culture, and thought. For example, Stephen Levinson's group at the Max Planck Institute in Nijmegen has studied language, culture and thought relations in the Yucatec Maya. Although these last two lines of research have not yet relied on video, this may be due to technological rather than conceptual barriers.

14. **Field linguistics**. The Shoebox program and other SIL linguistic software that have been developed under the leadership of Gary Simons have been used for data collection in hundreds of minority languages. Will Leben (Linguistics, Stanford) and colleagues have been amassing speech data from West Africa. Steven Bird (LDC, Penn) and Chris Manning (Linguistics, Sydney) have developed computational systems for linking lexical, syntactic, and auditory resources together over the Internet for the use of field linguists. The LDC has already begun to bring these developments into contact. However, the TalkBank framework will allow us to link these efforts with parallel ones in typology and anthropology.

15. **Speech Analysis**. Work in speech recognition, alignment, and generation has made considerable progress in the last decade. Following the lead of the LDC, many groups have turned to large speech corpora for the training and testing of speech analysis systems. For example, the Emu project, led by Jonathan Harrington (CS, Macquarie) supports corpus-based research in phonetics and phonology. Joe Picone (ECE, Mississippi State) directs a project that will make automatic alignment tools available to the broader research community. Mari Ostendorf (ECE, Boston University) and Mark Liberman (LDC, Penn) have developed detailed methods for the analysis of intonational contours. Dan Jurafsky (Linguistics, Colorado) has used the DAMSL discourse structure analysis to study the LDC SwitchBoard corpus with the goal of improving stochastic models of speech recognition. All of these researchers will participate in the TalkBank project and all of them have planned and managed large-scale efforts in the analysis of speech corpora.

16. **Semistructured data modeling.** Within the context of TalkBank, the Penn Database Group will investigate issues of fundamental importance for Computer Science. Existing database technology is inadequate for managing databases of text, speech, and video recordings, with evolving schemas, confidentiality protection, and embedded commentary. The Penn Database Group (Computer and Information Science, Penn) along with other groups at Stanford, AT&T, INRIA, the University of Washington, and the University of Toronto have made

important progress in the formulation and use of semistructured data and in query languages for XML. In related work, Henry Thompson (Computer Science, Edinburgh) has explored a related set of issues in his work on StandOff, as a support for annotation of corpora. These lines of work hold great promise for the mangagement of TalkBank data and the development of computational theory for database retrieval.

17. **Human-Computer Interaction**. The framework of TalkBank and Codon can be useful to several types of research in human-computer interaction. It can facilitate the comparison between actual human interactions and virtual interactions created by programmed animations (Justine Cassell, Media Laboratory, MIT). It can be used to evaluate the efficacy of multimodal interactions (Alex Waibel, Computer Science, CMU). It can be used to study interactions with tutorial systems (Kurt vanLehn, Computer Science, Pitt). Finally, the video components in TalkBank can be mined in order to extract metadata for automatic sorting and retrieval (Howard Wactlar, Computer Science, CMU).

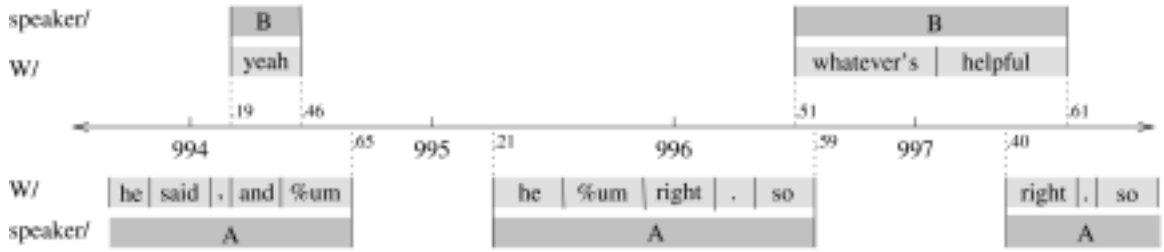
It is our goal to construct reciprocal dialogues between researchers from all of these communities and computer scientists interested in the fundamental computational issues posed by semistructured Codon data. In addition to the core expertise represented by the Penn database group, we will include computer scientists who will represent expertise in scientific discovery, speech technology, networking, and data mining. These meetings will also include interested parties from government and business, including the various corporations that have funded the work of the LDC. Although we have not yet established arrangements for corporate membership in TalkBank, we believe that several features of TalkBank will make this work quite interesting for businesses interested in transmitting video over the Internet.

Project 2: Formulating Codon

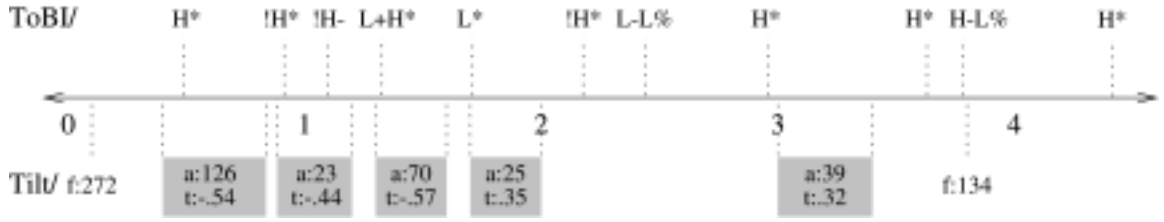
The major product of our working groups will be Codon – a system of standards, formats and tools for coding communicative interactions. Codon will foster the creation and adoption of open standards for all levels of annotation of communicative structure, including marking of movement, orientation, gesture, orthography, part-of-speech tagging, syntactic structure, syntactic class, coreference, phonetic segmentation, overlaps, disfluencies, code-switching, prosody, facial expression, situational background, and participant identity.

Bird and Liberman [65] have shown how current annotation formats, of which there are literally dozens, all involve the basic action of associating labels with stretches of recorded signal data. They are able to represent a disparate range of annotation formats in terms of labeled, acyclic digraphs having optional time references on the nodes. These “annotation graphs” (AGs) can specify the temporal alignment of linguistic units and the ways in which larger units are composed of smaller units on other levels. This work provides the algebraic foundation for inter-translatable formats and inter-operating tools. The intention is not to replace the formats and tools that have been accepted by any existing community of practice, but rather to make the descriptive and analytical practices, the formats, data and tools from each of the disciplines listed in Project 1 available to *all* of the disciplines. We view the Bird and Liberman model as a first-pass formulation of Codon. We will elaborate Codon by further consulting with practitioners from the different disciplines and examining the implications of widely used norms [9, 13, 24, 34, 37, 66].

The following example is one of many possible visualizations of a Codon annotation structure. It describes a fragment of an actual telephone conversation [65]. The diagram shows a timeline (in seconds) and the temporal locus of utterances made by speakers A and B. Each piece of annotation is represented as a shaded rectangle bearing a label, and written on a horizontal level corresponding to a particular type. Types are represented on the left and the W/ type used here is for “words”. The example shows the representation of overlapping hierarchical structures.



Another situation involving multiple annotations arises when two different theoretical models are used in the annotation of the same data. The following example shows annotation of English intonation in the ToBI [67] and Tilt [68] models. Codon will make it possible, for the first time, to undertake a wide-ranging comparison of the two models.



Similar comparisons could be made between two different speech act coding systems or even codes inserted by two different research assistants, as a way of checking for deviations from reliability.

Beyond the annotation of communicative data, Codon must provide the necessary structures for storing the web of annotations that corresponds to a particular database. It must also provide tools for creation, maintenance and query. Since the structure of annotation data is ill-suited to conventional data models, and since both the data and the structure are subject to continual revision during the time course of collection and analysis work, Codon will employ semistructured data models [69-77]. Under this approach to database design, the data is self-describing and there is no requirement to force annotations into the straightjacket of a relational schema. This move brings annotation structures into the realm of database technology, and provides the foundation for data exchange and transformation. The semistructured data model for annotation graphs is, in effect, an internal data structure for exchange of data between corpora, but having such a structure invites the idea of querying data in AG format directly.

XML is a natural "surface" representation for semistructured data, and we have adopted it as the primary exchange format for Codon. Import and export capabilities will be provided for current systems, such as CHAT [12], SignStream [42], and Emu [21]. Codon databases, and the multimodal information they store, will be accessible over the Internet. Various XML extensions, such as XML-Data [78], RDF [79] and XML-QL [80], will provide the starting point for our explorations of the representation and query of Codon structures within XML.

TalkBank will be configured as a consortium of allied databases rather than a central monolithic database. When users access a database, either locally or over the Internet, they will know that it subscribes to the Codon standards and can be manipulated with Codon tools. Of course, they will still need to understand the coding conventions of the particular sub-discipline in question. Creators of Codon-compliant databases will be able to run validation tools on their data, and these tools will provide summary statistics about the content and structure of the database. Creators of Codon databases would be encouraged to submit these statistics, along with prose and keyword descriptions of the database, to an Internet registry of Codon databases. This will be the index that allows people to locate the formats, tools and data most relevant to a particular problem.

Project 3: The construction of demonstration TalkBank data sets

In order to explore the use of Codon and related Codon tools, we will need to create several TalkBank demonstration data sets. Some of these data sets can be taken from earlier CHILDES and LDC corpora or from corpora in related fields. Others will be newly recorded data sets. The specific decisions regarding which corpora to include will be made at the conferences and workshops described above. However, in order to illustrate the shape of this demonstration database, we will list a few types of sample corpora we would like to include.

1. **Child-parent**. A set of videotapes of child language interactions will be included in the demonstration database. Some interesting candidates for inclusion would be Linda Acredolo's [81] tapes of "Baby Signs", Susan Goldin-Meadow's [82, 83] tapes of home signing by deaf children of hearing parents, or Marc Bornstein's huge cross-cultural study of infant socialization. We will also include data already in the CHILDES database such as the Bates [84] or New England [85] corpora. These data will be coded for both speech and gesture. As we proceed with project 3, the entire CHILDES database will eventually be integrated into TalkBank.
2. **Speech corpora**. We need to include samples of carefully transcribed speech corpora as ways of verifying the utility of Codon for the scientific study of speech (including e.g. phonetics, phonology, disfluencies, and dialect). For this, we can rely on corpora from the LDC.
3. **Human-computer interaction**. We would like to include videotapes of computer users learning to use a new program. One source of these tapes could be the current project evaluating the E-Prime system developed by PST. These data are now being analyzed in the ACT-R [86] framework by Chris Schunn [87] at George Mason.
4. **Animal behavior**. In this area, we would like to collect specific examples of natural communicative gestures in primate species of the type collected by Tomasello [88], Seyfarth, or others.
5. **Tutoring**. We will include detailed video studies of the process of tutoring by expert tutors and peer tutors. Kurt vanLehn has begun the formation of a database of this type.
6. **Classroom discourse**. Lauren Resnick has been conducting an ongoing project on the development of national standards for the use of spoken language by children in school contexts. The videotapes supporting this effort would be excellent candidates for inclusion in this demonstration data set.

Project 4: Confidentiality protection

As long as the CHILDES project dealt only with written transcripts, it was relatively easy to maintain confidentiality by using pseudonyms and eliminating last names and place names from transcripts. As we move into the era of multimodal data, it becomes more difficult to maintain confidentiality through the simple use of pseudonyms. As a result, researchers and subjects who would be happy to donate their transcript data to CHILDES might have serious second thoughts about donating the related audio or video data. How can we deal with legitimate and important concerns about speaker confidentiality and still promote international scientific collaboration for the study of verbal interaction? One approach that has been implemented by many local IRB committees focuses on specifying varying levels of confidentiality. In these systems, the most restrictive level provides no access at all and the least restrictive level allows full Internet access. These levels would typically be applied on a corpus-by-corpus basis, so that any given database within the distributed database system could contain corpora at each of these nine levels:

Level 1: Data are fully public (public speeches, public interviews, etc.) and generally viewable and copyable over the Internet, although they may still be copyrighted.

Level 2: Data are open to general viewing and listening by the public across the Internet, but watermarking and other techniques are used to block copying and redistribution.

Level 3: Transcript data with pseudonyms will be made publicly available. However, the corresponding audio or video data, for which anonymity is more difficult to preserve, will be made available on one of the next six, more restrictive levels.

Level 4: Data are only available to researchers who have signed a non-disclosure form. This form sets tight standards regarding avoidance of use of personal names when required. It allows some temporary copying or downloading of the data for local analysis, but requires that downloaded files be deleted after a specific period and never further copied or distributed. These requirements are enforced through watermarking and software blocks.

Level 5: Access is restricted to researchers who have signed non-disclosure forms. In addition, copying is disallowed.

Level 6: Data viewing requires explicit approval from the contributor of the data. This level would work much like a research laboratory that made copies of videotapes to send to other laboratories and required those laboratories to follow rules about non-distribution of data. However, unlike Level 6, this level would also include mechanisms for insuring that the data would not be copied or distributed.

Level 7: This level would only allow viewing and listening in controlled conditions under direct on-line supervision. This level is needed for data of a highly personal or revealing nature. This level has been used in the past for the viewing of material from psychiatric interviews.

Level 8: This level would only allow viewing and listening in controlled conditions under the direct, in person, supervision of the particular researcher. This level is needed for highly sensitive material.

Level 9: These data would not be viewable, but would be archived in the format of the general system for use by the original investigator only. This level allows the investigator to use the tools of the analysis system without actually "contributing" the data.

This system corresponds closely to procedures currently in use by Human Subjects review committees at the University of Minnesota and the University of California at Berkeley. In addition to protecting subject confidentiality, this system of varying levels can be used to support the academic interests of the original data collector. For example, if a researcher has not finished publishing the results of a study, access can be set to a more restrictive level. Once the research papers have been published, access can be changed to a less restrictive level.

Some aspects of this system of levels of confidentiality protection can benefit from the development of technical processes. For example, it is possible to create confidentiality by blurring audio and video images. This technology is generally unacceptable for the study of interactional processes, since facial expressions and intonation convey so many important components of communicative meaning. However, there are more sophisticated ways of morphing the face and the voice to images that are still communicatively adequate. Currently, the LDC is using audio morphing to preserve confidentiality in the Corpus of Spoken American English (CSAE) collected by researchers at UC Santa Barbara. Also, the technique of "watermarking" can be used to prevent or discourage the unauthorized copying of images [89-93].

TalkBank can succeed without posing a threat to confidentiality. In many cases, people will not want their data available publicly. However, even if data are analyzed off the Internet in the privacy of individual laboratories, the direct linkage of annotations to the data will greatly facilitate the process of scientific analysis. Over time, researchers and the wider community will adapt to the restrictions and possibilities offered by the Internet and learn to live within clearly established boundaries, while still achieving major scientific advances.

Project 5: Development of transcription and commentary tools

The development of data formats and annotated corpora must proceed hand-in-hand with the construction of tools for transcription and analysis. In accord with our pluralistic emphasis, we

will encourage the distributed construction of these tools, as well as the adaptation of existing tools. By adding import and export capabilities to existing tools, we can facilitate the development of interoperability between current tools. To seed this integration process, we will work with the authors of existing systems to get import/export capabilities for the Codon format (specified as an XML DTD, or in some other data structuring formalism that can be expressed in XML). We will also create our own open-source tools for the following tasks: transcription, commentary, alignment, browsing, and retrieval. Responsibility for constructing the tools will be divided between CMU and Penn. The CMU tasks are described here and the Penn tasks are described in Project 6. All new tools we construct will be informed by a critical review of existing annotation tools and by discussions with our participants group.

Transcription tool. We will provide a platform-independent Codon Editor implemented in Tcl/Tk or Java. This will provide a high-level interface looking somewhat like the displays in Project 2, and it will store annotation data as XML. Our experience with the CHILDES editor will provide a starting point for the design. We will provide multiple, customizable "views" of annotation and signal data.

Commentary tool. Annotation often references signal data directly. We will create a commentary tool, which permits "meta" annotations to reference existing annotations. This possibility for indirect annotation is already intrinsic to the annotation graph formalism (and to XML), and is sometimes termed "Standoff Markup" [47].

To illustrate the importance of commentary tools, consider a collection of articles edited by Mann and Thompson [94]. In this fascinating collection, 12 discourse analysts examine a two-page fund-raising letter mailed out in 1985 by the Zero Population Growth (ZPG) group after publication of their Urban Stress Test. A problem with this book-length presentation of the 12 analyses is that the reader finds it hard to compare analyses. If the comments of the analysts were structured with direct links to the text, we could immediately compare analyses in the context of the sentences being analyzed. Similarly, in a forthcoming issue of Discourse Processes, five researchers working from slightly different perspectives provide alternative analyses of a six-minute segment of problem-based learning (PBL) in a medical school context. A transcript is published in the journal and a digitized version of the video is included on a CD-ROM, but there is no linkage between the two media. As a result it is still difficult for readers of these articles to compare the analyses directly. If the TalkBank framework were available, these alternative analyses of the tutorial interaction could be directly linked to the audio and video and accessed over the Internet. Moreover, other researchers could then add further analyses and commentary.

Project 6: Development of tools for alignment, browsing, and query

Alignment tools. LDC, Informedia, and CHILDES have a variety of tools for the manual and automatic alignment of textual transcripts with audio data. These will be adapted to the Codon format and refined to handle a much more diverse range of user interface requirements. The underlying algorithms will be improved to handle a wider variety of recording conditions, alignment conventions, and analysis tasks.

Browsing and visualization tools. A variety of tools will be created for viewing data. The Codon Editor itself will be a flexible browsing tool and a web-browser plug-in version will be available. Existing annotation tools, equipped with a Codon import method, will provide a wealth of browsing and visualization possibilities. XSL, XQL, XML-QL or some other transformation tool will provide other ways to define views [95]. Another tool, inspired by Lerner's WorkBench, will allow users to create a derived database or "folio" consisting of data and annotation fragments selected from one or more existing databases. Yet another tool will allow the extraction and tabulation of summary statistics, to be processed by separate statistical software for visualization and analysis.

Query tools. While we intend to use XML as the standard for data exchange, this leaves open what application programming interfaces (APIs) and other interfaces (GUIs and query languages) should be provided. It is clear that the provision of such interfaces for the Talk Bank data will

greatly enhance its use and is likely to determine the success of the project. While a number of APIs for XML are under development, the only generally accepted API at the time of writing is the Document Object Model [96]. Of particular importance is the development of an efficient query language that will allow researchers to scan collections of annotations for features of interest. These will be relatively complex (e.g. parse tree structures) and involve temporal relationship across modalities.

The annotation framework proposed by Bird and Liberman [65] is essentially that of a labeled graph, and the direct representation of this in XML is, in its simplest form two relations consisting of a binary (node-id, node-label) node relation and a ternary (node-id, edge-label, node-id) edge relation. (Indeed, this is the representation used in [97] to construct a graph query language and arguably one of the first semistructured query languages.) Any non-trivial query against this representation involves joins, and the current crop of XML query languages differ greatly in their ability to express joins: at present only XML-QL provides for arbitrary joins, and even in that language a complex query will be quite cumbersome and probably inefficient. The problem is that XML query languages tend to be tuned to the tree-like structure of XML documents. In this case the tree is "flat". We are left with two avenues of investigation: (1) to augment the XML structure so that it is a better match to the query language, or (2) to consider alternative query languages. This will be one of our first investigations. It is likely that a combination of the two approaches will be needed. We shall also investigate the construction of other APIs designed to simplify the problem of programming with Talk Bank. These may make use of an embedded query language.

All of the computational tools developed by this project, as well as the source code, will be made freely available on the Internet, using the servers of the LDC and CHILDES projects. Subject to the permission of the authors and confidentiality considerations, the Codon annotations and the primary audio and video data will also be made available both on the Internet and through CD-ROM or DVD.

Project 7: Dissemination

See section 8 below for a discussion of our plans for dissemination.

4. How TalkBank will be used

In particle physics, the decision to build a new accelerator is difficult to justify a priori. The correctness of the choice can be judged only in retrospect after about a decade of operation [98]. The same is true of TalkBank. However, we can extrapolate from the successes of CHILDES and the LDC to provide a fairly reliable forecast of how TalkBank will be used. In the first section of this proposal, we said that the four basic types of uses would include cross-corpora comparisons, folio construction, single corpus studies, and collaborative commentary. In Project 1, we outlined why 17 different fields are interested in TalkBank. In this section, we will look at this same issue from a slightly different perspective by outlining the ways in which TalkBank will lead to scientific progress. In order to maintain consistency, all of these examples will be taken from the field of child language research. However, exactly the same set of examples could be elaborated for each of the 17 fields surveyed in Project 1.

The first impact of TalkBank will be on the linking of theory to data. To illustrate this effect, consider the debate regarding the relative roles of syntactic [99] and semantic [100] bootstrapping in the child's acquisition of verb argument structures. Siskind [101, 102] has presented an efficient algorithm for acquiring argument structures from noisy [103] contextual data. However, as Slobin [104] noted, researchers have often made excessively strong assumptions regarding the availability of contextual information to disambiguated verb frame interpretations. In order to compute exactly how much information is available, we need to record complete interactions which we then code for situational information. Siskind has begun work of this type. However, TalkBank would provide an ideal framework for the examination of this issue, since it provides direct linkage between video data and situational annotations inserted by coders. In fact, some

aspects of the situation could be automatically derived using technique currently under development by the Informedia project [105, 106].

The second impact of TalkBank will be on the integration of disciplinary perspectives. Consider work by Alibali, McNeil, and Evans [107] suggesting that children with Specific Language Impairment (SLI) may try to communicate gesturally when they are having trouble formulating their meanings verbally. This work integrates the study of language development, the study of gestural development, and the study of mathematics learning. To investigate this issue, Alibali et al set up a task that includes a challenge to mathematical thinking. This experimental interaction is then videotaped and subjected to microgenetic analysis [108, 109]. During the process of coding and analysis, the availability of TalkBank tools will greatly facilitate the speed of the analysis and the reliability of coding. It will also make it possible to tightly align gesture and speech in a way that was not possible in the methodology previously used by Alibali and colleagues.

This same example can be used to illustrate the third effect of TalkBank. Once the database of problem-solving sessions from Alibali et al. is made available, other researchers will be able to examine their evidentiary database to decide whether they wish to produce alternative accounts of the data. Although they may not be able to dispute the numerical analyses produced by the original investigators, some researchers may believe that cases of gesture-speech mismatch are really caused by some other process. For example, a conversation analyst (CA) might want to argue that children are using gesture because of problems in turn-taking. To further develop these competing analyses, the two groups will produce folios of evidence for their respective accounts. These folios will contrast gesture-speech mismatch in normal children, normal adults, and children with SLI. The direct confrontation between alternative viewpoints facilitated by TalkBank tools will lead to further research, experimentation, and progress.

The fourth impact of TalkBank will be on the education of young researchers and the larger society. TalkBank will make available materials on gesture-speech mismatch, early verb learning, and a myriad of other topics in the social sciences. It will be possible to find examples of primate communication, prosodic shifts in West African languages, or breakdowns in intercultural communication. Together, this rich database of interaction will help us teach students how to think about communication and will provide us with a dramatic way of communicating our research to the broader public.

5. Conclusion

In the movie “Field of Dreams”, the hero built a baseball field when voices came to him advising that “If you build it, he will come.” Our experiences with CHILDES and the LDC have taught us that the same message holds true for resources for the study of human communication. If we build these resources, there are thousands of researchers ready to make use of them.

The advent of new computational opportunities makes it possible to build a system that we could have only dreamed about ten years ago. We can build on the lessons and successes of the CHILDES and LDC projects to build a new system that will lead to a qualitative improvement in social science research on communicative interactions. It is important to begin this project now, before the proliferation of alternative formats blocks the possibility of effective collaboration across disciplinary boundaries.

6. Letters of Cooperation

We have received formal pledges of cooperation from these 57 researchers. This list of participants is not meant to be exhaustive or exclusive. Rather, this is a group of researchers with whom we were in immediate contact. We intend to broaden this group.

These participants have read the current proposal and provided comments.

Semi-structured data modeling and XML
- Peter Buneman (Computer Science, U Penn)
- Howard Wactlar (Computer Science, CMU)
- Dan Suciu (AT&T Research)
- Henry Thompson (HCRC, Edinburgh)
- Gary Simons (Summer Institute of Linguistics)
- Alex Waibel (Computer Science, CMU)

Classroom Discourse, Tutoring
- Lauren Resnick (Learning Research and Development Center, Pitt)
- James Stigler (Psychology, UCLA)
- Catherine Snow (Graduate School of Education, Harvard)
- Kurt VanLehn (CS, Pitt)
- Tim Koschmann (Medical Education, Southern Illinois)

Conversation Analysis and Sociolinguistics
- Emanuel Schegloff (Sociology, UCLA)
- Charles Goodwin (TESL / Applied Linguistics, UCLA)
- Gene Lerner (UCSB Sociology)
- Johannes Wagner (Linguistics, Odense)
- John Heritage (Sociology, UCLA)
- William Labov (Linguistics, Penn)
- Gillian Sankoff (Linguistics, Penn)
- Greg Guy (Linguistics, York)

Text and Discourse
- Wallace Chafe (Linguistics, UCSB)
- Herbert Clark (Psychology, Stanford)
- Arthur Graesser (Psychology, U Memphis)
- John DuBois (Linguistics, UCSB)

Second Language Learning
- Manfred Pienemann (Paderborn)
- Ryuichi Uemura (Information Engineering, Fukuoka Institute of Technology, Japan)
- Hal Schiffman (Penn Language Center)
- Roger Andersen (Applied Linguistics and TESOL, UCLA)

Corpus Linguistics
- Lou Burnard (Humanities Computing Unit, U Oxford)
- Geoffrey Sampson (Cognitive and Computing Sciences, Sussex)
- Michael Seadle (MSU Library, Gallery of the Spoken Word)

Speech and Hearing / Language Acquisition
- Nan Bernstein-Ratner (Hearing and Speech Sciences, Maryland)
- Julia Evans (Communicative Disorders, UW Madison)
- Brian MacWhinney (Psychology, CMU)
- Steven Gillis (Linguistics, University of Antwerp)
- Kim Oller (Psychology, U Maine)

Gesture and Signed Languages
- Justine Cassell (MIT Media Lab)
- David McNeill (Psychology, U Chicago)
- Carol Neidle (Linguistics, Boston University

Phyllis and Sherman Wilcox (Linguistics, UNM)
Adam Kendon (Emeritus, Penn)

Psychiatry, conflict resolution

Mardi Horowitz (Psychiatry, UCSF)
Preston Covey (Philosophy, CMU)
Laurie Weingart (Graduate School of Industrial Administration, CMU)

Ethology and observation

Christopher Clark (Ornithology, Cornell)
Lucas Noldus (Noldus IT)
Robert Seyfarth (Psychology, Penn)
Marc Bornstein (NICHD)

Anthropology

Brenda Farnell (Anthropology, U Chicago)
Stephen Levinson (Max Planck Institute for Psycholinguistics)
David Zeitlyn (Centre for Social Anthropology and Computing, UKC)

Corpus-Based Field Linguistics

Steven Bird (LDC, Penn)
Will Leben (Linguistics, Stanford)
Chris Manning (Linguistics, Sydney)

Speech Analysis

Jonathan Harrington (SHLRC, Macquarie)
Joe Picone (Electrical and Computer Engineering, Missisippi State)
Mari Ostendorf (Electrical and Computer Engineering, Boston University)
Dan Jurafsky (Linguistics, U Colorado Boulder)

Mark Liberman (LDC, Penn)

References

[1] J. Dore, ""Oh Them Sheriff": a pragmatic analysis of children's response to questions," in *Child Discourse*, S. Ervin-Tripp and C. Mitchell-Kernan, Eds. New York: Academic Press, 1977.
[2] R. Pittenger, C. Hockett, and J. Danehy, *The first five minutes*. Ithaca, NY: Martineau, 1960.
[3] S. Ervin-Tripp, "Children's verbal turn-taking," in *Developmental pragmatics*, E. Ochs and B. Schieffelin, Eds. New York: Academic Press, 1979.
[4] K. Adolph, "Psychophysical assessment of toddlers' ability to cope with slopes," *Journal of Experimental Psychology*, vol. 21, pp. 734-750, 1995.
[5] P. Sanderson, *MacSHAPA Manual*. Wright-Patterson Air Force Base: CSERIAC, 1994.
[6] http://www.mitre.org/technology/alembic-workbench/
[7] http://www.entropic.com/products/annotator/overview.html
[8] http://lacito.vjf.cnrs.fr/ARCHIVAG/ENGLISH.htm
[9] G. Jefferson, "Transcript notation," in *Structures of social interaction: Studies in conversation analysis*, J. Atkinson and J. Heritage, Eds. Cambridge: Cambridge University Press, 1984, pp. 134-162.
[10] http://www.mpi.nl/world/tg/CAVA/CAVA.html
[11] http://www.cs.vassar.edu/CES/
[12] http://childes.psy.cmu.edu
[13] M. Pienemann, "COALA - A computational system for interlanguage analysis," *Second Language Research*, vol. 8, pp. 59-92, 1992.
[14] S. Long and M. Fey, *Computerized profiling: User's manual*. San Antonio: The Psychological Corporation, 1993.
[15] http://cslu.cse.ogi.edu/toolkit/
[16] http://www.daisy.org/
[17] http://www.cs.rochester.edu/research/trains.html
[18] http://www.eloq.com/SuePap.htm
[19] http://www.digitallava.com/
[20] http://humanitas.ucsb.edu/depts/linguistics/research/csae/conventions/index.html
[21] http://www.shlrc.mq.edu.au/emu/
[22] http://www.cstr.ed.ac.uk/projects/festival/arch.html
[23] http://www.dcs.shef.ac.uk/research/groups/nlp/gate/overview.html
[24] K. Ehlich and J. Rehbein, "Halbinterpretative Arbeitstranskription (HIAT).," *Linguistische Berichte*, vol. 45, pp. 24-41, 1976.
[25] http://www.ldc.upenn.edu/hyperlex/
[26] www.informedia.cs.cmu.edu
[27] http://www.isip.msstate.edu/resources/software/swb_segmenter/
[28] www.ldc.upenn.edu
[29] http://www.gate.net/home/lipp.htm
[30] http://www.csse.swin.edu.au/macshapa/index.html
[31] http://mate.mip.ou.dk/
[32] http://www.mpi.nl/world/tg/CAVA/mt/MTandDB.html
[33] L. P. J. J. Noldus, R. J. H. Trienes, A. H. M. Hendriksen, H. Jansen, and R. G. Jansen, "The Observer Video-Pro: An integrated system for collection, management, analysis, and presentaiton of time-structured data from live observations, video tapes and digital media files," *Behavior Research Methods, Instruments, and Computers*, vol. 31, pp. 122-134, 1999.
[34] http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html
[35] http://fonsg3.let.uva.nl/praat
[36] http://www.bell-labs.com/project/tts/sable.html
[37] J. Miller and R. Chapman, *SALT: Systematic Analysis of Language Transcripts, User's Manual*. Madison, WI: University of Wisconsin Press, 1983.
[38] http://www.waisman.wisc.edu/salt/index.htm

[39] R. Bakeman and M. Quera, "The SDIS format," *Behavior Research Methods, Instruments, and Computers*, vol. 24, pp. 554-559, 1992.
[40] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Segmenter: A free tool for segmenting, labeling, and transcribing speech," presented at Proceedings, First International Conference on Language Resources and Evaluation, Granada, Spain, 1998.
[41] http://www.cs.helsinki.fi/~jjaakkol/sgrep.html
[42] http://www.bu.edu/asllrp/SignStream/index.html
[43] http://www.sil.org/computing/shoebox.html
[44] http://www.speech.kth.se/SNACK/index.html
[45] http://humanitas.ucsb.edu/depts/linguistics/lab/transcription.html
[46] http://www.jaars.org/icts/sa.htm
[47] http://www.ltg.ed.ac.uk/~ht/sgmleu97.html
[48] http://www.cogs.susx.ac.uk/users/geoffs/RSue.html
[49] http://www.sign-lang.uni-hamburg.de/software/syncWRITER/info.eng
[50] http://www-tei.uic.edu/orgs/tei/
[51] www.tipster.org
[52] http://www.cis.upenn.edu/~treebank/home.html
[53] http://morph.ldc.upenn.edu/mirror/Transcriber/
[54] http://www.ling.gu.se/SLSA/SLSAbody.html#Tools
[55] http://humanitas.ucsb.edu/depts/linguistics/lab/transcription.html
[56] http://www.ldc.upenn.edu/myl/utf.pdf
[57] N. Bernstein-Ratner, B. Rooney, and B. MacWhinney, "Analysis of stuttering using CHILDES and CLAN," *Clinical Linguistics and Phonetics*, vol. 10, pp. 169-187, 1996.
[58] L. Elbers and F. Wijnen, "Effort, production skill, and language learning," in *Phonological development*, C. Ferguson, L. Menn, and C. Stoel-Gammon, Eds. Timonium, MD: York, 1993, pp. 337-368.
[59] J. Evans and H. Craig, "Language sample collection and analysis: Interview compared to freeplay assessment contexts," *Journal of Speech and Hearing Research*, vol. 35, pp. 343-353, 1992.
[60] K. Boosman and G. Szagun, "Caretaker's speech input to young children with cochlear implants," in *New Neuroethology on the Move: Proceedings of the 26th Neurobiology Conference*, N. Elsner and R. Wehner, Eds. Stuttgart: Georg Thieme Verlag, 1998, pp. 214-215.
[61] J. G. Nicholas, "Sensory aid use and the development of communicative function," in *Effectiveness of cochlear implants and tactile aids for deaf children: A report of the CID study*, A. E. Geers and J. Moog, Eds. Washington D.C.: Alexander Graham Bell Association, in press.
[62] M. Horowitz, "Psychodynamics and cognition," . Chicago: University of Chicago Press, 1988.
[63] G. Kochanska, "Toward a synthesis of parental socialization and child temperament in early development of conscience," *Child Development*, vol. 64, pp. 325-347, 1993.
[64] G. Bateson and M. Meade, *Balinese character: A photographic analysis*. New York: New York Academy of Sciences, 1942.
[65] S. Bird and M. Liberman, "Towards a formal framework for linguistic annotations," presented at ICSLP Conference, Sydney, 1998.
[66] J. Edwards and M. Lampert, "Talking data: Transcription and coding in discourse research," . Hillsdale, NJ: Erlbaum, 1993.
[67] www.ling.ohio-state.edu/phonetics/E_ToBI/etobi_homepage.html
[68] P. Taylor, "The TILT intonation model," presented at ICSLP 98 Synd proceedings of the 5th international conference onspoken language processing, 1998.
[69] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener, "The Lorel query language for semistructured data," *International Journal on Digital Libraries*, vol. 1, pp. 68-88, 1997.
[70] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the web: From relations to semistructured data and XML*: Morgan Kaufmann, 1999.

[71] P. Buneman, *Semistructured data: Principles of database systems '97*. Tucson: Association for Computing Machinery, 1997.
[72] P. Buneman, W. Fan, and S. Weinstein, "Path constraints in semistructured and structured databases," presented at PODS '98, 1998.
[73] P. Buneman, A. Deutsch, and W.-C. Tan, "A deterministic model for semistructured data," presented at Workshop on query processing for semistructured data and non-standard data formats, 1999.
[74] P. Buneman and B. Pierce, "Union types for semistructured data," Department of Computer Science, University of Pennsylvania 1999.
[75] S. Chawathe, S. Abiteboul, and J. Widom, "Representing and querying changes in semistructured data," Department of Computer Science, Stanford University 1997.
[76] A. Deutsch, M. Fernandez, and D. Suciu, "Storing semistructured data with STORED," presented at Proceedings of the ACM SIGMOD international conference on management of data, 1999.
[77] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom, "Lorel: A database management system for semistructured data," *SIGMOD Record*, vol. 26, pp. 238-247, 1997.
[78] http://www.w3.org/TR/1998/NOTE-XML-data-0105
[79] http://w3c.org/TR/PR-rdf.syntax
[80] www.w3.org/TR/NOTE-xml-ql
[81] L. Acredolo and S. Goodwyn, "Symbolic gesturing in normal infants," *Child Development*, vol. 59, pp. 450-466, 1988.
[82] S. Goldin-Meadow, "The resilience of recursion: a study of a communication system without a conventional language model," in *Language acquisition: The state of the art*, E. Wanner and L. Gleitman, Eds. New York: Cambridge University Press, 1982.
[83] J. P. Morford and S. Goldin-Meadow, "From here and now to there and then: The development of displaced reference in Homesign and English," *Child Development*, vol. 68, pp. 420-435, 1997.
[84] E. Bates, B. O'Connell, and C. Shore, "Language and communication in infancy," in *Handbook of infant development*, J. Osofsky, Ed. New York: Wiley, 1987.
[85] C. E. Snow, R. Perlmann, and D. Nathan, "Why routines are different: Toward a multiple-factors model of the relation between input and language acquisition," in *Children's Language*, K. Nelson and A. Van Kleeck, Eds. Hillsdale, NJ: Erlbaum, 1994.
[86] J. Anderson, *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.
[87] C. Schunn and J. Anderson, "Scientific discovery," in *Atomic components of thought*, J. Anderson, Ed. Mahwah, NJ: Lawrence Erlbaum, 1997.
[88] M. Tomasello, J. Call, and A. Gluckman, "Comprehension of novel communicative signs by apes and human children," *Child Development*, vol. 68, pp. 1067-1080, 1997.
[89] J. Berghel and L. O' Gorman, "Protecting ownership rights through digital watermarking," *IEEE Computer*, vol. 29, pp. 101-103, 1996.
[90] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio, and video," *IEEE International Conference on Image Processing*, vol. 3, pp. 243-246, 1996.
[91] S. Craver, N. Memon, B.-L. Yeo, and M. Yeung, "On the invertibility of invisible watermarking techniques," *IEEE International Conference on Image Processing*, pp. 234-250, 1997.
[92] F. Hartung and B. Girod, "Copyright protection in video delivery networks by watermarking of pre-compressed video," in *Multimedia Applications, Services, and Techniques*, vol. 1242, S. Fdida and M. Morganti, Eds. Heidelberg: Springer, 1997, pp. 423-436.
[93] N. Shivakumar and H. Garcia-Molina, "Building a scalable and accurate copy detection mechanism," *Proceedings of the First ACM Conference on Digital Libraries*, 1996.
[94] W. C. Mann and S. A. Thompson, *Discourse description*. Amsterdam: John Benjamins, 1992.
[95] http://www.w3.org/TR/NOTE-XSL.html
[96] http://www.w3c.org

[97] M. P. Consens and A. O. Mendelzon, "The {G}+/GraphLog Visual Query System," presented at SIGMOD Conference, 1990.
[98] S. Myers and E. Picasso, "The LEP collider," *Scientific American*, vol. July, pp. 21-27, 1990.
[99] L. Gleitman, "The structural sources of verb meanings," *Language Acquisition*, vol. 1, pp. 3-55, 1990.
[100] S. Pinker, *Learnability and cognition: the acquisition of argument structure*. Cambridge: MIT Press, 1989.
[101] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, pp. 39-91, 1996.
[102] J. M. Siskind, "Learning word-to-meaning mappings," in *Cognitive Models of Language Acquisition*, J. Muure and P. Broeder, Eds. Cambridge: MIT Press, 1999.
[103] G. Marcus, "Negative evidence in language acquisition," *Cognition*, vol. 46, pp. 53-85, 1993.
[104] D. Slobin, "Universal and particular in the acquisition of language," in *Language acquisition: The state of the art*, E. Wanner and L. Gleitman, Eds. New York: Cambridge University Press, 1982.
[105] C. Faloutsos and K.-I. Lin, "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," *ACM SIGMOD*, vol. May 23-25, pp. 163-174, 1995.
[106] A. Biliris, C. Faloutsos, H. V. Jagadish, T. Johnson, N. D. Sidiropoulos, and B.-K. Yi, "Online data mining for co-evolving time sequences," , submitted.
[107] M. Alibali, N. McNeil, and J. Evans, "The role of gesture in children's language comprehension: Now they need it, now they don't," *Journal of Nonverbal Behavior*, under review.
[108] S. Goldin-Meadow, M. Alibali, and R. Breckinridge Church, "Transitions in concept acquisition: Using the hand to read the mind," *Psychological Review*, vol. 100, pp. 279-297, 1993.
[109] R. Siegler and K. Crowley, "The microgenetic method: A direct means for studying cognitive development," *American Psychologist*, vol. 46, pp. 606-620, 1991.
[110] D. Florescu and D. Kossmann, "A performance evaluation of alternative mapping schemes for storing {XML} data in a relational database," presented at INRIA, 1999.
[111] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk*, Second ed. Hillsdale, NJ: Lawrence Erlbaum, 1995.
[112] D. Crystal, P. Fletcher, and M. Garman, *The grammatical analysis of language disability*. London: Edward Arnold, 1976.
[113] P. Buneman, S. Davidson, and D. Suciu, "Programming constructs for unstructured data," presented at Proceedings of 5th international workshop on database programming languages, 1995.
[114] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu, "A query language and optimization techniques for unstructured data," presented at ACM-SIGMOD '96, Montreal, Canada, 1996.
[115] P. Buneman, S. B. Davidson, M. F. Fernandez, and D. Suciu, "Adding structure to unstructured data," *ICDT '97*, pp. 336-350, 1997.
[116] D. Suciu and V. Tannen, "Efficient compilation of high-level data parallel algorithms," presented at Proceedings of 6th ACM symposium on parallel algorithms and architectures, 1994.
[117] D. Suciu, "Parallel programming languages for collections," University of Pennsylvania, Institute for Research in Cognitive Science, Philadelphia, PA 1995.
[118] L. Labkin, R. Machlin, and L. Wong, "A query language for multidimensional arrays: Design, implementation, and optimization techniques," presented at Proceedings of ACM SIGMOD international conference on management of data, Montreal, Canada, 1996.
[119] P. Buneman, S. Nagyi, V. Tannen, and W. Limsoon, "Principles of programming with complex objects and collection types," *Theoretical Computer Science*, vol. 149, pp. 3-48, 1995.
[120] P. Buneman, L. Libkin, D. Suciu, V. Tannen, and L. Wong, "Comprehension syntax," *SIGMOD Record*, vol. 23, pp. 87-96, 1994.

[121] A. Kosky, “Types with extents: On transforming and querying self-referential data-structures,” : University of Pennsylvania, 1995.
[122] S. Davidson and A. Kosky, “WOL: A language for database transformations and constraints,” presented at Proceedings of the international conference of data engineering, 1997.
[123] P. Buneman, S. Davidson, K. Hart, C. Overton, and L. Wong, “A data transformation system for biological data sources,” presented at Proceedings of VLDB, '95, Zurich, Switzerland, 1995.
[124] S. Davidson, C. Overton, and P. Buneman, “Challenges in integrating biological data sources,” *Journal of Computational Biology*, vol. winter, pp. 557-572, 1995.