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 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-corpus 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.
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.
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 educational 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 comprehension. 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
TalkBank

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 management 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 tapes of “Baby Signs”, Susan Goldin-Meadow’s 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 or New England 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 framework by Chris Schunn 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, 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 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. Appropriateness for KDI and Roles of Project Personnel

Many scientific and industrial applications rely on multidimensional transcriptions, annotations and descriptions linked to audio and video recordings of communicative interactions. Computer hardware and software have reached the point where large amounts of annotated recordings can be created and manipulated in digital form. As a result, each of these various disciplines and applications faces the challenge of creating standards, algorithms and tools for creating, exploring, searching, visualizing, and maintaining these multidimensional annotations of multimodal recordings. Many of the disciplines involved are finding it difficult to meet this challenge effectively. Those who are succeeding are developing superficially divergent solutions, even when common solutions would have been as good or better. In consultation with leading representatives of the various disciplines involved, we will develop and test standards, algorithms and sample tools that meet these needs. We will use new computer technology to promote research within each discipline. We will create new science by cross-fertilization of ideas from at least 10 major disciplines. We will also make it possible for (expensive, publicly funded) interactional data to be shared with a wider community of researchers.

Solving these problems requires a series of fundamental new inventions in an area where there is a great deal of current research to draw on and to contribute to. The proposed linguistic databases are similar in purpose to a number of other shared scientific databases. However, compared with other scientific areas (e.g. molecular biology, crystallography), the work of establishing widely accepted schemas or ontologies for linguistic data largely remains to be done. Computer scientists involved in knowledge networking face a number of challenging issues here. First, these new systems need to represent the temporal relationships between annotations. Second, systems must represent recursive types such as parse trees. Third, they must deal with new and unexpected forms of annotation. Fourth, they must track the provenance of data extracted from distributed data sources (as involved in recently funded NSF grant to Penn on Data Provenance). The recent surge of research activity in semistructured data and in XML storage is timely [76, 110]. The edge-labeled graph representation suggested for Codon can be readily represented in a relational database as a binary (node, value) node relation and a ternary (node, label, node) edge relation [97]. This is also one of the possible implementations for semistructured data. Whether it can efficiently support a useful query language is a topic of considerable recent research activity, and it is an area to which we hope this project will contribute. We will study a variety of established techniques in this area, to determine whether we can rely on existing techniques or will have to supply new ones.

TalkBank addresses a strikingly diverse set of problems. This diversity of issues reflects the experiences, interests, and skills of the co-PIs and their colleagues. MacWhinney has working relations with a wide-ranging group of social and behavioral scientists. Buneman is a leading researcher in the area of semistructured databases and scientific databases. Liberman is the founder and director of an organization that supplies much of the annotated data now used in speech and language technology development. Bird is a computational linguist with significant experience in field linguistics, and a network of connections among linguists working on documentation of relatively undescribed languages. One of the graduate students at Penn will address the design and implementation of database indexing and search techniques, while the other will work on user interfaces for search, retrieval, display and analysis. Leonid Spektor, who has
programmed the CHILDES system, will develop the Codon editor. Two research assistants at CMU will be responsible for organizing, checking, and enlarging the TalkBank database.
7. Results from Prior NSF Support

A. The CHILDES Project – Brian MacWhinney

The Child Language Data Exchange System (CHILDES) is an international cooperative venture, involving over 800 child language researchers located in over 30 countries. Support for CHILDES came first from the MacArthur Foundation (1984-1988) and then from NIH (1989-1999). Recently, NSF has provided additional funding for the development of linguistic analysis in CHILDES through the NSF Linguistics Program. This funding began at the end of 1998. Although this support has only been in place for a few months, it has already funded the extension of the MOR morphological tagging system to Spanish and German and the writing of a probabilistic parser called POST that disambiguates part of speech codes produced by the MOR program.

The CHILDES system is composed of three major research tools. The first tool is the database. Nearly 90 research groups have contributed sets of computerized naturalistic transcript data to the CHILDES database. In over 1000 published articles, researchers have used these data as a way of testing a variety of predictions derived from linguistic and psycholinguistic theory. Researchers have been able to use the data both to test empirical hypotheses and to guide further data collection. The second CHILDES tool is the CHAT transcription system, which is codified in the manual for the CHILDES Project [111]. CHAT builds on the pioneering systems of Miller and Chapman [37] and Crystal, Fletcher, and Garman [112]. It is designed to maximize readability and expressive power, while also guaranteeing consistency for computational analyses. In particular, the system provides standard formats for transcribing words, utterances, prosodic contours, morphological analyses, syntactic analyses, overlaps, hesitations, speech acts, and other conversational features. In addition to the CHAT coding scheme, CHILDES now supports the CA coding scheme used in Conversation Analysis. The third CHILDES tool is the CLAN system of computer programs for Macintosh, Windows, and Unix. These programs facilitate a wide variety of frequency counts, word searches, and statistical analyses on the CHILDES data or on new transcripts coded in CHAT format.

Recently, the CHILDES Project has collaborated with the NSF Digital Libraries Informedia Project at CMU, directed by Howard Waetlar. This collaboration is designed to permit the inclusion of CHILDES data into the Informedia database and to allow the CHILDES project to benefit from the tools for video-transcript linkage and analysis developed by the Informedia Project. In the short term, this collaboration has led to improvements in the abilities of the CHILDES programs to play video and audio directly from the transcript. In the long run, we plan an ongoing collaboration that will include the CHILDES Project, the LDC, and the Informedia Project. All of these tools can be retrieved free over the Internet from childes.psy.cmu.edu. The form of these tools has been shaped by continual input from active members of the CHILDES system. Through electronic mail, workshops, and conference presentations, we continually seek input on ways to make these three tools more responsive to the needs of ongoing research projects.

Brian MacWhinney, Professor of Psychology at CMU, is a developmental psycholinguist. MacWhinney has directed the CHILDES Project since its inception in 1984. Apart from his work on CHILDES, MacWhinney has studied the use of neural networks as models of language
acquisition, brain organization for language processes in children with early focal lesions, and crosslinguistic effects in sentence processing.

B. Linguistic Data Consortium – Mark Liberman

The Linguistic Data Consortium, hosted at the University of Pennsylvania, is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases,lexicons, and other resources for research and development purposes. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems division of the National Science Foundation.

The LDC has built the world’s largest collection of spoken and written corpora. The text corpora include outputs from newspapers in German, Spanish, French, Japanese, Mandarin and more than ten other languages, as well as parallel dual- or multi-language texts from Canadian parliamentary records and United Nations archives. Many of the spoken corpora come from phone recordings through facilities such as CallHome (8 languages), CallFriend (13 languages), TIMIT (cellular telephone). Others come from radio and TV broadcasts in various languages. Although few of the corpora include video records, many of the spoken language corpora include conversational and interactional features that would make them useful for inclusion in TalkBank.

Since its foundation, the LDC has delivered data to 197 member institutions and 458 non-member institutions. An initial three-year grant from DARPA amplified the effect of contributions from this broad membership base, producing far more data than any member could afford to produce individually. Ongoing funding from the National Science Foundation has permitted creation of a web-searchable on-line repository for all LDC-data, and the development of new resources for several government-funded research projects. In addition to distributing previously-created databases, and managing the development of new ones, the LDC has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

The data made available through the LDC plays an important role in the “common task” method of the ARPA Human Language Technology (HLT) program. This approach begins each project by specifying a task, defining a formal, quantitative evaluation metric, and developing a large common database for training and testing purposes. Then each participant pursues solutions in an individual way, and all participants meet periodically to compare methods and results (including evaluation scores). Used since 1986, this technique has resulted in rapid performance improvements in several areas. For example, word error rate for speech recognition has been cut in half every two years for the past six years. Similarly, the performance of (text) message understanding and retrieval systems, measured in terms of metrics such as precision and recall, has improved at a rate between 20-50% per year. Common tasks have also been an effective method to engender cooperation and the productive exchange of ideas and techniques. Nearly all empirical work in speech recognition relies on data from the LDC.

The experience of the LDC in the construction of a shared database, internationalization of access, and linkage of the database to a common task method parallels the experiences of the CHILDES project. Although the area of speech technology is far more computationally sophisticated than the area of language acquisition research, both areas have benefited enormously by the ability to develop a collaborative research focus on a shared database. These two projects provide fundamental demonstrations regarding the importance of federal support for those aspects
of scientific infrastructure that maximize collaboration and communication in the context of shared data.

Mark Liberman, Professor of Linguistics and Computer and Information Science at the University of Pennsylvania, has directed the LDC since its inception in 1993. Steven Bird is a computational linguist who joined the LDC in 1998 as Associate Director. Recently, Bird and Liberman have collaborated in producing a fundamental formal analysis of linguistic annotation [65] that will illuminate core issues in the design of TalkBank.

C. Database Systems – Peter Buneman

Peter Buneman has been co-PI on two relevant NSF proposals: (a) CCR92-161122 “Collection Types in Programming Languages and Databases” (PI Val Tannen) and BIR-9402292 “Mediated Access to Biological Databases and Applications” (PI Susan Davidson). The first is most closely related to the contents of this proposal and is briefly summarized here. The period of this research for this project was from 08/01/94 to 07/31/97, and extended to 07/31/98. The proposal was to develop formalisms, languages and techniques for general collection types. It was on this contract that Buneman started work on semistructured data, which can be viewed as another collection type. This resulted in the development of the first query language for collection types based on structural recursion [113], its optimization [114] and proposals for ascribing structure to semistructured data [115]. Also developed on this project were the following:

1. Tannen and Suciu have worked on the development of efficient compilation of high-level data parallel algorithms [116, 117].
2. Libkin, Machlin and Wong have developed both an array query language and optimization techniques [118].
3. Buneman, Libkin, Suciu, Tannen and Wong continued the study of the use of collection comprehensions in database programming languages. The syntax of comprehensions is very close to the syntax of a number of practical database query languages and is, they believe, a better starting point than first-order logic for the development of database languages [119, 120].
4. In related research Davidson and Kosky have developed and implemented a system for declarative system for database transformations (Morphase) [121, 122] Davidson has also developed a number of techniques for biological database integration [123, 124]. Demonstrations of the software developed from this research and other prototypes may be found at: http://db.cis.upenn.edu and http://www.cis.upenn.edu/~sharker/morphase.
8. Dissemination of Results and Institutional Commitment

Our plans for the dissemination of the results of this project involve four components.

The first component is the substantive exchange of ideas on basic principles in the study of communicative interaction. Although the goals of the fields involved are very diverse, and in some cases the current communications bandwidth is very low, TalkBank will allow researchers to exchange ideas in the context of shared data sets. In to discussions among students of interaction, TalkBank will also promote constructive dialogue between computer scientists and behavioral scientists. For computer scientists, the TalkBank presents a set of interesting challenges in areas such as video data mining, metadata extraction, speech technology, data structures for online retrieval, and protection of confidentiality.

The second component of dissemination is publication of data, documentation and programs. The preferred mode of publication will be the Internet, although in some cases, publication on physical media such as CD-ROM and DVD-ROM will remain essential. The CHILDES and LDC web sites will provide access to Codon standards, manuals, and programs. Individual data sets will be published on physical media by the LDC (where the creators want to do this), and made available on the web in either a distributed or centralized fashion. In distributed access, the central site will have tables of links to data sets on machines across the world. Centralized access will use a large file server at the LDC, including a multi-terabyte robot tape unit for "near-line" storage. This device receives requests for data sets and will retrieve them within the space of a few minutes, if the requested data is not already cached on disk. These two modes of distribution have different advantages and disadvantages. The central store has advantages for efficient access, quality control, maintenance and updates. Sophisticated indexing will be easier for data in a single central store; and if and when coding standards change, these data can easily be updated. The advantage of distributed storage is that the creators of individual data sets have more direct control over access. We will provide software to help creators of such databases check and maintain data integrity and adherence to standards, and to contribute to various central indices if they wish to.

The third component of dissemination involves the creation of instructional materials to demonstrate the application of Codon conventions and the use of editing and retrieval programs. These materials will include hardcopy and electronic versions of standard manuals, as well as instructional videos distributed in DVD format. Instructional videos are crucial for this enterprise, since the application of particular codes to specific gestural, intonational, and conversational patterns must be illustrated in concrete ways by specific video cuts.

The fourth component of dissemination is the solicitation of feedback from users. We will seek out feedback about Codon, the computer tools, and the database. Much of this feedback will be provided through electronic bulletin boards and commentary that will be available from the beginning of the project. However, we will also conduct a broader phone sampling of university faculty in the social sciences to make sure that our attempts at dissemination have succeeded in making researchers aware of the availability of these resources.

In all four dissemination areas, the existing infrastructure at CMU and Penn provides both an effective mechanism for reaching researchers today, and assurance that the results of this project will continue to reach potential users in the future. The CHILDES project at CMU has been

24
distributing linguistic databases and programs to Child Language researchers since 1989. Data has been contributed by more than 100 researchers, and data and software have been distributed to more than 2000 researchers. The Linguistic Data Consortium at Penn has been publishing and distributing linguistic databases, primarily to speech and language engineers, since 1992. The LDC has published more than 150 databases, each consisting of up to 44 CD-ROMs, and more than 790 R&D institutions (companies, universities, or government laboratories) have received LDC data.

In addition to providing an ideal basis for dissemination of TalkBank databases, software, documentation, standards and ideas, the existing infrastructure at CMU and Penn provides all of the space and technical support required.
9. **Performance Goals**

The performance goals for the TalkBank Project are the same as the goals of the seven subprojects. They are:

1. Needs assessment through workshops and workgroups.
2. Establishment of Codon – a standard for data formatting and coding.
3. The construction of demonstration TalkBank data sets.
5. Construction of tools for creating, browsing, and searching TalkBank data sets.
6. Creation of tools for aligning and searching TalkBank data sets.
7. Dissemination.

For each of the three years of the project, we hope to make some progress along each of these seven fronts. However, it is important to be realistic about how much can be achieved during each year.

**Year 1.** The single most important goal during the first year is in project #1. Needs assessment will begin through email and phone contacts. In fact, that process has already begun in the work leading to the writing of this proposal. However, the workshop held during the first year will be the most significant forum for assessing these needs. This workshop will also allow us to make progress on projects 2, 5, 6, and 7. Our discussions will focus on exchange of information among the different disciplines involved, comparison and evaluation of existing solutions, and planning for needed TalkBank standards, tools and documentation. In preparation for the workshop, sample databases will be made available, along with first-order translations into Codon framework. Available tools will be put on display and discussed in comparative terms. Throughout Year 1, we will continue the process of needs assessment, tool evaluation, and the refinement of the Codon standard.

**Year 2.** The work in the second year will focus more specifically on tool creation and the production of a few initial, large Codon data sets. By the second workshop, we expect that the co-PIs and many of the external participants will have experimented with Codon at least as an exchange format, and some will have used this to develop or to use new tools for browsing, searching, editing, publishing, and so on. This should include both tools produced by the TalkBank project and also other tools made accessible by use of Codon as an exchange format. Some of these experiences will have been entirely positive ones, but most people will have experienced at least some problems, and some may have been entirely unsuccessful in reaching common ground. The goal of the second workshop will be to diagnose the reasons for these problems, and decide on solutions where feasible, whether these are new language features, improvements or fixes to tools, improved documentation, or help with particular one-time problems such as format conversions. We should come out of the second workshop with a sense of which applications or collaborations are really going to work, and which gaps are simply too wide to bridge at this time.

**Year 3.** By the time of the third workshop, some applications should be robustly successful. In particular, we are confident that the new database structures and new tools should be widely used by the communities of researchers in the co-PIs own areas of specialization. We hope that there will also be success stories in other areas, as well as some additional areas where results are promising enough to warrant continued efforts.


10. Management plan

Brian MacWhinney and Steven Bird are responsible for the management of TalkBank at their respective sites. MacWhinney will coordinate those aspects of the project that are assigned to CMU and Bird will coordinate those aspects of the project that are assigned to Penn. The CMU site will be responsible for the organization of the yearly work group meetings, the summarization and dissemination of the results of these meetings, outreach to new user groups, and the recruitment of new data sets across all areas of the social sciences. The Penn site will be responsible for the development of the computational tools supporting analysis, alignment, and formatting. Within this overall framework, Buneman and the graduate student working with him will focus on the development of tools for working with semistructured interactional data. Bird, Liberman, and the graduate student working with them will focus on the development of user interfaces for the alignment and retrieval tools. The CHILDES site will be responsible for the development of tools supporting data entry and visualization. The Research Assistants at CMU will be responsible for maintenance and reformatting of new additions to TalkBank.

In addition to these shared responsibilities, the two centers will pursue overlapping contact to assess the needs and usage levels within all user communities. Both centers will continue the process of opening up participation in TalkBank to all interested parties. We want to make sure that all participants understand that this is a fully collaborative enterprise. The collaboration initially involves CMU and Penn, but we hope to extend this circle of collaboration to many more institutions. In order to guarantee smooth coordination between the two sites, the four co-POs will meet three times each year. One of these meetings will be in Pittsburgh in conjunction with the yearly workshop; the other two will be in Philadelphia.
11. 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. They recognize that their primary obligation in terms of a participation in this project is to represent the needs and aspirations of the research communities with whom they identify. We were able to get faxed letters of participation from 35 of these people. Together with the four PI, this is a total of 36 formal letters. Because many people are traveling at this time, we are still waiting for faxes from the other 18 participants. However, these other 18 have all assured us of their interest either through email or over the phone. Scanned copies of the 35 faxes we have received are given after this list.

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 (NICHHD)

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, Mississippi State)
Mari Ostendorf (Electrical and Computer Engineering, Boston University)
Dan Jurafsky (Linguistics, U Colorado Boulder)
Mark Liberman (LDC, Penn)
May 11, 1999

Professor Brian MacWhinney
Psychology
Carnegie-Mellon University
Pittsburgh, PA

Dear Brian,

I am writing to acknowledge my willingness to participate in the TalkBank project. It will provide a remarkably rich cooperative effort that is much needed at this point.

Sincerely,

Roger W. Andersen
Professor of Applied Linguistics
Brian MacWhinney
Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

May 10, 1999

Dear Brian:

I have read your proposal for TalkBank with enormous interest. The CHILDES Project moved
the study of child language acquisition, language disorders and bilingualism into a new era and
allowed for unprecedented analyses of important questions in the field, while simultaneously
encouraging cross-institutional and cross-disciplinary research. Its contributions to the work I do
with children with developmental speech and language disorders have been immeasurable. The
same can be said for the contributions of the LDC to fields such as speech recognition.
TalkBank represents the logical and important next step in the evolution of data treatment. I am
most enthusiastic about this proposal. I am happy to participate in the project, support it
wholeheartedly, and fervently hope it gains funding support.

Sincerely,

[Signature]

Nan Bernstein Ratner, Chairman
Department of Hearing and Speech Sciences
University of Maryland, College Park
May 13, 1999

Brian MacWhinney
Carnegie Mellon University

Dear Brian,

Thank you for your invitation to contribute to the Gesture and Signed Language component of the TalkBank project. I am delighted to be part of the project and look forward to our future work.

Sincerely,

Justine Cassell
May 12, 1999

Dr. Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Dear Brian,

This is simply to say that I'll be very happy to participate in the TalkBank project. Your plans are exciting. I'm glad to be included, and will help however I can.

Sincerely,

Wallace Chafe
May 14, 1999

Dr. Brian MacWhinney
Carnegie Mellon University
Pittsburgh, PA

Dear Drs. Liberman, Bird, Buneman, and MacWhinney,

Thank you very much for reaching out to the broad community of researchers who are interested in advancing the science of acoustic communication in animals and man. I am particularly excited about the prospects of developing common mechanisms and conceptual frameworks. As a step in that direction, a few years ago we undertook a NSF sponsored project that resulted in a software analysis tool referred to as Canary. We have recently received funding to continue this effort which will build an advanced bioacoustics software tool in the NT Windows environment. Furthermore, Jack Bradbury will soon become Director of Cornell’s Library of Natural Sounds program, the world’s largest and most active collection of natural sound recordings. One of Jack’s goals is to make this rich and diverse collection available electronically and fully linked to the ancillary data associated with the recordings. In that respect, your project is a natural connection for anthropologically based collections and analyses processes.

In summary, we would be interested in working with your group toward the development of efficient mechanisms to facilitate the exposition and exploration of bioacoustics collections. I am convinced that this type of collaboration provides the most effective and most productive direction toward major advancements in acoustic communication.

Sincerely,

Christopher W. Clark, Ph.D.
I.P. Johnson Senior Scientist
Director, Bioacoustics Research Program
Cornell Laboratory of Ornithology

CWC/Mmc
May 14, 1999

Prof. Brian MacWhinney  
Psychology Dept.  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890  

Dear Brian,  

I would be very interested in participating in the TalkBank project as outlined in the proposal you sent to me. I feel the project is very important and timely, and the group you have assembled for it is excellent. I have been wanting to work on the "next generation" of some of the ideas and software that I've developed for multimedia aligned transcription, and this seems like just the right team to work on it with. I expect to be quite actively involved in the project. I will try to contribute a perspective that combines both linguistics and anthropology, as well as certain perspectives on multimedia representations.

Sincerely,

[Signature]

John W. Du Bois, Director  
Center for the Study of Discourse
May 12, 1999

Brian MacWhinney, Ph.D.
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Dear Brian,

I would like to add my enthusiastic support for your NSF TalkBank proposal. Given the impact your work has had in the area of language analysis and database development with the CHILDES project both with respect to normal aspects of language development, but also in the study of language disorders, I would be extremely pleased to be a part of this new project.

Sincerely,

Julia L. Reins, Ph.D.
Assistant Professor
Dear Dr. MacWhinney:

This is to state in more formal terms my willingness to participate in the TalkBank project. I studied the draft of the NSF KDI Proposal and I think a number of essential developments are described in which I am eager to participate.

Sincerely Yours,

Steven Gillis
Center for Dutch Language and Speech - CNTS
Department of Linguistics - GER
University of Antwerp
Universiteitsplein 1
B-2610 Wilrijk
Belgium
Brian MacWhinney
Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890

May 12, 1999

Dear Brian,

I very much look forward to participating the TalkBank project.

Sincerely,

Charles Goodwin
May 12, 1999

Dear Brian MacWhinney:

The TalkBank project is an extremely important contribution to disciplines that study language and discourse processing. I look forward to participating on the project.

Sincerely,

Arthur C. Graesser
Professor, Psychology and Mathematical Sciences
Brian MacWhinney  
CMU

Dear Brian,

Mark Liberman has contacted me about participating in your TalkBank project, including attending workshops, reading and commenting on material, etc. I appreciate the invitation and am happy to be included. I will contribute to the project any way that I can help.

Greg

Gregory R. Guy  
Professor of Linguistics  
DLLL, Ross S-561  
York University  
Toronto, ONT  
Canada M3J 1P3

416-736-5016
11 May 1999

Prof. Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh PA 15213-3890
fax: (412) 268-7251

Dear Brian:

I would be very interested in participating in the TalkBank project. Of all the projects that I could imagine to benefit researchers in human language, this sounds like the most important one by far.

With the access that has opened up through the web, we urgently need to develop a computational formalism that will permit organized access to others' collections of language data.

It is significant that your proposed project will involve a wide variety of scholars with expertise in different domains of language research, and I expect that you will find linguists in general very willing to cooperate in developing standards for data formats.

Yours truly,

William R. Leben
Professor of Linguistics
May 11, 1999

Brian MacWhinney
Department of Psychology
CMU

Dear Professor MacWhinney,

I am very interested in participating in the development of the TalkBank project.

Sincerely,

Gene H. Lerner
Dear Brian,

I, and some of my other colleagues at the University of Sydney, would be enthusiastic about the opportunity of participating in the TalkBank project. We have been very interested in the possibilities of using annotated digital video in linguistic fieldwork contexts (and have made a little use of the Max Planck Institute's MediaTagger?PPC), and are also working on building of computer tools for visualizing (primarily lexical) linguistic data.

Best,

Christopher Manning
May 12, 1999

Prof. Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA
FAX: 412-268-7251

Dear Brian,

I write to confirm my willingness to participate in the TalkBank Project, should it be funded.

Sincerely,

[Signature]

David McNeill
Boston University

Department of Modern Foreign Languages
and Literatures
718 Commonwealth Avenue
Boston, MA 02215

Carol Neidle
Associate Professor
Phone and fax: 617/353-6218
E-mail: carol@louis-xiv.bu.edu
URL: http://www.bu.edu/atlisp/carol.html

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh PA 15213

May 14, 1999

Dear Brian,

Thank you for the information you sent about the TalkBank project. Good luck with it. We would be happy to cooperate in the development of an interface with SignStream, and to share our experiences in dealing with video annotations, if this would be of help. We will be interested to see what kinds of data analysis tools your system will provide and to see how they might complement those available in SignStream.

With best wishes,

Carol Neidle
Dear Dr. MacWhinney,

Thank you very much for your letter of April 13, 1999 and the outline of your TalkBank project. Noldus Information Technology is highly interested in the proposed work. As you know, our core business is the development and distribution of tools for behavioral research. We are serving some 2,000 universities and research institutions worldwide. If future releases of our software become compatible with TalkBank, thousands of scientists who use our software (many of whom are psychologists, with an increasing number in linguistics) will be able to benefit from the results of the proposed project.

I envisage several ways in which we can contribute to your project. For instance, we can contribute directly to the demonstration database. We are in touch with many users of The Observer who use our software to analyze video recordings of behavioral processes, including human and animal communication. In the framework of the proposed project, we may be able to obtain video media with corresponding codes and annotations for inclusion in the TalkBank database. Furthermore, we can also play a role when it comes to dissemination of the project results. For instance, we can write an interface between our commercial software products and the TalkBank database, allowing thousands of scientists across the world to benefit from your data and media sets.

Please keep me informed about the status of your project proposal. I look forward to working with you.

With kind regards,

[Signature]

Managing Director
May 12, 1999

Dear Dr. MacWhinney:

I am pleased to indicate my willingness to participate in the TalkBank project. As you are aware, for about 10 years I have been working in the development of tools for coding and transcription of human interaction. The work has been especially focused in the area of phonetics through the development and use of a transcription environment called LIPP (Logical International Phonetics Programs). That computer-based transcription system is in the process of an extensive update that will greatly profit from the standards to be developed under the TalkBank project.

Sincerely,

D. Kimbrough Oller, Ph.D.
Professor and Chair
May 14, 1999

Dear Prof. MacWhinney,

I am writing to indicate my support of the TalkBank proposal. It will provide an important improvement in the research infrastructure for a large variety of problems, as well as provide a platform for bringing together researchers from diverse disciplines. I would also be interested in the data for my own work, in part because of the new research avenues that it would open up.

Sincerely,

Mari Ostendorf
Associate Professor
May 10, 1999

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Dear Dr. MacWhinney,

The Institute for Signal and Information Processing at Mississippi State University agrees to participate in the TalkBank project. We think this is a very important and exciting project. Our research currently focuses on the development of public domain speech recognition software and technology, and the support of these resources within the community. Hence, we are very interested in supporting projects that promote shareable resources.

Sincerely yours,

Dr. Joseph Picone, P.E.
Associate Professor, Dept. of Elect. and Comp. Eng.
Director, Institute for Signal and Information Processing
Mississippi State University
413 Simrell, Hardy Rd.
PO Box 93571
Mississippi State, MS, USA 39762
Phone: (662) 325-3149
Fax: (662) 325-3149
Email: picone@isip.msstate.edu
Dear Brian,

This is to confirm that I shall be glad to participate actively in the planned TalkBank research project.

With best regards

Yours sincerely

Geoffrey Sampson
Reader in Computer Science & Artificial Intelligence
UNIVERSITY of PENNSYLVANIA

PENN LANGUAGE CENTER
715-716 Williams Hall
Philadelphia, PA 19104-6205
Tel. 215-898-6039
Fax: 215-573-8139
plc@crat.sas.upenn.edu

May 14, 1999

To: Brian MacWhinney, Carnegie Mellon University
From: Harold Schiffrin, Penn Language Center,
University of Pennsylvania

Re: TalkBank

Fax to: (412) 268-7251

Dear Brian,

This is to confirm my interest in participating in the TalkBank Project you emailed me about. We are developing applications for multimedia teaching of languages here at Penn, and are very interested in having access to the kind of data and software that this will enable us to access. We hope very much to be able to incorporate this into pedagogical materials available via Internet. We are also involved in dictionary preparation here and will have an English-Tamil dictionary database to contribute to the database at some point.
12 May 1999

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3793
Fax 412-268-7251

Dear Brian,

I am writing to express my support for and willingness to participate in the TalkBank research project.

I am one of the Co-Principal Investigators on the National Gallery of the Spoken Word project, which seeks to digitize large amounts of speech, to make the sound searchable, and to provide an interface which a wide variety of researchers and students can use, including those in K-12 environments.

One of our critical needs is to understand human communications, particularly how people use and understand a mix of sound, text, and images. Our own site will provide a laboratory for studying this interaction with real users, and we will certainly benefit from the diverse perspectives of the researchers in this study.

As a librarian I see the potential for broad applications of this research as research and teaching methods become more and more dependent on information using digital technology and formats. As our capabilities in this area have grown, so have the number of choices. The more we understand how humans perceive, process, and react to different forms of communication, the better we will be able to function in the future as information system designers.

Sincerely,

Michael Seadle, PhD.

MSU is an Affirmative-action/Equal-opportunity Institution
May 13, 1999

Dear Dr. MacWhinney,

I have reviewed the proposal for your TalkBank project. I am happy to participate in this project and look forward to the results it will produce.

Yours,

Dr. Gary Simons
Director, Academic Computing
Brian MacWhinney  
Department of Psychology  
Carnegie Mellon University  
FAX: (412) 268-7251

Dear Brian:

I am very excited about the prospects of the TalkBank project and happy to participate to whatever extent I can be helpful. The access to a wide variety of data the project offers, and the incorporation into the basic design of language data from such a wide variety of contexts, is enormously promising. I am particularly pleased to see so much attention within the project to issues of classroom discourse. Best of luck with your search for funding.

Sincerely yours,

Catherine E. Snow  
Henry Lee Shattuck  
Professor of Education

12 May 1999
12th May 1999

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
+1 412 268-7251

Dear Brian

We are fully committed to participation in the TalkBank project.

We think the time is right for cooperation and coordination of effort worldwide in this area, and very much hope public funds will be forthcoming on both sides of the Atlantic to support this.

Yours sincerely

Henry S. Thompson
Reader in Artificial Intelligence and Cognitive Science
University of Edinburgh
Dear Dr. MacWhinney:

Thank you for your invitation to join TalkBank project. It is my great pleasure to be able to contribute to this challenging international project. I am ready to offer a substantial amount of Japanese L2 spoken data (including video, audio and transcribed texts) along with application of sophisticated technology to the development of learners' corpus. And I would like to gain knowledge and tools for retrieving multimedia objects in a given large-scale hypermedia corpus.

I am looking forward to your further information as soon as possible.

Best regards,

Ryuichi Uemura

Associate professor of Linguistics
Faculty of Information Engineering
Fukuoka Institute of Technology
Dear Brian,

I am happy to participate in the TalkBank project.

Sincerely,

Kurt VanLehn,
Director, CIRCLE
Johannes Wagner  
Institute for Language and Communication

To Prof. Brian MacWhinney  
Carnegie Mellon University  
Fax + 412 268-7251  

May 11, 1999

regarding the TalkBank project

It's my pleasure to state that I am happy to participate in the TalkBank project.

Yours sincerely

Johannes Wagner  
dr. phil
To Whom It May Concern:

I am happy to participate in the TalkBank project.

Sincerely,

Laurie R. Weingart
May 15, 1999

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Dear Brian MacWhinney,

Please accept my appreciation for being included in the establishment of the TalkBank grant proposal. I am definitely interested in the creation of a distributed web-based data archiving system for transcribed video data on communicative interactions. This will provide direct access to critical real-life data and will make a tremendous difference in research productivity. I look forward to hearing from you in the near future.

Sincerely,

Phyllis Perrin Wilcox
Department of Linguistics
Signed Language Interpreting Program
Max-Planck-Institut für Psycholinguistik
Technical Department

Directors:
Prof. Dr. Anne Cutler
Prof. Dr. Wolfgang Klein
Prof. Dr. Willem J.M. Levelt
Prof. Dr. Stephen Levinson

Wundtlaan 1
NL-6525 XD Nijmegen
31 - (0)24 - 3521-911

Bank:
ING 68.13.85.553
Postbank 1278751

Email: pawi@mpi.nl
Fax: 31-(0)24-3557950
Web: www.mpi.nl

Mailing address:
Pb 310
NL-6500 AH Nijmegen
Nijmegen, 15.5.1999

Dear Sirs,

With this letter we would like to formulate our great interest in a tight interaction with the TalkBank project as suggested by CMU and LDC. It is in the line of developments which are very important for many areas of research where linguistic resources form the basis material. Since modern linguistic resources contain the raw audio and video sources to give immediate access to multi-modal information to the researcher, we need a new type of unifying access and annotation tools. TalkBank is suggested exactly along these lines.

At the MPI we are developing such a set of new tools since about 2 years (LUDICO and BC - see our Web-site) with similar intentions as formulated for the TalkBank project. A tight synchronisation of ideas and perhaps even sharing of code would be helpful for both projects.

Yours sincerely

Peter Wittenburg
Technical Director
12 May 1999

Dear Brian MacWhinney,

I am writing to confirm my interest in participating in the TalkBank program - I believe it to be a far sighted and innovative interdisciplinary program which will stimulate much new research.

Yours sincerely

[Signature]

Dr David Zeitlyn
Senior Lecturer in Social Anthropology