10-1-2001

From CHILDES to TalkBank

Brian MacWhinney
Carnegie Mellon University, macw@cmu.edu

From CHILDES to TalkBank

Brian MacWhinney
Carnegie Mellon University

MacWhinney, B. (2001). New developments in CHILDES. In A. Do, L. Domínguez & A. Johansen (Eds.), *BUCLD 25: Proceedings of the 25th annual Boston University Conference on Language Development* (pp. 458-468). Somerville, MA: Cascadilla.

a similar article appeared as:

MacWhinney, B. (2001). From CHILDES to TalkBank. In M. Almgren, A. Barreña, M. Ezeizaberrena, I. Idiazabal & B. MacWhinney (Eds.), *Research on Child Language Acquisition* (pp. 17-34). Somerville, MA: Cascadilla.

Recent years have seen a phenomenal growth in computer power and connectivity. The computer on the desktop of the average academic researcher now has the power of room-size supercomputers of the 1980s. Using the Internet, we can connect in seconds to the other side of the world and transfer huge amounts of text, programs, audio and video. Our computers are equipped with programs that allow us to view, link, and modify this material without even having to think about programming. Nearly all of the major journals are now available in electronic form and the very nature of journals and publication is undergoing radical change.

These new trends have led to dramatic advances in the methodology of science and engineering. However, the social and behavioral sciences have not shared fully in these advances. In large part, this is because the data used in the social sciences are not well-structured patterns of DNA sequences or atomic collisions in super colliders. Much of our data is based on the messy, ill-structured behaviors of humans as they participate in social interactions. Categorizing and coding these behaviors is an enormous task in itself. Moving on to the next step of constructing a comprehensive database of human interactions in multimedia format is a goal that few of us have even dared to consider. Surprisingly enough, some of the most recent innovations in Internet and database technology are designed to address exactly this problem. Unlike the structured databases of relational database programs like Excel or Access, the new database formats are designed specifically to handle messy, ill-structured data such as that found in human communication. In particular, the new framework of XML, XSL, and XML-Schema that is being developed by the World Wide Web Consortium or W3C (http://w3c.org) can be applied to represent language data. This interlocking framework of programs and protocols allows us to build new systems for accessing and sharing human language data. At the same time, improvements in computer speed, disk storage, removable storage, and connectivity are making it easier and easier for users with only a modest investment in equipment to share in this revolution.

Among the many fields studying human communication, there are two that have already begun to make use of these new potentials. One of these fields is the child language acquisition community. Beginning in 1984, with help from the MacArthur Foundation, and later NIH and NSF, Catherine Snow and Brian MacWhinney developed a system for sharing language-learning data called the Child Language Data Exchange System (CHILDES). This system has been used extensively and forms the backbone of much of the research in child language of the last 15 years. A second field in which data-sharing has become the norm is the area of speech technology. There, with support from DARPA and a consortium of businesses and universities, Mark Liberman and Steven Bird have organized the Linguistic Data Consortium (LDC). The corpora of the LDC now also function as the backbone for the development and evaluation of technologies for automatic speech recognition and generation.

Recognizing the positive role of data sharing in these two fields, the National Science Foundation has recently provided funding for a major new data-sharing initiative in the social sciences. This new project is called TalkBank and it is a direct outgrowth of the CHILDES and LDC Projects. The goal of TalkBank is the creation of a distributed, web-based, data archiving system for transcribed video and audio data on communicative interactions. The work on this new project has its roots in CHILDES and the LDC. However, this new project seeks to construct a new set of tools and standards that will be responsive to the research needs of a still wider set of research communities. In order to understand where the TalkBank Project is heading, we need to step back a bit to take a look at how students of human behavior and communication have been analyzing their data up to now.

1. Transcription

In traditional societies, communication occurs exclusively in face-to-face encounters. These encounters can arise spontaneously, or they can involve highly scripted social formulas. In modern societies, conversations can also take place across phone lines and video connections. In addition to spoken interactions, there are interactions that use written forms as in letter writing and email. The focus of TalkBank is on the study of all forms of spoken or signed interactions, although written interactions are also of occasional interest. Whatever the specific format, each communicative interaction produces a complex pattern of linguistic, motoric, and autonomic behavior. In order to study these patterns, scientists produce transcripts that are designed to capture the raw behavior in terms of patterns of words and other codes. The construction of these transcripts is a difficult process that faces three major obstacles.

Lack of coding standards. The first major obstacle is the lack of established coding standards that can be quickly andreliably entered into computer files. The most complex set of codes are those devised by linguists. For transcribing sounds, linguists rely on systems such as the International Phonetic Alphabet. However, until very recently, there have been no standard ways of entering phonetic codes into the computer. For words, we all use the standard orthographic forms of our language. However, the match between standard word and the actual forms in colloquial usage is often inexact and misleading.

To code morphology and syntax, dozens of coding systems have been devised and none has yet emerged as standard, since the underlying theory in these areas continues to change. Similarly, in areas such as speech act analysis or intentional analysis, there are many detailed systems for coding, but no single standard. The superficial display form of a transcript and the way in which that form emphasizes certain aspects of the interaction is also a topic of much discussion (Edwards & Lampert, 1993; Ochs, 1979).

Indeterminacy. The second major problem that transcribers face is the difficulty of knowing exactly what people are saying. Anyone who has done transcription work understands that it is virtually impossible to produce a perfect transcription. When we retranscribe a passage we almost always find minor errors in our original transcription. Sometimes we mishear a word. In other cases, we may miss a pause or a retrace. Often we have to guess at the status of a word, particularly when it is mumbled or incomplete. Child language interactions present a particularly serious challenge, because it is often difficult to know what to count as an utterance or sentence. All of these issues in transcription have been discussed in detail in the CHILDES Manual (MacWhinney, 2000), but it is important to realize that some of these problems simply cannot be resolved. This means that we must accept of a certain level of indeterminacy in all transcription.

Tedium. The third problem that transcribers face is related to the second. Researchers often find that it takes over ten hours to produce a useable transcript of a single hour of interaction. Transcribing passages of babbling or conversations with high amounts of overlap can take up to 20 hours per hour or more. The time commitment involved here is considerable and can easily detract from other important academic goals. Sometimes, when teaching researchers how to use the transcription format of the CHILDES system, I am asked whether these programs will automatically generate a transcript. Would that life were so easy! The truth is that automatic speech recognition programs still struggle with the task of recognizing the words in the clear and non-overlapped speech of broadcast news. As soon as we start working with spontaneous speech in real conditions, any hope for automatic recognition is gone. It will be still several decades before we can achieve truly automatic transcription of natural dialogs.

Tedium also arises during the final phases of transcription and the process of data analysis. During these stages, researchers need to check their transcriptions and codes against the original audio or videotapes. The problem is that doing this involves a tedious process of rewinding the tape, trying to locate a specific passage, word, or action. Consider the example of a researcher, such as Adolph (1995), who is interested in observing and coding 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, 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 the 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. This legacy technology is extremely fragile, cumbersome, and unreliable.

A direct solution. There is now an effective way of dealing with the three-headed monster of indeterminacy, tedium, and lack of standards in transcription. The solution is to use programs that link transcripts and codes directly to the original audio or video data. The idea here is extremely simple. It involves an "end run" around the core problems in transcription. Since transcriptions and codes will never fully capture the reality of the original interaction, the best way for researchers to keep in contact with the data is to replay the audio or video after reading each utterance in the transcript. In the era of VHS video and casette-based audio, this solution was possible in principle, but extremely difficult in practice. However, linking of transcripts to audio and video is now extremely simple, once one learns the basics.

The first step in linking transcripts to video is to digitize the media. Researchers who are new to digitization can find descriptions of the procedures on the web at http://childes.psy.cmu.edu and at http://talkbank.org. Digitizing audio files is extremely easy. All one needs is a computer, a sound card, digitizing software such as SoundEdit or CoolEdit, and the proper cable connections. Once several hours of sound have been digitized, the output can be written from the hard disk to a recordable CD-ROM for storage and later transcription.

For video, the process is similar, but a bit more time-consuming and costly. An excellent current digital format is mini-DV. However, for data from older studies, we first have to convert VHS video to digital format. The JVC SR-VS10 dual-deck system provides a great way of both converting VHS to mini-DV, as well as providing smooth access to the computer through the IEEE or FireWire port. Digitization can be done within a variety of programs on both Macintosh and Windows computers. However, we are currently using Final Cut Pro for digitization and Media Cleaner with the Sorensen codec for compression. All of this technology is rapidly changing and new options will soon be available. What is important is simply the fact that all of the pieces for solving this problem are now in place for consumer-level machines at reasonable prices.

Audio digitization is far easier that video digitization. Digital audio files directly address the three core problems in transcription that we have mentioned. However, for certain types of interaction, researchers may feel that video is crucially necessary. If the researcher wants to pay close attention to the positions of the speakers, their gestures and facial expressions, and their use of external objects, then video is indispensable. The point I wish to make here is that both digital audio and digital video are excellent solutions to the core problems in transcription. Audio is easier to produce, but video is preferable for microanalytic studies of the details of interactions.

Linking. Once the recording has been digitized, we are ready to begin transcription. This process relies on special software to control a two-pass process in which transcription and linking are done within the same software application. The two pieces of software that can control this two-pass transcription process are Transcriber, a system developed by Claude Barras at LIMSI (http://www.etca.fr/CTA/gip/Projets/Transcriber/) and CLAN, the CHILDES editor program (http://childes.psy.cmu.edu). These two systems work in the same fashion, but I will describe the process for CLAN.

To begin the first pass of this process, you open a new blank file in CLAN, insert an @Begin line and an @Participants line for the speakers in the file. You then use the F5 key to locate a sound or video file. The sound or video file begins to play and you press the space bar at the end of each utterance. This automatically inserts a new line for the preceding utterance along with a bullet that contains the time codes that link each line of the transcript to a segment of the digitized audio or video. You listen through the whole digitized file completely, pressing the space bar at the end of each utterance. You will often encounter problems deciding when an utterance has ended, but try not to stop the process. You can correct these problems in the second pass. This first takes only one hour to segment one hour of dialog, since this is done in real time. Once you are finished with this first pass, you can display and then rehide the time marks using escape-A.

In the second pass, you use the bullets you entered as a way of replaying the audio or video. CLAN provides additional keys for several functions. You can replay a sound using command-click at the bullet. There are keys for moving up and down from bullets. You can use the keys in the Tiers menu to insert speaker codes. You use the normal text editor functions to transcribe the utterance. If you need to change the borders of the demarcated sound, there are keys for adjusting the front or the end of the sound segment. Using these new transcription methods, transcription time can be reduced by at least 40% from older approaches.

Linking the Existing Database. By linking transcripts to the original recordings, we have lifted a burden off of the shoulders of transcription. Without linkage, transcription is forced to fully represent all of the important details of the original interaction. With linkage, transcription serves as a key into the original recording that allows each researcher to add or modify codes as needed. If a phonetician does not agree with the transcription of a segment of babbling, then it is easy to provide an alternative transcription.

The linkage of transcripts to recordings opens up a whole new way of thinking about corpora and the process of data sharing. In the previous model, we could only share the computerized transcripts themselves. For some important child language corpora, such as the Brown corpus, the original recordings have been lost. For others, however, we have been able to locate the original reel-to-reel recordings and convert them to digital files. We have done this for the corpora from Hall, Wells, Peters, Bernstein, MacWhinney, Sachs, Feldman, and Korman. Hopefully, we will be able to digitize still other corpora in the future. For the Bernstein and MacWhinney corpora, we have used the first-pass linking process to create rough links between the existing transcripts and the newly digitized files. These new data are now available from http://childes.psy.cmu.edu and will eventually be distributed on DVD disks. In the future, many new contributions of data to CHILDES will already be linked, just as many of the core transcripts in the LDC database are already linked. The first contributed corpus that included links was Susanne Miyata's Tai corpus. In the near future, we look forward to including various new linked corpora, including data from the ESF second language project.

2. Collaborative Commentary

An important side effect of this new way of thinking about corpora is the possibility of collaborative commentary. The idea of providing alternative views of a single target is at the core of many areas of historical analysis and literary criticism. However, these fields deal with written discourse, rather than spoken discourse. The works of Shakespeare, Joyce and others have now been digitized and it is easy to refer to specific passages directly. But this was easy to do even in the period before the advent of computers. In the area of spoken discourse, direct reference to a corpus is far more difficult. However, there is now a precedent for this in the field of classroom discourse. This ground-breaking work was contained in a special issue in 1999 of Discourse Processes, edited by Tim Koschmann. This special issue focused on a 5-minute video of an interaction in a problem-based learning (PBL) classroom for medical education. The six students were attempting to diagnose the etiology of a case of an apraxic, amnesic, dysnomic. This interaction was digitized into MPEG format and included at the back of the special issue as a CD-ROM, along with a transcript in Conversation Analysis (CA) format. However, the transcript was not linked to the video and the five commentary articles made reference to the video only indirectly through the transcript. Despite these limitations, this special issue established a model in which researchers from differing theoretical positions could provide alternative views of the same piece of data. In the next iteration of this process, which is scheduled for a forthcoming special issue of the Journal of the Learning Sciences, a second group of researchers, directed by Anna Sfard and Kay McClain, will use a video segment that is linked to a CLAN transcript. The focus of this group is on students' understanding of graphic representations of numerical data. The CD-ROM will include copies of the articles in HTML format with links that directly play video segments through QuickTime and a browser.

These two initial experiments in collaborative commentary only begin to illustrate the ways in which shared, linked, digitized data can reshape the process of scientific investigation. Consider the application of this technology to the study of child language acquisition. One model uses relies on small clips from a larger transcript as the basis of commentary. For example, Ann Peters has contributed a set of illustrations of her subject Seth's use of fillers. Currently, these examples are provided as illustrations, rather than as evidence in support of a particular theory. However, it is clear that some of the examples could be subjected to multiple interpretations. For example, it appears that one of Seth's fillers may be simply a reduced form of the progressive –ing. If a reader of the CHILDES home pages wishes to add this observation to Ann's commentary, we will need to have a mechanism in the HTML pages for comment insertion.

Another approach relies not on small clips, but on larger collections of files or whole corpora. For example, researchers in childhood bilingualism are currently debating the extent to which there may be interlanguage effects in two- and three-year-old bilinguals. Examples of transfer between languages (Döpke, in press; Hulk & van der Linden, 1998) can also be interpreted as due to errors or incomplete learning of one of the languages. In order to resolve such issues, it would be very helpful to have complete access to all of the data involved, along with direct HTML links illustrating specific claims regarding examples of transfer. If the data were made available in this way, it would be possible to directly compare alternative accounts in terms of both qualitative and quantitative claims.

A third model for collaborative commentary involves even deeper coding and analysis of data. Currently, the CLAN programs provide only a limited set of tools for transcript coding. The main tool in this area is Coder's Editor, which allows the researcher to construct a set of codes that are then applied in lock-step fashion to each utterance in a transcript. Workers in the tradition of “qualitative analysis” have developed more sophisticated programs such as *NUDIST and NVivo which give the analyst more dynamic control over both the coding scheme and the way in which it is linked to transcripts. As we move toward a fuller understanding of the process of collaborative commentary, it will be necessary for us to support more powerful approaches of this type.

3. A Community of Disciplines

TalkBank seeks to provide a common framework for data sharing and analysis for each of the many disciplines that studies conversational interactions. The major disciplines involved include Psychology, Linguistics, Speech and Hearing, Education, Philosophy, Computer Science, Business, Communication, Modern Languages, Sociology, Ethology, Anthropology, and Psychiatry. Within each of these larger traditional disciplines, there are subdisciplines that concern themselves specifically with conversational interactions. For example, within the larger discipline of Education, there is the subdiscipline of Educational Psychology that studies classroom discourse. We have identified 16 such subdisciplines that are specifically concerned with the same basic issues in transcription and analysis that we have faced in child language. We are currently organizing meetings of researchers in each of these subdisciplines to collect a better understanding of their specific needs for transcription software and systems for data sharing. The original TalkBank proposal included a list of 50 researchers from these 16 fields. As we progress, we hope to expand this list to include a much fuller representation of each of the fields involved.

The first four meetings we have organized have focused on these four subdisciplines: classroom discourse, animal communication, field linguistics, and computational analysis. Let me summarize here what we have learned from these first four meetings. Detailed reports, ongoing activities, along with a list of the participants are available from http://talkbank.org .

1. **Classroom discourse**. Researchers in educational psychology have a long history of relying on videotape to study classroom interactions. It is clear that the technology we are developing will have a major impact on this field and there are now 12 new projects relying on new TalkBank technology. Despite this immense positive interest, it has been difficult to develop a system for data sharing in the area of classroom discourse. The major problem involves securing permission from children and teachers to open video recordings to scientific analysis. In some cases, teachers are concerned that they will be subject to unfair criticism and even job discrimination or litigation. In other cases, parents are unwilling to have their children filmed for fear that their learning will be criticized. Dealing with these problems will require the creation of special systems for data protection that we will discuss later. Classroom discourse also requires extremely detailed use of ethnographic methods for linking types of data relevant to instructional episodes. These data may include notebooks, room layouts, songs, graphs, diaries, homework, and a wide variety of other materials. TalkBank is committed to providing ways of digitizing records for all of

these formats. Workers in classroom discourse make use of a wide variety of display methods for their data. These include the standard transcript format of CHAT and CA, left-to-right viewers such as SyncWRiter, and spreadsheet formats with both columns and rows. By relying on XML for data storage, it will be relatively easy for TalkBank to display a core set of data in each of these alternative display forms as desired by the researcher.

2. **Animal communication**. The concept of data sharing would seem to be a natural for the area of animal communication. There is already an archive for bird song at the Cornell Laboratory of Ornithology. However, researchers in this field had not yet considered the possibility of developing a generally available archive of data from a wide variety of species. The major problems facing data sharing in this area are technical. First, researchers need to adapt a standard format for audio and video recordings and the linkage of these data to annotations. Most data in this field are best represented in spreadsheet format with rows indicating successive sounds ordered in time and columns representing changing aspects of the environment. We have already built three simple tools for entering data in this area. They have been designed specifically for meerkats, vervets, and dolphins. These systems are essentially alternative data-entry systems, since all the data are stored in a common underlying XML-based format. The second major problem facing this field is the fact that the data files are often huge. The problem is not one of storage, since disk space is now extremely inexpensive. Rather, the problem is one of transmitting huge files across the Internet. To deal with this, we have to rely on complete access to all files through XML-based tools.
3. **Field linguistics**. Linguists have always been concerned with studying the great diversity of languages that exists on our planet. However, many of the languages spoken by small groups of people are now under great pressure and will become extinct by the end of the century. One of the major goals of TalkBank is to develop effective tools for storing transcribed data from these many endangered languages, as well as the hundreds of other diverse languages that will survive into the next century. The community that studies these languages has already made important steps toward beginning a process of data sharing. One initiative, sponsored by a variety of groups summarized at http://www.ldc.upenn.edu/atlas involves the construction of a set of MetaData descriptors that will allow researchers to locate data on the Internet on specific languages. However, once these data are located, researchers will currently be faced with a diversity of formats and programs for data access and analysis. To overcome this problem, TalkBank will work in collaboration with groups such as the Summer Institute of Linguistics (SIL) to provide users and database developers with a uniform set of XML-based tools for constructing transcripts linked to audio, lexical databases, and grammars linked to examples. Finally, this field faces a series of concerns regarding the rights of the native peoples who have contributed the original data and the linguist who worked with these people. Protection of these rights will be an ongoing concern in this area.
4. **Computational analysis**. The fourth meeting that we convened during the past year focused not so much on a particular type of data, but on general methods for representing and processing data. This meeting reached a number of important conclusions regarding the shape of future program development for TalkBank. It

was agreed that the ATLAS format for annotation graphs (Bird *et al.*, 2000b) can function well as the basic format for data representation. However, it was also noted that much of the functionality of annotation graphs can already be represented in XML (Thompson & McKelvie, URL). Moreover, elaborations of annotation graphs for higher-dimensional spaces may not be necessary for the work of creating a linguistic database or querying that database (Bird *et al.*, 2000a). Given this, it makes sense to continue work using standard XML tools, while continually verifying that these tools work well with annotation graphs. This meeting also considered the issue of providing a standard user interface for TalkBank tools. This interface, which would be available in the open-source model, would rely primarily on C++ modules with a series of plug-ins that could be controlled by a scripting language such as Python.

During the 2001, we plan to meet with four additional groups:

5. **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. This field is just now beginning consideration of data sharing. Unfortunately, many of the transcripts that have served as models in this field cannot be shared because permissions were never obtained from the original participants. This means that TalkBank will need to focus on the collection of new data for CA analysis.
6. **Text and discourse**. Closely related to Conversation Analysis is the field of Text and Discourse that is loosely identified with the Society for Text and Discourse. Focusing on both spoken and written discourse, researchers in this field have emphasized structured systems for text comprehension and verbal problem solving. Unlike conversation analysis, the field of text and discourse emphasizes the construction of complex cognitive structures to account for the various features of conversation. For this group, the construction within the TalkBank framework of systems that can support multimedia analysis as well as qualitative analysis of codes as in NVivo will be important. This field has not yet established solid plans for data sharing and may face problems similar to those in the CA field.
7. **Gesture**. Researchers studying gestures have developed sophisticated schemes for coding the relations between language and gesture. For example, David McNeill and his students have shown how gesture and language can provide non-overlapping views of thought and learning processes. A number of laboratories have large databases of video recording of gestures and the introduction of data sharing could lead to major advances in this field.
8. **Signed Language**. There are several major groups studying the acquisition of signed languages. One group uses the CHAT-based Berkeley System of Transcription.

Other researchers use either the SignStream system developed by Carol Neidle or the Media Tagger system developed by Sotaru Kita. Other groups use adaptations of CHAT and SALT. Because each of these groups is heavily committed to its own current approach, it may be difficult to find a common method for data sharing. However, by relying on XML as an interlingua, it should be possible to store data from all of these formats in a way that will permit movement back and forth between systems. However, the details of this will need to be worked out in a meeting with the various groups involved.

In addition to the four meetings planned for 2001, we hope to work with at least 8 additional groups over the remaining years of the project:

9. **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 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. TalkBank will work to create a process of data sharing that will address both of these problems.
10. **Corpus linguistics**. Although a great deal of corpus linguistics focuses on written documents, there are 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 seek to involve dozens of researchers in this tradition. Many of these corpora are currently available to researchers. However, often the conditions of access are rather restricted. TalkBank will work to improve access to corpora and uniform methods for data access and analysis. Crucially, it will be necessary to convert all of these corpora to XML format. This will be a major task.
11. **Speech production, aphasia, language disorders, and disfluency**. The facilities provided by TalkBank are also relevant to 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. Researchers in these areas have been very open to data sharing for transcripts and even audio data. However, there are serious confidentiality concerns that arise once we consider sharing video data, since it is important that the children and patients involved not be subject to any perception of discrimination as a result of their disabilities.
12. **First language acquisition**. There are now over 1200 published studies of first language acquisition that have relied on the use of the CHILDES database. 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. As we progress with developments in TalkBank, it will be necessary to maintain ongoing communication

with the child language community to make sure that the new TalkBank software properly addresses its needs. Of particular concern are the use of video, the transcription of code switching, and the representation of non-Roman alphabets.

13. **Cultural anthropology**. The interests of cultural anthropologists often overlap those of field linguists. Howoever, the two groups use rather different methodologies. In particular, at the turn of the century, ethnographers pioneered the use of film documentaries to record the lives of non-Western peoples. Modern-day anthropology has continued its reliance on film and video to record aspects of other cultures. For this reason, we believe that the use of multimedia in TalkBank could be of particular interest to cultural anthropologists, as long as concern is taken to preserve the rights of the people's been recorded and the ethnographers doing the field work. Currently, the only major system for data sharing in this field is the Human Relations Area Files (HRAF) at Yale. However, this system is largely devoted to the archiving of field notes, rather than actual recordings of interactions.
14. **Psychiatry, conflict resolution**. Psychiatrists such as Horowitz (1988) 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 within Ethics. Currently, there are no systems for data sharing in these fields.
15. **Behavioral analyses**. Researchers in areas of developmental psychology such as attachment, emotion, and socialization have collected enormous archives of videotape material, much of it using longitudinal designs. This material, if it could be opened to data sharing, would constitute an invaluable source of information on human development and social interaction. There are several research groups who are willing to share their data. The major barrier to data-sharing here is economic. The sheer volume of video data would make it impossible to digitize the whole corpus for distribution on the web. It will probably be necessary in this field to organize an alternative system in which most of the data are digitally archived and only a subset of the data are available for on-line access.
16. **Human-Computer Interaction**. Computer scientists are becoming more and more interested in constructing computational agents that can interact in human ways with human computer users. Some laboratories are building animated faces and bodies that express human gestures and facial expressions. These researchers also need to trace the responses of computer users to these new agents. Computer scientists are also interested in constructing automatic representations of ongoing discourse to facilitate the accuracy of speech recognition. Workers in the area of data-mining are interested in extending their techniques to spoken interactions as well as written language. As video data become increasingly available on the web, new methods for data-mining will need to build methods for automatic face and scene recognition. All of these computational challenges can be furthered by the construction of the various TalkBank databases. Moreover, computer scientists themselves can often contribute data that they are collecting.

Our initial plan is to work with each of these 16 groups in a partially separate fashion. However, as the work progresses, we will see more and more interactions between these groups as they begin to work to analyze a shared database.

4. The Next Steps

In this section, I will outline our plans for TalkBank development activities for the next three years. It is important that workers in child language and related fields understand the shape of these activities, so they can make optimal use of the new tools that will be available. Work on the CHILDES project has already benefited to some degree from the spread of ideas between LDC, CHILDES, Informedia, the Wisconsin Center for Educational Research, and other groups. However, in the short term, progress on the core CHILDES tools will be slowed during 2001, as we develop the new TalkBank framework. The reformatting of the database into XML will not impede the additions of new corpora, although it will require some additional work on our end. However, the development of new features for the CLAN programs will be essentially frozen during 2001, as we build the new computational framework. Beginning in 2002, child language researchers will be able to make use of this new framework. In this section, I will outline the major new tools that will be created in the new framework.

Coder. One of the first tools we propose to create is a flexible tool for qualitative data analysis called Coder. Functioning much like *Nudist or NVivo, Coder will allow the user to create and modify a coding framework which can then be applied to various segments of the transcript. Because the underlying data will be represented in XML, we can view Coder as an XML editor in which tags are created on the fly. These tags will be represented in the X-Schema representation of the data. Users will not need to know anything about XML or X-Schema. What they will see is something much like a standard editor window with a separate window that displays the coding system. There will be extensive facilities for comments and linkages to programs for finding and tabulating codes.

Displays. A major limitation of the current CLAN programs is the lack of good facilities for building alternate displays of data. CLAN has a method for repressing dependent tiers, a program for adding line numbers called LINES, and two old and seldom used programs for formatting called COLUMNS and SLIDE. These last two have not been rewritten since the days of MS-DOS and 80-column windows. A major goal of our new initiative is the creation of flexible ways of displaying data. One method uses a sliding window, as in SignStream, Media Tagger, and SyncWriter. Another method uses columns as in MacShapa, Excel, or other home grown systems. For each of these display methods, users will want additional features, such as control of colors, scroll bars, and so on. In our new XML framework, developing these new features will be easier and will generalize better across platforms.

Profiles. With the current CLAN system, the construction of developmental profiles requires several steps. One has to select a group of files, impose a set of filters, run analysis programs, and ship the results off to statistical analysis. There are tools for doing all of this, but the options are opaque and the interface is difficult for a novice. New versions of the SALT program do a better job of allowing the user to filter data and compare against a standardized age-matched data set. We need to implement a similar, checklist approach to data analysis within the new TalkBank tools.

Queries. One of the major benefits of the movement to a structured XML database is the facility it gives us for constructing query interfaces. It will be relatively easy to create screens of check boxes that allow users to select specific data fields to be searched for particular strings. Eventually, this system will replace programs such as KWAL and COMBO. The results of queries will be collected in tab-delimited files that can be imported to Excel or other data analysis programs.

Codon. As TalkBank moves into a broader set of user communities, the need to translate between formats increases. Child language researchers have only needed to deal with the SALT and CHAT formats. However, outside of our field, particularly in the fields of speech technology and corpus linguistics there is a virtual Babel of formats. Fortunately, the annotation graph framework allows us to produce a basic translation between formats in terms of links to media. However, a fuller translation of formats requires the construction of semantic equivalencies. To do this, we will need to extend aspects of CHAT. For example, there are a few features of prosodic coding in CA transcription that are not well represented in CHAT. This means that these features need to be added to the more general Codon language. More importantly, CHAT seldom codes features on the phonetic level, so these features will need to be added to Codon. Sometimes coding systems will create largely incommensurate representations of data. For example as comparison of ToBI and Tilt models for coding English prosody relies on units that are not equivalent in terms of their time duration. Although both types of representation can be stored in Codon, this will require that Codon simply incorporate both systems as optional representations.

Distributed access. 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 need to know that it subscribes to the TalkBank standards and can be manipulated with TalkBank tools. This goal will be accomplished through XML validation tools and the construction of MetaData. Much of the video and audio data in this distributed database will be made available through streaming access. Currently, server support for streaming access cannot access segments within larger files. However, we hope that new XML technology will soon remove this limitation.

Confidentiality. 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 proposed series of levels will be further enhanced by technical tools for password protection, domain construction, face blurring, and sound masking.

Commentary. Earlier, we discussed the importance of opening up our data sets to collaborative commentary. In order to facilitate this process we will build web-based systems for introducing new coding lines into our XML database. Researchers will be able to tag either whole transcripts or individual lines for commentary. They will also be able to add commentary to TalkBank web pages, such as the Peter's filler pages. The final addition of commentary to the database will be subject to editorial control.

Teaching. The increased availability of TalkBank data will have important consequences for teaching. By providing examples of specific types of language phenomena, we can directly introduce students to the study of language behavior and analysis. TalkBank will make available materials on gesture-speech mismatch, fillers, code-switching, referential communication, learning of L2 prosody, vervet communication, parrot problem-solving, tonal patterns in African languages, prosody in motherese, phonological processes in SLI, persuasion in small groups, conflict resolution processes, breakdowns in intercultural communication, and a myriad of other topics in the social sciences. 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.

Community Control. Currently, the construction of CHILDES, the LDC database, and TalkBank are very much in the hands of a few individuals. Over the next few years, it is important that this system of control be given back to the community. To do this, we will need to establish links between professional societies and the databases. For example, in child language, there could be a committee of the International Association for the Study of Child Language (IASCL) that supervises additions to the database. In the field of discourse studies, this committee could be associated with the Society for Text and Discourse. Societies such as the LSA or SRCD could form similar groups. These groups would recommend corpora for addition and solicit contributions. They could also be responsible for giving awards for excellent contributions to the database and excellent empirical publications. Finally, they could work with journal editors and granting agencies to maximize contributions of new data to the shared database.

5. Conclusion

It is important that we begin the technical construction of TalkBank now, even as we work towards methods for community control. 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 ongoing proliferation of alternative formats and computational frameworks blocks the possibility of effective collaboration across disciplinary boundaries.

References

Adolph, K. (1995). Psychophysical assessment of toddlers' ability to cope with slopes. *Journal of Experimental Psychology, 21*, 734-750.

Bird, S., Buneman, P., & Tan, W. (2000a). *Towards a query language for annotation graphs*. Paper presented at the LREC 2000, Athens.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., & Liberman, M. (2000b). *ATLAS: A flexible and extensible architecture for linguistic annotation*. Paper presented at the LREC 2000, Athens.

Döpke, S. (in press). Generation of and retraction from cross-linguistically motivated structure in bilingual first language acquisition. In F. Genesee (Ed.), *Bilingualism, language, and cognition: Aspects of bilingual acquisition*. Cambridge: Cambridge University Press.

Edwards, J., & Lampert, M. (Eds.). (1993). *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum.

Horowitz, M. (Ed.). (1988). *Psychodynamics and cognition*. Chicago: University of Chicago Press.

Hulk, A. C. J., & van der Linden, E. (1998). Evidence for transfer in bilingual children? *Bilingualism: Language and Cognition, 1*(3), 177-180.

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

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics*. New York: Academic.

Thompson, H., & McKelvie, D. (URL). Hyperlink semantics for standoff markup of read-only documents. from http://www.ltg.ed.ac.uk/~ht/sgmleu97.html