THE TALKBANK PROJECT

BRIAN MACWHINNEY
Carnegie Mellon University, USA

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 behavioural 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 supercolliders. Much of our data is based on the messy, ill-structured behaviours of humans as they participate in social interactions. Categorizing and coding these behaviours 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. However, recent innovations in Internet and database technology provide excellent methods for building this new facility. 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. XML tools developed by the World Wide Web Consortium or W3C (http://w3c.org) can be applied to represent language data. The interlocking framework of XML programs and protocols allows us to build new systems for accessing and sharing spoken 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 opportunities. One of these fields is the child language acquisition community. Beginning in 1984, with help from the MacArthur Foundation, and later NIH and NSF, MacWhinney and Snow (1985) 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, and the need for improvement in infrastructure for the social sciences (http://vis.sdsc.edu/sbe/), the National Science Foundation provided funding for a new project called TalkBank (http://talkbank.org). The goal of the project is to support data-sharing and direct, community-wide access to naturalistic recordings and transcripts of human and animal communication. Talkbank has identified these seven shared needs:

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

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 behaviour and communication have been analyzing their data up to now.

1 Transcription

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 behaviour. In order to study these patterns, scientists produce transcripts that are designed to capture the raw behaviour in terms of patterns of words and other codes. The construction of these transcripts is a difficult process that faces three major obstacles.

1.1 Lack of coding standards

The first major obstacle is the lack of established coding standards that can be quickly and reliably 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 (International_Phonetic_Association, 1999). 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).

1.2 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 re-transcribe 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 a certain level of indeterminacy in all transcription.

1.3 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 behaviours 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 behaviours to categories. However, access to the full video database involves rewinding hours of tape to access and re-evaluate 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.

1.4 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 re-

searchers 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 cassette-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 (http://talkbank.org/da).

The first step in linking transcripts to video is to digitize the media. 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 iMovie for digitization and Media Cleaner with the Sorensen codec for compression (http://talkbank.org/dv). All of this technology is rapidly changing with new options continually becoming available. What is important is the fact that all of the pieces for solving this problem are now in place for consumer-level machines at reasonable prices.

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. 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.

1.5 Linking

Once the recording has been digitized, we are ready to begin transcription. This process relies on special software that allows the transcriber to link while transcribing. The three pieces of software that can control this two-pass transcription process are TransAna (http://transana.org), Transcriber (http://www.ldc.upenn.edu/mirror/Transcriber/), and CLAN (http://childes.psy.cmu.edu). These three systems work in the same basic way, 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 a @Begin line and a @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.

1.6 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 that we then link to the transcripts. We have done this for older corpora from Bates, Bernstein, Deuchar, Feldman, Hall, Korman, MacWhinney, Ornat, Peters, Sachs, and Snow. Other new corpora, such as those from Forrester, Brent-Siskind, Miyata, Ishii, Thai, FLLOC, as well as virtually all of the corpora in the TalkBank database, have been contributed in already linked form.

All of the child language corpora mentioned above, along with about 100 additional corpora, are available from http://childes.psy.cmu.edu. Full documentation with references can be downloaded in the form of individual electronic manual from http://childes.psy.cmu.edu/manuals. The TalkBank corpora from adults and school-age children are available from http://talkbank.org and the electronic manual for these datasets is available from that site too. Examples of major new TalkBank corpora include the Santa Barbara Corpus of Spoken American English (SBCSAE) and the SCOTUS corpus that includes 50 years of oral arguments at Supreme Court of the United States.

2 Collaborative Commentary

An important side effect of the availability of corpora linked to media is the opening of new opportunities for 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 (1999) which analyzed a 5-minute video of an interaction in a problem-based learning (PBL) classroom for medical education. The six students in the class were attempting to diagnose the aetiology 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 a further refinement of this process, Sfard and McClain (2002) edited a special issue of the Journal of the Learning Sciences based on a video segment linked to a CLAN transcript. The focus of the commentary in this special issue was on students' understanding of graphic representations of numerical data. The CD-ROM included with the special issue 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 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 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, 2000; 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 (http://www.qsrinternational.com/) 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.

In order to expose the CHILDES and TalkBank corpora to collaborative commentary over the web, we developed a series of new computational structures. First, we worked for several years to formulate a consistent XML Schema for all of the current corpora. This new schema was created by extending the CHAT format (MacWhinney, 2000) to include additional conventions from CA, SALT, Discourse Transcription and other coding systems, all focused on extracting a single, coherent underlying set of meaningful coding categories. We

then reformatted all of the CHILDES and TalkBank corpora to match the new standard. Then we built tools for checking the accuracy of the XML by converting CHAT to XML and then back to CHAT to verify accuracy by requiring a complete match across the roundtrip. We then created HTML from this verified XML. For media, we produced hinted streaming QuickTime movies available on our servers. Then, we built a Java Webstart program (http://childes.psy.cmu.edu/tbviewer) that could browse the transcript database. Currently, we are working to use adapt this viewer tool to support collaborative commentary.

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 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 have organized meetings of researchers in seven of these subdisciplines to collect a better understanding of their specific needs for transcription software and systems for data sharing. These meetings included groups in classroom discourse, animal communication, field linguistics, aphasia, child phonology, gesture, and computational analysis. Let us consider some of the current database needs in these fields.

3.1 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.

3.2 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 (http://www.birds.cornell.edu/). 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. Currently, TalkBank has developed datasets of this type for bird song, vervet calls (Seyfarth & Cheney, 1999), and meerkat calls. Other datasets will eventually be added.

3.3 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 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.

3.4 Conversation analysis

Conversation Analysis (CA) is a methodological and intellectual tradition stimulated by the ethnographic work of Garfinkel (1967)and systematized by Sacks, Schegloff, and Jefferson(1974)and others. Recently, workers in this field and the related field of text and discourse 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 (http://www.conversation-analysis.net), 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.

3.5 Gesture and Sign

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. There are also 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.

3.6 *Second language learning and bilingualism*

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. The database now has major corpora from learners of French, Czech, German, English, Japanese, and Spanish. In addition to these new corpora from older second language learners, there are several extensive new video studies of bilingual development in young children. Finally, there are six corpora documenting dual language interaction and code-switching in adult bilinguals.

3.7 *Aphasia*

The facilities provided by TalkBank are also relevant to the study of language disorders. We have now created a password-protected database of 15 corpora of conversations with aphasic patients. As we move to expand this initial database in the context of the AphasiaBank project, we will establish a standardized protocol that will maximize our ability to conduct comparative analyses across patients.

3.8 *First language acquisition*

The most fully developed component of TalkBank is the CHILDES database. There are now 100 CHILDES corpora and over 1500 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.

The second major new facility for child language researchers is the PHON program (http://childes.psy.cmu.edu/phon) developed by Yvan Rose and Greg Hedlund within the TalkBank framework. This program will allow students of child phonology to analyze segmental and prosodic patterns in great detail within and across languages.

3.9 *Cultural anthropology*

The interests of cultural anthropologists often overlap those of field linguists. However, 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, http://www.yale.edu/hraf/). However, this system is largely devoted to the archiving of field notes, rather than actual recordings of interactions.

3.10 *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.

3.11 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 (http://www.informedia.cs.cmu.edu), 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.

4 The Next Steps

In this section, I will outline plans for further developments in TalkBank. We have already discussed the construction of the TalkBank Viewer and the system for collaborative commentary. In addition to this new tool, we hope to eventually construct several additional systems.

4.1 *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.

4.2 *Alternative 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 colours, scroll bars, and so on. In our new XML framework, developing these new features will be easier and will generalize better across platforms.

4.3 *Profiles and Queries*

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. This facility should be linked to an increasingly powerful method for querying the database.

4.4 *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 behaviour 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.

4.5 *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. One solution here is technical. By providing methods for setting up local TalkBank databases, we can distribute control over the system across many research groups. In addition, we 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*

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.

Döpke, S. (Ed.). (2000). *Cross-linguistic structures in simultaneous bilingualism*. Philadelphia: John Benjamins.
Edwards, J., & Lampert, M. (Eds.). (1993). *Talking data: Transcription and coding in discourse research*. Hillsdale, NJ: Erlbaum.
Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice Hall.
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.
International_Phonetic_Association. (1999). *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.
Koschmann, T. (1999). Special Issue: Meaning making. *Discourse Processes, 27*(2), 98-167.
MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language, 12*, 271-295.
Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics*. New York: Academic.
Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.
Seyfarth, R., & Cheney, D. (1999). Production, usage, and response in nonhuman primate vocal development. In M. Hauser & M. Konishi (Eds.), *Neural mechanisms of communication* (pp. 57-83). Cambridge, MA: MIT Press.
Sfard, A., & McClain, K. (2002). Special Issue: Analyzing tools: Perspective on the role of designed artifacts in mathematics learning. *Journal of the Learning Sciences, 11*, 153-388.

On-line references

http://childes.psy.cmu.edu

http://talkbank.org

http://www.etca.fr/CTA/Projects/Transcriber/

http://www.ldc.upenn.edu/atlas

World Wide Web Consortium or W3C (http://w3c.org)