University of Colorado at Boulder ILL

ILLiad TN: 895811

Borrower: PMC

Lending String:
EAU,EEM,*COD,UMC,CSL,RRR,CLU,CUI,NUI,DGU

Patron:

Journal Title: Proceedings of the ... Annual Boston University Conference on Language Development.

Volume: **Issue:**
Month/Year: 2001 **Pages:** 458-468

Article Author: Brian MacWhinney

Article Title: New developments in CHILDES

Imprint: Somerville, Mass. : Cascadilla Press, 1995-

ILL Number: 184413194

Call #: P118 .B63 25th 2001 v.1 or v.2 (please check)

Location: Norlin Library--Stacks AVAILABLE

1080-692X

ODYSSEY ENABLED

Charge
Maxcost: 35.00IFM

Shipping Address:
Carnegie Mellon University-Hunt Library-ILL Office
4909 Frew St.
Hunt Library Room 305
Pittsburgh, Pennsylvania 15213
United States

Fax: (412)268-6944
Ariel:
128.2.21.4/128.2.21.96
Email: hl82@andrew.cmu.edu
Odyssey Address: 128.2.20.146

New Developments in CHILDES

Brian MacWhinney
Carnegie Mellon University

The Child Language Data Exchange System (CHILDES) is an important component of the research infrastructure in the field of language acquisition. Until recently, CHILDES data have been limited to written transcripts derived from audiotapes. In the last five years, the basic CHILDES tools have been extended in the areas of morphological analysis, automatic disambiguatioh, syntactic analysis, CA transcription, text-to-speech linkage, video linkage, and streaming audio and video. Most recently, we have begun to expand the CHILDES system into a more general system for processing language data known as TalkBank. In this paper, I survey these new developments, with particular focus on the TalkBank initiative.

Advances in computer power 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.

Recognizing the positive role of data sharing, 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 project (MacWhinney, 2000) and the database development of the Linguistic Data Consortium at Penn. 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

© 2001 Brian MacWhinney. *BUCLD 25 Proceedings*, ed. Anna H.-J. Do et al., 458-468.
Somerville, MA: Cascadilla Press.

project has its roots in the CHILDES Project. However, TalkBank 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 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. 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 (Ochs, 1979) (Edwards & Lampert, 1993).

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. 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, 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 (Hulk & van der Linden, 1998) (Dopke, in press) 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. 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. 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.
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. We have already built three simple tools for entering data in this area. They have been designed specifically for meerkats, vervets, and dolphins.
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.

4. **Computational analysis**. The fourth meeting that we convened during the past year focused on general methods for representing and processing data. This meeting reached a series of decisions regarding shared computational approaches and an open-source model for code.

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

5. **Conversation analysis**. Recently, workers in this field have begun to publish fragments of their transcripts over the Internet. 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.
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.
7. **Gesture**. Researchers studying gestures have developed sophisticated schemes for coding the relations between language and gesture. 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. 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.

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, including second language learning, corpus linguistics, speech production, aphasia, language disorders, and disfluency, first language acquisition, cultural anthropology, psychiatry, conflict resolution, behavioral analyses in human development, and human-computer interaction. 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. We have developed a series of levels of confidentiality protection designed to address these issues.

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.

5. Conclusion

It is important that we begin construction of TalkBank now. 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

Bird, S., Buneman, P., & Tan, W. (2000). 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. (2000). ATLAS: A flexible and extensible architecture for linguistic annotation. Paper presented at the LREC 2000, Athens.

Dopke, 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. Available: http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.