Introduction to the Database

There are now eight PDF documents that describe the CHILDES database. They are all available at http://childes.psy.cmu.edu/data/manual/ The eight guides are:

1. Intro: This current manual.
2. English: A guide to the English child language data.
3. Bilingual: A guide to the data on bilingual language acquisition.
4. Clinical: Data from adults and children with language disorders in various languages.
5. Narrative: A guide to data on narrations by adults and children in various languages.
7. Romance: A guide to data on the acquisition of Romance languages.
8. Other: A guide to data on the acquisition of other languages.

The data files themselves are available from http://childes.psy.cmu.edu. The database includes a rich variety of computerized transcripts from language learners. Most of these transcripts record spontaneous conversational interactions. The speakers involved are often young monolingual, normally developing children conversing with their parents or siblings. However, there are also transcripts from bilingual children, older school-aged children, adult second-language learners, children with various types of language disabilities, and aphasics who are trying to recover from language loss. The transcripts include data on the learning of 26 different languages.

Guidelines

The fact that so many students of language learning have freely contributed the hard-won results of such time-intensive work to this database is a moving testimony to the vibrancy and openness of the entire language learning community. These scholars deserve our thanks and respect for their important contributions. To express this respect, it is crucial that researchers obey the guidelines given for the CHILDES and TalkBank projects at http://talkbank.org. For the CHILDES project, the most important guidelines are provided in the current introduction.

Researchers who publish articles based on the use of these data must include citations to publications by the contributors of the corpora. The specific citations are listed at the end of the description of each corpus. It is absolutely imperative that publications include at least one reference to a listed article and more where appropriate. In addition, articles using CHILDES data should include a citation to this book: MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

Drafts and preprints of articles based on the use of corpora should be sent to the contributors using the addresses given at the beginning of the description of each corpus.
Introduction to the Database

It is extremely important that these procedures be followed exactly, in order to respect the rights of the contributors.

All participants in the corpora are identified with pseudonyms, except in cases where the participants are the children of the investigators. All participants have given permission to have their transcripts included in the database. We would like to request that all analyses of the data respect the dignity of the people being recorded, whether or not pseudonyms are provided.

Support for the construction of the database has come from the National Institute of Child Health and Human Development (NIH-NICHD) and the National Science Foundation Linguistics Program. To document ongoing requests for support, we need to show that the CHILDES tools are being used productively. This means that we need to receive detailed feedback regarding published articles that have used the programs and database. This information should be sent to: Brian MacWhinney, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA or to macw@cmu.edu through electronic mail.

Data-Sharing

The database can be of vital importance in understanding the process of language learning. These data can be used in a wide variety of ways. For example, a researcher may wish to examine the interaction between language type and pronoun omission in order to evaluate the claims of parameter-setting models. Another researcher may be interested in finding out whether children with Down syndrome are responsive to maternal requests. Another group of researchers may wish to determine the ways in which children first learn mental state verbs such as “remember” or “know.” For each of these problems, conducting an analysis on a small and unrepresentative sample may lead to incorrect conclusions. Because child language data are so time-consuming to collect and to process, many researchers may actually avoid using empirical data to test their theoretical predictions. Or they may try to find one or two sentences that illustrate their ideas, without considering the extent to which their predictions are important for the whole of the child’s language. In the case of studies of pronoun omission, early claims based on the use of a few examples were reversed when researchers took a broader look at larger quantities of transcript data.

Having access to a database in which many children are included also provides us with a way of generalizing our initial results. In some cases, conclusions about individual differences in child language have been based on analysis of as few as two children, and rarely on groups larger than 25. Because statistical tests based on three or four participants have very little power, researchers may avoid the use of statistics altogether in corpora-based studies. This problem arises in a particularly clear form when linguistic or psycholinguistic theory make predictions regarding the occurrence and distribution of rare events such as dative passives or certain types of NP-movement. Because of the rarity of such events, large amounts of data must be examined to find out exactly how often they occur in the input and in the child’s speech. In these and other cases, researchers who are trying to focus on theoretical analyses are faced with the dilemma of
having to commit their time to basic empirical work, rather than being able to focus on
the development of acquisitional theory.

There is now a realistic solution to this dilemma. Using the CHILDES database, a re-
searcher can access data from a number of research projects. The database includes a
wide variety of language samples from a wide range of ages and situations. Although
more than half of the data comes from English speakers, there is also a significant
component of non-English data. The total size of the database is now approximately 300
million characters (300 megabytes). The corpora are divided into five major directories:
English data, non-English data, narrative data, data from clinical populations, and data
from bilingualism and second-language acquisition. In addition, the database includes a
bibliography of work on child language and the MacArthur Communicative Development
Inventory (CDI) database.

**Confidentiality**

For conventional transcripts, it has been relatively easy to maintain confidentiality by
using pseudonyms and eliminating place names from transcripts. In the transcript
database, all of the subjects’ names are pseudonyms with a few important exceptions.
There are several children who were studied by their parents. For these children, it is
impossible to guarantee anonymity in any meaningful way. Therefore, we have asked
these children to give us permission to use their names. All have agreed.

As we move from transcript data to multimodal data, it becomes increasingly difficult to
maintain confidentiality through the simple use of pseudonyms. Researchers and subjects
who would be happy to donate their transcript data to CHILDES might have concerns
about donating the related audio or video data. One approach to this problem that has
been implemented by many local IRB (Institutional Review Board) committees in the
United States 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:

1. Data are fully public (public speeches, public interviews, and so forth) and gener-
   ally viewable and downloadable over the Internet, although they may still be
copyrighted.
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.
3. Transcript data with pseudonyms will be made publicly available. However, the
   corresponding audio or video data, for which anonymity is more difficult to pre-
   serve, will be made available on one of the next six, more restrictive levels.
4. Data are only available to researchers who have signed a nondisclosure form. This
   form sets tight standards regarding avoidance of use of personal names when re-
   quired. 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.

5. Access is restricted to researchers who have signed nondisclosure forms. In addition, copying is disallowed.

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 distribution of data. However, unlike Level 6, this level would also include mechanisms for insuring that the data would not be copied or distributed.

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.

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.

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

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

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

Documentation

With the exception of a few corpora of historical interest, all of the files in the CHILDES database are in chat format and have been run through the CHECK program to guarantee correct usage of the CHAT coding and transcription conventions. Many of the files have full documentation of the conversational context, although the documentation in others is far less complete. An attempt has been made to secure as much documentation as possible from the contributors. For editorial consistency, these descriptions have been edited into a common format, eliminating use of the first person and clarifying certain
points. However, none of the factual elements of these documentation files have been changed.

None of the corpora collected before 1987 were transcribed initially in chat. Instead, we used optical scanning and computer programs to reformat these earlier corpora into the new standard. After reformatting, the transcripts were checked by hand. Since 1987, most of the corpora added to the database have been transcribed directly in chat. These new corpora have certain obvious advantages over the older corpora. First, the corpora that have been transcribed directly into chat make full use of the various contrasts available in CHAT coding. Second, because the new data did not go through a process of optical scanning and reformatting, these processes did not introduce new errors.

In continuing the construction of the CHILDES database, the overall goal is to achieve a continually higher standard of transcription accuracy and contextual documentation. In a sense, the current shape of the database is a statement both about the impressive size of our current child language database and also about the many ways in which that database must be expanded and improved. Recently, we have digitized the audiotapes for about half of the corpora. If sufficient resources become available, these digitized files will be made available over the Internet in the near future. Eventually, we plan to link these audio files to individual utterances in the transcripts.

Users who are ready to contribute their files to the database should first make sure that all files pass the CHECK program and that the 00readme.cdc file is complete and accurate. When this is done, all of the files should be combined into a single zip file. The zip file can then be transferred to CMU by sending it as an e-mail attachment to Brian MacWhinney at macw@cmu.edu.