The Child Language Data Exchange System

Brian MacWhinney
Carnegie Mellon University

Catherine Snow
Harvard University

The CHILDES (Child Language Data Exchange System) project provides an international system for exchanging and analyzing child language transcript data. This system has developed three major tools for child language research: (1) the CHILDES database of transcripts, (2) the CHAT system for transcribing and coding data, and (3) the CLAN programs for analyzing CHAT files. Here we sketch out the current shape of these three major tools and the organizational form of the CHILDES system. A forthcoming book (MacWhinney, in press) documents these tools in detail.[1]

Child language research thrives on naturalistic data – data collected from spontaneous interactions in naturally occurring situations. However, the process of collecting, transcribing, and analyzing naturalistic data is extremely time-consuming and often quite unreliable. To improve this process, the Child Language Data Exchange System (CHILDES) has developed tools that facilitate the sharing of transcript data, increase the reliability of transcription, and automate the process of data analysis. These new tools are bringing about such significant changes in the way in which research is conducted in the child language

field that researchers who deal with naturalistic data will want to understand their nature.

Background

The dream of establishing an archive of child language transcript data has a long history, and there were several individual efforts along such lines early on. For example, Roger Brown's original Adam, Eve, and Sarah transcripts were typed onto stencils and mimeographed in multiple copies. The extra copies have been lent to and analyzed by a wide variety of researchers – some of them attempting to disprove the conclusions drawn from those data by Brown himself! In addition, of course, to the copies lent out or given away for use by other researchers, a master copy – never lent and in principle never marked on – has been retained in Roger Brown's files as the ultimate historical archive.

Such storing and lending of hard copies of transcripts formed an historical precedent for the establishment of a true, comprehensive, international, crosslinguistic child language data archive, but a revolution in the basic conception of such an archive was made possible by the emergence of computers as tools for storage, analysis, and communication. In the traditional model, everyone took his copy of the transcript home, developed his/her own coding scheme, applied it (usually by making pencil markings directly on the transcript), wrote a paper about the results and, if very polite, sent a copy to Roger. The original database remained untouched. The nature of each individual's coding scheme and the relationship among any set of different coding schemes could never be fully plumbed.

The dissemination of transcript data allowed us to see more clearly the limitations involved in our analytic techniques. As we began to compare hand-written and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability became more apparent. But, just as these new problems arose, a major technological opportunity also emerged. As microcomputer word-processing systems became increasingly available, researchers started to enter transcript

data into computer files which could then be easily duplicated, edited, and analyzed by standard data-processing techniques. Computer storage and exchange allow us not only to transcend the limitations of non-computerized analyses, but also to change the basic conception of an "archive." Rather than primarily serving as an historical record, or as a means of short-circuiting the painful, time-consuming process of transcribing for some researchers, a computer archive can become a constantly growing dataset, enriched by every user, since anyone who borrows from the system undertakes at the same time to contribute to the system.

The origin of the CHILDES system can be traced back to the summer of 1981 when Dan Slobin, Willem Levelt, Susan Ervin-Tripp, and Brian MacWhinney discussed the possibility of creating an archive for typed, hand-written, and computerized transcripts to be located at the Max-Planck Institut für Psycholinguistik in Nijmegen. In 1983, the MacArthur Foundation funded meetings of developmental researchers in which Elizabeth Bates, Brian MacWhinney, Catherine Snow and other child language researchers discussed the possibility of soliciting MacArthur funds to support a data exchange system. In January of 1984, the MacArthur Foundation awarded a two-year grant to Carnegie Mellon University for the establishment of the Child Language Data Exchange System with Brian MacWhinney and Catherine Snow as Principal Investigators. These funds provided for the entry of data into the system and for the convening of a meeting of an Advisory Board for the System. The early status of the system is described in MacWhinney and Snow (1985).

The reasons for developing a computerized exchange system for language data are immediately obvious to anyone who has produced or analyzed transcripts. With such a system, we can (1) widen our database, (2) exercise greater scientific precision in coding and transcription, and (3) automate the analysis of large amounts of conversational text. The CHILDES system has addressed each of these possibilities by developing three separate, but integrated, tools. The first tool is the Database itself, the second tool is the CHAT transcription and coding format, and

the third tool is the CLAN package of analysis programs. Let us look at the current status of each of these three tools.

The Database

The first major tool in the CHILDES workbench is the database itself. The importance of the database can perhaps best be understood by considering the dilemma facing a researcher who wishes to test a detailed theoretical prediction on naturalistic samples. Perhaps the researcher wants to examine the interaction between language type and pronoun omission in order to evaluate the claims of parameter-setting models. Gathering new data that are ideal for the testing of a hypothesis may require months or even years of work. However, conducting the analysis on a small and unrepresentative sample may lead to incorrect conclusions. Because childlanguage data are so time-consuming to collect and to process, it is simply not feasible to undertake certain kinds of studies of great potential theoretical interest. For example, studies of individual differences in the process of language acquisition require both an intensive longitudinal analysis and large numbers of subjects – a combination which is practically impossible for a single researcher or a small research team. As a result, conclusions about differences in child language have been based on analysis of as few as two children, and rarely on groups larger than 25. A similar problem arises when linguistic or psycholinguistic theory makes 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. However, there is now a realistic solution to this dilemma. Using the CHILDES database, a researcher can access data from a number of research projects that can be used to test a variety of hypotheses. The CHILDES database includes a wide variety of language samples

from a wide range of ages and situations. Although more than half of the data come from English speakers, there is also a significant component of non-English data. Many of the corpora have been formatted into the CHAT standard and we are now in the process of checking that formatting for syntactic accuracy. The total size of the database is now approximately 140 million characters (140 MB). The corpora are divided into six major directories: English, non-English, narratives, books, language impairments, and second language acquisition.

English Data

The directory of transcripts from normal English-speaking children constitutes about half of the total CHILDES database. The subdirectories are named for the contributors of the data. Except where noted, the data are from American children and are transcribed in CHAT format.

Bates: This subdirectory contains data collected by Elizabeth Bates from videotape recordings of play sessions with a group of 20 children first at 20 months and then at 28 months.

Bernstein-Ratner: These data were collected by Nan Bernstein-Ratner from nine children aged 1-1 to 1-11. There are three samples from each child at three time points, all transcribed from high-quality reel-to-reel audiotapes in UNIBET notation.

Bloom: This subdirectory contains the appendix to Bloom (1970) "One Word at a Time" with language samples from Lois Bloom's daughter Allison between ages 1-4 and 2-10. The subdirectory also contains a large corpus of longitudinal data from Bloom's subject Peter between ages 1-9 and 3-1.

Bohannon: This subdirectory contains transcripts collected by Neil Bohannon from one child aged 2-8 interacting with 17 different adults.

Brown: This subdirectory contains three large longitudinal corpora from Adam, Eve, and Sarah collected by Roger Brown and his students. Adam was studied from 2-3 to 4-10; Eve from 1-6 to 2-3; and Sarah from 2-3 to 5-1.

Clark: This subdirectory contains data from a longitudinal study of a child between age 2-2 and 3-2 by Eve Clark. The transcripts pay close attention to repetitions, hesitations, and retracings.

Evans: This subdirectory contains transcripts contributed by Mary Evans from 16 dyads of first graders at play.

Fawcett: This subdirectory contains data collected by Robin Fawcett from 96 British children aged 6 to 12. The data are accompanied by a full syntactic coding, but are not yet in CHAT format.

Fletcher: This subdirectory contains transcripts from 72 British children ages 3, 5, and 7. The data were collected by Paul Fletcher and are not yet in CHAT format.

Garvey: This subdirectory contains 48 files of dialogues between two children with no experimenter present. Each dyad is taken from a larger triad, so that there are files with A and B, B and C, and C and A from each triad. There are 16 triads in all. The triads range in age from 3-0 to 5-7. The transcriptions are exceptionally rich in situational commentary.

Gathercole: This subdirectory contains cross-sectional data from a total of 16 children divided into four age groups in the period between 2 and 6 years. The children were observed at school while eating lunch with an experimenter present. There is detailed description of actions and situational changes.

Gleason: This subdirectory contains data collected by Jean Berko-Gleason from 24 subjects aged 2-1 to 5-2. The children are recorded in interactions with (1) their mother, (2) their father, and (3) at the dinner table.

Hall: This subdirectory contains extensive data collected by Bill Hall from 38 four-year-olds in a variety of situations. The target children were from four groups: White working class, Black working class, White professional, and Black professional.

Higginson: This subdirectory contains data from 17 hours of early language interactions recorded by Roy Higginson. The children are aged 1-10 to 2-11, 0-11 to 0-11, and 1-3 to 1-9.

Howe: This subdirectory contains data from 16 Scottish mother-child pairs in their homes in Glasgow collected by Christine Howe. The ages of the children are around 1-17.

Korman: This subdirectory contains the speech of British mothers to infants during the first year. The data were contributed by Myron Korman and are not yet formatted in CHAT.

Kuczaj: This subdirectory contains data from a large longitudinal study of Stan Kuczaj's son Abe from 2-4 to 5-0.

MacWhinney: This subdirectory contains data from a longitudinal study of Brian MacWhinney's sons Ross and Mark from 1-2 to 5-0. Data were collected from 5-0 to 9-0, but they are not yet transcribed.

Sachs: This subdirectory contains a longitudinal study of Jacqueline Sachs' daughter Naomi from 1-2 to 4-9. Only partial data are available from 1-2 to 1-8.

Snow: This subdirectory contains a longitudinal study of Catherine Snow's son Nathaniel from 2-5 to 3-9.

Suppes: This subdirectory contains a longitudinal study Patrick Suppes' subject Nina from age 1-11 to 3-3.

VanHouten: This subdirectory contains data from Lori VanHouten comparing adolescent and older mothers and their children at ages 2 and 3.

VanKleeck: This subdirectory contains data from 37 children age 3 to 4 in a laboratory setting contributed by Anne VanKleeck.

Warren-Leubecker: This subdirectory contains data from 20 children interacting either with their mothers or their fathers. One group of children is aged 1-6 to 3-1 and the other group is aged 4-6 to 6-2. The data were contribued by Amye Warren-Leubecker.

Wells: This extensive corpus from Gordon Wells contains 299 files from 32 British children aged 1-6 to 5-0. The samples were recorded by taperecorders that turned on for 90 second intervals and then automatically turned off. The data are not yet in CHAT format.

Non-English Data

With the exception of the data from Afrikaans, Polish, and Tamil, the various non-English data sets have no English glosses or morphemic codings. Therefore, they are currently most useful to researchers who are familiar with the languages involved.

Afrikaans: Jan Vorster of the South African Human Sciences Research Council contributed a large syntactically-coded corpus of data from children between 18 and 42 months learning Afrikaans. The data do not have English glosses, but they are in CHAT format, and, given the extensive syntactic coding, they are well suited for syntactic analysis.

Danish: Kim Plunkett of the University of Aarhus contributed longitudinal data from four children learning Danish. The data are in CHAT format without English glosses.

Dutch: This subdirectory contains a longitudinal study of a single child from Steven Gillis of the University of Antwerp and another longitudinal study by Loekie Elbers and Frank Wijnen of the University of Utrecht. Both corpora are in CHAT.

French: This subdirectory contains a longitudinal study of a single child by Christian Champaud of the CNRS in Paris and another longitudinal study of a single child by Madeleine Leveillé of the CNRS in Paris.

German: This subdirectory contains four corpora. The first is a non-CHAT set of diary notes by Clara and Wilhelm Stern on the development of their three children. The second is a set of transcripts from 13 children between ages 1 and 14 from Klaus Wagner of the University of Dortmund. The third is a set of protocols taken from older children by Jürgen Weissenborn of the Max-Planck Institut in the context of experimental ellicitations of route descriptions. The fourth are transcripts of non-continuous interactions collected by Henning Wode of the University of Kiel from his children in German during a period when they are also learning English. None of the German data are yet in CHAT format.

Hebrew: Ruth Berman of Tel-Aviv University has contributed one longitudinal study of a Hebrew-learning child and cross-sectional transcripts for children from ages 1 to 6. All the data are in CHAT format.

Hungarian: Brian MacWhinney has entered transcripts of four Hungarian children studied for a 10 month period.

Italian: Elena Pizzuto of the CNR in Rome has contributed data in CHAT from a longitudinal study of a single child.

Polish: Richard Weist of SUNY Fredonia has contributed data in CHAT from four children learning Polish. The data are coded morphemically in a way that is very useful for comparative analysis.

Slobin: Dan Slobin of the University of California at Berkeley has contributed data from a comparative study of clausal semantic structures in English, Italian, Serbo-Croatian, and Turkish. Reformatting of the data into CHAT is not yet complete.

Spanish: Jose Linaza of the University of Madrid has contributed data from a longitudinal case study of a child between ages 2 and 4. The data are not yet in CHAT.

Tamil: R. Narasimhan and R. Vaidyanathan of the Tata Institute in Bombay have contributed a longitudinal study of a Tamil child between ages 9 and 33 months.

Narrative Data

The data in this directory are narratives, currently mostly derived from retellings of stories in books and movies.

Gopnik: The files in this directory were contributed by Myrna Gopnik. They are stories elicited by teachers from children between the ages of 2 and 5.

Hicks: The data in this subdirectory were contributed by Deborah Hicks. They were elicited by showing the silent film "The Red Balloon" to children in grades K through 2 and asking them to then tell the story in each of three different genres.

The data are transcribed in CHAT and coded for a variety of anaphoric devices.

Sulzby: The data in this subdirectory were contributed by Elizabeth Sulzby. They contain discussions with children aged 3 and 4 about their favorite books.

Books

The database also includes the complete text of several books and articles. We have obtained permissions from the publishers to include these books in the database. There is also an extensive computerized bibliography of research in child language development.

Carterette and Jones: This subdirectory contains the complete text of "Informal Speech" by Edward Carterette and Margaret Jones. Conversations with first, third, and fifth grade California school children and adults are transcribed both orthographically and in the CHILDES UNIBET phonemic notation. The files were entered from the original computer tape that was used to prepare the book; they are not reformatted into CHAT, but will be in the future.

CHILDES/Bib: With support from CHILDES, Roy Higginson of Iowa State University used a variety of existing resources to compile a rich computerized bibliography of research in child language development that can be searched with the CLAN program called BIBFIND. The status of this independent CHILDES tool is discussed in detail in the accompanying article in this issue by Higginson.

Haggerty: This subdirectory contains the text of an article from 1929 that reports the exact conversation carried on in the length of one day by the author's 31-month-old daughter. The file is not reformatted into CHAT, but will be eventually.

Isaacs: This subdirectory contains the complete text of "Intellectual Growth in Young Children" by Susan Isaacs (1930) and "Social Development in Young Children" by Isaacs (1933). The author records interesting interactions with upper-middle class British children, often in nearly verbatim form.

Weir: This subdirectory contains the phonetic transcriptions from the appendix to "Language in the Crib" by Weir (1970).

Language Impairments

In the next few years we plan to substantially increase the amount of data in the system on language disorders and impairments. Currently, these corpora are available:

CAP: This subdirectory contains transcripts gathered from 60 English, German, and Hungarian aphasics in the Comparative Aphasia Project directed by Elizabeth Bates. The transcripts are in CHAT format and large segments have full morphemic coding and error coding.

Bliss: This subdirectory contains a set of interviews with 7 language-impaired children and their matched normal controls collected by Lynn Bliss at Wayne State University and formatted in CHAT.

Conti-Ramsden: This subdirectory contains transcripts of five British specifically language-impaired preschool children interacting separately with their mothers, their fathers, and a normally developing MLU-matched younger sibling. The data are in CHAT and were contributed by Gina Conti-Ramsden of the University of Manchester. Control transcripts from the sibling interacting with the mother and the father are also included.

Feldman: This subdirectory contains a set of CHAT files collected by Heidi Feldman at Children's Hospital in Pittsburgh from 14 children suffering from various forms of early brain damage. The data are part of an ongoing project entitled "Foundation of Language Assessment" directed by Catherine Snow.

Hargrove: This subdirectory contains a set of interviews in CHAT format between a speech therapist and 6 language-impaired children in the age range of 3 to 6. The files were contributed by Patricia Hargrove of Mankato State University.

Holland: This subdirectory contains a set of interviews with 40 recovering stroke patients who are suffering aphasic symptoms. The data were contributed by Audrey Holland of the University of Pittsburgh and are in CHAT format.

Hooshyar: This subdirectory contains CHAT files collected by Nahid Hooshyar of the Southwest Family Institute from 30 Down Syndrome children between the ages of 4 and 8.

Japanese: This subdirectory contains adult normal Japanese speech error data transcribed in CHAT by Yasushi Terao of Tsukuba University.

Rondal: This subdirectory contains data collected from 21 Down syndrome children in Minnesota by Jean Rondal of the University of Liège. The data have not yet been reformatted into CHAT.

Second Language Acquisition Data

ESF: This subdirectory contains data from the large project on second language learning by immigrant workers directed by Wolfgang Klein at the Max-Planck Institut in Nijmegen. The data are not yet in CHAT format.

Guthrie: This subdirectory contains data in CHAT collected by Larry Guthrie of the Far West Laboratory from three first grade classrooms of immigrant children in San Francisco.

Snow: This subdirectory contains picture descriptions and word definitions in both English and Spanish from 190 Puerto-Rican children in second through sixth grade bilingual classrooms transcribed in minCHAT format. The picture descriptions are coded for explicitness and narrativity. Similar data from an additional 18 fifth graders who are not in bilingual programs, and from 14 third graders who are monolingual Spanish speakers are also included. These data have been contributed by Catherine Snow.

Further information on these corpora can be found in MacWhinney (in press) and in on-line documentation files available with most of the data sets. Researchers can request copies of segments

of the database on either MS-DOS or Macintosh floppies. Copies are sent out free of charge from the Center at Carnegie Mellon and users are asked to return the floppies after copying the data to their hard disk. Copies of data can also be secured from Helmut Feldweg at the Max-Planck Institut für Psycholinguistik in Nijmegen. If members wish to have a complete copy of the database, they need to request data on magnetic tape or in forms compatible with certain specific mass storage devices available for the IBM/XT/AT or the Macintosh.

CHAT

The second major tool in the CHILDES workbench is the CHAT system for transcription and coding. The most conceptually difficult task involved in developing the CHILDES workbench was the creation of the CHAT system. Several years of work with a variety of earlier coding schemes and a great deal of input from our colleagues has led to the formation of the system we call CHAT (Codes for Human Analysis of Transcripts). As discussed in detail in MacWhinney (in press), no coding or transcription system can ever fully satisfy all the needs of all researchers. Nor can any transcription system ever hope to fully capture the richness of interactional behavior. Despite these limitations, the availability of a lingua franca for transcription can facilitate data exchange, data analysis, and the growth of scientific precision.

The CHAT system is designed to function on at least two levels. The simplest form of CHAT is called minCHAT. Use of minCHAT requires a minimum of coding decisions. This type of transcription looks very much like the intuitive types of transcription generally in use in child language and discourse analysis. A fragment of a file in minCHAT looks like this:

@Begin
@Participants: ROS Ross Child BRI Brian Father
*ROS: why isn't Mommy coming?
%com: Mother usually picks Ross up around 4 PM.
*BRI: don't worry.
*BRI: she'll be here soon.
*ROS: good.
@END

There are several points to note about this fragment. First, all of the characters in this fragments are ASCII characters. The @Begin and @End lines are used to guarantee that the file was not destroyed or shortened during copying between systems. Each line begins with a three-letter speaker code, a colon, and then a tab. Each line has only one utterance. However, if the utterance is longer than one line, it may continue onto the next line. A new utterance must be given a new speaker code. Commentary lines and other coding lines are indicated by the % symbol.

Beyond the level of minCHAT, there are a variety of advanced options that allow the user to attain increasing levels of precision in transcription and coding. Some of the major specifications available in the full CHAT system are:

1. File headers. CHAT specifies a set of 24 standard file headers such as "Age of Child," "Birth of Child," "Participants," "Location," and "Date" that document a variety of facts about the participants and the recording.
2. Word forms. CHAT specifies particular ways of transcribing learner forms, unidentifiable material, and incomplete words. It also provides conventions for standardizing spellings of shortenings, assimilations, interactional markers, colloquial forms, baby talk, and certain dialectal variants.
3. Morphemes. CHAT provides a system for morphemicization of complex words. Without such morphemicization, mean length of utterance is computer based on words, as defined orthographically.
4. Tone Units: CHAT provides a system for marking tone units, pauses, and contours.
5. Terminators: CHAT provides a set of symbols for marking utterance terminations and conversational linkings.
6. Scoping: CHAT uses a scoping convention to indicate stretches of overlaps, metalinguistic reference, retracings, and other complex patterns.

7. Dependent Tiers: CHAT provides definitions for 14 coding tiers. Coding for three dependent tiers have been worked out in detail.
   a. Phonological Coding: CHAT provides a single-character phonemic transcription system for English and several other languages called UNIBET. It also provides an ASCII translation for the extended IPA symbol set called PHONASCII. These systems were devised by George Allen of Purdue University.
   b. Error Coding: CHAT provides a full system for coding speech errors.
   c. Morphemic Coding: CHAT provides a system for morphemic and syntactic coding or interlinear glossing.

The full CHAT system is discussed in MacWhinney (in press).

CLAN

The third major tool in the CHILDES workbench is the CLAN package of analysis programs. The CLAN (Child Language Analysis) programs were written in the C programming language by Leonid Spektor at Carnegie Mellon University. They can be compiled to run under MS-DOS, UNIX, VMS, XENIX, or Macintosh operating systems. The Center at Carnegie Mellon provides members with executable versions of CLAN on floppies and with a manual for the programs. Most users install the programs on a hard disk along with CHAT files either from their own research projects or from the CHILDES database.

In MS-DOS and UNIX, CLAN commands are issued as single line commands to the operating system. For example the command

freq -f *.cha

runs the FREQ program on all the files in a given directory with the ".cha" extension. The "-f" switch indicates that the output of each analysis should be written to a file on the disk. Unless specifically given a file extension name, the FREQ program will figure out names for the new files.

Each of the CLAN programs is started up and run separately. The search programs contains options that allow one to focus the analysis on a particular speaker or a particular coding tier. Most of the programs also allow the user to limit the analysis to a particular numerical range of utterances, such as the first 100 utterances. The most useful CLAN programs are:

Check: This program performs a thorough check for adherence to the syntactic specifications of CHAT. However, the user can short-circuit full error checking in a variety of ways.

ChString: This program replaces specific strings in files with other strings. Although such changes can also be done in most text editors, ChString can effect a whole series of changes on a whole collection of files with a single command. The strings to be changed can be specified in a file that is created by the user.

Combo: This program conducts Boolean searches using a variety of logical operators and wild card symbols. For example, using Combo, one can search for all utterances with a wh-word followed somewhere else in the text by a present tense auxiliary. The user can specify the extent of material to be included in the window around the matching search string.

Freq: This program computes a variety of frequency analyses for the words in a file or corpus. The analyses can be for all the words in a corpus or for only those words matching certain search strings. Search strings can be specified with wildcards in a variety of ways. The shape of words can be varied by changing the nature of the punctuation set. Freq is particulary useful useful in providing data summaries for codes added to a transcript, when options for including coding tiers and excluding text tiers are used. A wide variety of statistics can be obtained with this program as with several of the other search programs.

Gem: This program is designed to allow the user to place important passages into a file for later analysis. Using a text editor, the user marks the passages to be stored. Gem then

uses these marks to determine what should be excised and placed in the "gems" file.

KWAL: This program performs an analysis that is somewhat similar to the key-word-and-line analysis found in concordance packages. However, it is not designed to produce a printed concordance, but rather a record that can be used by a researcher who is interested in testing hypotheses against examples. The program can be used with a file of search strings of words of a certain type, such as all the personal pronouns in the language.

MLT: This program allows the user to define words and turns in a variety of ways to provide means and standard deviations for the mean length of turn.

MLU: This program allows the user to define words and morphemes in a variety of ways to obtain different values for the mean length of utterance. The user can also simply use the standard definition of MLU as a default.

Retrace: This program is useful for tracking the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. The program was written by Jeffrey Sokolov of Carnegie Mellon University.

WdLen: This program tabulates word and utterance lengths and prints a histogram of those lengths. Data can also be output for statistical analysis.

In addition to these general-purpose programs there are also a variety of programs for special needs. Special-purpose programs include:

BackW: This program matches tier line codes with the corresponding main line text.

BibFind: This program finds selected entries in an CHILDES/Bib file. See the article by Higginson in this issue.

CapWd: This program prints all capitalized words. Useful for working with proper nouns.

ClanMan: This program types out documentation on a given CLAN program.

Dist: This program lists average distances between words or codes. This program is particular useful for conducting analyses of chains of anaphoric reference or tense marking chains.

Flo: This program adds a "flow" line to a transcript to represent the conversation without any coding or special symbols as a simple string of words.

KeyMap: This program creates an immediate contingency table for a given key search string.

RevConc: This program creates a reverse concordance. Revconc must be run twice together with one run of the Uniq program.

SaltIn: This program takes file in SALT format and converts it to CHAT format.

Uniq: This program displays unique lines. Is most useful when used with RevConc or Wheel.

Wheel: This program "rolls" through text finding word clusters. If the size of the wheel is set to "3," the program will find all clusters of three words within a given utterance. In its current shape this program can do a simple distributional or cooccurrence analysis.

The CapWd, Freq, MaxWd, Wdlen, and Wheel programs use some of the programming concepts found in programs of the same name developed in the HUM package written by Bill Tuttle for producing concordances. The full CLAN system is discussed in MacWhinney (in press).

The Organization of the System

Administratively, the System has three components: the Advisory Board, the centers, and the members.

The Advisory Board

The first meeting of the Advisory Board for the System was held in Concord, Massachusetts between March 15 and March 18. The meeting was organized by Catherine Snow. The board members present were Elizabeth Bates, Ursula Bellugi, Lois

Bloom, Melissa Bowerman, Robin Chapman, Eve Clark, Jane Edwards, Susan Ervin-Tripp, Paul Fletcher, Willem Levelt, Brian MacWhinney, Jon Miller, Ann Peters, Dan Slobin, and Catherine Snow. At this meeting, the Board sketched out the organization of the system, the shape of the database, and the types of programs to be used. No specific decisions were reached regarding a standard transcription system, although a variety of possibilities were explored. It was agreed that, if funding were available, meetings of the Advisory Board should be held every other year. Unfortunately, because of difficulties in securing funding for such meetings, it was only possible to convene subsets of the Board in 1985 and 1987. However, a full meeting is scheduled for the Fall of 1989 with users of the system who are willing to contribute their time and effort to its development and improvement. In addition to the input provided by the Advisory Board, we solicit suggestions from all researchers regarding modifications to CHAT or CLAN and possible additions to the database.

The Centers

Currently, complete copies of CHAT, CLAN, and the database are located at Carnegie Mellon University in Pittsburgh, Harvard University in Boston, Aarhus University in Denmark, and the Max-Planck Institut für Psycholinguistik in Nijmegen, The Netherlands. The most up-to-date versions of CHAT, CLAN, and the database are those at Carnegie Mellon. The other centers receive updates about twice a year. Further centers can be established wherever there are sufficient computational resources to store and transfer the database. The basic functions and shape of the system are duplicated at each center. The centers keep in correspondence by computer mail, phone, and regular mail for updating of files and task sharing.

The Members

Membership in the System is open. However, members must agree to abide by the rules of the System, not to distribute

copies of programs or files without permission, to abide by the wishes of data contributors, and to properly acknowledge the contributors and the system. Any article that uses the data from a particular corpus must cite a reference from the contributor of that corpus. The exact reference is given in a file called 00readme.doc which is distributed along with each data set. Members are urged to support the progress of the System through contributions of data, programming expertise, or professional advice. It is important for all researchers to understand that further development of the CHILDES tools depends entirely on funding support from government agencies and private foundations. Currently, support for the system comes from the National Institute of Child Health and Human Development at the National Institutes of Health. The best way to argue for such support is to show that the CHILDES tools are being used productively. This means that we need to get feedback from users about articles that have been published using CHILDES data or projects which are underway using the CHAT and CLAN tools.

We attempt to keep researchers informed about the development of the system in a variety of ways. From 1984 to 1987, we mailed out a newsletter that reported on a variety of issues in transcript analysis. Beginning in 1988, we decided that it would be better to use our resources to send out frequent updates of the manuals for CHAT and CLAN. We have also established an electronic mailing list which can be used to discuss issues relating to CHILDES work or other topics in child language development. In 1988, we ran three small workshops at Carnegie Mellon designed to familiarize researchers with the use of CHAT and CLAN. In June 1989, we ran a somewhat larger workshop at Harvard University. Similar workshops are planned for 1990 for Boston and for the International Child Language Association Meetings in Hungary. The 1990 Harvard Workshop will be particularly designed to promote use of CHILDES by researchers working in language disorders. There have also been CHILDES workshops in The Netherlands, Italy, and Denmark. We have also delivered brief presentations of key aspects of the system at child language conferences in Stanford, Austin,

and Boston. We have also placed announcements of the system into seven journals.

The CHILDES system is not for everybody. There are many important parts of child language research that remain outside the scope of the CHILDES system. Comprehension data and data from standardized tests are ignored in our current focus on production data. Moreover, some researchers are asking questions that cannot be addressed with anything but entirely new data. For such researchers, only the CHAT and CLAN tools may be interesting. There are still other researchers for whom none of the CHILDES tools are appropriate. There is, of course, no reason that the CHILDES tools should prove to be equally useful to all researchers. However, the increasing use of CHILDES tools in published research over the last two years indicates the extent to which these tools have begun to play an important role in our field.

Future Directions

Although we have completed a great deal of work in the past six years, there is still an enormous amount to be done. Our plans for the future focus on these goals:

1. We hope to complete the reformatting and checking of the current CHILDES database by 1991. Beginning in 1990, we hope that all new data that are to be added to the database will already by in the correct CHAT format and will have correctly run through the CHECK program.
2. Over the next few years, we expect the database to grow beyond the current focus on first language acquisition by normal children. In the future, the database will include large components of second language acquisition data, adult interactional data, and a variety of data on language disorders. Eventually, we may wish to distinguish between the CHILDES system and a larger Language Data Exchange System (LANDES).
3. During 1990, we will publish the CHAT and CLAN manuals in book format. This volume will also include a description

of the database. The publication will be done in a format that will allow us to issue new editions every one or two years, much as is done for statistical packages such as SPSS or BMDP. Work is currenlty underway to develop options that will have the CLAN programs ouput files in a format useful for analysis by statistical packages such as SPSS, SAS, or SYSTAT. Currently, the CLAN programs MLU, MLT, and FREQ generate a small data ouput file for each transcript file anlayzed. The new options will allow the data from each of the separate analyses to be listed together with a subject and/or session indentification number (read off a header line) in a fixed format data matrix which can be used directly for statistical analyses.

4. During the next few years, we will focus increased attention on the development of a parser-tagger for the semi-automatic analysis of morphological and syntactic structure. A simple version of this system already exists, but much more work will be needed before a full version is ready.
5. We hope to develop a workbench for phonological analysis, probably using the Macintosh computer. Parts of this tool such as digitization, signal analysis, and IPA fonts are already available as off-the-shelf products. We hope to put these tools together in a form that will allow researchers and their students to produce reliable phonological transcriptions which can be analyzed automatically.
6. With the basic tools of CHAT and CLAN, we are working on new ways of assessing language development. Together, these new measures and analyses will provide surer foundations for language assessment.

We encourage other researchers to join us in these goals, to make full use of the current CHILDES tools, and to propose to us new directions and possible improvements to the system. Please address correspondence on CHILDES to Brian MacWhinney, Department of Psychology, Carnegie Mellon University, Pittsburgh PA 15212 USA or send electronic mail to brian@andrew.bitnet or brian@andrew.cmu.edu.

Note

1. Support for the CHILDES system and the preparation of this report was provided by NIH grants HD 23388 and HD 23998.

References

Carterette, E., & Jones, M. H. (1974) *Informal speech.* Berkeley: University of California Press.

Haggerty, L. C. G. (1930) What a two-and-one-half-year-old child said in one day. *Journal of Genetic Psychology*, 38, 75-100.

Isaacs, S. (1930). *Intellectual development in young children.* London: Routledge, Kegan, Paul.

Isaacs, S. (1933). *Social development in young children.* New York: Harcourt, Brace, & Co.

MacWhinney, B. (in press). *Computational tools for language analysis: the CHILDES system.* Hillsdale, N.J.: Lawrence Erlbaum.

MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language*, 12, 271-296.

Weir, R. (1970). *Language in the crib.* The Hague: Mouton.