THE CHILDES SYSTEM

Brian MacWhinney

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, described in detail in MacWhinney (1995), are bringing about significant changes in the way research is conducted in the field of child language. This chapter reviews the background to the formation of the CHILDES system, the shape of the basic CHILDES tools, and the relation of particular tools to particular research goals. It concludes with a presentation of extensions to the system that will be developed during the coming decade.

1. 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 (1973) transcripts from the children called Adam, Eve, and Sarah were typed onto stencils from which multiple copies were duplicated. The extra copies have been lent to and analyzed by a wide variety of researchers— some of them (Moerk, 1983) 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. In this traditional model, everyone took his copy of the transcript home, developed his or 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 Brown. 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 mimeographed and photocopied transcript data allowed us to see more clearly the limitations involved in our analytic techniques. As we began to compare handwritten and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability became more apparent. But, just as these new problems were arising, a major technological opportunity also was emerging. Microcomputer word-processing systems allowed researchers to enter transcript data into computer files, which could then be easily duplicated, edited, and analyzed by standard data-processing techniques. The possibility of utilizing shared transcription formats, shared codes, and shared analysis programs started to glimmer on the horizon. The idea of a data exchange system for the study of child language development slowly began to emerge. To track the emergence of this system, let us examine the changes that have occurred in the five major periods of child language research.

Period 1: Naive Speculation

The first attempt to understand the process of language development appears in a remarkable passage from the *Confessions* of Saint Augustine (Augustine, 397). In this passage, Augustine actually claims that he remembered how he had learned language:

> This I remember; and have since observed how I learned to speak. It was not that my elders taught me words (as, soon after, other learning) in any set method; but 1, longing by cries and broken accents and various motions of my limbs to express my thoughts, that so I might have my

> will, and yet unable to express all I willed or to whom I willed, did myself, by the understanding which Thou, my God, gavest me, practise the sounds in my memory. When they named anything, and as they spoke turned towards it, I saw and remembered that they called what they would point out by the name they uttered. And that they meant this thing, and no other, was plain from the motion of their body, the natural language, as it were, of all nations, expressed by the countenance, glances of the eye, gestures of the limbs, and tones of the voice, indicating the affections of the mind as it pursues, possesses, rejects, or shuns. And thus by constantly hearing words, as they occurred in various sentences, I collected gradually for what they stood; and, having broken in my mouth to these signs, I thereby gave utterance to my will. Thus I exchanged with those about me these current signs of our wills, and so launched deeper into the stormy intercourse of human life, yet depending on parental authority and the beck of elders (p. 4).

Augustine's fanciful recollection of his own language acquisition remained the high-water mark for child language studies through the Middle Ages and even the Enlightenment. However, Augustine's recollection technique is no longer of much interest to us, since few of us believe in the accuracy of recollections from infancy, even if they come from saints.

Period 2: Diaries and Biographies

The second major technique for the study of language production is the biographical method pioneered by Charles Darwin. Using note cards and field books to track the distribution of hundreds of species and subspecies in places like the Galapagos and Indonesia, Darwin was able to collect an impressive body of naturalistic data in support of his views on natural selection and evolution. In his study of gestural development in his son, Darwin (1877) showed how these same tools for naturalistic observation could be adapted to the study of human development. By taking detailed daily notes, Darwin showed how researchers could build diaries that could then be converted into biographies documenting virtually any aspect of human development. Following Darwin's lead, scholars such as Ament, Preyer, Gvozdev, Szuman, Stern, Ponyori, Kenyeres, and Leopold created monumental biographies detailing the language development of their own children.

Darwin's biographical technique also had its effects on the study of adult aphasia. Following this tradition, Low (1931), Pick (1913, 1971), Wernicke (1874), and many others presented biographical studies of the language of particular patients.

Period 3: Transcripts

The limits of the diary technique were always quite apparent. Even the most highly trained observer could not keep pace with the rapid flow of normal speech production. The emergence of the tape recorder in the 1950s provided a way around these limitations and ushered in the third period of observational studies. This period was characterized by projects in which groups of investigators collected large data sets of tape recordings from several subjects across a period of 2 or 3 years. As long as there was sufficient funding available, these tapes were transcribed either by hand or by typewriter. Typewritten copies were reproduced using ditto master, stencil, or mimeograph. Comments and tallies were written into the margins of these copies and new, even less legible, copies were then made by thermal production of new ditto masters. Each investigator devised a project-specific system of transcription and project-specific codes. As we began to compare handwritten and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability became more apparent.

Period 4: Computers

Just as these new problems were coming to light, a major technological opportunity was emerging in the shape of the powerful, affordable microcomputer. Microcomputer word-processing systems and database programs allowed researchers to enter transcript data into computer files, which could then be easily duplicated, edited, and analyzed by standard data-processing techniques. In 1981, when the CHILDES Project was first conceived, researchers basically thought of these computer systems as large notepads. Although researchers were aware of the ways in which databases could be searched and tabulated, the full analytic and comparative power of the computer systems themselves was not yet fully understood.

In 1984 a meeting of 16 child language researchers in Concord, Massachusetts, formally launched the CHILDES system with Brian MacWhinney and Catherine Snow as codirectors. The initial focus of the CHILDES project was on the collection of a nonstandard-ized database of computerized corpora. Between 1984 and 1986, our work focused on the assembly of a large computerized database of transcripts. As the database grew, it soon became apparent that researchers needed more than a disparate set of corpora transcribed in a confusing diversity of styles. They needed a consistent set of standards both for the analysis of old data and for the collection and transcription of new corpora. During the period from 1986 to 1991, the CHILDES system addressed these needs 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. These three tools are presented in detail in MacWhinney (1991) and illustrated through practical examples in Sokolov and Snow (forthcoming). Researchers who plan to make use of the CHILDES tools will want to consult both of these resources.

Period 5: Connectivity and Exploratory Reality

At the end of the 20th century, the world of computers went through a series of remarkable revolutions, each introducing new opportunities and challenges. The processing power of the home computer now dwarfs the power of the mainframe of the 1980s, new machines are now shipped with built-in audiovisual capabilities, and devices such as CD-ROMs, DAT tapes, and optical disks offer enormous storage capacity at reasonable prices. This new hardware has opened up the possibility for multimedia access to transcripts of aphasic language production. In effect, a transcript is now the starting point for a new Exploratory Reality in which the whole interaction is accessible through the transcript in terms of both full audio and video images. For those who are just now becoming familiar with this new technology, Table 1 summarizes some of the relevant pieces of hardware and software options.

Term	Explanation
Audiovisual / Multimedia	The use of computers for sound and video
CD-ROM	Device giving access to huge amounts of nonerasable data
CHAT	CHILDES format for transcription and coding
CLAN	The CHILDES data analysis programs
DAT tape	Tapes for storing digitized sound files
Digitized speech	Sound in a format that can be played on a computer
Electronic bulletin board	Computer program that allows discussions through email
E-mail	Personal messages sent between computers
FTP	File transfer protocol transferring data between computers
Hard drive	The device on a computer that stores files
HTML	The format for data on pages accessed over the Internet
Internet	The international set of connections between computers
Macintosh	Computers manufactured by Apple Computer
Unicode	The new format for characters in all the world's languages
XML	The more highly structured replacement for HTML

Most recently, microcomputers all across the world have become interconnected through a global high-speed network called the Internet that supports the movement of all sorts of information, including text, sound, and video. This connectivity between computers is matched by an increasing interactivity between the operating system and individual programs. The user can record a sound in one program, take it immediately to another for detailed acoustic analysis, and then to a third for database storage. Together, these new hardware and software developments have led to an enormous increase in interconnectivity between computers, between programs, and between researchers. We are just now beginning to understand the potential consequences of this connectivity for researchers.

From this quick survey of the development of tools for language analysis, we see that the possibilities for careful, detailed analysis of production data have markedly widened in the last few years. The methodological tools that are now available far exceed those of previous eras. In order to envision the possibilities that are open for constructing a database for aphasia, let us look at the shape of the current database for first and second language development and for language disorders.

As we noted earlier, the three major components of the CHILDES system are the database, the CHAT transcription systems, and the CLAN programs. The next three sections describe these three basic tools.

2. 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 nralistic 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 to test a specific 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 child language data are so time-consuming to collect and process, it may be unfeasible 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 that 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 the 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.

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. All of the major corpora have been formatted into the CHAT standard and have been checked 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, language impairments, bilingual acquisition, and books.

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.All of the English files have been checked for correct use of the CHAT transcrip-tion conventions through use of the Check program.

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. The children were recorded in three contexts in the laboratory with their mothers: free play, snack time, and storytelling.

Bemstein-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 iime points, all transcribed from high-quality reel-to-reel audiotapes in UNIBET notation. In addition to the transcripts with the children are interviews with the mothers.

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.

Braine: This subdirectory contains a short set of early utterances reported in Braine (1976). Three of the children are English speakers, one is a Hebrew speaker, and three are Samoan speakers.

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. We have also tagged these corpora for part of speech and the tagged version of the corpora will eventually be included in the database.

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 ortho-graphically on the main line and in UNIBET phonemic notation on the %pho line.

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.

Fletcher: This subdirectory contains transcripts from 72 British children ages 3, 5, and 7 collected by Paul Fletcher and colleagues.

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 children 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. The study contains a 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.

Haggerty: This file contains material from an article published in 1929 that reports the exact conversation carried on in the length of one day by the author's daughter at age 2;7.

Hall: This subdirectory contains a large corpus of data collected by Bill Hall from 38 4-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 three children are aged 1;10 to 2;11, 0;ll,and l;3to 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 between 1 ;6 and 2;2.

Korman: This subdirectory contains the speech of British mothers to infants during the first year.

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 also collected from 5;0 to 9;0, but these sessions are not yet transcribed.

Peters/Wilson. This subdirectory contains transcripts from Robert Wilson's son Seth from 1 ;7 to 4;1. Seth was visually disabled.

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 of Patrick Suppes's 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;0 and 3;0.

Warren-Leubecker: This subdirectory contains data collected by Amye War-ren-Leubecker 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.

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.

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. All of the data are in CHAT and have passed through the CHECK program without error, although a few of the datasets make use of additional codes found in a "OOdepadd" file.

Afrikaans: Jan Vorster of the South African Human Sciences Research Council contributed a large syntactically-coded corpus of data from children between 1;6 and 3;6 learning Afrikaans. The data do not have English glosses, but they do have extensive syntactic coding which makes them well suited for syntactic analysis.

Danish: Kim Plunkett of Oxford University contributed longitudinal data from two children learning Danish.

Dutch: This subdirectory contains two corpora. The first is a longitudinal study of a single child between 1 ;9 and 1; 11 from Steven Gillis of the University of Antwerp. Additional files from this child, covering a broader age range, will eventually be available. The second corpus is a longitudinal study of three children by Loekie Elbers and Frank Wijnen of the University of Utrecht. The two boys and one girl in this study are in the age range of 2;3 to 3;1.

French: This subdirectory contains a longitudinal study of a single child by Christian Champaud of the CNRS in Paris, a longitudinal study of a single child by Madeleine Leveille of the CNRS in Paris, and a group study by Jean Rondal of the University of Liege.

German: This subdirectory contains three corpora. The first is a set of transcripts from 13 children between ages 1;5 and 14;10 from Klaus Wagner of the University of Dortmund. The second is a set of protocols taken from older children by Jiirgen Weissenborn of the Max-Planck Institut in the context of experimental ellicitations of route descriptions. The third corpus includes transcripts of noncontinuous interactions collected by Henning Wode of the University of Kiel from his four children with a chief focus on his youngest son and daughter.

Hebrew: Ruth Herman of Tel-Aviv University has contributed a longitudinal study of a Hebrew-learning child named Naama between 1;7 and 2;6 and cross-sectional transcripts for children from ages 1 to 6. A third corpus from Dorit Ravid is a longitudinal study of her daughter Sivan from 1;11 to 6;11.

Hungarian: Brian MacWhinney has donated transcripts of five Hungarian children aged 1 ;5 to 3;3. The children were recorded in a free play context in their nursery school.

Italian: Elena Pizzuto of the CNR in Rome has contributed data in CHAT from a longitudinal study of a single child. The research team located in Pisa at the Center for Computational Linguistics and the Stella Maris Institute has donated data from two girls and one boy ages 1;8 to 2;11.

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

Spanish: Jose Linaza of the University of Madrid has contributed data from a longitudinal case study of a child between ages 2;0 and 4;0.

Tamil: R. Narasimhan and R. Vaidyanathan of the Tata Institute in Bombay have contributed a longitudinal study of a Tamil child between ages 0;9 and 2;9.

Turkish: There are two corpora for Turkish. One is a set of interviews from children aged 2;0 to 4;8 donated by Dan Slobin. The second is a set of "Frog Story" descriptions donated by Ayhan Aksu-Koc. The children in both data sets are cross-sectionally sampled.

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 coded for a variety of anaphoric devices.

MacBates 1, 2. These corpora from Elizabeth Bates and Brian MacWhinney were elicited from American, Hungarian, and Italian children using pictures and movies that the children were asked to describe.

Language Impairments

In the past few years, quite a few corpora on language disorders and impairments have been added to the database. All of these corpora are in CHAT and have passed through the Check program.

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

Brinton/Fujiki: This corpus has spontaneous language samples from 65 young-adult Down's syndrome subjects in a half-hour interview format with an adult male experimenter.

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.

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 corpus includes longitudinal data collected by Heidi Feldman for 58 children with brain damage (46 with PVL; 12 with infarctions) and 21 normal language controls. Of the 46 children with PVL, 38 have bilateral lesions, 4 have right-hemisphere lesions, and 4 have left-hemisphere lesions. For the 12 subjects with infarctions, 8 have right-hemisphere lesions and 4 have left-hemisphere lesions. From ages 15 to 48 months, data were collected every 3 months using a set of tasks designed to elicited spontaneous discourse with their parents. From ages 4 years to 7 years, data were collected yearly employing both freeplay tasks and narrative tasks. The language samples collected for ages 4 to 7 are with a female experimenter. The normal language controls for the 15- to 48-month data collection are nine children born prematurely with no underlying brain damage. The normal language controls for the 4- to 7-year data collection are the subject's siblings, Sibling data have been collected when the child is at an age that is a chronological age match for the subject. In addition to collecting spontaneous language samples, standardized cognitive and language measures have been administered yearly.

Hargrove: This subdirectory contains a set of interviews collected by Patricia Hargrove between a speech therapist and six language-impaired children in the age range of 3 to 6.

Holland: This subdirectory contains a set of interviews collected by Audrey Holland from 40 recovering stroke patients who are suffering aphasic symptoms.

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

Rondal: This subdirectory contains data collected from 21 Down's syndrome children in Minnesota by Jean Rondal of the University of Liege.

Tager-Flusberg: The subjects for Helen Tager-Flusberg's longitudinal corpus include six autistic children and six children with Down's syndrome matched on age and MLU at the start of the study. The autistic children were diagnosed based on Rutter's criteria, which was consistent with the DSM-III criteria. IQ scores were assessed using the Leiter International Performance Scale (Leiter, 1969). The IQ scores for five of the six autistic children fell in the normal to low-normal range. The Down's syndrome children were matched based on chronological age and language level as measured by MLU. Spontaneous language samples were collected with the child and his or her mother during bimonthly visits to the children's homes over a period of between 12 and 26 months.

Bilingual Acquisition Data

DeHouwer: Annick De Houwer of the University of Antwerp has donated a corpus of transcripts collected between ages 2;7 and 3;4 from a girl who learned English and Dutch simultaneously.

Deuchar: Margaret Deuchar of the University of Cambridge has donated a corpus of transcripts collected between ages 1 ;3 and 2;6 from a girl who learned English and Spanish simultaneously.

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

Hayashi: Mariko Hayashi of the University of Arhus has donated a corpus of transcripts collected between ages 1;0 and 2;5 from a girl who learned Japanese and Danish simultaneously.

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.

Access to the Database

Membership in CHILDES is open. However, members are asked to abide by the rules of the system. In particular, users should not distribute copies of programs or files without permission, they should abide by the stated wishes of the contributors of the data, and they should acknowledge properly all uses of the data and the programs. Any article that uses the data from a particular corpus must cite a reference from the contributor of that corpus.

There are two levels of membership in the system: passive membership and contributing membership. All members can have passive access to the database and tools through the Internet. Members who have contributed data or who plan to contribute data can receive further documentation and assistance in using the programs. To access the database on the Web go to http://childes.psy.cmu.edu.

Reformatting of the Database

All of the current corpora are in good CHAT format. However, only a few of the most recent corpora were entered directly into CHAT. The process of converting the older corpora into CHAT required years of careful work. Some corpora had to be scanned into computer files from typewritten sheets. Other corpora were already computerized, but in a wide variety of transcription systems. For each of these datasets, we had to translate the project-specific codes and formats into CHAT. Several sets of files were translated from SALT (Miller & Chapman, 1983) using the SALTIN program. For other corpora, special-purpose reformatting programs had to be written. The fact that these transcripts now all pass through CHECK without error means that all of the files now have correct headers, correct listings of participants, and correct matches of coding tiers to main lines. Each main line has only one utterance and every utterance ends with a legal terminator.

There are no incorrect symbols in the middle of words and all paired delimiters are correctly matched. However, we cannot yet guarantee that the files are consistently coded on the level of individual words. This level of consistency checking remains a goal for our future work with the database.

3. CHAT

The most conceptually difficult task we faced in developing the CHILDES system was the formulation of the CHAT transcription system. From 1984 to 1990, during the period which we now refer to as the period of proto-CHAT, we explored a variety of transcription forms. In 1990 we began to finalize the shape of CHAT until it reached the more stable form published in the manual (MacWhinney, 1995). Users have expressed happiness with the current status of CHAT and the fact that virtually no changes have been made to the basic conventions since 1990. The finalization of CHAT allowed us to sharpen the workings of the CLAN CHECK program so that it now constitutes a computational implementation of the whole CHAT system.

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 its inevitable limitations, the availability of CHAT as a *lingua franca* for transcription both within the Program Project and within the general field of child language research has already led to solid improvements in data exchange, data analysis, and scientific precision.

Key Features of CHAT

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
@Languages:     en
@Participants:  ROS Ross Child BRI Brian Father
*ROS:   why isn't Mommy coming?
%com:   Mother usually picks Ross up around 4PM
*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 the following:

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 of these dependent tiers have been worked out in detail.
    a. *Phonological coding:* CHAT uses Unicode to allow for direct input of IPA phonological characters.
    b. *Error coding:* CHAT provides a full system for coding speech errors.
    c. *Morphemic coding:* CHAT provides a system for morphemic and syntac tic coding or interlinear glossing.

The full CHAT system is covered in MacWhinney (2000).

How Much CHAT Does a User Need to Know?

CHILDES users fall into two groups. One group of researchers wants to examine corpora but has little interest in collecting and transcribing new data. Another group of researchers wants to collect and transcribe new data and may be only marginally interested in analyzing old data. Typically, linguists and computer scientists fall into the first category and developmental psychologists and students of language disorders fall into the second category. Researchers in the second group who are using CHAT to transcribe new data soon come to realize that they need to learn all of the core CHAT conventions. Although these users may begin by using minCHAT, they will eventually gain familiarity with all of the conventions used on the main line, as well as with those dependent tier codes relevant to their particular research goals.

Users who are focusing on the analysis of old data may think that they do not need to master all of CHAT. For example, a researcher who wants to track the development of personal pronouns in the Brown corpora may think it sufficient to simply look for strings such as *he* and *it.* However, this type of casual use of the database is fairly dangerous. For example, users need to understand that, in the Brown corpora, the forms *dem* and *dese* are often used as spelling variants for *them* and *these.* Failure to track such variants could lead to underestimates of early pronoun usage. There are dozens of correspondences of this type that researchers need to understand if they are to make accurate use of the database.

When users start to use more detailed analysis programs such as MLU or DSS, the need to understand symbols for omitted elements and repetitions becomes increasingly important. For example, users need to understand that some corpora have been morphemicized in accord with the standards of chapter 6 and that others have not. Users also need to understand how symbols for missing elements or retracings can affect both lexical and syntactic analyses. It is possible that, in many cases, users could reach correct conclusions without a full understanding of the core features of CHAT. However, it is impossible to guarantee that this will happen. It is clear that the best recommendation to CHILDES users is to try to learn as much of CHAT as possible. If a researcher makes erroneous use of the database, these errors cannot be attributed to the CHILDES system, but only to the researcher who has failed to fully learn the system. Reviewers of articles based on the use of CHILDES data need to be convinced that the researcher fully understood the shape of the database and the inevitable limitations of any empirical dataset.

IV 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. The programs benefited from work done by Jeffrey Sokolov, Bill Tuthill, and Mitzi Morris, as well as from the SALT systematization developed by Jon Miller and Robin Chapman (Miller & Chapman, 1983).

CLAN commands include the program name, a set of options, and the names of the files being analyzed. 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. Many of the programs have quite a few possible options. Each option is explained in detail in the manual. In addition, you can get a brief list of options for a program by just typing the name of that program with no further options. For example, if you *type freq,* these options will be displayed:

- Freq creates a frequency word count
- For complete documentation on freq, type: help freq
- Usage: freq [c o dN fS k m pF rN sS tS u y zN] filename(s)
- +c: find capitalized words only
- +o: sort output by descending frequency
- +d: outputs all selected words, corresponding frequencies, and line numbers
- +dl: outputs word with no frequency information, in KWAL or COMBO format
- +d2: sends output to a file for STATEFREQ. Must include speaker specifications
- +d3: sends statistics only to STATEFREQ. Must include speaker specifications
- +d4: outputs only type/token information
- +fS: send to file (program will derive filename)
- -f: send output to the screen or pipe
- +k: treat upper and lower case as different
- +m: store output file(s) in the directories of input file(s)
- +pF: define punctuation set according to file F
- +rN: if N = 1 then "get(s)" goes to "gets", 2 - "get(s), 3 - "get"
- +sS: either word S or words in file @S to search for in a given input file
- — sS: either word S or words in file @S to be exclude from a given input file
- +tS: include tier code S
- — tS: exclude tier code S
- +u: merge all specified files together.
- +y: work on non-CHAT format files (default CHAT format)
- +zN: compute statistics on a specified range of input data

The programs have been designed to support five basic types of linguistic analysis (Crystal, 1982; Crystal, Fletcher, & Garman, 1989): lexical analysis, morphological analysis, syntactic analysis, discourse analysis, and phonological analysis. Let us look at how CLAN can be used to test hypotheses in each of these five areas.

CLAN for Lexical Analysis

The easiest types of CLAN analyses are those which look at the frequencies and distributions of particular word forms. For example, it is a simple matter to trace the use of a word like *under* or a group of words such as the locative prepositions. The analysis can be done on either a single file or a group of files. For example, let us suppose that we want to trace the use of personal pronouns in the three children studied by Roger Brown. We would construct a file including all of the personal pronouns with one pronoun on each line and call this file "pronouns." We would then use the FREQ command to count the occurrences of the pronouns in a file with a command like this:

```
freq +spronouns +t*ADAadam01.cha
```

The switch +t*ADA is included in order to limit the tally to only the utterances spoken by the child. If we also want the frequencies of the words spoken by the mother, we would use this command:

```
freq +spronouns +t*MOT adamOl.cha
```

If we want to extend our analysis to all of the files in the directory, we can use the wild card:

```
freq+spronouns+t* ADA adam*.cha
```

If we want the collection of files to be treated as a single large file, we can add another switch:

```
freq +spronouns +t*ADA +u adam*.cha
```

The FREQ command is powerful and quite flexible, permitting a large number of possible analyses. The outputs of these analyses can be sent to either the screen or to files. The names of the output files can be controlled. For example, one might want to maintain a group of output files with the extension .mot for the frequencies of the mother's speech. These can be kept in a separate directory for further analysis.

The second major tool for conducting lexical analyses is the KWAL program, which outputs not merely the frequencies of matching items, but also all the full context of the item. For example, the KWAL command that searches for the word *chalk* in the sample.cha file will produce this output:

```
kwal + schalk sample.cha kwal is conducting
analyses on: ALL speaker tiers
* * * * * * * * * * * * * * * * * * * * * * * * * *     From     file
<sample.cha>

***File sample.cha. Line 39. Keyword: chalk *MOT: is there
any delicious chalk?

***File sample.cha. Line 51. Keyword: chalk *MOT: do-'nt
you eat wonderful chalk.

***File sample.cha. Line 54. Keyword: chalk *MOT: it's not
good to want chalk.
```

It is possible to include still further previous and following context using additional switches.

Frequency Analyses

With tools like FREQ and KWAL, one can easily construct frequency analyses for individual children at specified ages. Many such counts have been produced. However, it is more difficult to move up to the next level of generalization on which a frequency count is constructed across children and ages. First, one would want to tabulate frequency data for the speech of children separately from the speech of adults. And one would not want to automatically combine data from children of different ages. Moreover, we would not want to merge data from children with language disorders together with data from normally developing children. Differences in social class, gender, and educational level may lead one to make further separations. Also it is important to distinguish language used in different situational contexts. When one finishes looking at all the distinctions that could potentially be made, it becomes clear that one needs to think of the construction of a lexical database in very dynamic terms.

Such a database could be constructed using FREQ and other CLAN tools, but this work would be fairly tedious and slow. What we plan to do to address this problem is to build a file with every lexical item in the entire database and attach to each item a set of pointers to the position of the item in every file in which it occurs. These key files and pointer files will be stored along with the database on a CD-ROM. Using the pointers from the master word list to the individual occurrences of words, the user can construct specific probes of this database configured both on facts about the child and facts about the words being searched. The program that matches these searches to the pointer file will be called LEX. Using LEX, it will be possible, for example, to track the frequency of a group of evaluative words contained in a separate file in 2-year olds separated into males and females. And the same search can also yield the frequency values for these words in the adult input. Although we may want to publish hard-copy frequency counts based on some searches through this database, the definitive form of the lexical frequency analysis will be contained in the program itself.

Once the LEX tool is completed, the path will be open to the construction of three additional tools. The first of these is a simple extension of the current KWAL program. Currently, researchers who want to track down the exact occurrences of particular words must rely on the use of the +d option in FREQ or must make repeated analyses using KWAL and keep separate track of line numbers. With the new LEX system, instead of running through files sequentially, KWAL will be able to rely on the pointers in the master file to make direct access to items in the database.

Lexical Field Analyses

A second type of lexical research focuses attention not on the entire lexicon, but on particular lexical fields. Using the +s @file switch with FREQ and KWAL, such analyses can already be computed with the current version of CLAN. Completion of the LEX facility will further facilitate the analysis of lexical fields. For example, using the lexical database, we will be able to examine the development of selected lexical fields in the style of the PRISM analysis of Crystal (1982) and Crystal, Fletcher, and Garman (1976). This analysis tracks the child's developing use of content words in 239 lexical subfields. Examples of these fields include farm tools, units of weight measurement, and musical instruments. These 239 fields can be merged into a set of 61 categories which can, in turn, be merged into nine high-level fields. Bodin and Snow (1993) showed how analyses of this type can be conducted on the CHILDES database. Likely candidates for intensive examination include mental verbs, morality words, temporal adverbs, subordinating conjunctions, and complex verbs.

Other important semantic fields include closed class items such as pronouns, determiners, quantifiers, and modals. As Brown (1973), Lahey (1988), and many others have noted, these high-frequency, closed-class items each express important semantic and pragmatic functions that provide us with separate information about the state of the child's language and cognitive functioning. For example, Anti-nucci and Miller (1976), Cromer (1991), Slobin (1986), and Weist, Wysocka, Witkowska-Stadnik, Buczowska, & Konieczna (1984) argued that tense markings and temporal adverbs are not controlled until the child first masters the relevant conceptual categories.

It is also possible to track basic semantic relations (Bloom, 1975; Lahey, 1988; Leonard, 1976; Retherford, Schwartz, & Chapman, 1981) by studying the closed-class lexical items that mark these relations. In particular, we can follow these correspondences between semantic relations and lexical expressions by tracking grammatical words such as *not, with, some* or *my*.

Lexical Rarity Index

A third measure that can be developed through use of the LEX facility is the lexical rarity index (LRI). Currently, the major index of lexical diversity is the type-token ratio (TTR) of Templin (1957). A more interesting measure would focus on the relative dispersion in a transcript of words that are generally rare in some comparison data set. The more that a child uses words that are considered rare, the higher the lexical rarity index. If most of the words are common and frequent, the LRI will be low. To compute various forms of this index, the LRI program would rely on values provided by LEX.

Another easy way of tracking the emergence of longer words is to use the WDLEN program, which provides a simple histogram of word lengths in a file along with the location of the longest words.

CLAN for Morphological Analysis

Many of the most important questions in child language require the detailed study of specific morphosyntactic constructions. For example, the debate on the role of connectionist simulations of language learning (MacWhinney & Leinbach, 1991; MacWhinney, Leinbach, Taraban, & McDonald, 1989; Marcus, etal., 1991; Pinker & Prince, 1988; Plunkett & Sinha, 1992) has focused attention on early uses and overregularizations of the regular and irregular past tense markings in English. A full resolution of this debate will require intensive study of the acquisition of these markings not just in English, but in a wide variety of languages at a wide sampling of ages. Similarly, the testing of hypotheses about parameter setting within generative theory (Hyams, 1986; Pizzuto & Caselli, 1993; Valian, 1991; Wexler, 1986) often depends on a careful study of pronominal markings, reflexives, and wh-words.

During the earliest stages of language learning, the most obvious developmentsare those involving the acquisition of particular grammatical markings. The study of the acquisition of grammatical markers in English has been heavily shaped by Brown's (1973) intensive study of the acquisition of 14 grammatical morphemes in Adam, Eve, and Sarah and the cross-sectional follow-up by de Villiers and de Villiers (1973). Since Brown's original analysis, there have been scores of studies tracking these same morphemes in second-language learners, normally developing children, and children with language disorders. There have also been many studies of comparable sets of morphemes in other languages.

The 14 morphemes studied by Brown include the progressive, the plural, the regular past, the irregular past, *in, on,* the regular third-person singular, the irregular third-person singular, articles, the uncontracted copula, the contracted copula, the possessive, the contracted auxiliary, and the uncontracted auxiliary. Brown's framework for morpheme analysis has been extended in systems such as the LARSP procedure by Crystal, Fletcher, and Garman (1976), the DSS procedure by Lee (1974), the ASS procedure of Miller (1981), and the IPSyn procedure of Scarborough, Rescorla, Tager-Flusberg, Fowler, and Sudhalter (1991). Other markers tracked in LARSP, ASS, and DSS include the superlative, the comparative, the adverbial ending *-ly,* the uncontracted negative, the contracted negative, the regular past participle, the irregular past participle, and various nomi-nalizing suffixes.

Within the CHILDES framework, De Acedo (de Acedo, 1993) showed how CHAT and CLAN can be used to study the Spanish child's learning of grammatical markers. In the next two sections, we discuss several CLAN tools for the study of morphological structure.

Morphological Analysis from the Main Line

Chapter 6 of the CHILDES manual describes a system for coding the presence of grammatical markers on the main line. For example, the word *cats* can be coded as *cat-s.* This coding then allows users to search for all instances of the plural marker *-s.* The basic lexical tools of FREQ and KWAL can be used to do this. This system provides direct access to the basic grammatical morphemes of English. However, only grammatical markers are coded directly. In a more complex language, such as Italian or Hungarian, a simple system of this type will quickly break down. Moreover, many analyses of morphological and syntactic structure require the construction of a more complete morphological analysis.

MOR—Automatic Morphological Analysis

The more extensive coding needed for many projects requires a complete construction of a part-of-speech analysis for each word on the main line. This analysis is placed on a separate tier called the %mor tier. The coding of the %mor tier is done in accord with the guidelines specified in chapter 14 of the CHILDES manual. Once a complete %mor tier is available, a vast range of morphological and syntactic analyses become possible. However, hand coding of a %mor tier for the entire CHILDES database would require perhaps 20 years of work and would be extremely error-prone and non-correctable. If the standards for morphological coding changed in the middle of this project, the coder would have to start over again from the beginning. It would be difficult to imagine a more tedious and frustrating task—the hand coder's equivalent of Sisyphus and his stone.

The alternative to hand coding is automatic coding. Over the past three years, we have worked on the construction of an automatic coding program for CHAT files. This program, called MOR, was first developed in LISP by Roland Hausser (1990), modified for C by Carolyn Ellis, and then completely rewritten by Mitzi Morris with assistance from Leonid Spektor. Although the MOR system is designed to be transportable to all languages, it is currently only fully elaborated for English and German. The language-independent part of MOR is the core processing engine. All of the language-specific aspects of the systems are built into files that can be modified by the user. In the remarks that follow, we will first focus on ways in which a user can apply the system for English.

How to Run MOR

The MOR program takes a CHAT main line and automatically inserts a %mor line together with the appropriate morphological codes for each word on the main line. The basic MOR command is much like the other commands in CLAN. For example, you can run MOR in its default configuration with this type of command:

```
 mor sample.cha
```

However, MOR is unlike the other CLAN programs in one crucial regard. Although you can run MOR on any CLAN file, in order to get a well-formed %mor line, you often need to engage in significant *extra work.* We have tried to minimize the additional work you need to do when working with MOR, but it would misleading for us to suggest that no additional work is required. In particular, users of MOR will often need to spend a great deal of time engaging in the processes of (a) lexicon building and (b) ambiguity resolution.

Files Used by MOR

Before we examine ways of dealing with lexicon building and ambiguity resolution, let us take a quick look at the files that support a MOR analysis. For MOR to run successfully, four files must be present in either the library directory or the current working directory. Although you do not need to have a detailed understanding of the functioning of these files, it will help you to have a view of the shape of these basic building blocks. The default names for the four basic files are eng.ar, eng.cr, eng.lex, and eng.clo. These four files contain the following information:

1. *Allomorphic rules.* The rules that describe allomorphic variations are called *arules.*
2. *Concatenation rules.* The rules that describe allowable concatenations are called *crules.*
3. *Closed class items.* The eng.clo file contains the closed class words and suf fixes of English. Because this group forms such a tight closed set, the user will seldom have to modify this file.
4. *Open class items.* The default name of the open class lexicon for English is eng.lex. This file is what we call the *disk lexicon.* Words in the disk lexicon are listed in their canonical form, along with category information.

MOR uses the eng.ar, eng.lex, and eng.clo files to produce a *run-time lexicon,* which is significantly more complete than eng.lex alone. When analyzing input files, MOR uses the run-time lexicon together with the eng.cr file. As a user, you do not need to concern yourself with the actual shape of the run-time lexicon, and you will usually not have to touch either the arules, crules, or drules. Your main concern will be with the process of adding or removing entries from the main open class lexicon file — eng.lex. If all of the words in your files can be located in the eng.lex file, running of MOR is totally trivial. You simply run:

```
mor filename
```

But matters are seldom this simple, because most files will have many words that are not found in eng.lex and you will need to refine the eng.lex file until all missing words are inserted. Therefore, the main task involved for most users of MOR is the building of the lexicon file.

Lexicon Building — Finding Missing Words

To see whether MOR correctly recognizes all of the words in your transcripts, you can first run MOR on all of your files and then run this KWAL command on the .mor files you have produced: kwal +t%mor

*.morIf KWAL finds no question marks on the %mor line, then you know that all the words have been recognized by MOR. If there are question marks in your *.mor output files, you will probably want to correct this problem by running MOR in the interactive update mode. If you know from the outset that your file includes many words that will not be found in eng.lex, you can directly begin

the process of lexicon building by running this FREQ command: freq +dl +u +k *.cha > output.frq

This produces a simple list of all the words in your transcript in a form useful for interactive MOR lexicon building. The +dl option outputs words without frequencies. The +k option is needed to distinguish between *Bill* and *bill.* The redirection arrow sends the output to a file we have called output.frq. Next you can run MOR again using the +s option with a filename added, as in this example:

```
mor +xl output.frq
```

MOR will use eng.lex to attempt to analyze each word in the output.frq file. If it cannot analyze the word, it will enter it in a output file of lexical entry templates with the name output.ulx. Then you need to look at the words in the output.ulx file, using an editor. Some may be misspellings and will have to be corrected in the original file. Others will be new words for which you will have to enter a part-of-speech characterization. When you are finished, you should rename the output.ulx file to output.lex. Then you can run MOR again in this form:

```
mor +loutput *.cha
```

If all has gone smoothly, MOR will now be able to enter a part-of-speech characterization for every word in the transcripts.

The Structure of eng.lex

Users of MOR may want to understand the way in which entries in the disk lexicon (eng.lex) are structured. The disk lexicon contains truly irregular forms of a word, as well as citation forms. For example, the verb *go* is stored in eng.lex, along with the past tense *went,* because this form is a suppletive form and is not subject to regular rules. The disk lexicon contains any number of lexical entries, stored at most one entry per line. A lexical entry may be broken across several lines by placing the continuation character backslash (\) at the end of the line. The lexicon may be annotated with comments, which will not be processed. A comment begins with the percent sign % and ends with a newline.

A lexical entry consists of the surface form of the word, followed by category information about the word expressed as a set of feature-value pairs. Each feature-value pair is enclosed in square brackets and the full set of feature-value pairs is enclosed in curly braces. All entries must contain a feature-value pair that identifies the syntactic category to which the word belongs, consisting of the feature *scat* with an appropriate value. Words that belong to several categories will be followed by several sets of feature structures, each separated by a backslash. Category information is optionally followed by information about the stem. If the surface form of the word is not the citation form of the word, then the citation form, surrounded by quotes, should follow the category information. If the word contains fused morphemes, these should be given as well, using the & symbol as the morpheme separator. The following are examples of lexical entries:

```
can     {[scat v:aux])\
        {[scatn}
a       {[scatdet]}
an      {[scat del]}        "a"
go      {[scatv][ir+]}
went    {[scat v][tense past]}    "go&PAST"
```

When adding new entries to eng.lex it is usually sufficient to enter the citation form of the word, along with the syntactic category information.

Ambiguity Resolution

MOR automatically generates a %mor tier of the type described in chapter 14. As stipulated in chapter 14, retraced material, comments, and excluded words are not coded on the %mor line produced by MOR. Words are labeled by their syntactic category, followed by the separator "|," followed by the word itself, broken down into its constituent morphemes.

*CHI: the people are making cakes.
%mor: del the n|people v:aux|be&PRES v|make-ING n|cake-PL.

In this particular example, none of the words have ambiguous forms. However, it is often the case that some of the basic words in English have two or more part-of-speech readings. For example, the word *back* can be a noun, a verb, a preposition, an adjective, or an adverb. The " ^{A}" character denotes the alternative readings for each word on the main tier.

*CHI: I want to go back,
%mor: pro|I v|want inf|to^prep|to v|go adv|back^n|back^v|back.

The entries in the eng.clo file maintain these ambiguities. However, open class words in the eng.lex file are only coded in their most common part-of-speech form. The problem of noun-verb ambiguity will eventually be addressed through use of the POST program. Those ambiguities that remain in a MOR transcript after the POST program has operated can be removed by using MOR in its ambiguity resolution mode. The program locates each of the various ambiguous words one by one and asks the user to select one of the possible meanings.

MOR for Other Languages

To maximize the portability of the MOR system to other languages, we have developed a general scheme for representing arules and crules. This means that a researcher can adapt MOR for a new language without doing any programming at all. However, the researcher/linguist needs to construct (a) a list of the stems of the language with their parts of speech, (b) a set of arules for allomorphic variations in spelling, and (c) a set of crules for possible combinations of stems with affixes. Building these files will require a major one-time dedication of effort from at least one researcher for every language. Once the basic work of constructing the rules files and the core lexicon files is done, then further work with MOR in that language will be no more difficult than it currently is for English. However, construction of new rules files is an extremely complex process, and construction of a closed class and open class lexicon will also take a great deal of time. Although no programming is required, the linguist building these files must have a thorough understanding of the MOR program and the morphology of the language involved. Complete documentation for the construction of the rules files is available from Carnegie Mellon and will also be included in the next edition of the CHILDES manual.

CLAN for Syntactic Analysis

Once a %mor line has been constructed, either through use of MOR or through hand coding, a variety of additional morphosyntactic analyses are then available. Instead of analyzing the lexical items on the main line, programs can now analyze the fuller morphosyntactic representation on the %mor line. The simpler forms of analysis can still be done using FREQ and KWAL. However, several other programs add additional power for morphosyntactic analysis. The MLU program can compute the basic mean length of utterance index in a variety of ways (Rollins, 1993). COMBO extends the power of FREQ and KWAL by permitting more complete Boolean string matching. For example, if the user wants to search for all instances of a relative pronoun followed eventually by an auxiliary verb, it is possible to compose this search string in COMBO. Certain types of matching between the main line and the %mor line can be achieved using the +s switch in the MODREP program. In addition, the COOCCUR program can be used to tabulate sequences of syntactic structures appearing on the %mor line.

Once we have collected a large database of transcripts in other languages and created a full %mor tier encoding, we can ask some of the basic questions in crosslinguistic analyses. Are there underlying similarities in the distribution of semantic relations and grammatical markings used by children at the beginning of language learning? Exactly which markings show the greatest language-specific divergences from the general pattern? How are grammatical relations marked as ergative in one language handled in another language? Under what circumstances do children tend to omit subject pronouns, articles, and other grammatical markers?

CLAN for Discourse and Interactional Analyses

Many researchers want to track the ways in which discourse influences the expression of topic, anaphora, tense, mood, narrative voice, ellipsis, embedding, and word order (Halliday & Hasan, 1976; MacWhinney, 1985). To do this, researchers need to track shifts in narrative voice, transitions between discourse blocks, and foreground-background relations in discourse. They are also interested in the ways in which particular speech acts from one participant give rise to responsive or non-responsive speech acts in the other participant. CLAN provides several powerful tools for examining the structures of interactions and narrations.

Coder's Editor

The most important CLAN tool for data coding is the Coder's Editor, which is a new program in CLAN 2.0. CED can lead to truly remarkable improvements in the accuracy, reliability, and efficiency of transcript coding. If you have ever spent a significant amount of time-coding transcripts or if you plan to do such coding in the future, you should definitely consider using CED.

CED provides the user with not only a complete text editor, but also a systematic way of entering user-determined codes into dependent tiers in CHAT files. The program works in two modes: coder mode and editor mode. Initially, you are in editor mode, and you can stay in this mode until you learn the basic editing commands. The basic commands have been configured so that both WordPerfect and EMACS keystroke equivalents are available. If you prefer some other set of key-strokes, the commands can be rebound.

In the coding mode, CED relies on a codes.1st file created by the user to set up a hierarchical coding menu. It then moves through the file line by line asking the coder to select a set of codes for each utterance. For example, a codes.lst list such as

```
$MOT :POS
     :Que
     :Res :NEG $CHI
```

would be a shorter way of specifying the following codes:

```
$MOT:POS:Que
$MOT:POS:Res
$MOT:NEG:Que
$MOT:NEG:Res
$CHI:POS:Que
$CHI:POS:Res
$CHI:NEG:Que
$CHI:NEG:Res
```

This coding system would require the coder to make three quick cursor movements for eacutterance in order to compose a code such as $CHI:NEG:Res.

Chains and Sequences

Once a file has been fully coded in CED, a variety of additional analyses become possible. The standard tools of FREQ, KWAL, and COMBO can be used to trace frequencies of particular codes. However, it is also possible to use the CHAINS, DIST, and KEYMAP programs to trace out sequences of particular codes. For example, KEYMAP will create a contingency table for all the types of codes that follow some specified code or group of codes. It can be used, for example, to trace the extent to which a mother's question is followed by an answer from the child, as opposed to some irrelevant utterance or no response at all. DIST lists the average distances between words or codes. CHAINS looks at sequences of codes across utterances. Typically, the chains being tracked are between and within speaker sequences of speech acts, reference types, or topics. The output is a table that maps, for example, chains in which there is no shift of topic and places where the topic shifts. Wolf, Moreton, and Camp (1993) apply chains to transcripts that have been coded for discourse units. Yet another perspective on the shape of the discourse can be computed by using the MLT program, which computes the mean length of the turn for each speaker.

Recasts

Currently there is only one CLAN program that focuses on the lexical and syntactic match between successive utterances. The is the CHIP program developed by Jeffrey Sokolov and Leonid Spektor. CHIP is useful for tracking the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. Sokolov and Moreton (1993) and Post (1993) have used it successfully to demonstrate the availability of useful instructional feedback to the language-learning child.

Discourse Display

There is more to a transcript than a series of codes and symbols. The superficial form of a transcript can also lead us to adopt a particular perspective on an interaction and to entertain particular hypotheses regarding developments in communicative strategies. For example, if we code our data in columns with the child on the left, we come to think of the child as driving or directing the conversation. If we decide instead to place the parents' utterances in the left column, we then tend to view the child as more reactive or scaffolded. Ochs (1979) noted that such apparently simple decisions as the placement of a speaker into a particular column can both reflect and shape the nature of our theories of language development. Because it is important for the analyst to be able to see a single transcript in many different ways, we have written three new CLAN programs that provide alternative views onto the data. The basic principle underlying these data display programs is the motto of "different files for different styles." The display programs—COLUMNS, LINES, and SLIDE—are designed to facilitate the alternative ways of viewing turns and overlaps.

The COLUMNS program produces CHAT files in a multicolumn form that is useful for explorations of turn taking, scaffolding, and sequencing. The COLUMNS program allows the user to break up the one-column format of standard CHAT into several smaller columns. For example, the standard 80-character column could be broken up into four columns of 20 characters each. One column could be used for the child, one for the parent, one for situational descriptions, and one for coding. The user has control over the assignment of tiers to columns, the placement of the columns, and the width of each separate column. As in the case of files produced by SLIDE, files produced by COLUMNS are useful for exploratory purposes, but are no longer legal CHAT files and cannot be reliably used with the CLAN programs.

CLAN for Phonological Analyses

Despite all the care that has gone into the formulation of CHAT, transcription of child language data remains a fairly imprecise business. No matter how carefully one tries to capture the child's utterances in a standardized transcription system, something is always missing. The CHAT main line induces the transcriber to view utterances in terms of standard lexical items. However, this emphasis tends to force the interpretation of nonstandard child-based forms in terms of standard adult lexical items. By including a rich phonological transcript on the %pho line, this imbalance can be corrected. As Peters, Fahn, Glover, Harley, Sawyer, and Shimura (1990) have argued, the inclusion of a complete CHAT %pho line is the best way to convey the actual content of the child's utterances, particularly at the youngest ages. CHAT provides two systems for transcribing utterances on the %pho tier. For coarse transcription, researchers can use the UNIBET systems that have been developed for English and several other languages. For a finer level of phonetic transcription, researchers can use the PHONASCII rendition of IPA notation.

Analysis of the %pho Line

The construction of a complete %pho tier for even a few hours of data is a formidable task. Verification of the reliability of that transcription is an even bigger problem. However, once this tier is produced, CLAN provides several programs to facilitate analysis. The two programs that are most adapted to this analysis are PHONFREQ, which computes the frequencies of various segments, separating out consonants and vowels by their various syllable positions, and MODREP, which matches %pho tier symbols with the corresponding main line text. For more precise control of MODREP, it is possible to create a separate %mod line, in which each segment on the %pho corresponds to exactly one segment on the %mod line.

Linking to a Digitized Record

Although inclusion of a complete %pho line is a powerful tool, even this form of two-tier transcription misrepresents the full dynamics of the actual audio record. If the original audiotapes are still in good condition, one can use them to continue to verify utterances. But there is no way to quickly access a particular point on an audiotape for a particular utterance. Instead, one has to either listen through a whole tape from beginning to end or else try to use tape markings and fast-forward buttons to track down an utterance. The same situation arises when the interaction is on videotape.

Computer technology now provides us with a dramatic new way of creating a direct, immediately accessible link between the audio recording and the CHAT transcript. The system we have developed at Carnegie Mellon, called Talking Transcripts, uses fast optical erasable disks, a 16-bit digitizer board, and the Macintosh operating system to forge these direct links. Once a large sound file has been written to disk, we use CED to control access to the file during the coding of a new transcript. Although this process requires some additional time setting up the basic digitization, this investment pays for itself in facilitating high-quality transcription. Each utterance can be played back exactly and immediately without having to use a reverse button or foot pedal.

As a user of this new system, I have found that having the actual audio record directly available gave me a much enhanced sense of an immediate relation between the transcript and the actual interaction. It is difficult to describe verbally the vivid quality of this immediate link, but the impact on the transcriber is quite dramatic. Having the actual sound directly available does not diminish the importance of accurate transcription, because the CLAN programs must still continue to rely on the CHAT transcript. However, the immediate availability of the sound frees the transcriber from the fear of making irrevocable mistakes, because the ongoing availability of the audio record means that codes can always be rechecked for reliability.

Phonological Analysis with a Digitized Record

Now that we have a link between the CHAT %pho line and digitized speech, the way is now open for us to design an entire Phonologist's Workbench grounded on the immediate availability of actual sound. The new programs for phonological analysis that we now plan to write include the following:

1. *Inventory analysis:* We will extend the PHONFREQ program so that it can compute the numbers of uses of a segment across either types or tokens of strings on the %pho line. The program will also be structured so that the in ventories can be grouped by distinctive features such as place, manner, ar ticulation, or groups such as consonants versus vowels. The ratio of conso nants to vowels will be computed. Summary statistics will include raw frequencies and percentage frequency of occurrence for individual seg ments. Nonoccurrences in a transcript of any of the standard segments of English will be flagged.
2. *Length:* The MLU program will be used to compute mean length of utter ance in syllables. This can be done from the %pho line, using syllable

boundaries as delimiters.

3. *Variability:* The MOD REP program will be made to compute the types and tokens of the various phonetic realizations for a single target word, a single target phoneme, or a single target cluster. For example, for all the target words with the segment /p/, the program will list the corresponding child forms. Conversely, the researcher can look at all the child forms containing a /p/ and find the target forms from which they derive.
4. *Homonymy:* Homonymy refers to a child's use of a single phonetic string to refer to a large number of target words. For example, the child may say "bo" for *bow, boat, boy, bone,* and so on. The MODREP program will calculate the degree of homonymy observed by comparing the child's string types coded on the %pho tier with the corresponding target forms coded on the %mod tier.
5. *Correctness:* To determine correctness, the child's pronunciation (%pho line) must be compared with the target (%mod line). The MODREP program will be modified to compute the number of correct productions of the adult target word, segment, or cluster. For example, the percentage consonants correct (PCC) will be computed in this way.
6. *Phonetic product per utterance:* This index (Bauer, 1988; Nelson & Bauer, 1991) will be computed by a new CLAN program called PHOP. The index computes the phonetic complexity of the utterance as a function of the number of place of articulation contrasts realized. This index is low if everything is at one place of articulation; it is high if all points of articulation are used.
7. *Phonological process analysis:* Phonological process analyses search for systematic patterns of sound omission, substitution, and word formation that children make in their simplified productions of adult speech. Thus, such processes refers to classes of sounds rather than individual sounds. Process analysis must be based on the comparison of the %pho and %mod tiers. The Clan Analysis of Phonology, or CAP, will examine rates of con sonant deletion, voicing changes, gliding, stopping, cluster simplification, and syllable deletion. In addition, nondevelopmental error will be identi fied and calculated (Shriberg, 1990).
8. *PHONASCII and UNIBETcode modifications:* PHONASCII and UNIBET codes will be modified or elaborated to enable cross-tier analysis.
9. *Automatic phonetic transcription of high-frequency words:* To facilitate phonetic and phonological transcription of corpora, we will develop an on line users reference to provide automatic phonological coding of the 2000 most frequently used words in the English language to facilitate phonetic transcription of naturalistic speech data (e.g., words such as *and* and *the* will not have to be redundantly transcribed each time they occur.
10. *Transcription playback:* The same phonological database used by the Pho nologist's Reference can be used to playback the sounds of candidate tran scriptions.

Alongside the development of programs to support these analyses, we will also be working to broaden the CHILDES database of phonological transcripts. Very few computerized transcripts are currently available, so we can reasonably start from scratch in this area. Because we are starting from scratch, we can require that all transcripts in the CHILDES phonological database be accompanied by good-quality tape recordings, which will be digitized at Carnegie Mellon and then distributed through CD-ROM.

Utilities

CLAN also includes a variety of features to make life easier when manipulating files and transcripts. The CHSTRING program can be used to make simple string replacements across collections of files. The CAPWD program prints all capitalized words. The GEM 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. A good example of the use of GEM with FREQ can be found in Post (1993). For users working with files from SALT, the SALTIN program helps in the conversion to CHAT. The LINES program allows users to mark their CHAT files with line numbers. DATES can be used to compute a child's age, given the current date and the child's birthdate.

CHECK

It is difficult to overestimate the importance of the CHECK program. Although CHECK performs no analysis and computes no numbers, it is perhaps the most important of the CLAN programs, particularly for researchers who are entering new data. By running new transcripts through CHECK, transcribers can avoid errors and guarantee adherence to the CHAT standards.

It is important to use CHECK early in the transcription process. If it is used early, error types will be noted before they begin to replicate. If CHECK produces too many errors, the +dl option can be used to cut down warnings to one of each type. When using the +dl option, it is helpful to know that there are 46 different error messages produced by CHECK. What the +dl option guarantees is that you will only get one complaint for each of these types of errors: missing line beginnings, missing tabs, missing colons, missing @Begin, missing @End, missing ©Participants line, nonstandard participant roles, missing roles, incorrect tier names, duplicate speaker declarations, missing speaker identifications, delimiters in words, unmatched paired delimiters, missing main tiers, undeclared codes, illegal date entries, illegal time entries, multiple utterances per line, undeclared prefixes, undeclared suffixes, duplicate coding tiers, missing terminators, extra terminators, and incorrect pairings of @Bg and @Eg markers.

Ideally, CHECK only needs to rely on the standard depfile. CHECK uses the codes in the depfile as its guide to understanding what CHAT codes should be permitted on which tiers. The depfile we distribute is, in effect, a summary of the CHAT system given in the manual. Sometimes users have good reasons for making exceptions to CHAT conventions. To override the definitions given in the depfile without having to tinker with that file, we have added the capacity to create a OOdepadd file. This file then also provides an overt record of additions or modifications to CHAT required for particular corpora. For example, if you need to allow for equals signs on the ©Comment line and for words with suffixes on the @Bgd line, you could create a OOdepadd file with these two lines:

```
@Bgd:*-*
©Comment:  =
```

If the depfile has a code that is too permissive, such as $*, you will want to remove this before entering the more specific codes in your OOdepadd file. In general, it I still best to focus on using CHECK early in the process of transcription, before you begin to accumulate errors. Whenever possible, it is best to use only the standard depfile, but sometimes there will be reasons for extending CHAT by using a OOdepadd file.

CHECK only examines files for their compliance to the syntactic specifications of CHAT. An important second type of checking can be achieved by using FREQ to create a unified frequency count for an entire corpus. This is best done with this command:

```
freq +u +f*.cha
```

This command will produce a single file with all the words you used on the main lines of all your files. You can then go over these words to check for spelling errors and other inconsistencies. A useful clue in looking for spelling errors is to search for words with a frequency of 1. If you use FREQ with the +o option, you can immediately find all the words with a frequency of 1 together at the end of the printout. Once your preliminary cleanup is done, you may want to repeat the same analysis using the +t% and —t* options so you can check for errors in the codes on the dependent tiers. Alternatively, you can provide CHECK with a complete listing of your codes by creating a OOdepadd file.

Group	Program	Description
Lexical search	FREQ	Tracks the frequency of each word used
	FREQMERG	Merges outputs from several runs of FREQ
	KWAL	Searches for a specific word or group of words
	STATFREQ	Sends the output of FREQ to a statistical program
Block search	GEM	Searches for premarked blocks of interaction
	GEMFREQ	Does a FREQ analysis on a particular block type
	GEMLIST	Profiles the types of blocks found in a file
Discourse/Interaction	CHAINS	Displays "runs" or "chains" of speech acts
	CHIP	Tracks imitations, repetitions, lexical overlap
	DIST	Tracks the distance between particular codes
	KEYMAP	Looks at the variety of speech acts following a given act
	TIMEDUR	Computes overlap and pause duration
	PAUSE	Computes speaking, pause, and overlap times
Morphosyntax	COMBO	Searches for combinations of words or types of words
	COOCCUR	Tabulates pairwise co-occurrence frequency
	KWAL	Searches for a specific word or group of words
	MOR	Performs a full morphological analysis using rules
	POSFREQ	Does a FREQ analysis by sentence position
Phonology	MODREP	Matches phonological forms to their corresponding words
	PHONFREQ	Tabulates the frequency of each phoneme or cluster
	Sonic CHAT	Uses the CED editor to link the transcript to actual sound
Coding tools	CED	A multipurpose editor for CHAT files
	RELY	Compares two sets of codes to compute reliability
Measures	GDI DB	A database of early maternal reports on lexical growth
	DSS	Computes the Developmental Sentence Score
	MAXWD	Lists the longest words and longest utterances in a file
	MLU	Computes mean length of utterance
	MLT	Computes mean length of turn
	FREQ	Includes computation of the type-token ratio
	WDLEN	A frequency distribution by word and sentence length
File display	COLUMNS	Displays CHAT files in the old "column" format
	FLO	Removes complex codes from a CHAT file
	LINES	Adds line numbers to a CHAT file
	SALTIN	Converts data from SALT to CHAT
	SLIDE	Puts a file onto one line that can be scrolled horizontally
Utilities	CHIBIB	A bibliographic access system with 14,000 references
	CHECK	Examines CHAT files for syntactic accuracy
	CHSTRING	Converts strings
	DATES	Computes a child's age for a given date
	TEXTIN	Takes simple unmarked text data and outputs a CHAT file

V. THE FUTURE

If the CHILDES database is to continue to grow, we must continue to receive extensive cooperation from individual scientists. Researchers who use the CHILDES tools to collect new data have the responsibility to contribute these new data to the database. In particular, researchers whose work has benefited from government support have a clear obligation to contribute to scientific progress by adding their data to the database. In fields such as the sequencing of proteins in DNA, researchers, journals, and the government have set the requirement that only data which are publicly available in the Human Genome database can be published. A similar policy for language development studies would ensure the stable and continued development of the CHILDES database. Until such a policy is developed, voluntary acceptance of these responsibilities will guarantee continued growth of the database.

We expect that new additions to the CHILDES database will no longer require reformatting, because they will be transcribed in CHAT from the start. Already, we are starting to receive most new data files in CHAT format. Soon we expect this to become the norm. The database will also continue to grow beyond its original scope. The first corpora included in the database were on first language acquisition by normal English-speaking children. In the future, the database will grow to include large components of second language acquisition data, adult interactional data, and a variety of data from children with language disorders. The numbers of languages represented will continue to grow. As the database grows, it will be important to distinguish between the CHILDES system, the Aphasia Language Data Exchange System (ALDES), the Second Language Acquisition Data Exchange System (SLADES), and the overall Language Data Exchange System (LANDES).

As multimedia computational resources become increasingly available, the CHILDES database will shift from its current concentration on ASCII transcripts to a focus on transcripts accompanied by digitized audio and video. Links between events in the audio and video records will be tied to an increasingly rich set of links in the transcript. These "hot" links will be increasingly dynamic, allowing the user to move around through the audio and video records using the transcript as the navigational map. The full digitization of the interaction will allow the observer to enter into the interaction as an explorer. This is not the virtual reality of video adventures. The scientist is not seeking to change reality or to interact with reality. Instead, the goal is to explore reality by viewing an interaction repeatedly from many different perspectives. These new ways of viewing a transcript will be important for phonological and grammatical analyses, but their most important impact will be on the analysis of interactional structure and discourse. Having full video and audio immediately available from the transcript will draw increased attention to codes for marking synchronies between intonational patterns, gestural markings, and lexical expressions in ongoing interactional relations.

The construction of this new multimedia transcript world would allow us to begin work on the successor to the CHILDES project. This is the Human Speech Genome project. One of the first goals of the Speech Genome project would be the collection, digitization, transcription, parsing, and coding of complete speech records for all the verbal interaction of a set of perhaps a dozen young children from differing language backgrounds. They might include, for example, a child learning ASL, a child with early focal lesions, a child growing up bilingual, and children with varying family situations. The multimedia records will allow us to fully characterize and explore all of the linguistic input to these children during the crucial years for language learning. We will then be in a position to know exactly what happens during the normal course of language acquisition. We can examine exactly how differences in the input to the child lead to differences in the patterns of language development. We will have precise data on the first uses of forms and how those first uses blend into regular control. We will be able to track all types of errors and first usages with great precision.

Alongside this rich new observational database, the increased power of computational simulations will allow us to construct computational models of the language learning process that embody a variety of theoretical ideas. By testing these models against the facts of language learning embodied in the speech genome, we can both refine the models and guide the search for new empirical data to be included in the multimedia database of the future.

ACKNOWLEDGMENTS

This work was supported from 1984 to 1988 by grants from the John D. and Catherine T. MacArthur Foundation, the National Science Foundation, and the National Institutes of Health. Since 1987, the CHILDES project has been supported by grants from the National Institutes of Health (NICHHD). For full acknowledgments and thanks to the dozens of researchers who have helped on this project, please consult pages viii and ix of MacWhinney, 1991. The CLAN programs were developed by Leonid Spektor. Mitzi Morris wrote the DSS and MOR programs with extensive help from Julia Evans, Leonid Spektor, Roland Haus-ser, Carolyn Ellis, and Kim Plunkett. Nan Bernstein-Ratner collaborated in the development of guidelines for the creation of a phonological analysis system. Ideas regarding the Talking Transcripts project came from Helmut Feldweg and Sven Stromqvist. Joy More-ton, Catherine Snow, Barbara Pan, and Lowry Hemphill helped test and design the CHAINS and CED programs. Important suggestions for modifications of CHAT

coding came from Judi Fenson, Frank Wijnen, Giuseppe Cappelli, Mary MacWhinney, Shanley Allen, and Julia Evans. Roy Higginson was the chief compiler of the CHILDES/BIB system.

REFERENCES

Achenbach, T. (1978). The child behavior profile: I. Boys aged 6-11. Journal of Consulting and Clinical Psychology, 46, 478-488.

Antinucci, F., & Miller, R. (1976). How children talk about what happened. City: 167-189.

Bauer, H. (1988). The ethologic model of phonetic development: I. Phonetic contrast estimators. Clinical Linguistics and Phonetics, 2, 347-380.

Bloom, L. (1970). Language development: Form and function in emerging grammars. Cambridge, MA: MIT Press.

Bloom, L. (1975). Language development. In F. Horowitz (Ed.), Review of child development research. Chicago: University of Chicago Press.

Bodin, L., & Snow, C. (1993). What kind of a birdie is this? Learning to use superordinates. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.

Braine, M. D. S. (1976). Children's first word combinations. Monographs of the Society for Research in Child Development, 41.

Brown, R. (1973). A first language: The early stages. Cambridge. MA: Harvard.

Cromer, R. (1991). Language and thought in normal and handicapped children. Oxford: Black well.

Crystal, D. (1982). Profiling linguistic disability. London: Edward Arnold.

Crystal, D., Fletcher, P., & Carman, M. (1976). The grammatical analysis of language disability. London: Edward Arnold.

Crystal, D., Fletcher, P., & Garman, M. (1989). The grammatical analysis of language disability. (2nd ed.). London: Cole and Whurr.

de Acedo, B. (1993). Early morphological development: The acquisition of articles in Spanish. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CH1LDES. Hillsdale, NJ: Erlbaum.

de Villiers, J., & de Villiers, P. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. Journal of Psycholinguistic Research, 2, 267-278.

Ervin-Tripp, S. (1979). Children's verbal turn-taking. In E. Ochs & B. Schieffelin (Eds.), Developmental pragmatics. New York: Academic Press.

Halliday, M., & Hasan, R. (1976). Cohesion in English. London: Longman.

Hausser, R. (1990). Principles of computational morphology. Computational Linguistics, 47.

Hayes, D. P. (1988). Speaking and writing: Distinct patterns of word choice. Journal of Memory and Language, 27, 572-585.

Higginson, R., & MacWhinney, B. (1990). CHILDES/BIB: An annotated bibliography of child language and language disorders. Hillsdale, NJ: Erlbaum.

Hollingshead, A. (1975). Four Factor Index of Social Status. Unpublished manuscript, University of Michigan.

Hyams, N. (1986). Language acquisition and the theory of parameters. Dordrecht: D. Rei-del.

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

Isaacs, S. (1933). Social development in young children. London: Routledge and Kegan Paul.

Lahey, M. (1988). Language disorders and language development. New York: Macmillan.

Lee, L. (1974). Developmental sentence analysis. Evanston, IL: Northwestern University Press.

Leiter, R. G. (1969). The Leiter International Performance Scale. Chicago: Stoelting.

Leonard, L. (1976). Meaning in child language. New York: Grune and Stratton.

MacWhinney, B. (1985). Grammatical devices for sharing points. In R. Schiefelbusch (Ed.), Communicative competence: Acquisition and intervention. Baltimore: University Park Press.

MacWhinney, B. (1992). The CH1LDESDatabase (2nded.). Dublin, OH: Discovery Systems.

MacWhinney, B. (1995). The CHILDES project: Tools for analyzing talk. Hillsdale, NJ: Erlbaum.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. Cognition, 29, 121-157.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? Journal of Memory and Language, 28, 255-277'.

Marcus, G., Oilman, M., Pinker, S., Hollander, M., Rosen, T., & Xu, F. (1991). Overregularization. Monographs of the Society for Research in Child Development.

Miller, J. (1981). Assessing language production in children: Experimental procedures. Baltimore: University Park Press.

Miller, J., & Chapman, R. (1983). SALT: Systematic Analysis of Language Transcripts, User's Manual. Madison, WI: University of Wisconsin Press.

Moerk, E. (1983). The mother of Eve—as a first language teacher. Norwood, NJ: ABLEX.

Nelson, L., & Bauer, H. (1991). Speech and language production at age 2: Evidence for tradeoffs between linguistic and phonetic processing. Journal of Speech and Hearing Research, 34, 879-892.

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

Peters, A., Fahn, R., Glover, G., Harley, H., Sawyer, M., & Shimura, A. (1990). Keeping close to the data: A two-tier computer-coding schema for the analysis of morphological development. Unpublished manuscript, University of Hawaii, HA.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition, 29, 73-193.

Pizzuto, E., & Caselli, M. (1993). The acquisition of Italian morphology: Implications for models of language development. Journal of Child Language.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. British Journal of Development Psychology, 10, 209-254.

Post, K. (1993). Negative evidence. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.

Retherford, K., Schwartz, B., & Chapman, R. (1981). Semantic roles and residual grammatical categories in mother and child speech: Who tunes into whom? Journal of Child Language, 8, 583-608.

Rollins, P. (1993). Language profiles of children with specific language impairment. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.

Rutter, M. (1978). Diagnosis and definition. In M. Rutter & E. Schopler (Eds.), Autism: A reappraisal of concepts and treatment. New York: Plenum.

Scarborough, H., Rescorla, L., Tager-Flusberg, H., Fowler, A., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. Applied Psycholinguistics, 12, 23-45.

Shriberg, L. (1990). Programs to examine phonetic and phonologic evaluation records. Hillsdale, NJ: Erlbaum.

Slobin, D. (1986). Crosslinguistic evidence for the language-making capacity. In D. Slobin (Ed.), The crosslinguistic study of language acquisition. Volume 2: Theoretical issues. Hillsdale, NJ: Erlbaum.

Sokolov, J., & Moreton, J. (1993). Individual differences in linguistic imitativeness. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.

Sokolov, J., & Snow, C. (forthcoming). Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.

Stark, R., & Tallal, P. (1981). Selection of children with specific language deficits. Journal of Speech and Hearing Disorders, 46, 114-133.

Templin, M. (1957). Certain language skills in children. Minneapolis, MN: University of Minnesota Press.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. Cognition, 40,21-81.

Weir, R. (1962). Language in the crib. The Hague: Mouton.

Weist, R., Wysocka, H., Witkowska-Stadnik, K., Buczowska, E., & Konieczna, E. (1984). The defective tense hypothesis: On the emergence of tense and aspect in child Polish. Journal of Child Language, 11, 347-374.

Wexler, K. (1986). Parameter-setting in language acquisition. In B. MacWhinney (Ed.), Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum.

Wolf, D., Moreton, J., & Camp, L. (1993). Children's acquisition of different kinds of narrative discourse: Genres and lines of talk. In J. Sokolov & C. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale, NJ: Erlbaum.