The CHILDES System

Brian MacWhinney

From: American Journal of Speech-Language Pathology, vol. 5, 1996, pp. 5-14

The Child Language Data Exchange System (CHILDES) is an international database organized for the study of first and second language acquisition. The project has been directed by Brian MacWhinney in collaboration with Catherine Snow of Harvard University. From 1984 to 1988 support came from the MacArthur Foundation. Since 1987, support has come from NIH and NSF. CHILDES includes three integrated components:

1. **CHAT** is the system for discourse notation and coding. This system includes detailed conventions for marking all sorts of conversational features, such as false starts, drawling, overlaps, interruptions, errors, and so on. This system was developed over a course of six years with continual input from language researchers. This standard transcription system is used for all the data in the database.
2. **CLAN** is the set of computer programs for searching and manipulating the database. Rather than focusing on canned analyses or rigid clinical packages, these programs provide the user with a toolkit of analytic possibilities to fit a specific research agenda. Most recently, the programs also provide tools for linking transcripts to digitized audio and video records.
3. **The Database**. The database has been donated to the language community from over sixty major projects in English and additional data from Cantonese, Danish, Dutch, French, German, Greek, Hebrew, Hungarian, Italian, Japanese, Mambila, Mandarin, Polish, Portuguese, Russian, Swedish, Tamil, Turkish, and Ukrainian. Along with data from normally-developing children, there are data from children with language disorders, adult aphasics, second language learners, and early childhood bilinguals. In essence, the database is simply a set of standard text files of transcripts of conversational interactions. In a few cases, the computerized transcript is accompanied by digitized audio and even video, but the vast majority of the corpora have only transcripts without audio or video.

CHILDES has led to the publication of over 1300 published research studies in the areas of language disorders, aphasia, second language learning, computational linguistics, literacy development, narrative structures, formal linguistic theory, and adult sociolinguistics.

Researchers use CHILDES in two major modes. The first mode focuses on the examination of patterns in the existing database. Researchers operating in this first mode need to learn the basic functions of the CLAN programs for searching across corpora. However, they are mostly interested in understanding the shape of the database and the nature of the various existing corpora. They may be interested in studying the development of specific syntactic constructions or parts of speech, such as questions, prepositions, plurals, or demonstratives. To study these issues, they typically use the basic search and tabulation programs in CLAN. Because there are fewer data on child language disorders and even fewer still from adult aphasics, this mode of research is somewhat less attractive currently for the areas of developmental language disorders and aphasiology.

The second mode of research uses the CLAN programs and the CHAT transcript format to transcribe and analyze new data. Workers operating in this second mode usually develop their own coding schemes and analysis routines designed to address project-specific questions. When researchers have completed their work, they then contribute their transcripts as new corpora for the database. Researchers operating in this mode are particularly interested in understanding the ways in which the various CLAN programs can help them address their current research needs. In order to maximize their use of the CLAN programs, they also need to understand the various alternative ways in which one can use the CHAT transcription system.

2. The Database

The first major tool in the CHILDES workbench is the database itself. Through CD-ROM or FTP, researchers now have access to the results of nearly a hundred major research projects in 20 languages. Using this database, a researcher can test a vast range of empirical hypotheses directly against either the whole database or some logically defined subset. The database includes a wide variety of language samples from a wide range of ages and situations. Almost all of the data represent real spontaneous interactions in natural contexts, rather than some simple list of sentences or test results. Although more than half of the data come from English speakers, there is also a significant component of non-English data.

Until 1989, nearly all of the data in the CHILDES database were from normally developing children. However, recent additions to the database have included several major corpora from children with language disorders. These include data from Down Syndrome contributed by Nahid Hooshyar, Jean Rondal, and Helen Tager-Flusberg; data from autistic children contributed by Helen Tager Flusberg; data from SLI (Specific Language Impairment) contributed by Lynn Bliss, Patricia Hargrove, Gina Conti-Ramsden, and Larry

Leonard; and data from children with articulatory disorders contributed by Susan Fosnot-Meyers and the Ulm University Clinic.

All of the major corpora are in the CHAT standard and have been checked for syntactic accuracy. The total size of the database is now approximately 180 million characters (180 MB). The corpora are organized into directories by language type and material type. In addition to the basic texts on language acquisition, there is a bibliographic database for Child Language studies (Higginson & MacWhinney, 1990).

Membership in CHILDES is open. Members receive electronic messages through the info-childes@mail.talkbank.org electronic bulletin board. In order to be officially included in the info-childes electronic mailing list and database, researchers should send email to macw@cmu.edu with their computer address, postal address, affiliations, and phone number. Users should abide by the rules of the System. In particular, they should abide by the stated wishes of the contributors of the data. Any article that uses the data from a particular corpus must cite a reference from the contributor of that corpus. The exact reference is given in the CHILDES manual.

All of the CHILDES materials can be downloaded without charge from http://childes.psy.cmu.edu.

3. CHAT

All of the files in the database use a standard transcription format called CHAT. This system is designed to accommodate a large variety of levels of analysis, while still permitting a bare bones form of transcription for those research projects in which additional levels of detail are not needed. Here is a brief example of segment of a transcript from a Broca's aphasic transcribed in CHAT. The file begins with these 10 lines of identifying material, or "headers."

```
@Begin
@Participants: PAT Patient, INV Investigator
@Age of PAT: 47;0.
@Sex of PAT: male
@SES of PAT: middle
@Date: 22-MAY-1978
@Comment: Group is Broca
@Filename: B72
@Coder: JMF
@Situation: Given/New task
```

After the headers, the actual transcript begins. This is a picture description task and each picture is identified with an @g marker to facilitate later retrieval. In the first three @g segment, the patient is describing a set of three pictures used in Bates, Hamby, and Zurif (1983) and MacWhinney and Bates (1978). In this first set, various animals are all eating bananas. In its "raw" form, what the patient said was simply, "rabbits, squirrel, monkeys." Here is how this is transcribed:

```
@g:     3c = bunny is eating banana
*PAT:   rabbits [*].
%mor:   DET|0 N|rabbit-*PL
%err:   rabbits = rabbit $SUB;
@g:     3b = squirrel eating banana
*PAT:   squirrel.
%mor:   DET|0 N|squirrel
@g:     3a = monkey eating banana
*PAT:   monkeys [*].
%mor:   DET|0 N|monkey-*PL.
%err:   monkeys = monkey $SUB ;
```

Here, the *PAT line conveys the simple shape of the patients description of the three pictures - "rabbits, squirrel, monkeys". We can notice several things about this transcription. First, the %err or "error" lines code the fact that plurals are used for two of the pictures when, in fact, only a single animal appears in each. The locus of these errors is marked in the main line or *PAT line with the symbol [*]. The %mor line is designed to indicate the morphological shape of the words on the main line. This line is used to study the use of different parts-of-speech and syntactic constructions. In this example, the %mor also provides a backup to the %err line, since both lines code for errors of omission and commission. The %mor line is intended to have a one-to-one correspondence with the main line, but when an item is marked as missing on the %mor line, it does not need to be present on the main line. For example, the code "DET|0" indicates that the determiner is missing on the main line. The code "N|monkey-*PL" indicates that the patient used the noun "monkey" in the plural, but that the use of the plural was an error in this case. The advantage of the elaborate coding on the %mor line is that it provides a more systematic structure for search programs that tabulate missing items by part-of-speech.

Let us look at one more segment from the same patient in the same study. Here the picture involves the dative verb "give". It is "raw" form, what the patient said was simply, "boy, girl, school, rat, boy no girl, girl truck girl." Here is how this is transcribed:

```
@g:     8a = lady giving present to girl
*PAT:   boy [*] [//] girl # school [*].
%mor:   DET|0 N|girl N|*xxx .
%err:   boy = girl $SUB ; school = [?] $SUB ;
@g:     8c = lady giving mouse to girl
*PAT:   rat .
%mor:   DET|0 N|rat
@g:     8b = lady giving truck to girl
*PAT:   <boy [*] no>[//] girl [/] girl truck # girl +...
%mor:   DET|0 N|girl N|truck N|girl.
%err:   boy = girl $SUB ;
```

In this example, we see several additional features. In description for picture 8a, the self-correction or retracing of "boy" by "girl" is marked by [//]. The repetition of the word "girl" is marked by [/]. Pauses are marked by # and the trailing off of the last sentence for picture 8b is marked by +... In the description for picture 8c, there is no %err line, since the characterization of the "mouse" as a "rat" is not judged to be so far off the mark as to constitute an error.

These two examples illustrate only a few of the many symbols and conventions available in the CHAT system. The system provides many options, but the transcriber only needs to select out those options that are relevant to the particular case. The simpler the transcription, the better, as long as it still captures the important aspects of the aphasic production.

The examples we have looked at illustrate some of the basic principles of the CHAT transcription system. Three of the most fundamental aspects of the system are:

1. Each utterance is transcribed as a separate entry. Even in cases when a speaker continues for several utterances, each new utterance must begin a new entry.
2. Coding information is separated out from the basic transcription and placed on separate "dependent tiers" below the main line. The CHILDES manual presents coding systems for phonology, speech acts, speech errors, morphology, and syntax. The user can created additional coding systems to serve special needs.
3. On the main line, transcription is designed to enter a set of standard language word forms that correspond as directly as possible to the forms produced by the learner. Of course, learner forms differ from the standard language in many ways and there

are techniques in the CHAT system for notating these divergences, while still maintaining the listing of word forms to facilitate computer retrieval.

For full examples of the coding system and its many options, the reader should consult the CHILDES manual.

4. CLAN

For the last few years, the main emphasis of new developments in the CHILDES system has been on the writing of new computer programs. Currently, there are two major components of the CHILDES programs. The first is the set of programs for searching and string comparison called **CLAN** (Child Language Analysis). The second is a set of facilities built up around the editor.

The **CLAN** programs support four basic types of linguistic analysis (Crystal, 1982; Crystal, Fletcher, & Garman, 1989): lexical analysis, morphosyntactic analysis, discourse analysis, and phonological analysis. In addition, there are programs for file display, automation of coding, measure computation, and additional utilities. The following table lists the programs by type.

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
Morphosyntax	COMBO	Searches for combinations of words or types of words
	COOCCUR	Tabulates pairwise cooccurrence 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	CDI 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

Lexical analyses. The programs for lexical analysis like **FREQ** and **KWAL** focus on ways of searching for particular strings. The strings to be located can be entered in a command line, one at a time, or put together in a master file. The strings can contain wild cards and words can be combined using Boolean operators such as "and", "not", and "or". Together, these various capabilities give the user virtually complete control over the nature of the patterns to be located, the files to be searched, and the way in which the results of the search should be combined into files or even reduced into data for statistical analysis. Scores of studies have appeared in the published literature using these techniques to track the development of lexical fields, such as morality, kinship, gender terminology, mental states, causative verbs, and modal auxiliaries. It is also possible to track the use of words of a given length or a given lexical frequency. **FREQ** outputs a complete frequency analysis for a single file or for groups of files. Here is an example of a **FREQ** frequency count for a single small file with only the Mother's utterances being analyzed.

```
freq sid.cha +f +t*MOT
Sun Jul 16 01:31:13 1995
freq (21-NOV-94) is conducting analyses on:
 ONLY speaker main tiers matching: *MOT;
****************************************
From file <sid.cha> to file <sid.fr0>
 13 a
 2 about
 1 ah
 4 all
 1 all+right
 1 ambulance
 7 and
 7 are
 1 are-'nt
```

```
2 back
2 be
1 because
1 bet
3 big
1 bought
3 boy
1 bring-ing
1 build
1 building
1 can
2 clever
2 come
1 crash
1 daddy
1 dear
1 did
7 do
5 do-'nt
```

In this analysis we see that the Mother used the word "big" three times. If we want to look more closely at these usages, we can use **KWAL** and we will get this output:

```
kwal +t*MOT +sbig sid.cha
Sun Jul 16 01:33:11 1995
kwal (21-NOV-94) is conducting analyses on:
 ONLY speaker main tiers matching: *MOT;
****************************************
From file <sid.cha>
----------------------------------------
*** File sid.cha. Line 336. Keyword: big
*MOT: is it go-ing to be a big ship ?
----------------------------------------
*** File sid.cha. Line 344. Keyword: big
*MOT: and that-'is go-ing to be a big ship .
----------------------------------------
*** File sid.cha. Line 379. Keyword: big
*MOT: that-'is <all the small lego> [//] all the big lego@ you-'ve got .
```

Each of these programs has many options that can allow the user to vary the shape of the input, the shape of the output, and the type of analysis that is being conducted.

Morphosyntactic analyses. Many of the most important questions in child language require the detailed study of specific morphosyntactic features and constructions. Typically, this type of analysis can be supported by the coding of a complete %mor line 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 twenty 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 coders 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.

To address this problem, we have built an automatic coding program for CHAT files, called **MOR**. Although the system is designed to be transportable to all languages, it is currently only fully elaborated for English, Japanese, Dutch, 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 which 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. 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. 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**.** In particular, users of **MOR** will often need to spend a great deal of time engaging in the processes of lexicon building and ambiguity resolution. To facilitate lexicon building, there are several options in **MOR** to check for unrecognized lexemes and to add new items. To facilitate ambiguity resolution, we have integrated a system for sense selection into the **CED** editor.

Construction of a full %mor line using **MOR** also makes possible several additional forms of analysis. One is the automatic running of the **DSS** program that computes the Developmental Sentence Score profile of Lee (1974). Parallel systems of analysis will eventually be developed for systems such as IPSYN (Scarborough, 1990) or LARSP (Crystal et al., 1989). The %mor line can also be used as the basis for CLAN programs such as cooccur which examines local syntactic structures and **CHIP** which examines recasts, imitations, and structural reductions.

Because of the importance of agrammatism in the study of aphasia, it would seem that the **MOR** program would be of particular interest to aphasiologists. However, the presence of large numbers of lexical, phonological, and syntactic errors in aphasic speech makes automatic application of the **MOR** program more difficult. Despite these difficulties, this is an area of great potential interest for work on language disorders.

Discourse and narrative. The most important CLAN tool for discourse analysis is the system for data coding inside the **CLAN** editor. The editor 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. In the coding mode, the editor allows the user to establish a predetermined set of codes and then to march through the file line by line making simple key stroke movements that enter the correct codes for each utterance selected.

Once a file has been fully coded, a variety of additional analyses become possible. The standard search 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 track 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 which maps, for example, chains in which there is no shift of topic and places where the topic shifts. Wolf, Moreton, and Camp (1994) 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.

Phonological analyses. Currently, phonological analysis is a bit of a step-child in **CLAN**, but we have plans to correct this situation. These plans involve two types of developments. One is the amplification of standard programs for inventory analysis, phonological process analysis, model-and-replica analysis, and other standard frameworks for phonological investigation. Currently, the two programs adapted to phonological 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.

The second set of plans for improving our ability to do phonological analysis focus on the use of digitized sound within the **CED** editor. On the Macintosh, the **CED** editor allows the transcriber direct access to digitized audio records that have been stored using an application such as Sound Edit 16. In the next year, we hope to implement a similar utility for the Windows platform. Using this system that we call "sonic CHAT", one can simply double-click on an utterance and it will play back in full CD quality audio. Moreover, the exact beginning and end points of the utterance are coded in milliseconds and the **PAUSE** program can use these data to compute total speaker time, time in pausing between utterances, and overlap duration time. A sample of a file coded in sonic CHAT with a wave form displayed at the bottom of the window is given in Figure 1. In this file, the numbers on the %snd tier refer to absolute time in milliseconds from the beginning to the end of a particular utterance.

The basic CLAN programs like **FREQ** and **KWAL** are easy to use and understand. They work on a simple MS-DOS type command line and one can often get the basic answers to important research questions without understand any of the more arcane uses of some of the

less common CLAN programs. In addition, users can rely on a well-tested manual that is now in its second edition and there are additional support resources available over the InterNet.

The most recent additions to CLAN focus on the linkage of transcripts to video. This facility uses QuickTime on both Macintosh and Windows. As we advance in the development of this technology, CHILDES becomes increasingly a multimedia database.

Beginning in 1999, CHILDES became the first organized component of the new, larger international database TalkBank system at http://talkbank.org.

References

Bates, E., Hamby, S., & Zurif, E. (1983). The effects of focal brain damage on pragmatic expression. *Canadian Journal of Psychology, 37*, 59-84.

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

Crystal, D., Fletcher, P., & Garman, M. (1989). *The grammatical analysis of language disability. Second Edition*. London: Cole and Whurr.

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

Lee, L. (1974). *Developmental Sentence Analysis*. Evanston, IL: Northwestern University Press.

MacWhinney, B., & Bates, E. (1978). Sentential devices for conveying givenness and newness: A cross-cultural developmental study. *Journal of Verbal Learning and Verbal Behavior, 17*, 539-558.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1-22.

Wolf, D., Moreton, J., & Camp, L. (1994). 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* (pp. 174-209). Hillsdale, NJ: Lawrence Erlbaum Associates.