Child and Adult Spoken Languages Resources: The CHILDES System

Brian MacWhinney
Carnegie Mellon University
Pittsburgh, PA, USA
[macw@cmu.edu]

Steven Gillis
University of Antwerp - UIA
2610 Antwerp, BELGIUM
[Steven.Gillis@uia.ua.ac.be]

Abstract

The CHILDES system consists of a large multilingual database of spontaneous child and adult speech, a system for discourse notation and coding (CHAT), and a set of computer programs (CLAN). These three integrated components of the CHILDES System are introduced.

Introduction

The study of spontaneous language samples involves an enormous time commitment to data collection, transcription, and analysis. In this paper we will discuss a system that can facilitate the process of free speech analysis. This is the system of programs and codes developed by the Child Language Data Exchange System (or CHILDES) Project (MacWhinney 1995, Sokolov & Snow 1994). The CHILDES system involves three integrated components: (1) a system for discourse notation and coding called CHAT, (2) a set of computer programs called CLAN, and (3) a large database of language transcripts formatted in CHAT.

A number of features distinguish the CHILDES system. Perhaps the most important is the linkage of the CHILDES programs to a large, internationally recognized, database of language transcripts. These transcripts include data from over forty major projects in English and additional data from 19 other languages. The additional languages are Chinese (Mandarin and Cantonese), Danish, Dutch, French, German, Greek, Hebrew, Hungarian, Italian, Japanese, Mambila, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Turkish, and Welsh. The highly crosslinguistic nature of the database and the involvement of language (acquisition) researchers in over 50 countries on every continent lends a decidedly international quality to this endeavor.

In the area of language and language acquisition studies, the attempt to include a crosslinguistic focus is not just an optional conceptual nicety; it is absolutely central to evaluating the core issues in language and language acquisition. The most prominent linguistic theories either make serious claims about universals of grammar (Chomsky 1965) and conceptual structures (Bickerton 1984, Slobin 1985), or else emphasize the role of language differences in terms of cue distribution (MacWhinney & Bates 1989) or social interaction (Schieffelin & Ochs 1987). In order to understand both the universals and the particulars of language and language acquisition, it is crucial to adopt this international, crosslinguistic perspective.

The database is not only international in scope, it also cuts across a variety of disciplinary boundaries. Included in the database are data from normally-developing children, children with language disorders, adults with aphasia, learners of second languages, bilinguals who have been exposed to language in early childhood, etc. Although most users of the CHILDES system are members of the child language research community, the system is also used extensively by students of child language disorders, aphasia, second language learning, computational linguistics, literacy development, narrative structures, and adult sociolinguistics.

The Database

The first major tool in the CHILDES workbench is the database itself. Through CD-ROM or WWW, researchers now have access to the results of nearly a hundred major research projects in 20 languages. Using this database, a researcher can directly test a vast range of empirical hypotheses against either this 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.

At first, nearly all of the data in the CHILDES database were from normally-developing children interacting with their (primary) caretaker(s). This yields an important stock of data from children's language (at various ages), caretakers language as well as spontaneous adult-adult interactions. Since most corpora in the CHILDES database consist of everyday conversations of adults and children, the data provide a rich source of everyday spoken language - as opposed to written language or language use in 'formal' contexts such as radio or television newsreports.

Table 1 presents an overview of the available resources based on the 1997 release of the CHILDES CD. Monolingual and bilingual CHILDES corpora of spontaneous interactions are distinguished. Corpora with children's narrations and corpora from language impaired subjects are not included into this overview. For each language represented in the database, Table 1 stipulates the number of children (up to age 12) and the number of adults whose speech was recorded and transcribed, as well as the total number of utterances and the total number of words tokens.

English is best represented in the database: language productions of 276 children and 1233 adults are available. For the children this amounts to an average number of 2,279 utterances (range: 1 - 46,480) and 8,297 word tokens (range 1 - 163,814) and for the adults, an average number of 686 utterances (range 1 - 47,456) and 3,181 word tokens (range (1 - 233,118). In the area of bilingualism, there are corpora of the following language

Language		Children			Adults		
		#	# Utterances	# Words	#	# Utterances	# Words
Monolingual							
	Cantonese	8	69,347	185,330	32	0	0
	Catalan	15	59,794	144,152	95	83,771	328,529
	Danish	2	30,351	63,274	12	42,398	156,453
	Dutch	79	180,996	451,597	52	226,994	926,007
	English	276	628,959	2,290,050	1,233	845,606	3,922,144
	French	77	58,511	211,096	11	75,351	410,708
	German	220	65,579	349,905	113	38,591	178,032
	Greek	4	7,213	16,120	13	12,087	35,827
	Hebrew	92	24,280	83,532	54	18,371	65,258
	Hungarian	31	19,804	59,352	10	14,729	59,242
	Italian	7	23,440	60,675	30	33,360	158,771
	Japanese	1	21,676	41,362	7	26,537	75,386
	Mambila	8	311	1,718	6	987	6,287
	Mandarin	5	17,108	35,972	24	58,005	225,932
	Polish	3	4,789	17,918	11	8,198	17,918
	Portuguese	99	6,647	42,438	5	8,657	52,221
	Russian	1	5,359	17,251	7	5,038	22,611
	Spanish	13	21,656	71,062	40	25,594	111,779
	Swedish	3	25,487	60,426	9	29,102	151,833
	Tamil	1	2,023	3,780	3	4,260	11,569
	Turkish	34	14,485	63,608	11	10,901	35,480
	Welsh	8	34,609	72,581	17	37,201	212,735
Bilingual		98	21,321	74,817	295	116,103	551,089
Total		1085	1,343,745	4,418,016	2,090	1,721,841	7,715,811

Table 1: Overview of monolingual and bilingual corpora in the CHILDES database.

pairs: English-Cantonese, English-Polish, English-French, English-Hebrew, English-Dutch, English-Spanish, English-Italian, Englis-Punjabi, French-Greek, and Japanese-Danish.

On the whole, spontaneous speech of 1,085 children and 2,090 adults is represented in the database, which amounts to 3,227,823 utterances and 12,828,076 word tokens. Recent additions to the database have included several major corpora from children with language disorders. These include data from Down Syndrome, from autistic children, from SLI (Specific Language Impairment) and from children with articulatory disorders. Also in the area of adult aphasia, the database includes several large corpora. These corpora are of prime importance to psycholinguists, but they may also constitute the ultimate test for robust language understanding systems.

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 160 million characters (160 MB). In addition to the basic texts on language acquisition, there is a database from the Communicative Development Inventory (Dale, Bates, Reznick & Morisset, 1989) and a bibliographic database for Child Language studies (Higginson & MacWhinney, 1990).

CHAT

All of the files in the database use a standard transcription format called CHAT (Codes for the Human Analysis of Transcripts). This system is designed to accommodate a large variety of levels of analysis, while still permitting a barebones form of transcription when additional levels of detail are not needed. The CHAT system is grounded on three basic principles.

1. Each utterance is transcribed as a separate entry in the system. Even in cases when a speaker continues for several utterances, we ask the transcriber to enter each new utterance on a new line. This is important, since it greatly facilitates the matching of additional information to the "main line".
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 create additional coding systems to serve special needs.
3. On the main line, the main goal of the transcription is 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 a wide variety of techniques in the CHAT system for notating these divergences, while still maintaining the listing of word forms to facilitate computer retrieval.

The example (1), taken from a Dutch corpus, shows the general format of a transcript. The main line is prefixed with an asterix followed by an abbreviation of the speaker' name (*ANN, *JOL). The main line contains a transcription in standard Dutch orthography. The dependent tiers are prefixed with a '%' followed by an abbreviated type description of the coded material: %pho for a phonemic transcription, %mor for a morpho-syntactic coding, %eng for an English gloss. A one-to-one relationship between the main line and the dependent tiers is maintained so that multiple tier manipulations are possible.

(1)
*ANN: de olifant.
%pho: də o:lifant.
%mor: det:deflde nlolifant.
%eng: the elephant.
*JOL: (h)et olifantje en (d)e tar [: kar].
%pho: ət o:li:fantʃə ɛn də tar.
%mor: det:deflhet nlolifant-DIM conj:coorlen det:deflde nlkar.
%eng: the elephant-DIM and the cart.

The CHILDES manual (MacWhinney, 1994) provides detailed guidelines for transcribing spoken language (including idiosyncracies of children's speech), conventions for marking discourse related phenomena (such as overlaps between utterances, retracings, etc.), guidelines for relating the spoken material to contextual information, etc. Full examples of the coding system and its many options are also provided in the CHILDES manual.

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 an editor called **CED** (CHILDES Editor).

The CLAN programs have been designed to 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.

In order to appreciate the scope and design of the CLAN programs, it should be kept in mind that the programs are meant to be used by non-specialists who are unaware of less dedicated though more powerful tools like AWK, or GREP. However the scores of studies which have appeared in the published literature using these programs for various tasks proves their usefulness for the purported audience. For this audience a tutorial introduction [Sokolov & Snow, 1994] to the use of the CLAN programs, illustrating their use in a functional, project oriented way, was published.

Lexical Analysis

Tools for lexical analysis focus on ways of searching for particular strings (much like AWK or GREP). The strings to be located can be entered in a command line or put together in a master file. The strings can contain wild cards and can be combined using Boolean operators. 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 can be combined in files or even reduced for statistical purposes.

Morphosyntactic analysis

An analysis of specific morphosyntactic features and constructions is supported by the coding of a complete %mor tier (as exemplified in (1)). Hand-coding of a %mor tier with a vast range of morphological and syntactic information is an extremely time consuming, error-prone and non-correctable undertaking. To address this problem, an automatic coding program has been developed for CHAT files, called MOR, which has now been fully elaborated for English, Japanese, Dutch and German.

CLAN tools that can be used for morphosyntactic analysis, include tools for combinatorial, Boolean serach, tools for tracking concordances and cooccurrence patterns, etc.

Discourse and narrative analysis

The most important CLAN tool for discourse analysis is the system for data coding inside the CED editor. **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. In the coding mode, **CED** 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.

For the actual discourse analysis, the **CHAINS**, **DIST**, and **KEYMAP** programs can be used 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. The most useful program for discourse analysis is the **CHAINS** program which looks at sequences of codes across utterances.

Phonological analysis

For phonological analysis there are standard programs for inventory analysis, phonological process analysis, model-and-replica analysis. 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.

Within the CED editor, phonological analysis can make use of digitized sound. The **CED** editor allows the transcriber direct access to digitized audio records that have been stored using an application such as SoundEdit16. Using this system - "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

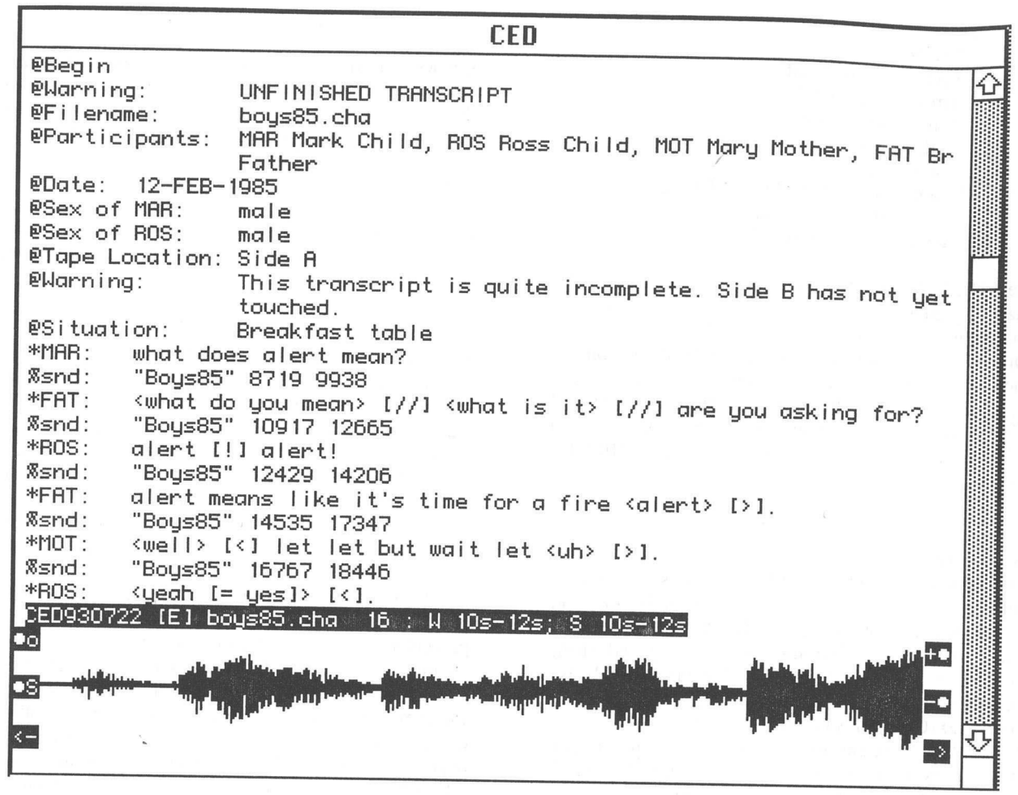

Figure 1: A sample file displayed in sonic CHAT with a waveform at the bottom.

milliseconds from the beginning to the end of a particular utterance.

The Future

Our plans for the future development of the CHILDES system state as a first priority the exploitation of the facilities of the World-Wide Web (WWW) to provide multimedia access to all CHILDES resources.

Equally important is the growth of connectivity between programs on a single computer. An example of the type of development we are currently supporting is the linkage of the CED editor to high-level speech analysis tools such as Signalyze on the Macintosh or WAVES on UNIX.

The Glossome. The emergent connectivity of the InterNet has opened up an even more exciting prospect that few have yet appreciated. This is the potential for the establishment of the Glossome Database. Much like the Human Genome Database, the creation of a set of standards for data transcription and transmission will allow us to access a wide variety of data on not just language learners, but also adult conversations, huge databases of written texts, phone conversations, schoolroom lessons, and all manner of human language production by all types of speakers in all languages. Of course, we will never encode the full contents of the Human Glossome, but we can divise tools that will allow us to understand the patterns involved in the enormous diversities of behaviors that we can human language.

Successful formation of this important new resource will require an overt commitment from researchers acting as individuals and through their professional societies and journals. 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 insure the stable and continued development of the CHILDES database and the gradual emergence of the Glossome Database. Until such a policy is developed and accepted, the voluntary acceptance of these responsibilities that has characterized the child language field will guarantee continued growth of the database.

Acknowledgements

Preparation of this paper was supported by CLIF (Computational Linguistics in Flanders] and by a VNC grant (contract number G.2201.96).

References

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

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.

Dale, P., Bates, E., Reznick, S. & Morisset, C. (1989). The validity of a parent report instrument. *Journal of Child Language*, 16, 239-249.

Higginson, R. & MacWhinney, B. (1994). *CHILDES/BIB 1994 Supplement*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1994). *The CHILDES Project: Tools for Analyzing Talk (Second Edition)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. & Bates, E. (Eds.), (1989). *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.

Schieffelin, B., & Ochs, E. (1987). *Language acquisition across cultures*. New York: Cambridge University Press.

Slobin, D. (1985). Crosslinguistic evidence for the language-making capacity. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition. Volume 2: Theoretical issues*. Hillsdale, N. J.: Lawrence Erlbaum.

Sokolov, J., & Snow, C. (Eds.), (1994). *Handbook of research in language development using CHILDES*. Hillsdale, NJ: Erlbaum.