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 newswcasts.

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

<table>
<thead>
<tr>
<th>Language</th>
<th>Children # Utterances</th>
<th>Children # Words</th>
<th>Adults # Utterances</th>
<th>Adults # Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>8</td>
<td>69,347</td>
<td>185,330</td>
<td>32</td>
</tr>
<tr>
<td>Catalan</td>
<td>15</td>
<td>59,794</td>
<td>144,152</td>
<td>95</td>
</tr>
<tr>
<td>Danish</td>
<td>30</td>
<td>65,274</td>
<td>222,907</td>
<td>12</td>
</tr>
<tr>
<td>Dutch</td>
<td>78</td>
<td>260,906</td>
<td>415,597</td>
<td>22</td>
</tr>
<tr>
<td>English</td>
<td>276</td>
<td>628,959</td>
<td>2,290,050</td>
<td>1,233</td>
</tr>
<tr>
<td>French</td>
<td>77</td>
<td>58,511</td>
<td>211,096</td>
<td>11</td>
</tr>
<tr>
<td>German</td>
<td>220</td>
<td>65,579</td>
<td>349,905</td>
<td>113</td>
</tr>
<tr>
<td>Greek</td>
<td>7</td>
<td>21,167</td>
<td>60,426</td>
<td>30</td>
</tr>
<tr>
<td>Hebrew</td>
<td>9</td>
<td>24,280</td>
<td>53,532</td>
<td>54</td>
</tr>
<tr>
<td>Hungarian</td>
<td>19</td>
<td>39,552</td>
<td>60,675</td>
<td>30</td>
</tr>
<tr>
<td>Italian</td>
<td>27</td>
<td>24,605</td>
<td>60,752</td>
<td>30</td>
</tr>
<tr>
<td>Japanese</td>
<td>1</td>
<td>21,676</td>
<td>41,362</td>
<td>7</td>
</tr>
<tr>
<td>Mandarin</td>
<td>5</td>
<td>17,108</td>
<td>35,927</td>
<td>24</td>
</tr>
<tr>
<td>Portuguese</td>
<td>6</td>
<td>4,857</td>
<td>62,483</td>
<td>10</td>
</tr>
<tr>
<td>Russian</td>
<td>1</td>
<td>3,539</td>
<td>17,251</td>
<td>7</td>
</tr>
<tr>
<td>Spanish</td>
<td>3</td>
<td>7,198</td>
<td>11,181</td>
<td>11</td>
</tr>
<tr>
<td>Swedish</td>
<td>25</td>
<td>27,567</td>
<td>30,340</td>
<td>30</td>
</tr>
<tr>
<td>Tamil</td>
<td>1</td>
<td>2,032</td>
<td>3,780</td>
<td>3</td>
</tr>
<tr>
<td>Turkish</td>
<td>34</td>
<td>14,485</td>
<td>63,608</td>
<td>11</td>
</tr>
<tr>
<td>Welsh</td>
<td>36</td>
<td>21,895</td>
<td>142,781</td>
<td>113</td>
</tr>
<tr>
<td>Bilingual</td>
<td>8</td>
<td>21,321</td>
<td>34,679</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>1,085</td>
<td>1,343,745</td>
<td>4,418,016</td>
<td>2,090</td>
</tr>
</tbody>
</table>

**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 number 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 major purpose 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 corresponding standard forms and thus require specific coding in the form of %mor, %ann, %pho, %det, etc.

**CLAN**

The CHILDES manual (MacWhinney, 1994) provides detailed guidelines for transcribing spoken language (including idiosyncrasies of children’s speech), conventions for marking discursive 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.

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 a set of programs for searching and string comparison called CLAN (Child Language Analysis). The second is a set of programs built up around an editor called CED (CHILDES Editor).

The CLAN programs have been designed to support four basic types of queries. (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-experts 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-transportable 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 search, tools for tracking concordances and cooccurrence patterns, etc.

**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 %mod tier symbols with the corresponding main line text. For more precise control of MODREP, it is possible to create a separate %mod tier 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 - called “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 simple sample of a file coded in sonic CHAT with a wave form displayed at the bottom of the window is shown in Figure 1. In this file, the numbers on the %mod tier refer to absolute time 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 simple sample of a file coded in sonic CHAT with a wave form displayed at the bottom of the window is shown in Figure 1. In this file, the numbers on the %mod tier refer to absolute time in milliseconds.
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 divine 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.2001.96).

References


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