

Out of the Baby Book and Into the Computer: Child Language Research Comes of Age

The CHILDES Project: Tools for Analyzing Talk, Vol. 1: Transcription Format and Programs (3rd ed.)

by Brian MacWhinney

Mahwah, NJ: Erlbaum, 2000. 159 pp. ISBN 0-8058-2995-4. \$75.00

The CHILDES Project: Tools for Analyzing Talk, Vol. 2: The Database (3rd ed.)

by Brian MacWhinney

Mahwah, NJ: Erlbaum, 2000. 418 pp. ISBN 0-8058-3572-5. \$75.00

Review by Jean Berko Gleason and R. Bruce Thompson

If you build it, they will come" is Brian MacWhinney's favorite quotation. And his field of dreams is the Child Language Data Exchange System (CHILDES), which has profoundly altered the way we study children's acquisition of language. Developmental psycholinguistics has become a science rather than a matter of philosophical inquiry within just the past 50 years. The transformation occurred because of new theories in the fields of linguistics and psychol-

MacWhinney's field of dreams is the Child Language Data Exchange System (CHILDES), which has profoundly altered the way we study children's acquisition of language.

ogy, the availability of data on many children, and, most recently, remarkable new electronic analytical tools. The author of these two volumes, Brian MacWhinney, has been the moving force behind much of the new technology, and these books and the accompanying compact disk provide a working introduction to CHILDES. The system is made up of three main parts: a standardized way of transcribing spoken language that makes it accessible to computer analysis; computer programs that can do such things as instantly list every word used by a child; and a database of transcribed language that has been contributed from research projects in many countries. The books come as a package along with a compact disk that contains the database, the computer programs and various other useful software, as well as the material that is also in the printed volumes. Computer-literate researchers can thus install the programs on their own

equipment and set about transcribing and analyzing their own data, or they can pose questions that they can answer by investigating the data already collected by others.

Language Research Before CHILDES

Before the middle of the 20th century, most studies of children's emerging language were conducted, as the author notes, in a rather private way: Parents kept diaries, for instance. They noted what was interesting about their child's speech. Much of the focus was on the emerging intellect, but linguistic questions were also considered, especially the development of vocabulary. Darwin published a study of one of his own sons in 1877. In the United States, G. Stanley Hall studied the "contents of children's minds" and inspired a whole early 20th-century school of American child language researchers. The research was almost always in the form of a baby book or diary study of the researcher's own child, notable exceptions being a few studies of feral or terribly abused children. Diary studies conducted by linguists were often extensive and insightful. For instance, the Russian linguist A. N. Gvozdev followed his son Zhenya's development until the boy was nine years old, documenting the appearance of sentence types, parts of speech, the sound system, and many other formal elements of the Russian language.

In the late 1950s a number of forces, both intellectual and technological (however primitive by today's standards), converged to help move child language study into a more public and accessible form. Advances in the fields of linguistics and psy-

BKIAN MACWHINNEY, Department of Psychology, Carnegie Mellon University.

JEAN BERKO GLEASON, Department of Psychology, Boston University, 64 Cummington Street, Boston, Massachusetts 02215. E-mail: gleason@bu.edu

R. BRUCE THOMPSON, Department of Psychology, University of Southern Maine, Portland, Maine 04104. E-mail: bttiomps@usm.maine.edu

chology led researchers to ask new questions and to revisit some ancient ones as well. For instance, one of the most compelling and enduring questions has been whether major aspects of children's language acquisition are invariant and universal, or whether language reflects a constellation of abilities that are highly sensitive to culture, class, and other social factors. On the technological side, we had tape recorders and spirit duplicators. Descriptive linguistics provided insights into the componential nature of language (separate systems for meaning, sound, and grammar), and the powerful theories of Noam Chomsky inspired researchers to investigate children's emerging syntax. At the same time, we were on the eve of the cognitive revolution, and psychologists became interested in the mental representations of the systems described by linguists. The field of psycholinguistics was thus born, and the three major areas of inquiry in this new field became language comprehension, language production, and the development of language by children.

In the explosion of research that followed, child language study went from the idiosyncratic, handcrafted style of the diary to a veritable cottage industry. At Harvard University, Roger Brown conducted pioneering longitudinal studies of language development in three children, not his own, called Adam, Eve, and Sarah. Graduate assistants went to their homes on a periodic basis, made recordings of the child and parents in typical interactions, and brought the tapes back to the laboratory, where they were transcribed. The resulting transcripts were duplicated and passed around to interested colleagues and to Brown's graduate seminars, where the data became the stuff of many dissertations. The transcripts were thus shared locally, and unlike diaries, relatively public, so researchers could ask their own questions or check on what others had found. Work of this nature proliferated and became the norm in observational studies over the ensuing two decades. There was, of course,

no general archive of duplicated transcripts, and comparability among projects was nonexistent. Transcription standards and coding conventions were developed at each research site in response to the kinds of questions being asked, and the published results presented only high-level analyses based on the criteria established by the research team.

The Tools

It was fundamentally the emergence of the microcomputer that made possible, in principle, everything that CHILDES was to become. The early 1980s brought word processing systems and database programs that could be used to enter language transcripts and extract information from them. A number of developmental psycholinguists, including MacWhinney, began to envision a system of shared data, transcription formats, and analytic programs. In 1983, with this goal in mind, the MacArthur Foundation funded a series of meetings of developmental researchers that led to the establishment of CHILDES. The following year CHILDES was put in place at Carnegie Mellon University with MacWhinney as principal investigator and Catherine Snow of Harvard as coprincipal investigator. Snow has remained a vigorous supporter, and has helped introduce the system internationally by conducting workshops at child language meetings. MacWhinney engaged Leonid Spektor, who began work on the development of the new analytic programs, and who remains as the primary software developer. CHILDES, from its inception, has been free, in the monetary sense, and open, in the sense that the entire research community has been invited to participate in its development. All of the data and programs can be downloaded from the CHILDES Web site, and the project maintains technical and general online discussion groups with heavy participation from the child language community. A question posted to the Info-CHILDES portion of the CHILDES Web site is liable to lead to responses from the best-

known names in the field. The third edition of the CHILDES Project volumes reflects this ongoing relation with the researchers who use the tools, as well as changes in the system made possible by new technology.

Transcription Rules

The first half of Volume 1 covers transcription and coding. The transcription system itself is called Codes for the Human Analysis of Transcripts (CHAT). The basic aim of CHAT is to provide a standardized way of preparing digitized transcripts of spoken or signed language that can then be subjected to computer analysis. Virtually all of the transcripts now in the database are in CHAT format, and currently about 60 research projects around the world are using this system. The system is very flexible, and can do as little or as much as the investigator wishes. A minimal transcript simply follows a few conventions, such as having a marked beginning and ending, and having each speaker identified by an asterisk and a three-letter code. The speaker's utterance follows. A simple transcript (p. 17) might read:

```
ftsbegin
^Participants: ROS Ross Child, BR1 Brian Father
*ROS: why isn't mommy coming? com:
Mother usually picks Ross up around 4p.m.
*BRI: don't worry. *BRI:
she'll be here soon. *ROS:
good, (send
```

Although there are approximately 200 pages of instructions for preparing transcripts that even include notation for American Sign Language conversations, this portion of the book is very well organized and accessible. It begins with instructions for the minimal requirements noted above. The novice user learns to make a small file and then run a program called CHECK that ensures that the analytic programs will run on the file. From this point on, remarkably sophisticated possibilities open up. For instance, coding can be added to the transcript (as in the comment line in the transcript shown earlier), and

there are sections in the manual on specialized codes for many purposes in a large variety of languages.

Computer Analysis

The analytic programs, called Computerized Language Analysis (CLAN), are covered in the second half of Volume 1, again including nearly 200 pages. There are instructions for installing the programs on the user's platform, a tutorial, and detailed descriptions of the function and purpose of about 40 different programs that can be run on the data. This portion of the book constitutes the CLAN manual, and anyone contemplating serious use of the programs had best have it at hand, because the online help within the programs is limited and in the typical UNIX format consisting of minimal English and a list of commands and possible switches. It is a rare researcher who actually knows all the CLAN programs. Rather, investigators learn to invoke just those routines that will be most helpful in accomplishing their current aims.

The manual begins with the most commonly used programs, and their use is shown in the tutorials. Some of the most useful programs are those that will show all words used, along with their frequencies, or search for particular words or word combinations and display them in context. Other common programs compute basic statistics on the language of one or many speakers, such as the mean length of a speaker's utterances, or the type-token ratio (a measure of lexical diversity.). Beyond the basics, CLAN provides capabilities for every level of analysis, from phonology to discourse. Research studies as disparate as the acquisition of individual sounds and the development of conversational competence can be accommodated by the system. Of course the power of such programs is such that they can be run on one child's speech, or the speech of 100 children at once, and produce results almost instantaneously.

The manual is clear, in the ways that a software manual can be. It is possible to read it, while in front of the computer, and work through the

limited exercises to get an idea of how the programs work. But the programs are not easy to learn this way. It takes a good deal of attention and care to master them, and even for advanced graduate students, there is a steep learning curve. Researchers will find it most useful to learn the basics in a workshop, or from an experienced colleague, and then to have the book as an invaluable reference.

Data in 25 Languages

Volume 2 is devoted to descriptions of the material in the database, including information on the circumstances under which the original recordings were made and relevant demographics. MacWhinney provides guidelines for the shared use of the data, as well as notes on confidentiality and other ethical concerns. The computerized database contains transcripts from approximately 100 major independent research projects around the world, in 25 different languages. It includes contemporary work, but also includes the classics in the field, such as Brown's transcripts of Adam, Eve, and Sarah, and some very interesting older data. For instance, there is a directory with a study of everything that one young child said in an entire day, from the time she arose in the morning until she went to bed at night. The copious notes on the child Helen were taken by her mother, a faculty member at the University of Minnesota, in 1912. When we discovered this, it was all we could do to keep writing this review, rather than crank up the computer and start trying to see how Helen compares with two year olds today.

There are corpora in English and the major European languages as well as in Mandarin, Tamil, Hebrew, Japanese, and Mambila (an African language spoken in Nigeria and Cameroon). Although the database contains primarily children's language, there are also studies of adult aphasic speech. The child data include clinical populations with varying types of language problems. Volume 2 discusses and organizes the

data by language and other criteria, but it does not contain any of the transcripts, which are available on the compact disk or on the World Wide Web in either Macintosh or Windows format.

Technology and Science

The CHILDES system outlined in these books has many obvious merits and advantages. It has given the child language community data on many children, ways of producing standardized transcriptions, and a set of computerized programs that can accomplish in seconds tasks that would take weeks to do by hand. The large numbers of languages inspire cross-cultural work. This is essential for the development of linguistic theory, which, too often, has been built on observations of a few Western languages. For instance, it was long thought that children everywhere find nouns much easier to process than verbs. Studies of children acquiring Asian languages have now shown that whether nouns or verbs are easier to process very much depends on the language being learned, rather than on some innate principle in the child's mind.

The availability of an archive of already collected data may, of course, shape the direction of inquiry: Researchers faced with the labor-intensive task of recruiting participants, recording them, and transcribing what was said may opt for using what is already there, and tailoring their questions accordingly. This problem is not new, and it is not unique to the study of language acquisition. The Harvard historian of science Peter Galison (1997) has documented similar concerns in the field of experimental physics. Galison notes that when leading physicists met in Karlsruhe in 1964 to discuss changes in their discipline, one physicist talked about the dangers of automation. Another expressed the fear that

in a few years... one would not go to start a new experiment but one would just go into the archives, get a few magnetic tapes, and start to scan the tapes from a new point of view ... that would be the experiment. (Galison, 1997, p. 1)

Related concerns can, in theory, be raised about transcription systems and analysis programs. The units of transcription imply a prior analysis of the language that may not be agreed on by all. The programs may tend to canonize certain procedures and discourage further consideration of alternate strategies. For instance, the mean length of utterance, called the MLU, has become a standard measure of the syntactic complexity of a child's emerging language. The average length of a child's utterances serves as an index of the child's stage of grammatical development. Perhaps there are better measures that could be developed, but it is easier to run the MLU program. There is no simple solution to the inherent danger in any science that new technology may mold the discipline, rather than vice versa. CHILDES confronts these problems dynamically through constant evolution and revision in response to the research community.

The Future

Work is under way now to overcome one of the major limitations in research that relies on transcripts: The problem of seeing and hearing real children talk. MacWhinney is developing an audio and video database that will be integrated with the text-based corpora. This new project, called Talkbank, allows researchers to access sound, video, and the transcript all at once, with links from the words in the transcript to video clips of the interaction under study. The CHILDES system itself is moving increasingly from the personal computer to the international arena of the World Wide Web, with mirror sites in Belgium and Japan.

Although MacWhinney has made substantive contributions to the field of child language research, and is himself a prominent connectionist theorist, the CHILDES system does not reflect a theoretical bias, except insofar as it is predicated on the notion that the analysis of actual data is required to understand how children learn language. The very fact that concerns about the system are so similar to those raised among physi-

cists is testament to how far the field of developmental psycholinguistics has come as a science. This does not mean that CHILDES is the only way we should be studying language. There will always be a need for careful observation, for insights derived from watching actual children, and for experimentation. What we

have now in CHILDES are the major data of our field, and a way of moving forward as a scientific community. Q

Reference

- Galison, P. (1997). *Image & logic: A material culture of physics*. Chicago: University of Chicago Press.