HANDBOOK OF
Neurolinguistics

Edited by

BRIGITTE STEMMER
Centre de Recherche du Centre Hospitalier
Côtes-des-Neiges, Montréal, Canada
and
Lurija Institute for Rehabilitation and Health Sciences at the University of
Konstanz, Kliniken Schmieder, Allensbach, Germany

HARRY A. WHITAKER
Department of Psychology
Northern Michigan University
Marquette, Michigan

1998

ACADEMIC PRESS
San Diego London New York Boston Sydney Tokyo Toronto

CHAPTER 44

Computational Transcript Analysis and Language Disorders

Brian MacWhinney
Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

The errors and omissions found in aphasic language production are a rich source of information about how language is processed in the brain. However, in order to fully exploit this type of data, we need a consistent methodology for elicitation, recording, transcription, and analysis. One such framework is provided by tools developed for the CHILDES (Child Language Data Exchange System) Project. This paper examines those tools in the light of the development of research methodology from the precomputer period into the current period of connectivity and exploratory reality. Although these tools were originally developed for the analysis of language acquisition data, they can be readily adapted to the study of language disorders.

When asked to describe simple pictures or to recite simple narratives, aphasics often illustrate a wide variety of paraphasias, word-finding difficulties, phonological disfluencies, and grammatical errors. These errors and omissions provide us with two important windows into the functioning of language in the brain. First, by studying the various types of errors and omissions in psychological and linguistic terms and by comparing their relative frequencies, we can learn a great deal about how aphasia affects the basic mechanisms of language processing. For example, the study of omissions of markers like the plural on the noun or the past tense on the verb has

HANDBOOK OF NEUROLINGUISTICS

been useful in developing our understanding of production processes in aphasia (MacWhinney & Osman-Sági, 1991; Menn & Obler, 1990).

Error patterns also offer us potential information about clinical groupings. By looking at differential patterns of errors and omissions, we can distinguish the telegraphic production patterns found in agrammatism or Broca's aphasia, the more verbose and error-prone patterns characterizing Wernicke's aphasia, the exclusively lexical error patterns characterizing anomia, and the more exclusively phonetic error patterns found in dysarthria and apraxia of speech. These patterns in aphasia can also be compared with possibly similar language patterns in the speech of people with schizophrenia, right-hemisphere damage, frontal lobe damage, Alzheimer's disease, and other neural disorders. We can then couple this information with additional information from comprehension, neural imaging, and other methodologies to advance claims regarding the structuring of language in the brain.

Unfortunately, the study of production errors and clinical patterns in production has often been a rather hit-or-miss endeavor. Because there is no standardized reference database of production data, it is difficult to evaluate the relative position of new clinical samples and data from single subjects. There are a number of standard diagnostic tests available for the study of aphasia, but the actual transcripts from these tests have not been collected in a publicly available repository and organized in a fashion that permits easy comparison between individual subjects and the larger database.

There are some fairly good reasons why this database has not yet been developed. Although it is extremely easy to collect production data, it is much more difficult to analyze these data in a scientifically consistent manner. Many laboratories have dozens and dozens of transcripts of disordered speech sitting in paper repositories or stored on computer disks. It is easy to turn on a tape recorder or videotape recorder and build up a huge library of hundreds of hours of tapes. However, transcribing, coding, and analyzing hours upon hours of recordings involves an enormous time commitment. If the work spent doing this transcription is to be meaningful, we need to set standards to guarantee the comparability of data across subjects, laboratories, protocols, and transcription formats.

44-1. THE CHILD LANGUAGE DATA EXCHANGE (CHILDES) SYSTEM

Fortunately, there already is a well-developed framework for the process of database formation and analysis that can be directly applied to the study of aphasia. This is the system of programs and codes developed by the Child Language Data Exchange (CHILDES) Project (Higginson & MacWhinney, 1990, 1994; MacWhinney, 1991a, 1991b, 1993, 1994b, 1994c, 1995, in press; MacWhinney & Snow, 1985; 1990). The CHILDES project has been directed by the author in collaboration with Catherine Snow of Harvard University and has been supported since 1987 by the National Institute of Child Health and Human Development (NICHD).

It is important to realize that, although the CHILDES system was originally developed for use in the study of child language data, we took care early on to make sure that the system would also be applicable to the analysis of aphasic language data. We did this by emphasizing the development of tools for the analysis of speech errors and by including data from both children and adults with language disorders. At this point, the acronym CHILDES is largely a historical relic, since the system is being used not only by child language researchers, but also by discourse analysts, sociolinguists, speech pathologists, aphasiologists, computer scientists, and applied linguists studying second language acquisition.

Sometimes researchers who are unfamiliar with CHILDES think of it only as a database. However, in order to develop a useful and consistent database, we had to construct a system based on 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 the course of 6 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 tool kit of analytic possibilities that can be combined to fit a specific research agenda. Most recently, the programs have been extended to provide tools for linking transcripts to digitized audio and video records.
3. **Database.** Finally, the system includes the database itself, with data donated to the language community from more than 40 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.

The system has been used as the basis of nearly 400 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.

There are two major modes in which researchers have used the CHILDES system. 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 on 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.

Each of the three components of the CHILDES system was designed from the beginning to be useful across languages. The crosslinguistic focus of the CHILDES system is not just an optional methodological nicety; it is conceptually central. And this centrality holds equally well for both child language and aphasiology. In both of these fields, certain prominent theories make strong claims about universals of grammar (Chomsky, 1965), conceptual structure (Bickerton, 1984; Slobin, 1985), or sentence processing (Frazier, 1987). Proponents of these universalist theories often argue that these abilities are located in specific brain modules that can be damaged in specific ways (Friederici & Frazier, 1992; Grodzinsky, 1990; Warrington & McCarthy, 1987). In the simplest case, these accounts would claim that all language is organized in the same basic way and that the patterns of dissociation we find across languages (Paradis & Lebben, 1987) should be basically the same. An alternative view stresses the importance of language differences in determining patterns of errors and omissions in aphasia (Bates, Wulfeck, & MacWhinney, 1991). These between-language differences are then attributed to variation in cue distribution (MacWhinney & Bates, 1989) or social interaction (Schieffelin & Ochs, 1987). In order to understand how both universals and particulars interact in language learning and language loss, researchers must adopt a crosslinguistic perspective. Ideally, we will also want to maximize comparability across languages by settling on a standardized set of pictures and other tasks for the elicitation of language production.

Before we begin a more detailed examination of the current status of the CHILDES system, it may be useful to step back a bit to look at the ways in which the methodology for the study of spontaneous language production has evolved historically. This historical overview can help us gain some perspective on the status of our current methodological advantages and the ways in which changes in methodology are linked to changes in theory. Starting in ancient times and continuing up through the present, we can distinguish five major periods. In each of these periods, our understanding of the nature of language has been closely linked to the nature of the methodology that

has been available for studying language performance. During each of these periods, the methodology used for the study of language acquisition has been essentially the same as the methodology used for the study of language disorders.

44-2. FIVE METHODOLOGICAL PERIODS

44-2.1. 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 I, 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.

44-2.2 Period 2: Diaries and Biographies

The second major technique for the study of language production was 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, studies of the language of particular patients have been presented by Low (1931), Pick (1913, 1971), Wernicke (1874), and many others.

44-2.3. 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 by 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.

44-2.4. 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.

44-2.5. Period 5: Connectivity and Exploratory Reality

Since 1981, the world of computers has gone 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.

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. What we lack now in the field of aphasiology are not the conceptual or computational tools, but the organizational commitment that will be

TABLE 1
Some Computer Terminology

Term	Explanation
Audiovisual/AV	Computer that can control sound and video
CD-ROM	Removable disk that gives access to huge amounts of nonerasable data
CHAT	CHILDES transcription and coding format
CLAN	CHILDES programs for data analysis
DAT tape	Inexpensive way of archiving large amounts of data
digitized speech	Storage of sound in a form that can be played by the computer
electronic bulletin board	A forum for the discussion of issues through computer mail For CHILDES this is info-childes@andrew.cmu.edu
E-mail	Electronic mail that operates over the Internet
FTP	File transfer protocol—a program for moving data between computers
hard drive	Built-in device that gives access to large amounts of erasable data
Internet	System of electronic links that allows computers to transfer data
Macintosh	An operating system designed to machine user-friendliness
MS-DOS	A common and easily controlled operating system for microcomputers
optical disk	Removable disk that gives access to huge amounts of erasable data
poppy.psy.cmu.edu	The machine that makes CHILDES data and programs available by FTP
TAR	A program that puts many files into one (like Zip or Compactor)
UNIX	A powerful, but sometimes difficult, operating system
World Wide Web	Software that facilitates use of the Internet for conceptual connections

needed to push forward the development of a standardized database. 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 child language disorders.

44-3. 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's 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.

In the area of adult aphasia, the database includes two large corpora. The first is a set of conversational interviews with 42 aphasic patients during the period of recovery from stroke donated by Audrey Holland. The second is a collection of interview and picture description data from aphasic speakers of English, German, Hungarian, Chinese, and Italian donated by Elizabeth Bates and her colleagues. One of the major priorities for the CHILDES project is the inclusion of additional data on both childhood language disorders and aphasia during the coming years. We are aware of a variety of additional computerized corpora in the area of adult aphasia (Menn & Obler, 1990; Paradis & Lebben, 1987) and we hope to be able to convince researchers in aphasiology of the importance of making these data sets publicly available.

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 180 million characters (180 MB). The corpora are divided into six major directories: English, non-English, narratives, books, language impairments, and bilingual acquisition. 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).

Membership in CHILDES is open. Members are listed in a standard database and receive electronic messages through the info-childes@andrew.cmu.edu electronic

bulletin board. In order to be officially included in the info-childes electronic mailing list and database, researchers should send E-mail to childes@cmu.edu with their computer address, postal address, affiliations, and phone number. Users are asked to 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 (MacWhinney, 1991b). In addition, researchers should cite the 1991 version of the manual, since this allows us to track references in the literature.

All of the CHILDES materials can be obtained without charge by anonymous FTP to childes.psy.cmu.edu in Pittsburgh and atila-ftp.uia.ac.be in Antwerp. Our address on the World Wide Web is http://childes.psy.cmu.edu. For users without access to the Internet, as well as for those who want a convenient way of storing the database, we have published (MacWhinney, 1994a) a CD-ROM that can be read by Macintosh, UNIX, and MS-DOS machines that have a CD-ROM reader. The disk contains the database, the programs, and the CHILDES/BIB system. One directory contains the materials in Macintosh format and the other contains the materials in UNIX/DOS format. The CD-ROM, the printed manual, and the research guide are available at nominal cost through Lawrence Erlbaum Associates.

44-4. 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 segments, 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 the following:

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 create 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 a variety of 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.

44-5. CLAN

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. Table 2 lists the full set of programs by type.

44-5.1. 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

TABLE 2
CLAN Programs and Their Function

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	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

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.

44-5.2. 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 becomes 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 noncorrectable. 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 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. 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, which 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.

44-5.3. Discourse and Narrative

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 keystroke movements that enter the correct codes for each utterance selected.

Once a file has been fully coded in CED, 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 that 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 that computes the mean length of the turn for each speaker.

44-5.4. Phonological Analyses

Currently, phonological analysis is a bit of a stepchild 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 focuses 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. We hope to implement a similar utility for the Windows platform. Using this system, which 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 waveform displayed at the bottom of the window

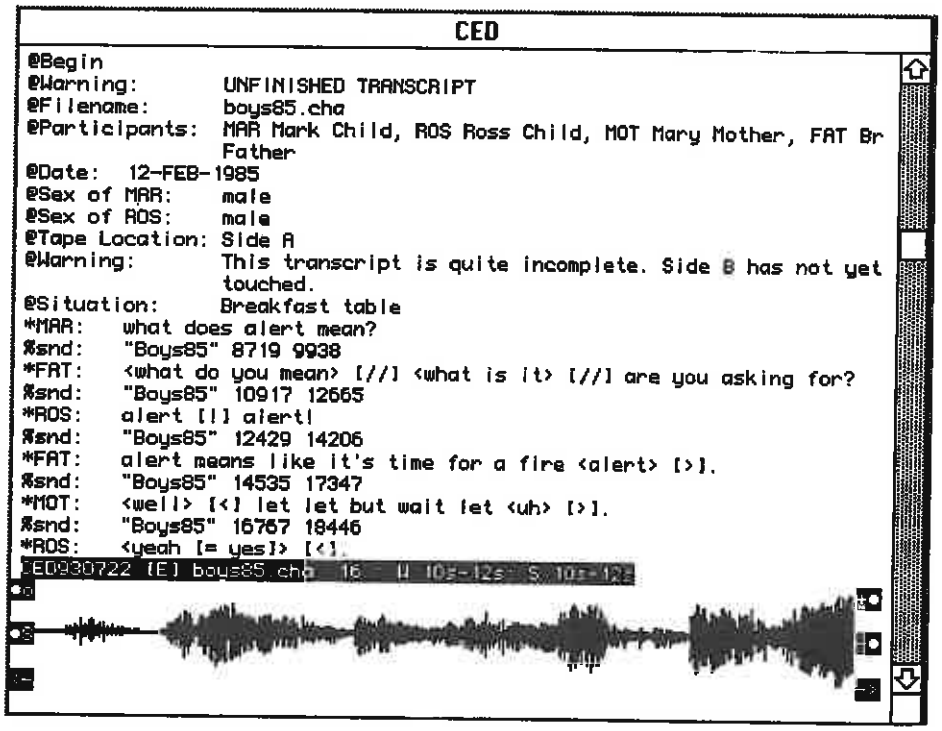

FIGURE 1 A sample file displayed in sonic CHAT with a waveform at the bottom.

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 extremely 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 an understanding of 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.

44-6. CONCLUSION

Earlier we looked at four periods in the growth of observational studies of language development. We are now entering the fifth period of methodological development. Our plans for the future development of the CHILDES system are based on the view of the fifth stage of observational research as being the period of electronic connectivity and exploratory reality. Our first priority for this period is to make full use of the facilities of the World Wide Web (WWW) to provide multimedia access to the database, the bibliographic system, and the manual. Using currently available tools such as Netscape, Macintosh AV facilities, and HTML formatting programs, it is now possible for a user to use a sequence of mouse clicks to open up pages of the CHILDES manual, search for particular files in particular corpora, open up those files, and hear the sounds in each. It is even possible to have pictures of the children and parents accessible over the net.

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. We also plan to have access to a reference database of IPA sounds, as well as audio examples of specific uses of CHAT symbols and codes.

44-6.1. The Glossome

The emergent connectivity of the Internet has opened up an exciting prospect that few researchers have yet considered. This is the potential for the establishment of a Glossome Database. Much like the Human Genome Database, the Glossome Database would be supported by data entry over the Internet. The creation of a set of standards for data transcription and transmission will allow us to store and access a wide variety of data from a wide variety of normal and disordered populations.

In order to make successful use of these new opportunities, we will need to develop a higher level of consciousness in both the adult aphasia research community and the child language disorders research community. In each of these areas, the strong commitment to patients' rights must be protected and encouraged. However, researchers often cite patients' rights as a motivation for not sharing their data with the broader

research community. This interpretation of personal rights does damage to the progress of the very field that is dedicated to improving the condition of the aphasic patient. The only way to counter these protectionist sentiments is for major figures in the field to lead by contributing new data to the database and by encouraging younger researchers to follow their lead.

Currently, we have had much more success in convincing students of child language disorders than students of adult aphasia to enter their data into CHILDES. Given the fairly advanced state of methodology in the CHILDES system and the small amount of aphasic data currently in CHAT format, it may now make more sense to focus our efforts on collecting new sets of well-transcribed data that are accompanied with full digitized audio records that could be accessed directly over the Internet. Ideally, we would like to see a large body of consistently transcribed data for comparable tasks, which could provide us with a consistent basis for comparison. Although the transcription standards and analytic programs are already in place, there must be a period of further dialogue regarding elicitation tasks and related issues. We would like to work together with workers in the field of adult aphasia to build a solid empirical database for studies of disordered language production in both adults and children.