J. Child Lang. **12** (1985), 271–296. Printed in Great Britain

The child language data exchange system*

BRIAN MACWHINNEY
Carnegie-Mellon University

AND CATHERINE SNOW
Harvard University

(*Received* 1 *June* 1984)

ABSTRACT

The study of language acquisition underwent a major revolution in the late 1950s as a result of the dissemination of technology permitting high-quality tape-recording of children in the family setting. This new technology led to major breakthroughs in the quality of both data and theory. The field is now at the threshold of a possible second major breakthrough stimulated by the dissemination of personal computing. Researchers are now able to transcribe tape-recorded data into computer files. With this new medium it is easy to conduct global searches for word combinations across collections of files. It is also possible to enter new codings of the basic text line. Because of the speed and accuracy with which computer files can be copied, it is now much easier to share data between researchers. To foster this sharing of computerized data, a group of child language researchers has established the Child Language Data Exchange System (CHILDES). This article details the formation of the CHILDES, the governance of the system, the nature of the database, the shape of the coding conventions, and the types of computer programs being developed.

Background and motivation for the system

The rapid expansion of child language studies during the period from the late 1950s to the present has been fuelled, in large part, by the introduction of

[*] This research was carried out with support from the John D. and Catherine T. MacArthur Foundation, Milton Grodsky Program Director. Among those who have participated in the discussions that have led to the formulation of these proposals and policies are Elizabeth Bates, Ursula Bellugi, Marjorie Beeghly-Smith, Lois Bloom, Melissa Bowerman, Robin Chapman, Eve Clark, Phillip Dale, Jane Edwards, Peter Eimas, Paul Fletcher, William Hall, Karen Hardy-Brown, Judith Johnston, Jerome Kagan, Patricia Kuhl, Willem Levelt, Laurence Leonard, Michael Maratsos, Joanne Miller, Jon Miller, Jane Morrison, Ann Peters, Steve Reznick, Dave Shucard, Dan Slobin, Tom Roeper, Peter Wittenberg, Kenneth Wexler, and Dennis Wolf. Address for correspondence: Dr. Brian MacWhinney, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 15213.

devices for audio and video tape-recording. These instruments permitted the collection of large corpora of spontaneous and elicited speech of children interacting with their peers, parents, and visiting experimenters. With the introduction of the audio tape-recorder and the video tape-recorder it became possible to make records of child language productions that were far more complete than the diaries that had been compiled from pencil and paper notes. With these new data, the field was able to advance well beyond the level of anecdotal summaries that had characterized the first 100 years of child language research. The new methodology permitted full recording of not just the child, but also those who interacted with the child. This new possibility slowly turned the field's attention to studies of the input to the language-learning child. During the 1960s and 1970s, researchers continually improved their use of recording technology, using directional microphones, wireless microphones, well-justified sampling strategies, non-intrusive recording methods, and detailed systems for coding aspects of the communicative context.

The proliferation of transcript data has led to immense advances in specificity and accuracy of our science. It has also allowed us to see more clearly the limitations involved in current analytic techniques. As we begin to compare hand-written and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability have become more apparent. But, just as these new problems have arisen, a new major technological opportunity has emerged. This opportunity is provided by the proliferation that is now occurring of microcomputers and microcomputer software. With microcomputer word-processing systems readily available to all researchers, it is easy to enter transcript data into computer files which may then be easily duplicated, edited, and analysed by standard data-processing techniques. The numbers of researchers entering their data in this fashion has grown exponentially over the last few years. However, it has also become apparent that this new opportunity might slip away if researchers could not move quickly to establish conventions, standardizations, and policies for the sharing and comparison of computerized transcript data.

Early in 1984, support to Carnegie-Mellon University from the John D. and Catherine T. MacArthur Foundation enabled us to begin work on the establishment of a system for the exchange of such transcript data. The goal of this system is to bring about major improvements in the organizational structure and accessibility of the empirical base for child language theory. We believe that there are three basic needs to be addressed: (1) the need for greater efficiency in data sharing, (2) the need for increased precision in data collection, coding, and analysis, and (3) the need for increased automation of coding and analysis. The new systems, which we have called the Child Language Data Exchange System (CHILDES), will utilize a variety of recent

advances in data-processing hardware and software to address these three specific needs and to work towards the goal of improving structure and accessibility of child language data. We will explain the organization of the exchange system below, but first we must consider in greater detail the three major needs we are addressing.

The need for greater efficiency in data sharing. Child language data – transcripts of conversations between children and interlocutors – are notoriously time-consuming to produce. Moreover, any given set of transcripts is typically greatly underused and undershared. For example, taping and transcribing one child on a weekly basis requires one hour of observation, 6–10 hours of primary transcription, another 1–3 hours of transcript checking, editing, and typing, all before any coding or analysis can be carried out. Such a corpus may then be used to ask questions about one aspect of language acquisition (i.e. the priority of tense vs. aspect) and then filed away on the shelf. The sad fact is that almost any corpus that is thus filed away is potentially useful for a large range of analyses and the non-availability of such data to the scientific community at large is a great loss to the field as a whole.

The need for greater precision in data collection and analysis. The reuse of transcript data for new purposes often forces the analyst to consider additional aspects of the process of data collection, transcription, coding, and analysis. Thus, the data sharing that will be promoted by the establishment of the CHILDES is likely to lead to an acceleration in the development of interest in data collection as a process. From the point-of-view of the long-term advance of the field, this would certainly be a welcome development.

Perhaps the most important outcome of this process will be the development of new standards for data collection, transcription, and analysis. The sharing of data that have already been collected may also lead to analytic advances. With files that are currently computerized, it is relatively easy to supplement the basic text line with additional coding features, thus (*a*) giving richer and more complete descriptions of each data set, and (*b*) enabling competing hypotheses to be tested against each other on a single data set. A recurrent issue in the analysis of child language data is how to determine the 'right' set of categories to use in the description; researchers have typically argued for the correctness of their own categories from their own newly collected data only, without the possibility of comparing them directly to other descriptions. For example, in the 1960s and early 1970s when semantic categories such as AGENT and ACTION were used extensively by researchers, the pragmatic notions of TOPIC and COMMENT were not applied. However, it is perfectly feasible to go back to earlier transcripts and add categories like topic and comment to the existent semantic coding schemes, and then to compare

directly the power of the pragmatic vs. the semantic categories. Moreover, whenever researchers have preserved the original audiotapes, it is possible to supplement transcripts with full segmental phonetic and prosodic coding.

Another example of the possibilities for better explanation opened by an accretional coding system involves adding interactional features to the coding, in an attempt to explain child language characteristics. For example, Folger & Chapman (1978) concluded from their analysis of six children that individual differences in children's tendency to imitate may relate to parental imitativeness, and that certain parental speech acts were much more likely to be imitated by children than others. If a central data exchange system had been available to Folger & Chapman, they could have added their speech act analysis to the transcripts for the six children whose imitations had been extensively studied by Bloom, Lightbown, & Hood (1975), to test directly the hypothesis that it is the shape of parental speech acts rather than the state of the child's knowledge that predicts children's imitation.

The need for increased automation in analysis. The advent of the tape-recorder led to the first automatization in the study of child language. The coming of accessible personal computing has opened up the door to a second and potentially even more important automatization. With handwritten or typewritten transcripts, scanning for the occurrences of particular events, word types, syntactic constructions, or speech act types is tedious and inaccurate, and adding additional coding categories is messy and error-prone. However, with increasingly accessible word-processing and database management programs, the tasks of entering, coding, and analysing speech transcripts is becoming simple, pleasant, and reliable. With data entry facilities of the type provided in programs such as CALARSP and PEPPER, or with larger DBMS systems such as Datatrieve, RS-1, and INGRES, the task of adding coding categories becomes feasible and relatively straight-forward.

We have found that the use of computational processing opens up new vistas for child language research, in much the way that the advent of recording techniques opened up the field in the 1950s and 1960s. However, there is a major and interesting difference between these two revolutions. In the first revolution, each researcher could scrape together money for a few tape-recorders and microphones, and, if s/he was lucky the budget might even accommodate a system for video-taping. In the case of the new computer technology, individual researchers may be able to purchase microcomputers for entry of their own data. However, they will often find that they cannot afford the major costs involved in main-frame processing designed to make analyses across multiple data sets. Thus, it appears that the researcher is in a position where s/he can only make full use of the new technology by engaging in some form of sharing of data, hardware, and software with other researchers. Unfortunately, the scarcity of funds for collaborative enterprises

has slowed the application of the new technology to the data. Only by establishing a well-organized system for the support of data exchange and the maintenance of software analyses can this new prospect in child language research become a reality.

A glance at a fairly representative publication in child language – the annual Stanford Papers and Reports on Child Language Development – indicates the extent to which failure to use automated schemes for searching is slowing progress in the field. In just the 20 papers in the 1983 volume from Stanford there are three papers based on paper-and-pencil counts of numbers of occurrences of different word types in the corpus of data collected by Roger Brown. Moreover, there are three other papers based on analyses of corpora other than that of Brown that could have utilized automated search and analysis techniques. By not using such techniques, researchers are not only wasting their time, but also increasing the likelihood of error and decreasing the explicitness of their analytic technique.

The formation of the system

In the summer of 1981, Dan Slobin, Willem Levelt, Susan Ervin-Tripp and Brian MacWhinney began discussions regarding the possibility of creating an Archive for typed and hand-written transcripts at Nijmegen. The possibility of adding computerized transcripts to the database was also considered. In January of 1984, the MacArthur Foundation awarded a two-year grant to Carnegie-Mellon University for the establishment of the Child Language Data Exchange System with Brian MacWhinney and Catherine Snow as Principal Investigators. These funds provided for the entry of data into the system management of correspondence, and the convening of a meeting of the Advisory Board of the System.

During the period between January and July, the following already-computerized corpora were entered into the database at CMU:

(1) the original computer files for *Informal Speech* by Carterette and Jones;
(2) the data on children gathered by Susan Ervin-Tripp and Wick Miller in the early 1960s;
(3) Bill Hall's data on school-aged lower-class children;
(4) two lists based on Francis and Kucera (1982) – one in alphabetical order and one in order of frequency;
(5) Brian MacWhinney's tape-recorded-diary of his two sons' development;
(6) Brian MacWhinney's and Elizabeth Bates' data on sentence production in 3-, 6-, and 10-year-olds in English, Italian, Hungarian, and Serbo-Croatian;
(7) Dan Slobin's study of English, Italian, and Turkish along with full semantic codings and English glosses;

(8) Anne van Kleeck's data from normal 3-year-olds;
(9) Jan Vorster's glossed and coded data from Afrikaans;
(10) Gordon Well's time-sampled data from British children.

By Kurzweil optical scanning of printed and typed materials, we have also entered:

(11) Roger Brown's data from Adam, Eve, and Sarah;
(12) Jacqueline Sachs' detailed study of her daughter Naomi;
(13) the transcripts appended to Lois Bloom's *One Word at a Time*;
(14) passages from Susan Isaac's *Intellectual Growth in Young Children.*

For all of the published materials, permission to copy has been secured from the publishers and authors. Because of their historical importance, we also plan to enter the hand-written transcripts of Bloom and diary notes of Werner Leopold. Books by Miller, Gvozdev, Leopold, and others will be entered by scanning in the near future.

The first meeting of the Advisory Board for the System was held in Concord, Massachusetts between March 15 and March 18. The meeting was organized by Catherine Snow. The Board members present were Elizabeth Bates, Ursula Bellugi, Lois Bloom, Melissa Bowerman, Robin Chapman, Eve Clark, Jane Edwards, Susan Ervin-Tripp, Paul Fletcher, Willem Levelt, Brian MacWhinney, Jon Miller, Ann Peters, Dan Slobin, and Catherine Snow. At this meeting, the Board sketched out the shape of this new System. These decisions delineated (1) the organization of the System, (2) the shape of the database, and (3) programming for the use of the database.

The organization of the System

The System has six components: the Advisory Board, the Centres, the database, programs, users, and contributors.

The Advisory Board. The Advisory Board is composed of child language researchers actively involved in the use of computers with transcript data. Its goal is to provide the Centres with technical advice and assistance in four areas. For each of the four areas, the Board constituted a working committee to form policies and decisions. The committees are as follows:

(1) The *Traffic Control Committee* decides which corpora should be entered into the database and in what priority.
(2) The *Transcription and Coding Committee* develops standards for codes on the various tiers (see below).
(3) The *Programming Committee* decides which programs should be written or purchased and in what priority.
(4) The *Project Development Committee* proposes major uses of the database and extensions of the System.

The Centres. Currently, the two Centres of the System are located at Carnegie-Mellon University in Pittsburgh and the Max-Planck Institut für Psycholinguistik in Nijmegen, Holland. Further Centres can be established at those sites that are able to dedicate resources to the maintenance and development of the System. The basic functions and shape of the System are duplicated at each Centre. The Centres keep in correspondence by computer mail, phone, and regular mail for updating of files and task sharing.

Users. Membership in the System is open. However, members must agree to abide by the rules of the System, not to distribute copies of programs or files without permission, to abide by the wishes of data contributors, and to properly acknowledge the contributors and the System. Members are urged to support the progress of the System through contributions of data, programming expertise, or professional advice. The Centres have made announcements regarding the System at the child language conference held during 1984 in Austin, Boston, and Stanford, and announcements have been placed in three journals. Announcements will also be made at the meetings of the American Speech and Hearing Association and the Society for Research in Child Development. Further announcements will be made at Boston, SRCD, and several European meetings. Announcements have been mailed to the membership of ASHA, AERA, NECLA, ICLA, and the Stanford Child Language Forum.

Members can make use of the database by (1) requesting tape copies of files, (2) logging on to the CMU-PSY-A computer by Telenet or Telnet, (3) sending jobs for analysis to CMU via ARPANET, BITNET or UUCP, or (4) coming to one of the Centres to log on locally. Only alternatives (1) and (4) are currently implemented; implementation of alternatives (2) and (3) must await further funding. *Transcript Analysis*, the newsletter of the CHILDES, will regularly brief members on the status of the database, the use of programs, and issues in analysis and coding.

Contributors. The Centres are responsible for soliciting contributions of data from members and others. This is done in accord with the suggestions of the traffic control committee. Contributors are asked to complete a contribution release form. This form gives permission to the Centre to circulate copies of the files. It also allows the contributors to set specific limitations and conditions on the use of their data. The Centres attempt to secure users' compliance with these requests of the contributors.

The database. The Centres will enter data sets into the System. Some data sets are already computerized; others must be typed in or entered by optical scanning. In such cases the correctness of the entry must be verified and some reformatting may be needed. Data sets will be entered into the system even

if they are not in compliance with the new transcription standards. However, potential contributors will be encouraged to transcribe new data according to the standards set by the Transcription and Coding Committee of the Advisory Board.

Programs. The Centres are currently designing a series of programs for analysis and database management. These programs are being designed to maximize accessibility to all users, while preserving proper administration of the System. These programs are described in detail below. Members are urged to contribute programs that they have developed.

The shape of the database

The System will store (1) transcript materials as received from contributors or as entered from materials furnished by contributors, (2) transcript files actively undergoing reformatting and the addition of new codes, (3) files with programs for database maintenance and transcript analysis, and (4) administrative files.

The archive. The first type of file will be referred to as the 'archive'. After initial entry and error checking, the original copies of these files will never be modified in any way. Users can request copies of archived files. However, these files will be in a wide variety of formats, and the Centres will not support programs for the analysis of these archived files. Preserving files in this form is important both for historical purposes and to monitor for possible errors in later forms of data conversion.

The active database. The Coding Committee of the CHILDES has developed a proposed normalized form for transcript data. A series of programs have been and will be developed to structure the files in the Archive into the normalized or standard format. Once a file has been reformatted, it is placed into the active database. Because files in the active database will have a consistent structure, it is possible to design a wide variety of analytic and administrative functions that operate upon these files. Some of the functions are:

(1) *Bibliography.* A program will allow the user to list bibliography and commentary relevant to particular files.
(2) *Usage.* A program will allow the user to list information on the usage of particular files.
(3) *Membership.* Information on the members of the CHILDES, their interests, contributions to the System, and their usage of the System will be available.
(4) *Annotation.* Users will be encouraged to add a variety of data to the active database. Entire tiers of coding (see below) can be added to files

or corpora. It is likely that the system will use INGRES as its DBMS (Database Management System) program. Within this program there are good facilities for adding levels of either hierarchical or non-hierarchical coding.

(5) *Updating.* There will be general facilities for updating aspects of the database.

(6) *Cross-Centre comparison.* There will be routines for checking to see that the shape of the database is consistent across Centres.

(7) *Outputting.* There will be facilities for generating a variety of output formats for transcripts and parts of transcripts.

(8) *Initial user interface.* There will be facilities for introducing users to the database and informing them of the nature of the data contained therein and its appropriateness for various sorts of analyses.

(9) *Analysis.* Facilities for data analysis will be discussed below.

Within the files of the active database, there is a five-level hierarchy of file information. There will be several partitions of the database by subject matter. Within each partition, data will be grouped by PROJECTS. A project is a major body of data donated by a contributor or team. Within a project, the data for a single informant will constitute a CORPUS. Thus for Roger Brown's child language project there are three corpora; Adam, Eve, and Sarah. Within each corpus, there are a series of FILES with each file corresponding to a recording session. Finally within each file there may be any number of EPISODES as defined by the contributor.

Files will be received from and sent to members in standard ASCII form. The preferred form of transmission will be IBM-compatible 9-track 800 or 1600 bpi ANSII-labelled magnetic tapes. The default recordlength for tapes produced by the Centres is 2048 bytes and the default blocksize is also 2048 bytes, but other values can be accepted. Tapes are made on a VAX 750 or 780 running VMS. The VMS copy command is used and the tape drive is mounted for an ANSI label.[1]

When sending tapes to the System, it would be helpful to have tapes prepared in this same way. However, it is usually possible to read tapes that are made in other ways on other systems.

Headers. Each of the five higher levels of the database (partition, project, corpus, file, episode) can be notated by headers. Header information is moved

---

[1] The exact commands issued to VMS are:

```
$ allocate mtaO:
$ initialize mtaO: trans
$ mount mtaO: trans
$ copy/log [ChiLDES...]*.* mtaO:*.*
```

The label 'mtaO:' is the name for the first tape drive on the CMU system and the tape label 'trans' is an arbitrary tape label name.

up to the highest possible level. For example, information that applies to all files in a corpus is stored not as a header on each file but as a corpus header. Contributors and the system will be responsible for encoding header information. The types of information to be coded are given below in order of relative generality. This information early in this list will probably apply to whole databases, whereas information late in the list will apply to episodes within files.

(1) *Basic understandings.* The System sets certain basic conditions on usage of any material. These include agreement to not distribute files and to acknowledge data sources.

(2) *Permissions and restrictions.* Contributors can set particular restrictions on the use of their data. To date no such restrictions have been set by any contributor.

(3) *Warnings on the label.* Contributors and the System may warn the user about particular limitations in the use of data from projects. For example, if an investigator paid no attention to correct transcription of speech errors, this may be noted as a project header.

(4) *Human subjects.* Headers will include information on whether informants gave release for use of their data and whether pseudonyms have been used to preserve informant anonymity.

(5) *History.* There should be detailed information on the history of the project. How was funding obtained? What were the goals of the project? How was data collected? What was the sampling procedure? How was transcription done? What was ignored in transcription? Were transcribers trained? Was reliability checked? Was coding done? What codes were used? Was the material computerized? How?

(6) *Codes.* If project-particular codes are being used, these should be described.

(7) *Biographical data.* Where possible, extensive demographic, dialectological, and psychometric data should be provided for each informant. There should be information on topics such as age, sex, siblings, schooling, occupation, previous residences, language background of the family, religion, interests, friends, etc.

(8) *Table of contents.* There should be a brief index to the contents of the corpora.

(9) *Situational descriptions.* On the level of the file and episode there should be complete header information that should allow the user to reconstruct the situation as much as possible. Who is present? What is the layout of the room or other space? What is the social role of those present, e.g. who is usually the caregiver? What activity is in progress? Is the activity routinized and, if so, what is the nature of the routine? Is the routine occurring in its standard time, place, and personnel configuration? What objects are present which affect or assist the interaction?

It will also be important to include relevant ethnographic information which would make the interaction interpretable to the user of the database (e.g. if the text is parent-child interaction before an observer, what is the culture's evaluation of behaviours such as silence, talking a lot, displaying formulaic skills, defending against challenges, etc.?).

Line headers

Below the five levels that can be given headers are two further levels of data. Within files the transcript is coded in terms of a series of UTTERANCES. Within each utterance, data may be transcribed in terms of a number of TIERS. In order to uniquely identify utterances and tiers, we place an asterisk before each utterance and before each tier line. This means that the standard orthographic or textual tier line begins with two asterisks, since it always comes first. The test line is marked by the speaker's name followed by a colon, and the other lines are marked with an abbreviation and a colon. Note that this scheme for line identification does not require a fixed format. It would be best to place the asterisks in the first column for clarity. However, long lines may be continued without repeating the line identifier. The occurrence of an asterisk is sufficient to mark the end of a tier of coding, and the occurrence of double asterisks marks the end of all tiers of coding for a single utterance.

Coding within the basic text line. A major aspect of the coding system is the extensive reliance on lines parallel to the basic text line to code additional comments and information. These additional lines are called TIERS. The basic line of text is called the TEXT TIER. In the simplest case, this is the series of words the speaker produces, with each work separated as in normal English text by spaces. If other tiers make reference to not just words but actually morphemes, the morphemes should be preceded by hyphens. Thus *dogs* can be coded on the text line as either *dogs* or *dog-s*. But prefixes and suffixes are preceded by the hyphen. In order to properly distinguish prefixes from stems, morphemically-segmented text lines should have stems preceded by number signs (#) when there are prefixes. Thus, the morphemic segmentation for *undecipherable* would be *-un # decipher-able*. When there are no prefixes, the number sign can be omitted. If a morpheme has been omitted from some obligatory context (Brown 1973), the omission can be marked by a percent sign. Thus, the utterance *that go here* could be notated as: that go-%3S here. And the utterance *he here* with *is* missing could be notated as: he %is here. A few additional conventions for notating the text line are as follows:

x	a syllable of nondecipherable material
pcf	a phonologically consistent form
.	the end of an unmarked (declarative) utterance

?	the end of a question
!	the end of an exclamation
^	the end of an incomplete utterance
'	contraction
,	intonational pause
<	break missing where required ('latching of turns')
>	normal ceding of a turn

Any comment-like material may be inserted into the text line at any point by placing the comment into square brackets. For example, coughing can be marked in the following way.

**Bill: I need a [coughs] milkie.

After the left square bracket there may be a code for the comment type. For example, [P1.800] means an 1800 millisecond pause, i.e. a pause of 1 second and 800 milliseconds. If a pause lasts into the minutes, this is recorded before a colon. Thus a pause of 2 minutes and 20 seconds would be [P2:20] with no milliseconds recorded. Turn-internal pauses may be marked as [IP...] and pauses between turns can be marked as [EP...]. The total duration of episodes and sessions can be marked in the headers.

A major problem facing those who wish to search out all occurrences of a given morpheme is that the spellings of morphemes are often inconsistent. Inconsistencies are produced by spelling rules, phonological processes, dialectal variation and so on. Full standardization of the spellings of morphemes may not be desirable. Researchers often wish to enter pseudo-phonetic characterizations into the main text line to mark certain colloquialisms and stylistic features. For example, the SALT analysis program (Miller & Chapman 1983) recognizes these spelling conventions for English:

ain't	atta (= that's a)	betcha	doctor (=Dr.)
gonna	gotta	hafta	hey
hi	huh	liketa	lookit
mhm	mister (= Mr.)	Mrs (= Mrs.)	Ms (=Ms.)
nope	opps	ok (= okay)	oughta
psst	sposta	trynta	uhhuh (= no)
wanna	whatcha	ya (= you)	yeah (= yes)
yep (= yes)	cuz (= because)		

Closely related to this problem of spelling conventions for colloquialisms is the introduction of variation by spelling rules and allomorphy. Thus the morpheme *have* occurs as *have* in *have*, but as *hav* in *having* and the morpheme *wife* appears as *wive* in *wives*. Our solution to this problem is to rely upon the computer to remind us about allomorphy, spelling rules, and colloquial pronunciations. Connected to the search mechanism within the

overall format of a system like GRITS, we will have a dictionary of allomorphs and orthomorphs. When a user decides to search for all occurrences of *wife*, the system will send a message advising the user that there is an additional spelling that could be searched: *wive*. The same facility will inform the user about possible homonymies. It will advise the user of ways of dealing with homonymy. For example, in order to distinguish the verbal suffix *-es* from the plural suffix *-es*, the user can select an automatic look-up of the part-of-speech of the previous stem in the part-of-speech table. However, even this heuristic will not solve all cases. For example, the stem *hoe* in *hoes* could be either a noun or a verb. In order to fully distinguish homophones, a comment field can be used in the main text line. Thus, *goes* would by *go-es*[*3S*] and *dogs* would be *dog-s*[*PL*].

If material within the comment refers to a particular stretch of speech, the stretch can be marked by angle brackets. For example, speaker overlap can be indicated as follows:

**Mary: But I ⟨wanted to go⟩ [overlap 182]
**Fred: ⟨No, you⟩ [overlap 181] can't.

Or the stressing of a string of words can be marked in the following way:

**Mary: Why ⟨can't I have cake?⟩ [stressed]

Possible alternative transcriptions can also be coded as comments:

**Tom: I think he ⟨did a little dance⟩ [Alt/didn't take a chance]

Coding on tiers. Because many researchers are interested in supplementing the basic code line with a variety of additional detailed codes that are more extensive than comments in the text line, it is important to provide a uniform way of entering this additional material. We propose to do this by developing a set of codes for tiers that are placed in relation to the main text line much in the way that parts of a musical score are placed in relation to the melodic line. For each tier recognized by the system, a set of codes will be developed on the basis of advice provided by researchers who specialize in the relevant subject area. In some areas, these codes have already undergone some standardization. In other areas, full standardization is a long way off and the best we can achieve is dissemination of possible codings. Where necessary, small working meetings will be held to settle on conventions.

In the actual files, all qualifiers or codes will be marked with a dollar sign and will be abbreviated as much as possible. Thus the code for 'intentionality' will be $INTEN. Here, however, we give only the unabbreviated qualifiers. Coding tiers that we currently recognize include:

(1) *Time marking.* Researchers who have equipment that permits full time-stamping from tape or videotape will add exact hours, minutes, and seconds on this line. Those who do not have time-stamping

equipment may make use of stopwatch times or the feet or metres registered by counters on tape-recorders. Entry of information on this tier is helpful, but not obligatory, since temporal relations can be judged by pegging markings on secondary tiers to words on the main text line.

(2) *Morphemic semantics.* Some standardization has been achieved in working out a coding system for the grammatical/semantic features of individual morphemes in different languages. Lehmann (1982) has published a proposal for standardization in this area, and that proposal has been adopted by Dan Slobin in his set of edited papers on the cross-linguistic study of language acquisition (Slobin 1984). These codes serve to provide an accurate morpheme-by-morpheme glossing for non-English transcripts. In order to ensure that the codes correctly match the morphemes of the text line, the morphemes of the text line and the codings of the morpheme semantics line must both be separated by hyphens. If more than one semantic code applies to a morpheme, the codes should be concatenated by slashes – thus, $PL/$ACC/$MASC/$FIRST codes for the first declension masculine accusative plural morpheme *-i* in Latin. When there is overt morphological marking of agreement, this marking is to be coded at each site. In addition to semantic codes, a set of morphophonological comment codes may be included on the morphosemantic line. These comment terms include: oblique, compound, contraction, metathesization, etc. For ASL morphophonological commenting, one can use terms such as: invented, cooccurrent, fingerspelling, and pantomime. If comments refer to a single morpheme, they are placed into square brackets following the morpheme to which they refer. For example, to note that the stem *wive* in the form *wives* is oblique we would have

**Keith: Seven wive-s?
*Mor: $NUM $COMMON[oblique]-$PL

Some of the semantic codes of the Slobin system refer to whole clauses. These codes will not be used on the morphemic semantic tier. Rather, they will be entered on the clausal relations tier. These morphemic codings go a long way towards representing many of the most widely studied semantic intentions. However, it will often occur that a child uses a sentence to express one of these semantic intentions even when there is no morphemic form on the surface to code the intention. In such cases, it is necessary to insert dummy or empty morphemes on the text line. This amounts to claiming that some material has been ellipsed from the text. This same mechanism can be used to code for the semantics underlying ellipsis in adults.

(3) *Structural coding.* A large number of coding categories refer not to individual morphemes, but to high groupings such as phrases, clauses, and sentences. In order to properly relate these structural codes to the units to which they refer, one needs a way of encoding hierarchical relations into the database. The most economical way of notating syntactic structure simply uses parentheses with subscripts as in:

(((MY)adj (DOG)noun)np ((HAS)verb (FLEAS)np)vp)s

A more flexible and more readable system would make use of a hierarchical database structure such as INGRES or Datatrieve to enter words on the text line into multiple fields in a relational database. This is the approach currently being taken by Gavin in his CALARSP system. Datatrieve and INGRES provide full query-language facilities for data entry and output facilities for analysis of such hierarchical structures. For each clause or utterance, the user sees the full hierarchical structure displayed on the screen and can edit this structure in order to insert the required codes. With this facility, codings such as subject, object, topic, comment, coordinate clause, subordinate clause, foregrounded clause, backgrounded clause, appositive, and so on can be entered on any level defined by the basic principles of syntactic bracketing.

(4) *Prosody.* No commitment to a system for coding prosody has been made. However, the scheme in Crystal (1969) might be adaptable for this purpose.

(5) *Paralinguistics.* Here, again, Crystal (1969) presents a possible scheme. Major phrase boundaries will be coded on the text tier with the symbols [,] and [.]. Hesitation phenomena will be coded on a separate tier. Stress and pitch modulations will be coded on the prosodic tier.

(6) *Phonetics.* IPA notation can be used for this purpose. However, a conversion to ASCII codes must be devised. A working session to resolve some of these issues is currently being organized.

(7) *Explanation of speaker's meaning.* This line is used to clarify unclear and missing referents in the child's speech. Some researchers refer to these explanations as 'glosses'. Ellipsed or omitted material may be explained on this line.

(8) *Alternative transcription.* If the transcriber is unsure of the correct transcription of a passage, an alternative transcription can be entered on this tier.

(9) *Situational contextual coding.* As noted earlier, major situational/contextual information should be noted in the headers to corpora, files, and episodes. Information that changes more rapidly should be coded on the situational/contextual tier. This information should allow the

user to reconstruct the ongoing flow of activities. Note, however, that some of the most detailed situational changes may also be coded on the gestural and proxemic tiers.

(10) *Speech acts.* Coding on this line will utilize a set of qualifiers in angle brackets. Some of the most important speech act/interactional qualifiers are: question command, request, invitation, prompt, suggestion, repetition, expansion, elaboration, break-down, rephrasing, completion, response, imitations, affirmative answer, negative answer, answer to yes-no question, answer to wh-question, request for repetition, compliance, denial, refusal, noncompliance, and choice among alternatives. A recent attempt to codify some of these categories is developed in Ninio & Wheeler (1984). Although it is far too early to reach full codification in this area, the development of an awareness of abbreviations and qualifiers will be useful.

(11) *Gesture and proxemics.* A variety of gestural and proxemic codes have been developed by workers in Los Angeles, Chicago, and elsewhere. Work on ASL notation is also relevant. A working meeting will be convened to support some standardization in this area. Among items to be coded are: gaze direction, reaching towards objects, manipulation of objects, examination of objects, postural changes, muscle tensings, distances between speakers, etc.

(12) *Free translation.* For transcripts in languages other than English, a free English translation can be provided as a utility to those not familiar with the language.

(13) *Errors.* A separate tier is devoted to errors. Some qualifiers for that tier include: secondary stem overgeneralization, primary stem overgeneralization, consonant assimilation error, harmony error, selection error, sandhi error, tone error, segmentation error, superfluity, contradiction, redundancy, overanalysis, neologism, incorrect morpheme order, agreement error, affix semantic extension, stem semantic extension, blend, malapropism, anticipation, perseveration, exchange, omission, and stranding.

(14) *Hesitations.* The coding of hesitations will include: retraced false start, word repetition, syllable repetition, drawling, consonant repetition, filled pausing, and incompletions. The types of fillers for filled pauses will be notated, and the durations of pauses will be coded. Overlapping of speakers will be coded directly by comments on the text tier. Latching (additions which follow the rhythm of the partner's speech, allowing no pause between speakers) will be coded on the main text line with the < symbol. The same symbol should be repeated on this line.

(15) *Comments.* General comments can be placed on the comment tier.

In order to map the relation of comment tiers to the text tier, we use numbers in square brackets to mark the words on the text line where the

comment starts. The codings themselves are placed in angle brackets. For example, if the phonetic line refers to words 2, 3, 4, 6, and 7 of the text line, the following coding is used.

**Ned: He wanted to go to the store, Mom.
*Phon: ⟨wahntəd tʌ go⟩ [2–4] ⟨ð stʌwr⟩ [6–7]

Programs

The Centres are currently involved in the acquisition and development of several types of computer program for the System. These programs are designed to facilitate utilization and updating of the database. Currently available packaged programs include SALT (Systematic Analysis of Language Transcripts), OCP (Oxford Concordance Programs), and the HUM UNIX C Routines, The SALT package at the Centre at CMU only runs on an Apple II and cannot be run on the mainframe. Source code is available for OCP and HUM so that they can be elaborated. In addition several Pascal utilities have been written at CMU for the System. The EMACS text editor provides many basic functions in an environment of user-definable functions. Particularly useful is the capacity for recursive calls to the editor which permits the user to edit a group of files interactively. For less interactive editing, the UNIX facilities are appropriate. Together, these programs allow the user to perform these analyses.

(1) *Search.* The most powerful search function is that provided by EMACS. The search string is any UNIX regular search expression. Users will be able to search for specific words, codes with number signs, comments in square brackets, line identifiers, or any combination of any of the above with any type of intervening material. SALT also provides highly general search facilities.

(2) *Frequency tables.* The various packages all perform frequency counts by word or string types. The tables generated can be of these types:

(*a*) *Complete frequency analyses of a text.* The numbers of occurrences of all words can be output in numerical or alphabetical order. Here words can be characterized as strings preceded by spaces.

(*b*) *Frequency analyses by line header type.* Using the codes in the line headers, programs like OCP can generate complete alphabetical and/or numerical frequency analyses for any given speaker or combination of speakers, language or combination of languages, hearer or combination of hearers.

(*c*) *Morpheme-based tables.* When morphemic structure has been coded on the main text line by use of hyphens, it will be possible to generate tables for only stems (items preceded by a space or a number sign), only prefixes (items followed by the number sign), or only suffixes (items preceded by hyphens).

(*d*) *Summary statistics.* The frequency lists can be accompanied by

summary statistics such as type-token ratios, total number of tokens, total number of types.

(3) *Mean length of utterance.* The programs will be able to calculate the number of utterances, the number of morpheme tokens and then return the ratio of morphemes divided by utterances.

(4) *Pause number and percentage.* If pausing has been coded in seconds, the entries can be counted and summed to yield total number and time of pausing.

(5) *Concordances.* All of the packages can create KWIC (Key Word In Context) indices. These are essentially searches in which a specified amount of material before and after the matching string is included in an output file.

(6) *Productivity analysis.* From concordance-like data one can construct productivity analyses such as those proposed by Ingram and Dale.

(7) *Dumps and listings.* Users may often want to dump out the whole file in some new format. For example, they may wish to extract from a dyadic interaction only the lines spoken by the parent. Or they may want to reformat data so that the mother's speech is in a column on the left and the child's in a column on the right. The Centres will attempt to provide general utilities for generating outputs of this type.

(8) *Assisted data entry.* A LISP program written at CMU is now being used to automatize the process of binding coding to tiers and/or adding new tiers for coding. Datatrieve offers a way of linking such coding to a full database structure.

(9) *Automatic tagging.* Systems such as that described by Marshall (1983) and Johansson (1982) for coding part of speech in major corpora such as the Brown and LOB corpora will be implemented for the CHILDES database.

In the long run the programs for analysing the database will be integrated into a single environment. An example design for such an environment is provided by Tom Peters' GRITS. This system will allow the user to move back and forth from searching, to dictionary checking, to advice-seeking from expert programs, to statistical analysis, and printing. GRITS is implemented within UNIX in C.

The value of CHILDES to the field

When fully established, the CHILDES should be able to assist the development of the field of language acquisition research in two major ways. First, the process of systematizing the database for the field will yield a variety of methodological contributions. Secondly, analysis based upon processing of the database should be able to address a wide variety of empirical issues that can lead to the advance of theory construction.

Data processing or methodological contributions. The CHILDES will in effect provide a catalogue of the types of child language data that are available. That catalogue will reveal what types of children, (e.g. middle-class, first-born, 2–3-year-old, standard English speakers) have received the most attention, and what groups, defined by socioeconomics, dialect language, age, birth order, health status, etc., have been underrepresented in the data pool. Thus, new data collection efforts will, by virtue of the availability of stored transcripts, automatically be concentrated in areas where existing data are sparse.

Furthermore, the CHILDES will generate a standard for how data should be collected. By virtue of using transcripts derived from different sources, it will become clear how well or how poorly observations collected in certain ways serve various purposes. Furthermore the fact that data will be collected with the likelihood that they will be shared will ensure better adherence to the highest standards of care and accuracy.

The development and organization of the CHILDES will also focus attention upon fundamental issues in the methodology and/or theory of transcription. Although there are a few widely relied upon guidelines for transcription, (e.g. Bloom & Lahey 1978), the specifics of notation (e.g. layout; treatment of partly incomprehensible utterances; use of ? & ! to indicate intonational vs. syntactic forms; treatment of elided forms such as *dontcha*, *whyntcha* and *wanna*; representation of auditory information other than the child's speech) vary widely, as do decisions regarding matters that are on the border between transcription and coding (e.g. treatment of contractions and compounds as one morpheme or two; amount of nonverbal contextual information provided; notation of gesture which may disambiguate speech act categorization). Absolute uniformity in such matters is neither feasible nor desirable (Ochs 1979), since different aspects of coding and transcription must be elaborated for different purposes, and elaborating all aspects on all transcripts would be unacceptably labour-intensive. Nonetheless, certain decisions can be taken to eliminate unresolvable incompatibilities among different transcription and coding systems, and to promote mutual intelligibility of transcribed material. The importance of working towards the establishment of standards is not unique to language acquisition research. In general, history shows that the introduction of standards has been an important source of progress in both science and technology.

An explicit goal of the CHILDES is to improve coding procedures at the most important tiers of the system. Considerable gains in efficiency and quality of coding will, of course, emerge simply from access to other coding schemes. The best way to understand and to learn to use a reliable coding scheme developed by someone else is to see how it was employed by the original investigator; the sharing of coded transcripts will make this possible,

and will ensure that the most valid and comprehensible systems are borrowed and perhaps improved upon. Without this kind of open access to other researchers' coding schemes, it would appear that students of language acquisition would be condemned to continually reinventing coding schemes every time they begin a new project.

Because of the time-consuming nature of any data processing for child language, the field has tended to be very conservative in its development of methods. For instance, since first used as a developmental index by Brown, Cazden & Bellugi (1968), mean length of utterance (MLU) has been a ubiquitous and multipurpose measure, despite general recognition of its limitations (it fails to discriminate well after about $3\frac{1}{2}$ years, it may not work effectively cross-linguistically as a basis for matching, it is inappropriate for children with certain kinds of language disorder, etc.). The availability of large data sets and automated analysis systems would for the first time enable researchers to consider alternate developmental indices (e.g. number of content words per utterance; a certain level of discourse competency; type-token ratio). For example, matching children on indices other than MLU may be of special importance in assessing the development of language-disordered children, whose language level is often seriously underestimated by MLU.

Contributions of the CHILDES to research and theory-building. In evaluating the potential contribution of the new System to the field, we consulted with several colleagues, asking them how their current research projects could benefit from use of a large quantity of computerized transcript data. The researchers we contacted were Elizabeth Bates, Ursula Bellugi, Lois Bloom, Robin Chapman, Eve Clark, Phillip Dale, Judith Johnston, William Hall, Laurance Leonard, Michael Maratsos, Jon Miller, Ann Peters, Tom Roeper, and Kenneth Wexler. We have used these written responses as a way of characterizing possible uses of this database. Of course, the uses we will mention are meant to be suggestive and are in no way intended as a complete or exclusive listing of possible applications and contributions.

There are four ways in which the availability of the CHILDES can be expected to have a major impact on research. As we will see, some types of research are essentially impossible without some data-sharing system; others are only very difficult and time-consuming.

(*i*) *Research requiring large samples.* The classic child language study since the early sixties has been the case study, sometimes replicated on a few additional children, but essentially remaining at the level of description and hypothesis generation rather than hypothesis testing. As Bloom says in her evaluation of the potential utility of the CHILDES, 'We have always considered these results tentative, awaiting confirmation from studies of other populations of

children'. Now, with the advent of the Data Exchange system, we have the opportunity to extend our analyses with a substantially larger data base that is, moreover, available for computer processing.

A first and major contribution of the CHILDES will be the possibility to test the generalizability of hypotheses generated from case studies. For example, Peters proposes that one could test the hypothesis that two verb constructions (*want/like/have to V*) emerge from utterances with the structure (I–semi-auxiliary–V–NP). In addition, Bloom proposes to test the generalizability of the sequences of development she has found in small samples for complex sentence structures and connective relations, and to test the hypothesis that acquisition of complementation is matrix-verb specific.

Variability and generalizability estimates are the truly valid basis for norms of spoken language development which are sorely needed to support research in various atypical populations as well as to support both research efforts and intervention with language-disordered children. Specific example proposals to use the CHILDES in these ways have come from Clark, who wishes to study affixes and use of compound word forms across a large sample of children; from Johnston, who proposes to compare patterns of predicate use in normal children with her own data on learning-disabled children; from Bellugi, who wishes to derive normative information about the development of expressions of temporal aspects, numerosity, temporal focus, manner and degree for comparison with signing children's development of these notions; and from Maratsos, who is interested in norms and variability in the development of prepositions and of verb structure.

Empirical study of individual differences in style of acquisition has been severely limited by the fact that individual researchers typically have access to data from only a few children. Proposing the existence of a few distinct 'styles' of language acquisition requires finding correlations among a fairly large number of variables. Multivariate research with large numbers of variables requires even larger numbers of subjects. Thus, both Bates and Dale argue that only by pooling transcripts from different investigators can one hope to perform the analyses that will determine whether the apparent relationships among (*a*) reliance on content words, (*b*) overgeneralization, (*c*) consistency, and (*d*) resistance to imitation are in fact real, and whether similar relationships are found among learners of languages other than English.

Limitations on sample size have similarly restricted the generalizability of conclusions about the nature of input to children. Although studies of input have typically involved 10–25 subjects, the power of any input analysis is greatly increased by larger sample size. Analyses of effects such as the match between input complexity and the child's language level require extensive data bases. A large database is also needed to reach reasonable conclusions about the input frequency of certain classes of items of particular interest,

e.g. motion verbs (Chapman & Miller), affixes and components (Clark), and certain kinds of errors and omissions (Roeper).

(*ii*) *Rare events*. The study of large amounts of data is also important in those cases where theories make predictions regarding certain rare events. Events which occur very rarely can be of great importance and interest to child language theorists. Errors of word use can be said to reveal nonstandard semantic analyses; speech errors are said to reveal storage and processing procedures; and lexical class errors are important in the debate regarding the functional bases of language learning. Here it is important to study the relative frequency of certain rare structures in both the input and in the child's productions. Maratsos, Roeper, and Wexler note that events of particular importance for syntactic theory include the possible production of certain nominalizations, subjectless sentences and transitive prepositions without objects that are ungrammatical in English but common in other languages, and the correct use of infrequent constructions such as passives, relatives, mental verbs, and left dislocations.

Large databases are crucial in these analyses precisely because the key issues revolve around productivity, i.e. the possibility that in early use such structures are severely restricted in terms of the verbs that control them, the syntactic context in which they can occur, and the meanings they express. Furthermore, analyses which depend on co-occurrence of two or more items are difficult or impossible with small data bases because the co-occurrence requirements diminish frequency. For example, Roeper proposes to analyze the co-occurrence of specific question types with different classes of verbs and to describe the development of binding rules by looking for co-occurrences of quantifiers and indefinite articles. To do this one must have enough text to find large numbers of questions with each verb type.

The study of rare events allows one to compare the importance of hypothesized learning mechanisms with the importance of interactional or tutorial techniques that may greatly facilitate learning. Hypotheses about the degree of robustness and innate preprogramming in the language aquisition system have until now been generated without any reliance upon data about how often incorrect utterances are heard by children, or about the relative frequency with which various sorts of explicit adult correction of child error occur. Availability of large and disparate samples of adult speech will, therefore, provide for the first time an empirical basis for the development of theories of learnability.

Finally, an extreme case of the rare event which without a very large database has remained intractable to study is the predicted non-occurrence of certain events. As Roeper states, 'linguistic theory claims that many logically possible forms are eliminated by constraints on universal grammar. It will be very interesting to establish that certain forms which are predicted

to not appear do not in fact appear'. An example of such a predicted non-occurrence are ergatives with by-phrases (**the boat sinks by Bill*). Furthermore, much has been made of the non-occurrence in child language of certain errors that would seem to be quite easy to make, e.g. the use of action adjectives with verb endings, such as *the dog was very snoopying* (Maratsos). Search of a large database could confirm that such very important errors do not in fact occur.

(*iii*) *Comparing experimentally generated with spontaneous data.* Roeper notes that 'the history of acquisition research reveals that there is virtually nothing as persuasive as naturalistic data'. However, many structures of interest to the child linguist have until now been studied primarily through laboratory techniques which promote production and assess comprehension. The researchers who have developed and used such experimental techniques all concede the superiority of natural speech data, but have resorted to experiment because they need a large sample of utterances, and because the natural occurrence of the structure of interest is infrequent. The CHILDES will contribute to research by allowing for conclusions from laboratory studies of such structures to be tested against spontaneous speech data. Specific proposals to do such analyses have been presented for pronoun use, passives, mental verbs (Maratsos), and dimensional adjectives (Johnston).

(*iv*) *Crosslinguistic analyses.* Crosslinguistic analysis has long been recognized as a crucial test for theories of child language development, but few researchers have the resources to conduct extensive cross-language studies or analyses. The inclusion in the CHILDES of non-English corpora supplied with English gloss tiers at two levels (morpheme-by-morpheme as well as best 'normal translation') will make the language analyses accessible to even those researchers who have only a beginning knowledge of the language. Virtually any question that can be asked within a single language can be asked cross-linguistically. One may look at phonetic codings of transcript data for evidence regarding the universality of phonological processes, or one may examine data from several languages in an attempt to determine whether the child controls tense before aspect. One may look at the extent to which parents demand imitations from their children or the ways in which they respond to their children's questions. What is interesting in the cross-linguistic comparisons is the degree to which either language structure or cultural differences or both influence any differences that are observed.

CONCLUSION

The formation of CHILDES has been stimulated by the increasing accessibility of data-processing equipment. If this new opportunity for child

language studies is to attain its full potential, it will be important for researchers in all aspects of child language study to become involved in the process of data exchange and the analysis of a shared database. There is much work to be done, and the field stands to benefit quite generally by encouraging this enterprise.

REFERENCES

Bloom, L. & Lahey, M. (1978). *Language development and language disorders*. New York: Wiley.

Bloom, L, Lightbown, P. & Hood, L. (1975). Structure and variation in child language. *MonogrSocResChDevel* **40**. No. 2.

Brown, R. (1973). *A first language: the early stages*. Cambridge, Mass.: Harvard University Press.

Brown, R., Cazden, C. & Bellugi, U. (1968). The child's grammar from I to III. In J. P. Hill (ed.), *Minnesota symposia on child development*. Minneapolis: University of Minnesota Press.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: C.U.P.

Folger, J. & Chapman, R. (1978). A pragmatic analysis of spontaneous imitations. *JChLang* **5**. 25–38.

Francis, W. N. & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Johansson, S. (1982). *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities.

Lehmann, C. (1982). Directions for interlinear morphemic translations. *FolLing* **16**. 119–224.

Marshall, I. (1983). Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. *Computers and the Humanities* **17**. 139–50.

Miller, J. & Chapman, R. (1983). *SALT: Systematic Analysis of Language Transcripts, User's Manual*. Madison: University of Wisconsin-Madison.

Ninio, A. & Wheeler, P. (1984). A manual for classifying verbal communicative acts in mother-infant interaction. *Working Papers in Developmental Psychology: Hebrew University* **1**. No. 1.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (eds), *Developmental pragmatics*. New York: Academic Press.

Slobin, D. I. (1984). *Cross-linguistic studies of language development*. Hillsdale, N. J.: Erlbaum.

APPENDIX

Survey for Child Language Data Exchange System

Name _____

Title/Position _____

Department _____

Institution _____

Street Address _____

City/State/ZipCode _____

Telephone number _____

(1) Please provide a general description of the nature of any child language transcript data that you would be willing to contribute to the System. Please give the size of the corpora, numbers of subjects, ages of subjects, nature of communicative context, method of transcription, etc.

(2) Please describe the nature of the media in which these data are coded, i.e. handwritten, typed, published, or computer-coded.

(3) Are you willing to share these data with other researchers? What conditions of copyright or authorship would you wish to place upon their use?

(4) Are there specific types of analyses that would be appropriate for the data (i.e. phonological, lexical, syntactic, semantic, speech act, discourse, individual differences, etc.)?

(5) What kinds of data would you like to receive from the exchange system? Are you interested in particular types of transcripts? Particular types of codings? etc.

(6) What computer systems do you use? What data formats can you read and write? Do you have access to ARPANET, BITNET, or CSNET? What programs do you use and/or need to process child language data?

(7) In general, do you have suggestions regarding policies and directions for the exchange system?

Please return to: Dr Brian MacWhinney, Child Language Data Exchange System, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA 15213. Computer address: macwhinney@cmu-psy-a. Phone: (412) 578-2656.