J. Child Lang. **19** (1992), 459–471. Printed in Great Britain

NOTES AND DISCUSSION

The wheat and the chaff: or four confusions regarding CHILDES*

BRIAN MACWHINNEY
Carnegie Mellon University

AND

CATHERINE SNOW
Harvard University

(*Received* 29 *July* 1991. *Revised* 27 *August* 1991)

ABSTRACT

Edwards (1992) presents a set of examples from the Child Language Data Exchange System (CHILDES) as prototypes of bad transcription practice. Her discussion is based upon four basic confusions. First, Edwards confuses old and discarded versions of CHAT with current CHAT. Second, she confuses the relation between CHAT standards with the implementation of these standards during the process of reformatting older corpora. Third, she confuses transcription for automatic analysis with transcription for documentation. Fourth, she confuses the CHAT guidelines with the larger CHILDES system. We argue that these confusions have misled Edwards into developing an overly rigid set of principles for data analysis which, if followed literally, could choke off progress in the analysis of spontaneous language samples.

INTRODUCTION

Edwards (1992) has presented a number of principles designed to guide research on archived language data. Like Edwards, we have devoted considerable attention to these issues and are much concerned with the optimal utilization of computerized transcript data. We have no basic quarrel

[*] Address for correspondence: Brian MacWhinney, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA. e-mail: brian@andrew.cmu.edu.

with her list of principles, and we have consistently subscribed to the basic goals she has laid out. These goals involve promoting data accountability and usability without introducing methodological problems or errors. Given our shared goals, it was discouraging to find that the only examples of bad transcription practices cited by Edwards were those deriving from the CHILDES system. If her goal had been to criticize transcription systems in general, one would have expected to see examples of bad practices drawn from many different transcription systems. But this is not what one finds. Rather, Edwards has trained her sights directly on the CHILDES system, ignoring a variety of other potential targets. Though Edwards has denied that her article constitutes an attack on CHILDES (in a message to the Info-CHILDES Electronic Bulletin Board, 17 June 1991), the actual illocutionary force of her comments has been widely interpreted as identical with that of a direct attack.

Over the last several years we have relied extensively on user suggestions as ways of improving the system. These comments have come from dozens of scholars around the world. In earlier critiques, Jane Edwards provided feedback that was uniformly helpful and constructive. Unfortunately, the current article by Edwards does not have this same constructive quality. Instead, it presents a view that could be interpreted in a way that would discourage child language researchers from the prospect of using any computerized archive. To avoid this danger, it is important to clarify how CHILDES in fact lives up to the principles articulated by Edwards, and to document the specific confusions that led her to her incorrect and counter-productive formulations.

It is crucial to point out in advance that CHILDES is much more than simply a data archive. CHILDES is, in fact, a system designed with the following goals:

(1) to enable child language researchers to access others' transcript data,

(2) to provide guidelines for transcription and coding of newly collected corpora, and

(3) to provide a package of computational tools for the analysis of transcript data.

In order to meet these various goals, the activities of CHILDES over the last several years have focused on three processes.

(1) *CHAT standardization.* In the Fall of 1990 we completed the process of developing, testing, and improving guidelines for transcription and coding, referred to as CHAT. The current CHAT guidelines have been published in a book we will refer to here as The Manual (MacWhinney, 1991).

(2) *Database improvement.* We are still engaged in the process of upgrading donated datasets to match the CHAT guidelines.

(3) *CLAN development*. We are also still engaged in the process of writing and refining the CLAN programs (also documented in the Manual) for analysis of CHAT-formatted data.

In assessing Edwards' criticisms of CHILDES, it is important to distinguish those comments which relate to the process of CHAT standardization from those comments which relate to the process of database improvement. Although a fully upgraded archived database is an important ultimate goal, we feel that criticisms aimed at the basic structure of CHAT are the ones that deserve the most careful consideration by potential CHILDES users.

We believe that Edwards' critique contains a mix of wheat and chaff. It will be our task to winnow out the chaff, so that we can focus on the wheat. Many of the claims that Edwards makes regarding CHAT and CHILDES are now simply wrong, because they are based on old versions of the coding system and old versions of the database. Some of her statements regarding general principles make good common sense. However, the conclusions drawn from these general principles are often rigid and unproductive. We will attempt in our comments to point out some basic confusions that may have been the source of the various misrepresentations found in Edwards' article.

Confusion 1: *proto-CHAT and CHAT* 1.0

The most serious problem in Edwards' critique of CHAT is the fact that it is seriously out of date. Edwards' criticisms only make sense if one turns back the clock two, three or even four years to a time when the CHAT system was a set of loose proposals being tested out in pilot studies. This was the time of proto-CHAT. It was also a time when, as a researcher at the Max-Planck Institute in Nijmegen, Edwards was directly involved in the CHILDES Project. In an article circulated in 1986, she made a series of useful suggestions. We publicly thanked her for her suggestions and used many of them for changes which we made in CHAT during 1988. Because the CHAT system is tightly linked to a set of programs, and because changes in CHAT must eventually be reflected by changes in the codes in the database, we could not implement each suggestion immediately. However, over time, we brought the system into correspondence with most of her suggestions, as well as those we received from dozens of other researchers. During the period from 1986 to 1990, we kept Edwards continually informed through electronic mail and new versions of the Manual regarding the implementations of these changes.

In early 1991, we published the Manual (MacWhinney, 1991) which contained the current form of the CHAT system, along with a manual for the CLAN programs and the CHILDES database. In the version of CHAT and CLAN published in the manual, all of her objections regarding the CHAT system were answered. This current form of CHAT called CHAT 1.0 was

sent to her in late 1990. Unfortunately, Edwards has chosen to ignore the published CHAT 1.0 version and to criticize instead the unpublished and now entirely irrelevant drafts of proto-CHAT from two, three and even four years ago.

Let us go through each of Edwards' criticisms of CHAT in her section on readability and minimal bias. In each case, what is involved is some flaw in proto-CHAT which was corrected during the period from 1988 to 1990.

(1) *Prosodics.* Edwards claims that CHAT only permits transcription of prosody on a separate line. This is not true; it was only true in proto-CHAT. Chapters 6 and 7 of the Manual present a system for annotating prosodics and contour on the main line. This system of prosodic coding in CHAT has now been fully integrated with the CLAN programs and is easily extensible if a researcher so desires.

(2) *Comments on the main line.* Edwards claims that comments cannot be coded on the main line in CHAT. Again, this was only true in proto-CHAT. Chapter 8 (p. 49) of the CHAT Manual provides an explicit way of adding comments directly to the main line, just as Edwards requests.

(3) *Metacomments in codes.* Edwards criticizes forms like 'mm/thinking@q' both for readability and for the insertion of 'thinking' as a metacomment. This ignores the fact that such forms have been absent from both CHAT and the database since the end of the era of proto-CHAT.

(4) *Marking omitted words.* Edwards inveighs against marking omitted words with the ∅ symbol. In fact, this symbol virtually never appears in the CHILDES database. Pages 26 and 36 of the CHAT Manual issue dire warnings against the use of these symbols. Chapter 14 (p. 96) of the Manual presents a more rigorous way of coding omissions on the %mor line, but Edwards fails to mention this superior CHAT alternative.

(5) *Lexical commitments.* Edwards objects to the coding of the reduced form ''em' as '(th)em', claiming that if forces the coder to make a commitment to a particular lexical item. Her criticism ignores the fact that CHAT 1.0 provides the option of avoiding a commitment to a particular lexical item by coding this sequence as '&em'.

(6) *Letters.* Edwards objects to the use of '@l' as a marker for letters in CHAT, suggesting that this is the only mechanism for spelling letters in CHAT. She ignores the fact that section 4.5.1. (p. 28) of the Manual presents a system for spelling out letters that is even more readable and reliable than the alternative she suggests. The CHAT convention that is suggested is simply to spell 'b' as 'b'.

(7) *Overlap markers.* Edwards objects to the use of forms such as '[overlap above]'. But this form disappeared with proto-CHAT. CHAT 1.0 uses the nonobjectionable form [<] to point in the direction of the overlap.

(8) *Glossing in codes.* Throughout her paper Edwards objects to the now-defunct 'ya/you@e' type of notation. However, like the 'mm/thinking@q'

notation discussed above, this notation was purged long ago. In its place there is the notation 'ya [: you]' which allows for greater separation and readability, just as Edwards requests. Equally importantly, this new replacement notation links up directly to the CLAN programs and allows one either to make the replacement in lexical searches or to suppress the replacement.

Perhaps we are missing some kernels of truth mixed in with the chaff. On the level of general principles, Edwards argues that human readability is a primary goal in formulating transcript conventions. This goal sounds good enough. Certainly, we always want to do whatever we can to maximize readability. But exactly what is readability and what do we do when readability conflicts with other goals? Edwards treats readability as a single uniform dimension that is somehow constant across transcripts and human readers. But reading is a dynamic process involving an interaction between the researcher and the transcript. Moreover, this process can be facilitated on the computer by changes in the display. Just as the new generation of statistical packages on microcomputers promotes exploratory analysis of a data set, the CLAN programs also promote exploratory variations in the displays of CHAT transcripts. The programs called COLUMNS, GEM, FLO, LINES, and SLIDE are all designed to allow the user to view transcripts in different ways to facilitate creative views of the data. These programs and others that are being planned all rely on the fact that CHAT is well defined computationally and supports a wide variety of display options.

In the area of transcripts and readability it is surely the case that one man's meat is another man's poison. To researchers used to transcript formats that provide a column for each speaker, the CHAT system is initially difficult to read. On the other hand, long-time CHAT users find the column format difficult. Edwards finds the Gumperz-Berenz style of transcript highly readable (see her example, p. 440), but this is surely because she has had considerable practice with these particular conventions. In the end, disputes about the exact shape of transcript displays are much like disputes over the quality of wines. A gracious host may provide a variety of different bottles of wine in the hope of pleasing each guest. With transcript formats we can do even better – we can write a program to allow each researcher to see the data in a different way. We have written several programs that do this and we are willing to write more. If a group of researchers wants to view a CHAT transcript in a particular way, they can define that format and the CHILDES system can implement a program to display CHAT in that form. Doing this allows us to escape from the Latin adage *de gustibus non disputandum est* by invoking the computational principle: 'different files for different styles'.

Confusion 2: the status of the database

In her second section, Edwards trains her sights on alleged errors in the CHILDES database. Criticisms of the shape of the database are potentially very important, but they must be stated in quite precise terms. Unfortunately, Edwards' critique confuses two different issues: (i) Does the CHAT coding system introduce fundamental inconsistencies that will impede data retrieval? And (ii) Are there inconsistencies in the way particular forms are transcribed in the CHILDES database? Our answer to the first question is, 'No'. Our answer to the second question is, 'Of course, there are'. What Edwards does is to use admitted inconsistencies in older corpora in the database to somehow argue against aspects of proto-CHAT. Here, again, we need some history.

The CHILDES database is a collection of transcripts derived mostly from projects that were completed long before the finalization of the CHAT coding system. Before the emergence of CHAT coding, there was no coding standard for the field of child language. In fact, if one looks across the original versions of the 50 corpora in the CHILDES database, one will quickly realize that no computer program could be devised that would be able to process so many differing forms of transcription. At this point, we and others in the field simply thought of the CHILDES data as an archive. However, we soon realized that certain basic transformations on the formatting of the data would allow one to ask simple questions across data sets. Over time, we have brought the database into closer and closer conformity with CHAT 1.0. Most recently, we have written a program called CHECK which runs through a CHILDES file to guarantee that it accurately subscribes to the CHAT 1.0 coding standards. CHECK is a precise instrument which detects deviations from CHAT format as well as many mistypings, miscodings, and incorrect usages. The Manual also provides guidelines for using FREQ to examine a transcript for lexical anomalies and typographical errors. We believe that the use of CHECK along with other recommended procedures for checking transcripts allows us now to attain a much higher standard of data accountability.

Detecting errors with CHECK is quite fast, but correcting those errors is a very laborious process. Since 1988, two research assistants have been working full time at Carnegie Mellon to bring the CHILDES database into increasingly close accord with CHAT 1.0. Their checking has gone beyond the filter provided by CHECK to include a general upgrading of proto-CHAT codes. We do not believe that Edwards has accorded proper recognition to the importance of their careful work. In particular, this work has made the two major criticisms in the second section of Edwards' paper already obsolete.

(1) *Special learner form markers.* Our extensive consistency checks on both the Brown and MacWhinney corpora have made the analysis that Edwards provides in Table 1 irrelevant. Symbols such as @e, @h, @m, and @q are no longer a part of CHAT coding as it is formulated in the CHILDES Manual (MacWhinney, 1991, p. 23), nor are these symbols present in these corpora. We have recently completed removal of these symbols from all of the corpora.

(2) *Pronominal forms.* Edwards claims that there are nine variants of *you* in the MacWhinney corpus. Long ago, we attached a specific warning to the documentation file for the MacWhinney corpus which notes that neither transcription nor checking for this corpus is complete. This warning is repeated on page 253 of the Manual. As checking of the MacWhinney corpus has advanced over the last year, the details of Edwards' critique of this corpus have become largely out of date. For example, the removal of proto-CHAT forms for the MacWhinney corpus has left only *ya* and *you*. These are the two forms of the pronoun which are codified in Chapter 4 (p. 34) of the Manual. In addition, Chapter 4 (p. 31) of the Manual presents further codifications of common pronoun–auxiliary assimilations.

These observations raise a general issue. Any researcher interested in working with lexical forms in the CHILDES database needs to read the 14 pages in Chapter 4 of the Manual quite carefully. This warning applies particularly to anyone interested in studying pronouns in English, since they are involved in so many assimilations. Eventually, we hope to be able to introduce replacement forms such as 'dya [: did you]' throughout the database. However, for the moment, it is important to be aware that pronouns take on the various forms specified in Chapter 4.

Although Edwards' observations regarding the Brown corpus are now of only historical interest, some of the errors she found in the MacWhinney corpus were still there when we received her article. For the record, these were the remaining errors.

(1) The form 'hmm [:thinking]' should have been 'hmm [= thinking]'. This error occurred five times in the file boys61.cha, but nowhere else in the corpus. We were unable to find any other similar errors. We have corrected the errors.

(2) We located the form '&ei(ther)' in the file boys74.cha in line 1638. This was a coding error. We found 12 further errors of this type in the 'boys' files, but none in the 'ross' files. We have corrected this error and have changed the CHECK program to catch such errors in the future.

Until the entire database is brought into complete correspondence with CHAT 1.0, it is important to remind users of potential inconsistencies. Indeed, child language researchers in general have shown themselves to be generous and helpful in pointing out problems in corpora they are using, to everyone's mutual benefit. We have developed a mechanism for the fast and

efficient dissemination of such information using the info-CHILDES electronic bulletin board (Info-CHILDES@andrew.cmu.edu). Notices of problems with data files or with CLAN programs can be posted to this board. This will allow the staff responsible for correcting errors to make improvements quickly and other CHILDES users will benefit from rapid correction of errors.

Yes, consistency is important. Striving toward greater consistency is a constant goal of the CHILDES Project. But let us not overestimate the impact of a small number of inconsistencies in the database. Edwards has located a few inconsistencies scattered about a collection of 140 million characters. She then goes on to suggest that these minor inconsistencies could, by themselves, lead to incorrect conclusions in particular empirical studies. However, this form of argumentation confuses the notion of numerical inaccuracy with the notion of incorrectness of conclusions. If a study of, say, morality words misses two uses of the word *good* because of some miscoding or misspelling, the actual empirical report of the numbers of these forms used in particular corpora will be inaccurate by some small percentage. However, an error in exact percentages is unlikely to be so great that it will lead to incorrect conclusions. If Edwards believes that these errors are indeed so frequent that they can actually lead to wrong conclusions, she will have to demonstrate this through concrete examples in several major corpora. If she could do this for several important domains, she would be making an important contribution to our understanding of corpus analysis.

It is misleading to claim that the presence of even one inconsistency is enough to destroy our ability to form any sort of empirical generalization from transcripts. Doing this enforces a level of data accuracy that may be correct for measurements in quantum mechanics, but which is quite out of place for child language research. Until we have complete transcripts for all the speech spoken to a child and by a child over the complete course of language development, we will have to accept some level of uncertainty in our sampling and measurement. Even attaining an absolutely complete record for one child would only be a beginning. We would need eventually to have compete records for scores of children from scores of different sociolinguistic environments. This is a reasonable goal for the CHILDES system in the next century, but it is an unreasonable requirement for child language research in 1992. Insistence on artificially high and unattainable standards for data perfection can only serve to choke off creative research. If, on the other hand, some large collection of data transcribed in a way that was demonstrably more consistent than the CHILDES database were made publicly available to the child language community, we would be the first to rejoice.

Edwards' comments on these matters are not all chaff, but they must be winnowed carefully. The issue of data consistency is a deep one that will be

with us for a long time. It will never be the case that corpora collected in different laboratories, using different protocols, by different researchers asking different research questions will be totally comparable with one another. Transcripts can be brought up to certain standards of consistency concerning spelling of words, marking of morphemes, and indications of variants, but they cannot be made consistent after-the-fact in their attention to various sorts of detail (articulation, prosody, pausing, voice quality, etc.). Thus, attention to consistency must go hand in hand with documentation for areas in which consistency cannot be obtained. Corpora in the CHILDES system are accompanied by 'warnings on the label' telling potential users what aspects of the language system received most careful attention by the original transcribers, as well as the sorts of analyses that would not be supported by the transcripts. Given the great inconsistency in human affairs, it is not clear that the data collector or archiver can do any more than this. It is the end-user who must decide what can and what cannot be compared. Thus, Edwards' fourth goal, comparability of datasets, is one which places responsibility primarily on researchers selecting corpora to compare, rather than on the original researcher or the archiver. This is an important kernel of wheat.

Confusion 3: *automatic analysis vs. documentation*

In section 3 of her paper, Edwards expresses concerns regarding the looseness with which certain CHAT codes are applied to the database. These worries tend to grow from the same confusion regarding the status of the database that was prevalent throughout Edwards' section 2. In particular, she worries about looseness in the use of three types of codes.

(1) *The comma* Edwards correctly notes that the use of the comma in the Kuczaj corpus does not correspond to its uses in either CHAT or other corpora in the database. Here Edwards has harvested a grain of truth. Sokolov & MacWhinney (1990) ran into this problem when conducting MLU and CHIP analyses on the Kuczaj corpus. We have discussed this matter with Kuczaj and he plans at some point to go through his data to clarify the ways in which the comma is being used. We have posted a note on this matter to the info-CHILDES bulletin board.

(2) *Special form markers* Here we find less grain and more chaff. As we noted above, the offending special form markers were removed years ago and it is not productive to continue to discuss codes that are of only historical interest.

(3) *Coding tiers* Here we see most clearly that Edwards has confused coding for automatic data analysis and coding for documentation. In Chapter 9 of the Manual, we explain that the system for naming dependent tiers is intended to be open-ended. There are two types of dependent tiers. Tiers such as the '%err', '%pho', and '%mor' line have been tightly specified and

can serve as the basis for automatic analysis. Tiers such as '%gpx' or '%int' are intended simply to provide a loose framework for documentation of interesting aspects of a transcript. As long as a researcher understands the difference between coding for automatic analysis and coding for documentation, we see no danger in allowing for project-specific coding to be determined by the needs of the project. In regard to the older data sets in CHILDES, almost none of the dependent tiers can yet be used as the basis for automatic analysis. This is not a surprising fact and we have never encountered a researcher who had somehow assumed that the codings on the '%gpx' tier were rigorous and exact. This is not to say that some rigorous way of describing proxemics cannot eventually be found. However, until it is, insisting on rigour where none was imagined seems counter-productive.

Confusion 4: misjudging the role of computational technology

Much of the confusion in the Edwards critique involves her failure to keep up to date on past developments in the coding system and the database. There is little evidence that she correctly understands the computational tools being developed by CHILDES. Let us consider four particular ways in which the availability of particular computational tools affects Edwards' analysis.

(1) *Data display*. Edwards fails to understand the extent to which the CLAN programs provide tools for changing the display of CHAT files to suit various viewing needs and preferences. Citing Ochs (1979), Edwards points to the importance of being able to represent data in columns with the child's utterances on the left. Within CLAN the COLUMNS program allows users to output files with a column for the child at either the left margin or some other position. In effect, this program converts a theoretical issue regarding transcription to a simple selection of a computer program. Here, again, we can think of this in terms of the principle of 'different files for different styles'. A CHAT file can be displayed in a novel way that makes the representation of overlaps entirely veridical. This is done with the SLIDE program which converts a CHAT file into a set of single lines for each speaker which can be moved across the computer screen from left to right. At any point in time, only 80 columns are displayed, but the user can rapidly move to any other point in this single left-right line by using the cursor keys. When two speakers overlap in a conversation, SLIDE displays the overlapped portions on top of each other. SLIDE can also be used to get accurate placement of material otherwise indicated by '<aft>' and '<bef>' and to provide correct placement of morphemes on a '%mor' line with corresponding words on the main line as required for interlinear morphemicization. This form of display provides far better time-space iconicity than any previous form of display used in our field. Of course, this display cannot be captured on the printed page; it is only available on the computer screen with

its capacity to scroll almost limitlessly left to right. An earlier non-computerized prototype for SLIDE can be found in Ervin-Tripp (1979). Many other forms of display variation are available in CLAN, including the suppression of 'unwanted' codes by the FLO program and the exclusion of specified tiers by the KWAL program.

(2) *Standard Generalized Mark-up Language.* Edwards cites with approval the development of a new standard for transcription within the framework of the Text Encoding Initiative (TEI) of the Association for Computational Linguistics (ACL). We have also co-operated with that initiative and consider it important to continue to seek better standards. However, the underlying syntax of the TEI is Standard Generalized Mark-up Language or SGML. This system requires that all fields be marked by paired delimiters and annotations for data types. The resultant text structures contain nested expressions of the type found in programming languages such as LISP. Readability is certainly not a virtue of SGML. Given this, it is surprising to find Edwards looking to SGML as the core of a transcription system that can satisfy her four basic principles. Once a full SGML has been formulated, we will certainly want to write a program to output a SGML file from a CHAT file; there are no conceptual barriers to such a translation.

(3) *Digitized sound.* We have always thought of transcription as a fairly imprecise, albeit important, business. In the end, we all want to know what the child really said and neither CHAT, nor SGML, nor the unpublished Edwards Transcription System can fully satisfy this need. This is why the CHILDES database of the future will be linked to digitized auditory records with links between the transcript and segments of the digitized playlist. We have already begun discussion through the info-CHILDES bulletin board of ways to develop standards for digitized speech using either DAT or optical erasable cartridges. Eventually, development of this technology will make the choices such as the one between '(th)em' and '&em' less pivotal to the development of language acquisition theory. Development of this technology will also support the creation of a complete '%pho' line particularly for transcripts from our youngest children. As Peters, Fahn, Glover, Harley, Swayer & Shimura (1990) have argued, the inclusion of a complete CHAT '%pho' line does much to guarantee real data accountability, particularly at the youngest ages.

(4) *CLAN.* It is not clear that Edwards understands how CHAT is linked to CLAN. The Edwards Transcription System provides the user with no computational tools for automatic analysis. The CHAT system, on the other hand, is linked to CLAN programs that can automatically compute frequencies, as well as the MLU, TTR and DSS indices. The CLAN programs include CHAINS for computing discourse profiles, COMBO for Boolean searches, COOCCUR for co-occurrence analysis, CHIP for the analysis of conversational interaction and imitation, or MODREP for item-replica

analysis and cross-tier matching for morphosyntactic analysis. Each of these programs depends closely on particular aspects of CHAT. In her critique of CHAT, Edwards has ignored all these linkages between CHAT and CLAN. However, these links are crucial to many child language researchers, who are seeking not just well-transcribed data, but also the tools to analyse these data automatically. Edwards fails to discuss these crucial linkages between CHAT and CLAN and presents no alternative tools for analysing data within the Edwards Transcription System.

There would be many serious problems in implementing a set of programs for the Edwards Transcription System. For example, searches for morphemes would have to filter out the word-internal prosodic markets Edwards recommends. Doing this within CLAN was a major computational challenge. Solving any single challenge of this type may not be difficult, but addressing all of the computational issues raised in the context of the Edwards Transcription System could be a very tough job. Until Edwards has fully articulated her coding and transcription system, applied it to major data sets, devised a set of programs that work with her system, and shared these tools with the entire research community, it will be impossible to evaluate her proposals.

In her final section, Edwards presents four summary conclusions, each related to one of the principles she endorses. Unfortunately, her conclusions are much more constraining than simple adherence to the principles would require, perhaps because her use of proto-CHAT as a straw man in elucidating those principles has misled her into assuming an unnecessarily rigid position. Rather than simply endorsing the value of readability, she attempts to impose her own personal preferences concerning transcript format on the field as a whole. She seems to claim that empirical generalizations require perfect datasets – a demand that is out of touch with the realities of empirical work in child language or any other field, and which ignores the sampling and testing procedures designed to tolerate small amounts of error. Her hopes for a field-wide standard are laudable, but in the meantime she seems to imply that individual researchers are not competent to assess the meaning of the data they are working with, and will be easily misled into superficial analyses. In contrast, CHAT and CLAN have been developed in an attempt to offer researchers a toolkit of possible coding systems and analytic strategies, rather than a set of prescriptions for coding or analysis, because ultimately, of course, researchers are themselves responsible for the intellectual and methodological rigour of their own work. Finally, Edwards' warning that data from different datasets should not be combined unless totally comparable would seem to impose quite an unnecessary level of pseudo-rigour on the researcher. It makes sense to urge researchers to pay attention to certain crucial criteria when deciding to include a particular dataset in their analysis and to be aware of possible

sources of noncomparability. But this is quite different from excluding at the outset any sort of cross-project comparison, even of extremely robust phenomena like lexical choice or morphological markings.

We believe there is some wheat in Edwards' analysis. If the chaff is winnowed out, the valuable grains that are left might be usefully sown so that the field can harvest a crop of fresh ideas. The field of child language will benefit more from Edwards' development of her four basic principles and empirical verification of their limits than it will from her development of critical treatments of already superseded transcription systems and already corrected versions of particular data sets.

REFERENCES

Edwards, J. (1992). Computer methods in child language research: four principles for the archived data. *Journal of Child Language* **19**, 435–58.

Ervin-Tripp, S. (1979). Children's verbal turn-taking. In E. Ochs & B. Schieffelin (eds), *Developmental pragmatics*. New York: Academic Press.

MacWhinney, B. (1991). *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

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

Peters, A., Fahn, R., Glover, G., Harley, H., Sawyer, M. & Shimura, A. (1990). Keeping close to the data: a two-tier computer-coding schema for the analysis of morphological development. Unpublished manuscript, University of Hawaii, Honolulu.

Sokolov, J. & MacWhinney, B. (1990). The CHIP framework: automatic coding and analysis of parent–child conversational interaction. *Behavioral Research Methods, Instruments, and Computers* **22**, 151–61.