The Child Language Data Exchange System in research on second language acquisition

Authors: William Rutherford, Margaret Thomas

Persistent link: http://hdl.handle.net/2345/3228

This work is posted on eScholarship@BC, Boston College University Libraries.


These materials are made available for use in research, teaching and private study, pursuant to U.S. Copyright Law. The user must assume full responsibility for any use of the materials, including but not limited to, infringement of copyright and publication rights of reproduced materials. Any materials used for academic research or otherwise should be fully credited with the source. The publisher or original authors may retain copyright to the materials.
Review article

The Child Language Data Exchange System in research on second language acquisition


With the appearance of this third edition, Brian MacWhinney’s introduction to the CHILDES project now fills two separate volumes with an accompanying CD-ROM. Our purpose in this review article is first to provide a brief overview of the Child Language Data Exchange System (CHILDES), and to point out its latest features as displayed in the third edition of the users’ manual (cf. MacWhinney 1991, 1995). We then discuss ongoing efforts to exploit the potential of the CHILDES project in research on second language acquisition, in which field it is our sense that the computational tools developed for CHILDES have so far not been adequately put to use. The CHILDES project could, and should, be adapted to the needs of second language acquisition researchers, especially where such research demands recourse to massive quantities of L2 data and sophisticated coding mechanisms.

I The CHILDES project
The CHILDES project was begun in the early 1980s and has been supported over the years by several funding sources based in the United States: the MacArthur Foundation, the National Science Foundation, and the National Institutes of Health. Brian MacWhinney, along with Catherine Snow, are the principal developers of CHILDES, but it has been from early on an international, interdisciplinary, and collaborative venture involving many people. The project aims to make accessible to researchers on child language acquisition a very large (now approximately 300 megabytes) and diversified database of child speech, which has been transcribed and coded in a standardized format. That coding system renders the data searchable, by computer, using a conventionalized set of commands.

In MacWhinney’s (2000: 4–5) words, three ‘coarticulated’ tools together support CHILDES, like the legs of a three-legged stool. First, there is the database, made up of many hours of raw data contributed by many researchers. The prototypical source of CHILDES data is an audiotaped spontaneous exchange between an adult caretaker and a child. In the 20 years since CHILDES was initiated, the database has expanded to include data from language learners of various ages, learning various languages, recorded in various social contexts. Some corpora derive from observational studies, some experimental; some are longitudinal, some cross-sectional; some involve normally-developing learners, some exceptional learners; some learners are acquiring oral languages, some signed. The tapes which make up the database have been transcribed using the second CHILDES tool, a set of transcription and coding conventions known as CHAT (‘Codes for Human Analysis of Transcripts’). CHAT provides a standardized system for converting recorded speech into typed text, including conventions for analyzing and labeling the phonetic, prosodic, morphological, and syntactic features of speech as perceived by the transcriber. CHAT also provides a standardized system of coding salient information about the context of the speech and the behavior or attributes of the interlocutors. The third CHILDES tool is a broad array of computer programs, collectively known as CLAN (‘Computerized Language Analysis’) programs. Using CLAN programs a researcher can, for example, search the coded transcripts for specific linguistic or conversational features, or calculate the frequency
of co-occurrence of one feature with another feature. CLAN then displays the results of those operations, in a range of customizable ways depending on the goals of the researcher.

The CHILDES system thus opens up a vast storehouse of data—far more than any one person could gather in a lifetime—and provides a means of efficiently bringing those data to bear on specific research questions. For scholars whose object of interest is (a) addressed in the content of the collected data; and (b) can be identified within the CHAT codes; and (c) can be measured (or otherwise manipulated) through the CLAN programs, the conception of CHILDES was an act of visionary creativity. Its implementation brought a significant change to their work. Linguists, psychologists, clinicians, and educators have all used CHILDES to the great advantage of research in their fields.1

II MacWhinney (2000)

MacWhinney (2000) is the third edition of a how-to manual for users of CHILDES. Its intended readership includes novices, for whom it provides patient, clear, step-by-step instructions. There are, for example, exercises designed to lead beginners through elementary CLAN analyses (Volume I, Part 2, pp. 144-153).2 The text is also a relatively user-friendly reference manual for readers already experienced with CHILDES, and for researchers preparing transcripts to be included in the database. It contains, for example, advice about how to do audio and video digitizing (Volume I, Part 1, pp. 186-191), and guidelines for resolving CHAT coding dilemmas (distributed throughout Volume I, Part 1; invaluable for researchers who are transcribing new corpora). In the discussion of MacWhinney (2000) which follows, we assume some familiarity with earlier editions of the manual, and highlight what is new in this latest version.

The third edition of MacWhinney’s manual displays in a number of ways the growth of CHILDES since the project’s inception. Most obviously, MacWhinney (2000) is more than twice as long as the first edition, MacWhinney (1991). The third edition is formally separated
into three parts, each dedicated to one of the constitutive tools of CHILDES. The first two parts are bound together as Volume I.3 After a short introduction to the whole project, Part 1 of Volume I addresses CHAT, the transcription and coding system. This section preserves the order and content of earlier editions of the manual, except that there is generally more detail, representing the on-going development of CHAT. New sub-sections provide information about the integration of two alternative transcription and coding systems into CHILDES. One is called ‘Conversational Analysis’, and is designed for the study of conversational turn-taking. The other is the ‘Berkeley Transcription System’, created to analyze signed languages. CLAN programs have been modified to operate on both. In addition, the third edition of the manual moves into Part 1 a section called ‘Word Lists’, which had been buried inside discussion of CLAN in the first and second editions of the manual. Its purpose seems to be to establish a uniform terminology for identifying particular English words. MacWhinney has also revised references to certain features of CHAT which have now been abandoned or backgrounded. Part 1 of Volume I ends with a list of references, an alphabetic index, and a useful separate index of symbols used in CHAT transcriptions.

Part 2 of Volume I addresses CLAN, the computer program which analyzes transcripts encoded in CHAT. Compared to the first edition of the manual, there is a large jump in the number of commands which CLAN makes available, from 23 to 38 (second edition: 34). Presentation of the individual commands has been augmented, and in some cases reorganized. CLAN commands can be customized by the use of optional switches; in MacWhinney’s third edition, discussion of switches unique to each command has been separated from discussion of switches which operate over CLAN commands in general. As in Part 1, certain abandoned features of CHILDES have been removed and new features added. CLAN now runs on either Macintosh or Windows machines, instead of the earlier MS-DOS operating system, so discussion of DOS has therefore been dropped. A significant enhancement of the new version of CLAN is that it can link certain transcripts to digitized audio or video files. Using these new ‘Sonic Mode’ and ‘Video Mode’ capacities, a researcher can not only view a coded transcript, but listen
to (and in some cases, view) its recorded version. Part 2 of Volume I ends with an index encompassing general features of CLAN programs, the individual commands, and each of the switches.

Elaboration and refinement of the capacities of CHAT and CLAN are thus evident throughout Volume I. But it is in Volume II that the growth of the CHILDES project becomes most evident. This portion of the text addresses the database component, enumerating and providing background information about each of the corpora which have been contributed to CHILDES and made accessible to analysis via CLAN. What covered not quite 100 pages in MacWhinney (1991) now stretches over more than 400 pages, representing 134 corpora, in 24 languages. New data have been contributed pretty much across the board since 1995, augmenting established categories of data and initiating new categories. There are double the number of corpora in Spanish; six new ones in English; and entries for the first time in Cantonese, Estonian, Irish, Japanese, Mambila, Mandarin, and Polish. There has also been significant growth in bilingual and clinical CHILDES data, the latter category covering data from children and adults with diverse impairments, including autism, Specific Language Impairment, Down Syndrome, and aphasia. The business of Volume II is to document the background of each of these corpora, to whatever extent information is available. In some cases, MacWhinney provides extensive information about the provenance of the data, the participants, the circumstances of collection, the socio-cultural context, and any idiosyncrasies of transcription. In other cases, background information is sparse. Volume II closes with an index listing the corpora by name and language.

Accompanying MacWhinney (2000) is a CD-ROM which contains the entire CHILDES database to date plus the documentation contained in Volume II, the CHAT manual (Volume I, Part 1), the CLAN manual (Volume I, Part 2), and the CLAN programs (in both Macintosh and Windows). In addition, the CD includes a copy of a 30,000-item bibliography on child language acquisition, a revision of the 1991 CHILDES / BIB compiled by Roy Higginson and MacWhinney.5 For the very computer-comfortable, this CD may be all that is needed to gain
entry into CHILDES; for others, the CD is still essential because it contains the database and
CLAN programs. Yet nothing electronic replaces the sensory pleasure of new books, and that
pleasure is enhanced by the careful design and generally high production standards of
MacWhinney (2000).6

This then is the status of the CHILDES project, as it now stands. The project was
originally constituted for research on child first-language acquisition, but its tools have now been
applied to at least one adjunct field, the study of language disorders. Could CHILDES also serve
the study of adult second language acquisition? Volume II of third edition of the manual
includes a sub-section entitled ‘Bilingual Corpora’. Thirteen of the 15 corpora it documents
address child language learning. Of the remaining two corpora, one is a study of 16-year-old
learners of French, and the other of 34 Greek-speaking learners of French, aged 14 to 29,
described as ‘early bilinguals’ (p. 145). It is notable that the first edition of the manual contained
including four corpora. Two of those, both focused on elementary-school aged bilinguals, have
been reassigned to the Bilingual corpora sub-section in MacWhinney (2000). Of the other two,
one was also a study of childhood bilingualism, and the other the massive European Science
Foundation’s project on naturalistic L2 acquisition by adult immigrants.7 No special L2 section
appears in either the second or the third editions of the manual. This absence is rather startling,
given that utilization of CHILDES for L2 data handling was touted almost a decade ago in
Eubank and Earley (1992: 451): ‘. . . more child as well as adult L2 data are finding their way
into the . . . CHILDES archive, thereby permitting widespread access to the data.’ More
astounding is a statement made eight years later by Perdue (2000: 300), that ‘One spin-off from
the [European Science Foundation] project was the computerization of a sizable ESF Second
Language Data Bank, which continues to provide material for analysis. . . and which is now part
of the CHILDES database (MacWhinney, 1991)’. This is simply false. Clearly, the best of
intentions notwithstanding, the ESF data were never integrated into the CHILDES database, nor
were any other data from adult L2 learners. Global, complete non-utilization of CHILDES for
purposes of L2 encoding, storage, and retrieval is nowhere better illustrated than in the edited
volume, *Learner English on Computer* (Granger, 1998). Granger presents an overview of ‘the
new and rapidly expanding field of corpus-based research into learner language’ without making
any reference at all to CHILDES. As of the time of publication of MacWhinney (2000) then, the
potential of CHILDES to provide public access to massive amounts of data, keyed to convenient
and flexible analytic tools, has remained unrealized within research on L2 acquisition. Given the
apparent ambitious plans early in the last decade for CHILDES to become a repository for L2
learner data, one is therefore left to wonder what has happened.

### III The CHILDES project and research on second language acquisition

Coincidentally, the third edition of CHILDES went into production at the same time as another
project was getting underway for adaptation of the CHILDES computational tools in the creation
of an analogous L2 database, an enterprise that MacWhinney himself has supported from the
start and with whose leaders he has since been in frequent contact. The database project in
question is ACSL (Annotated Corpora of Second-Language Learners), initiated by William
Rutherford and Maria Luisa Zubizarreta in the Department of Linguistics, University of Southern
California, and designed to serve the field of Foreign/Second Language Acquisition as well as of
classroom research.8 ACSL, with its utilization of the CHILDES computational tools, is planned
as the eventual repository of coded spoken and written data for any number of target languages
produced by learners of any native language. In its current very early stage of development,
however, ACSL consists of a small but expanding computerized corpus of coded L2 English
written essays produced by two groups of native speakers: Romance (Spanish, Italian, French,
Portuguese, Catalan) and East Asian (Mandarin, Japanese, Korean).9 At this juncture, the units
within the University of Southern California which administers courses in English to non-native
speakers (the American Language Institute (ALI), and its affiliate, the Language Academy) serve
as the source of the data. Those data consist of written production by matriculating students whose English proficiency is assessed at levels below that of a native speaker. Transcription, coding, and analysis of the data have adapted the CHILDES computerized resources of CHAT and CLAN. Data sources will subsequently increase with the establishment of a consortium of researchers/consultants at other academic institutions who are associated with the project.

1 CHILDES and a repository for L2 written data

CHILDES, as a spell-out of its acronym implies, is conceived as a project for the computerized storage, coding, and retrieval of data produced by children and their interlocutors. It is prima facie a repository for spoken-language data, and its CHAT transcription/coding procedures and CLAN computer tools were originally conceived primarily for that purpose. Yet the same tools can, with little adjustment, serve to transcribe and code non-native spoken-language adult data, notwithstanding the mystery of their non-appearance in the CHILDES corpora so far. What then of L2 written data? Subsequent expansion, development, and refinement of CHILDES over the years has extended the range of language modes to be accommodated, as mentioned earlier. Chapter 12 of Volume I, Part 1 (MacWhinney 2000), titled ‘Extending CHAT’, outlines adaptations and modifications for the transcription of code-switching (12.1), elicited narratives (12.2), written language (12.3), and language disfluency (12.4). The section on written language specifies procedures for transcription of printed text and of the written data of school children, where the latter can be accomplished within the overall framework of CHAT together with certain ‘modifications and extensions’. This is essentially the route that ACSL has taken for entry of post-pubescent L2 written data.

a Adaptation of CHILDES for L2: Utilization of CHILDES in transcribing and coding L2 data for ACSL has led to a number of additions to the category of HEADERS and a few alterations in
the labeling of grammatical categories recognized by the parser. Headers—three obligatory, the rest optional—appear prefixed by ‘@’ at the beginning of every file and denote the relevant non-linguistic features or learner characteristics. The list of headers appearing in a sample ACSL file (the three obligatory headers in boldface, some of the others devised specifically for L2) might be as shown in Table 1. Data from this file will serve to illustrate coding options in Section 2.

< Insert Table 1 about here >

b Proficiency: Among the many learner characteristics that are important for research in language acquisition are some that relate only to the post-pubescent phase. One of these factors is that part of language assessment termed ‘level of proficiency’. Although it is absolutely essential for the usefulness of L2 corpora that learner data ultimately be coded for something recognizable as ‘proficiency’, the difficulty in devising a viable procedure for doing so has long been recognized. Thomas (1994: 307), in a critical analysis of L2 proficiency measures, documents ‘four major conventions’ which L2 researchers have resorted to: ‘institutional status’, ‘impressionistic judgment’, ‘research-internal or in-house measures’ and ‘standardized test scores’. Institutional status serves as a proxy for proficiency in many of the L2 corpora cited in Granger (1998). Since the comparatively small amount of ACSL data coded so far have been supplied by the on-campus ALI, all proficiency rating have been assigned holistically and impressionistically by locally-trained staff.13 What ACSL, or any potentially global L2 database, will need to have is a reasonably valid and reliable measure that can serve researchers contributing written data from elsewhere. Eventually, that measure would need to be adapted to the assessment of L2 proficiency in languages other than English. None of the four conventions identified by Thomas (1994) would seem to suffice in this instance. It is possible that, ultimately, such an adequate measure of proficiency might have to turn on multivariate correlations among a yet-to-be-identified subset of learner linguistic and non-linguistic characteristics.14
2 CHILDES and coding for L2

The standard view is that second language acquisition as a field of research came into being in the early 1970s. A convenient point of inception might well be the Selinker (1972) seminal paper ‘Interlanguage’ (IL), the coinage in whose title has ever since been generally adopted as the technical term in reference to an adult language learner’s interim grammar at any chosen point in time. Observation of the nature of those grammars, which typically embody features found neither in the L1 nor the L2, gave rise to the ‘IL Hypothesis’, the proposal that ILs are systematic at any given stage of development. It would then follow that ILs are to be analyzed sui generis, as would of course any natural language. It has been argued, however, that whether nonnative language is natural language—i.e. ‘whether native-speaker linguistic knowledge and nonnative-speaker linguistic knowledge are epistemologically equivalent’ (Schwartz 1999: 648)—has to be determined empirically and cannot simply be assumed. The epistemological status of interlanguage, together with the IL Hypothesis, therefore pose an interesting dilemma for any L2 database project in which learner data are to be not only transcribed but also coded.

The dilemma is the following: Computerized transcription and coding of L2 learner data force one to make repeated judgments as to where there is learner ‘deviation’ from target ‘norms’. For those who assume that ILs are natural language, then the question is how, if the oft-cited IL Hypothesis and ‘comparative fallacy’ (Bley-Vroman 1983) are to be taken seriously, coding of this kind can be justified. For those who hold that assumed native language–non-native language epistemological equivalence is what remains to be demonstrated, then the question is how, if the linguistic vocabulary for drawing IL representations is the same as that for the native language, coding of this kind can be justified. In other words, linguistic coding of L2 learner data thus seems to contravene both opposed theoretical positions. Should coding then be abandoned? The answer is ‘no’, for reasons both of practicality and of the options associated with (the vaguely termed) ‘coding’ itself.
Learner errors, or whatever else one may wish to call them, need somehow to be identified and rendered amenable to computerized retrieval, at least to some extent, in any corpus designed to serve L2 researchers working within a linguistic framework. This would apply to the formal study of L2 acquisition as well as to that of classroom learning. Not to proceed in this way would provide a database devoid of accessibility to the kinds of data needed by many of the very researchers whom the database is designed to serve. If this points to coding, then what kind and to what extent?

It is well to note that choice of coding procedures within CHILDES will be largely dictated by the architecture of the CHAT and CLAN programs themselves. For example, some aspects of coding are actually required in the CHAT transcription mode—e.g. correction of spelling after running TEXTIN, essential for word recognition in the operation of the MORPHOLOGICAL PARSER—and are thus of course not options for the coder. Another aspect is coding relegated to an ERROR TIER, designed optimally for treating lexical material. Still another would be assignment of grammatical category by the parser. CHILDES embodies a number of coding stages extending from initial verbatim transcription of raw data (including material crossed out by the learner but readable) in CHAT mode to identification of special lexical items (e.g. unaccusatives) through matching with appropriate lists in a TEXT file with the CLAN program. For illustration, a brief learner text in ACSL will be taken through the stages in question.

a Main file before TEXTIN:  Transcription of raw data for the file whose headers were displayed in Table 1 would appear as it does in Table 2. (Crossed out but legible material is coded as <of> [/], illegible material as xxx; all lexical items must appear here in lower case, then to be adjusted with coding.)
b Coding in CHAT: After coding, the file displayed in Table 2, together with the headers in Table 1, would appear as in Table 3.

< Insert Table 3 about here >

Three tiers are now displayed: MAIN TIER (STU:), ERROR TIER (%err:), and COMMENT TIER (%com:). In addition to coding of spelling errors, the (obligatory) main tier serves to identify an error site [*] and error type (omission: 0 [*]; commission: ITEM [*]). On the error tier the tagged errors are coded as equivalences: error form = target form. E.g. work = works; 0 = are; be = 0, etc. Coder uncertainty as to the equivalent target form would be entered as ITEM = [?]. The comment tier is reserved for calling attention at any point to something that the coder may deem noteworthy for research purposes: e.g. ‘note topic-comment construction’.

c Morphological parser: In addition to the error and comment tiers, CHILDES provides the means—viz. the MORPHOLOGICAL PARSER—for automatic generation (from the main tier, in CHAT mode) of a tier for coding part-of-speech and inflectional morphology: the %MOR tier.19 Engagement of the parser results in brute-force labeling of all possible grammatical categories for each lexico-grammatical item taken in isolation. The massive technical ambiguity created at this stage is then resolved by context and partially by hand, leaving coded texts amenable especially to research in the development of L2 grammatical marking. Illustrated in Table 4 is the last segment of the L2 learner text from Table 3 after generation of a %mor tier, but before disambiguation:

< Insert Table 4 about here >

Note that the parser, oblivious to context or collocation, has blindly coded young man, for example, as n|young^adj|young v|man^n|man^co|man (where CO is the CHAT label for ‘communicator’). Shown in Table 5 is the %mor tier with this same text fragment, after disambiguation:

< Insert Table 5 about here >
ACSL has modified the CHAT part-of-speech inventory in small ways not shown in these Tables. The frequent necessity for such modification is in fact recognized by the CHILDES program: ‘Individual researchers will need to define a system of part-of-speech codes that correctly reflects their own research interests and theoretical commitments’ (MacWhinney 2000, Vol I, Part 1, p.157).

**d Coding and proficiency:** Deployment of coding procedures for learner data running the gamut of L2 proficiency means that the intensity of main-line coding will likely co-vary with proficiency level. The error coding shown in Table 3, given the relatively low level, identifies redundancies, omissions, aberrations, mismatches, etc., with single items (both lexical and functional) and small clusters. The text does not reach a level of sophistication sufficient to clearly reveal what the learner is trying to express and to warrant coding attention, for example, to tense and aspect, morpho-syntactic relationships, lexical collocations, etc. In any event, decisions as to how far to take coding will usually have to be something of a judgment call grounded to a great extent in theoretical perspective and research agenda. It should be noted, however, that users of the ACSL corpora, as with the CHILDES corpora, can always, if desired, access a particular transcription in its original form (before any coding).20

3 **CHAT and CLAN in the service of an L2 research issue**

Large computerized corpora have potentially the great research value for L2 that CHILDES has long had for L1. One of the rare instances of massive amounts of computerized coded data in pursuit of an L2 research question is the Oshita (1997) [also cite Oct 2000 SLR article?] study of L2 split intransitivity. Oshita’s data source was the enormous Longman Learners Corpus (LLC; 1993 version) which consists of L2 essays from around the world, but is devoid of any linguistic coding. The loose status of proficiency, assigned independently by each contributor to the LLC, was all there was to work with. Absent any means of computerized retrieval of linguistic
features, the utility of the LLC depended upon creation of such means by the researchers themselves. Oshita and his colleague, J. Allen, therefore had to create a search program for identification of twenty common verbs with one argument—ten unaccusatives and ten unergatives—which yielded a total of 1,581 tokens. Of these, 941 were unaccusative and 640 unergative. But crucially, the surface construction for 851 of the unaccusatives and all but one of the 640 unergatives was NP V21, leading Oshita to surmise that all intransitives are initially analyzed by the L2 learner as unergative. The validity of the claim aside, it could not have arisen at all without recourse to an extremely large computerized corpus of data coded for linguistic features. But it was left to the researchers to create the very computer tools needed to make such use of that corpus.

CHILDES has long provided various means for retrieval of learner data in terms of lexical and structural category. The option chosen for ACSL data can be illustrated for unaccusatives as an additional stage with reference to the development of L2 learner text shown in Tables 1–5. Retrieval of unaccusatives (or any identifiable lexical class) requires the compilation of a list of such verbs in a specially created TEXT file. From one of the appendices in Levin & Rappaport Hovav (1995) was drawn a list of fifty-three fairly common non-alternating unaccusatives for placement in the file. A CLAN program called COMBO then searched and located the utterances containing those tokens and provided for each one the file and line numbers. With reference again to the brief learner text illustrated in the preceding tables, the output of COMBO in a search for non-alternating unaccusatives would therefore yield the verb die, as follows:

< Insert Table 6 about here >

If the research interest were L2 passivized unaccusatives (as with die in the sample text under discussion), presumably the researcher would want to document any possibly occurring passivized unergatives as well. All instances of passive morphology itself could be retrieved from the %mor tier, even without disambiguation, with another COMBO sub-command. Although unaccusatives have been the focus here, it should be evident that the computer tools
provided in CLAN are designed to enable the identification, access, and retrieval of limitless lexical and structural material, and over vast amounts of data.

IV Conclusion

The updated and expanded exposition of CHILDES in MacWhinney (2000) makes the enormous effect of the project on the field of child first language acquisition unmistakable. There is now widespread recognition that creation and subsequent evolution over almost two decades of the CHILDES enterpriseWhereas the potential value of the CHILDES mechanisms for research in post-pubescent language acquisition has always been with us, it has (inexplicably) not been demonstrated until now. It is clear that the rich CHILDES resources refined over considerable time in CHAT and CLAN can eminently serve a wide array of research pursuits in adult language acquisition. We therefore have every reason to believe that the impact CHILDES has had upon the shape of L1 acquisition research will in time be replicated for L2.
Acknowledgements

Thanks go to Monica Cabrera for helpful comments on part of the draft of this review.
References


Meunier, F. 1998: Computer tools for the analysis of learner corpora. In Granger (ed.).


Pye, C. 1994: Review of *The CHILDES project: Tools for analyzing talk: 1st Edition. Language* 70, 152–156. [Not in text; was a reference to this work deleted?]


*William Rutherford*
*University of Southern California*

*Margaret Thomas*
*Boston College*
Footnotes

1 MacWhinney (1995, pp. 421-436) includes a list of published research up to 1995 which employed CHILDES. Unfortunately, MacWhinney (2000) neither updates nor reprints this list.

2 More such exercises, with answers in the back of the book, are available in Snow and Sokolov (1994).

3 Annoyingly, the two parts of Volume I are paginated separately, as if Erlbaum had originally intended to publish the third edition as three physically autonomous texts. There is no apparent advantage to binding Parts 1 and 2 together without repaginating them; in fact, lack of continuous pagination makes use of the volume’s two indices very clumsy.

4 For example, MacWhinney (1991, 1995) relied on a locally created form of UNIBET, ‘PHONASCII’, for phonetic transcription of ASCII codes. The third edition of the manual abandons PHONASCII in favor of a more widely used, more cross-linguistically consistent, version of ASCII encoding. But the overall dependence of CHILDES on ASCII encoding is reduced in the latest version of CHAT because the text editor now allows transcription in the International Phonetic Alphabet, and some CHILDES transcripts have been linked to digitized sound recordings. Nevertheless, ASCII encoding is still employed in some of the corpora, so the third edition of the manual contains tables listing UNIBET codes for English and seven other languages.

5 For a review of the CHILDES / BIB see Falk (1992).

6 [[Footnote about availability of the whole CHILDES thing over the Internet. Is it possible to get this stuff without buying the books? [yes; that’s what we did for ACSL] [WR: I’m trying to reach someone from Erlbaum to talk about this issue. More to follow] If so, why might one buy them?]]

7 For references to publications produced by the EFS project and an assessment of it, see Carroll (1999).

8 The project has been supported [by The Dean of the University of Southern California Graduate School of Arts and Sciences (or whatever) since its inception [in 199x] in the form of a graduate-student research assistantship.

9 ACSL also contains a very small corpus consisting of nineteen transcribed and coded Spanish L2 essays produced by first-semester students in the Univeristy of Southern California Department of Spanish and Portuguese. Attention will soon be turned to expansion of this corpus.

10 CHAT specifies four obligatory headers; see MacWhinney (2000), Volume I, Part 1, Chapter 4 (4.1).
11 The ‘training’ header reads ‘years of study’ on the actual ALI form, although interpretation of the entry may not be very reliable. For some learners ‘years of study’ can mean the year in which they began studying English. Others may not understand the entry as worded. Still others will be ashamed to cite a lengthy period of study with little to show for it. Moreover, no entry will reveal anything about the intensity of study. (These observations were offered by Paul Bruthiaux.)

12 The ‘date’ header indicates the date on which the sample was obtained.

13 ALI’s holistic/impressionistic assessment of ‘3’ shown in Table 1 indicates ‘that the student’s English is fairly weak, but sufficient to survive in a regular university class that would make minimal writing demands. A paper in this category exhibits basic competence in writing but exhibits several of the following characteristics:
   a. fairly limited vocabulary with fairly frequent word choice errors
   b. fairly frequent syntactic errors
   c. little syntactic variety with some coordination but little or no subordination
   d. some development (use of detail to support or illustrate points)
   e. frequent breakdowns in coherence and progression
   f. few cases of incomprehensible due to one or more of the above features’

14 This direction, as one possibility, has been suggested by Lyle Bachman (personal communication).

15 Schwartz (1999: 48) put it this way: ‘The very act of adopting the technicalities of generative linguistic theory to analyze interlanguage data, by definition . . ., presumes that interlanguage representations are drawn by the very same psychological hand that draws native-language representations.’

16 Error coding is apparently not very widespread: ‘Very few corpora have been annotated for errors and those that have are not usually available’ (Meunier 1998: 26).

17 In the words of Meunier (1998:26), ‘The analysis of error-tagged corpora will improve current pedagogical tools—dictionaries, grammar books, writing books, and style and grammar checkers’.

18 Note the value for research purposes of having a record of L2 learner ‘afterthought’, as it were, in the form of emended text. On the written portion of its entrance exam, ALI has thus for some years stipulated the use of pen instead of pencil, with emendation by crossing out and/or carat-insertion.

19 The details are spelled out in Volume I, Part 1, Chapter 16. To date, computerized parsing in CHILDES has been worked out for English, Dutch, German, Japanese, and Spanish.

20 Hand-written essays in the ACSL corpus will be made available in scanned format.
The remaining 90 unaccusatives displayed a range of syntactic distribution—passive, existential, expletive, V NP, etc.—over the four L1 learner groups studied that formed the essence of Oshita’s research.
Table 1  Headers in a sample ACSL file

<table>
<thead>
<tr>
<th>Header</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>@Begin</td>
<td></td>
</tr>
<tr>
<td>@Participants:</td>
<td>STU Student</td>
</tr>
<tr>
<td>@Sex of STU:</td>
<td>Male</td>
</tr>
<tr>
<td>@Filename:</td>
<td>STU KOREAN 09</td>
</tr>
<tr>
<td>@Language 1:</td>
<td>Korean</td>
</tr>
<tr>
<td>@Language 2:</td>
<td>English</td>
</tr>
<tr>
<td>@Status:</td>
<td>Second Language Learner  [cf. Foreign Language Learner]</td>
</tr>
<tr>
<td>@Training:</td>
<td>5 years11</td>
</tr>
<tr>
<td>@Proficiency:</td>
<td>3  [see footnote 12]</td>
</tr>
<tr>
<td>@Location:</td>
<td>Los Angeles, CA, USA</td>
</tr>
<tr>
<td>@Coder:</td>
<td>[name]</td>
</tr>
<tr>
<td>@Task:</td>
<td>Timed writing (30 minutes)</td>
</tr>
<tr>
<td>@Comment:</td>
<td>Topic of the writing sample: Working too hard</td>
</tr>
<tr>
<td>@Date:</td>
<td>05-JAN-199312</td>
</tr>
<tr>
<td>@End</td>
<td></td>
</tr>
</tbody>
</table>
Table 2  Transcription of raw data

morden time is that every body work to do something and, if some one grown up to more higher level <of> [/] insociety, he have to working harder. generaly, many people want to better life so, today many people working harder but people work too hard. specially, korea and japan are many young man working too hard for example, xxx about forty years old man going to be die a hundred thousand per year in japan because of they work to hard.
Table 3  Coding, with error and comment tiers

@Begin
@Participants: STU Student
@Sex of STU: Male
@Filename: STU KOREAN 09
@Language 1: Korean
@Language 2: English
@Status: Second Language Learner
@Training: 5 years
@Proficiency: 3
@Location: Los Angeles, CA, USA
@Coder: [name]
@Task: Timed Writing (30 minutes)
@Comment: Topic of the Composition: Working too hard
@Date: 05-JAN-1993

*STU: morden [: modern] time is that <every body> [: everybody] work
[*] to do something and, if <some one> [: someone] <grown up>
[*] to 0 [*] more [*] higher level <of> [//] insocieny
[: in society], he <have to working> [*] harder.

%err: work = works; grown up = grows up; 0 = a; more = 0; have to working = has to work

STU: generaly [: generally], many people want to [*] better life so,
today many people 0 [*] working harder but people work too hard.

%err: to = a; 0 = are
Table 3, continued

<table>
<thead>
<tr>
<th>STU:</th>
<th>specially [: especially], 0 [<em>] Korea and Japan are many young man [</em>] working too hard for example, xxx about &lt;forty+years+old&gt; [<em>] man [</em>] 0 [<em>] going to be [</em>] die a hundred thousand per year in Japan because of [*] they work to [: too] hard.</th>
</tr>
</thead>
<tbody>
<tr>
<td>%err:</td>
<td>0 = in; man = men; forty+years+old = forty+year+old; man = men; 0 = are; $UNA be = 0; of = 0</td>
</tr>
<tr>
<td>%com:</td>
<td>note topic-comment construction; note passivized unaccusative ‘die’</td>
</tr>
</tbody>
</table>

@end
Table 4  The %mor tier before disambiguation

STU:  specially [: especially], 0 [*] Korea and Japan are many young man [*] working too hard for example, xxx about <forty+years+old> [*] man [*] 0 [*] going to be [*] die a hundred thousand per year in Japan because of [*] they work to [: too] hard.

%mor:  adv|especial-LY n:prop|Korea conj:coo|and n:prop|Japan v|be&PRES^v:aux|be&PRES qn|many^v|man-Y^n|mane-Y n|young^adj|young v|man^n|man^co|man part|work-PRES_PART^v|work-PROG adv|too adj|hard prep|for^conj:subor|for n|example prep|about^ptl|about^adv|about n:comp|forty+years+old v|man^n|man^co|man part|go-PRES_PART^v|go-PROG prep|to^inf|to v|be^v:aux|be v|die^n|die det|a det:num|hundred det:num|thousand prep|per n|year prep|in^ptl|in^n|in n:prop|Japan conj:subor|because prep|of pro|they v|work^n|work adv|too adj|hard.
Table 5  The %mor tier after disambiguation

<table>
<thead>
<tr>
<th>%mor:</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv</td>
</tr>
<tr>
<td>qn</td>
</tr>
<tr>
<td>n</td>
</tr>
<tr>
<td>v:aux</td>
</tr>
<tr>
<td>n</td>
</tr>
<tr>
<td>v</td>
</tr>
</tbody>
</table>
### Table 6 The search for non-alternating unaccusatives with COMBO + sub-commands

```plaintext
> combo +s@NAunaccusatives.cut @
((seem*+appear*+aris*+arose+awak*+come+coming+came+dawn*+emanat*+emerg*+
erupt*+evolv*+flow*+gush*+issu*+materializ*+plop*+result*+ris*+rose*+steal*+
stoile*+stem*+stream*+superven*+surge*+wax*+die*+dying+disappear*+expir*+laps*
+perish*+vanish*+ensu*+eventuat*+happen*+occur*+recur*+transpir*))
```