Child Language Data Exchange System

The Child Language Data Exchange System (abbreviated: CHILDES) was founded in 1984 by Brian MacWhinney (Carnegie Mellon University) and Catherine Snow (Harvard University) as an effort to bring together child language corpora and to make them publicly available for researchers. Since its inception, CHILDES has developed into a rich archive of corpora with speech data of many children and their interlocutors. Most corpora are available as written transcripts, but for a growing number of them also the audio and/or video recordings are available. The speech materials have all been transcribed and coded in a uniform way, using specialized conventions (CHAT), and various software tools (CLAN) have been developed which can assist the researcher in transcribing, coding and analyzing data. In the last 30 years, CHILDES has become well known in the research community and its resources are frequently used, judging only from the more than 3,000 published papers in which available corpora are studied. The CHILDES corpora and tools, as well as the documentation manuals, are freely available on the world wide web (http://childes.psy.cmu.edu/) for child language researchers.

Two elements are quintessential in appreciating and evaluating CHILDES. First of all, CHILDES functions as a repository. The corpora archived are donated by individual researchers and research groups. Hence, they are not the result of a coordinated effort, but are the product of past and present research projects from all over the world, each with its own aims and scope. This explains, for instance, differences in the amount of data between corpora. Secondly, for a repository a technical infrastructure is required to make the archive publicly available, to secure a standard format of all the corpora, to keep the data compatible with evolving international standards for corpus transcription and annotation, but also to keep the infrastructure in line with continuous developments in the world of computers and custom electronics. For instance, the computational tools provided by CHILDES are developed and updated for different platforms (Windows, MacOS, Unix), which implies that with an upgrade of an operating system the tools have to be updated too. The repository is reachable using different web browsers (Firefox, Safari, Explorer, …) which all have their own peculiarities that have to be accommodated. Moreover, the CHILDES infrastructure has to keep pace with the continuously changing market of personal computers. A recent example is the introduction of tablets next to desktop and laptop devices. All these tasks are coordinated and directed by Brian MacWhinney in collaboration with the programmers Leonid Spektor and Franklin Chen.

1. The database: child language corpora

The corpora in the CHILDES database can be broadly divided into four main categories: (1) monolingual corpora with longitudinal (and some cross-sectional) data from monolingual children; (2) bilingual corpora, with data from concurrent and sequential acquisition of two or more languages, (3) narrative corpora, a subdivision containing retellings of stories, such as the often used frog story, (4) clinical corpora, with data from language disordered children. Each corpus in the four subsections is documented in the database manuals that can be consulted at the CHILDES website (http://childes.psy.cmu.edu/manuals/). The documentation provides detailed metadata about the participants (ages, languages spoken, setting of the recordings, etc.), the

data files that can be consulted, as well as corpus specific details with respect to transcription and coding.

The CHILDES database comprises transcriptions of children's and their interlocutors' speech, derived from audio and/or video recordings of spontaneous adult – child interactions. Although in most corpora parent-child conversations in various settings were recorded, some corpora consist of child-child interactions, and still others are classroom interactions between children and their (kindergarten) teachers. Moreover, in the clinical corpora the conversational partner is sometimes a language and speech therapist.

At present, 39 languages are represented in the database. Table 1 provides an overview of the languages (in alphabetical order), in which the following language families occur: Germanic languages (Afrikaans, Danish, Dutch, English, German, Norwegian, Swedish), Romance languages (Catalan, French, Italian, Portuguese, Romanian, Spanish), Slavic languages (Croatian, Russian, Serbian, Slovenian), Celtic languages (Irish, Welsh), Afroasiatic languages (Arabic, Berber, Hebrew), Sino-Tibetan languages (Cantonese, Mandarin Chinese, Taiwanese, Thai), Altaic languages (Turkish, Japanese, Korean), Dravidian languages (Tamil), Uralic languages (Estonian, Hungarian), Indo-Iranian languages (Farsi), Austronesian (Indonesian), the Hellenic languages (Greec), Algonquian languages (Cree), and Basque. For each language the number of utterances and the number of word tokens were counted. These counts were done separately for the "target child" (as indicated in the database manual) and for all "non-target" participants.

Language	Target Child		Adult		Total	
	# Utterances	# Words	# Utterances	# Words	# Utterances	# Words
Monolingual						
Afrikaans	3,892	12,974	3,008	13,020	6,900	25,994
Arabic	9,564	9,792	0	0	9,564	9,792
Basque	30,709	79,959	40,882	143,707	71,591	223,666
Berber	3,827	10,006	0	0	3,827	10,006
Cantonese	81,038	207,184	147,882	657,602	228,920	864,786
Catalan	23,601	51,046	43,269	178,146	66,870	229,192
Chinese	69,226	218,491	150,241	570,647	219,467	789,138
Cree	3,056	4,922	3,751	10,343	6,807	15,265
Croatian	36,010	86,417	55,830	215,267	91,840	301,684
Danish	34,977	82,466	48,939	192,504	83,916	274,970
Dutch	180,670	444,933	271,278	1,130,194	451,948	1,575,127
English	1,460,992	4,320,926	2,917,501	13,772,058	4,378,493	18,092,984
Estonian	43,902	148,592	49,562	225,420	93,464	374,012
Farsi	20,007	61,009	53,679	196,081	73,686	257,090
French	227,006	720,222	423,832	2,019,858	650,838	2,740,080
German	474,258	1,430,003	703,980	3,334,142	1,178,238	4,764,145
Greek	13,642	22,753	11807	35055	25449	57808
Hebrew	50,572	155,145	101,969	385,247	152,541	540,392
Hungarian	21,213	56,685	32,907	123,798	54,120	180,483
Indonesian	270,930	739,721	540,750	1,606,552	811,680	2,346,273
Irish	12,293	20,983	20,025	99,737	32,318	120,720
Italian	26,306	62,181	69,523	289,636	95,829	351,817
Japanese	288,528	791,590	329,372	1,000,938	617,900	1,792,528
Korean	6,440	9,664	10,600	26,639	17,040	36,303

Norwegian	2,385	7,916	8,587	47,338	10,972	55,254
Polish	133,420	654,053	96,200	460,285	229,620	1,114,338
Portuguese	28,634	62,443	27,378	173,067	56,012	235,510
Romanian	11,680	15,723	16,543	60,496	28,223	76,219
Russian	7,369	21,764	7,780	32,430	15,149	54,194
Serbian	95,143	219,260	226,853	807,587	321,996	1,026,847
Sesotho	28,791	84,863	40,683	149,308	69,474	234,171
Slovenian	1,023	5,299	375	2,663	1,398	7,962
Spanish	244,559	856,765	330,918	1,406,854	575,477	2,263,619
Swedish	64,591	168,320	78,122	357,652	142,713	525,972
Taiwanese	21,207	50,530	41,782	153,196	62,989	203,726
Tamil	1,999	3,808	4,244	11,662	6,243	15,470
Thai	4,655	8,850	42,320	206,106	46,975	214,956
Turkish	14,816	35,580	14,457	43,968	29,273	79,548
Welsh	180,105	634,888	34,700	194,607	214,805	829,495
Bilingual	391,415	1,410,953	581,080	2,422,625	972,495	3,833,578
Narratives	77,160	528,398	40,376	257,211	117,536	785,609
Clinical	224,308	671,129	530,295	2,168,969	754,603	2,840,098
Total	4,925,919	15,188,206	8,153,280	35,182,615	13,079,199	50,370,821

Table 1: Overview of the CHILDES corpora per language and kind

Table 1 shows that the complete CHILDES database contains at present (May 2013) more than 13 million utterances consisting of more than 50 million word tokens. The monolingual corpora take up the largest part: 43 million word tokens in approximately 11 million utterances. In this group, English is best represented (18 million word tokens), followed by German (4,7 million), and French (2,7 million). It should be noted that in some corpora only the target children's utterances were transcribed, as in for instance Arabic and Berber.

The bilingual corpora consist of almost 4 million words in approximately 1 million utterances. These corpora comprise the language pairs Arabic-Dutch, Turkish-Dutch, Spanish-English, Russian-German, French-English, Chinese-English, Japanese-Danish, Russian-English, Polish-English, Portuguese-Swedish, English-French, Catalan-Spanish, Italian-Dutch, Cantonese-English, and the triple Persian-English-Hungarian. In many cases these corpora also include monolingual, age-matched controls.

The clinical corpora comprise almost 3 million word tokens. They contain data from atypically developing children with SLI (Specific Language Impairment), Down's syndrome, hearing impairment (children with traditional hearing aids and with cochlear implants), Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS), different levels of mental retardation, autism, epilepsy, apraxia, children who stutter, children exposed to cocaine in utero, and late talkers. The range of languages is fairly restricted: Dutch, English, French, German, Hebrew, and Spanish.

Finally, the narrative corpora consist of approximately 750 thousand words. An interesting parallel subcorpus comprises the "frog story" narratives in which a particular picture book is used to elicit narratives from children acquiring English, French, German, Hebrew, Italian, Russian, Spanish, Thai and Turkish.

The age ranges covered in the various corpora differ greatly: from children in the prelexical stage to adolescents. However, when charting out the various ages represented in CHILDES against the number of words available for each age, it strikes the eye that especially children in their third and fourth year of life are well represented. For older children the amount of data steadily decreases. In Figure 1 that distribution is plotted for the Dutch monolingual corpora. The distribution shows that from approximately 1;10 to 3;06 the Dutch CHILDES data have at least 10,000 word tokens per month of life. Younger children are scarcely represented and from 3;06 onwards the number of word tokens decreases rapidly.

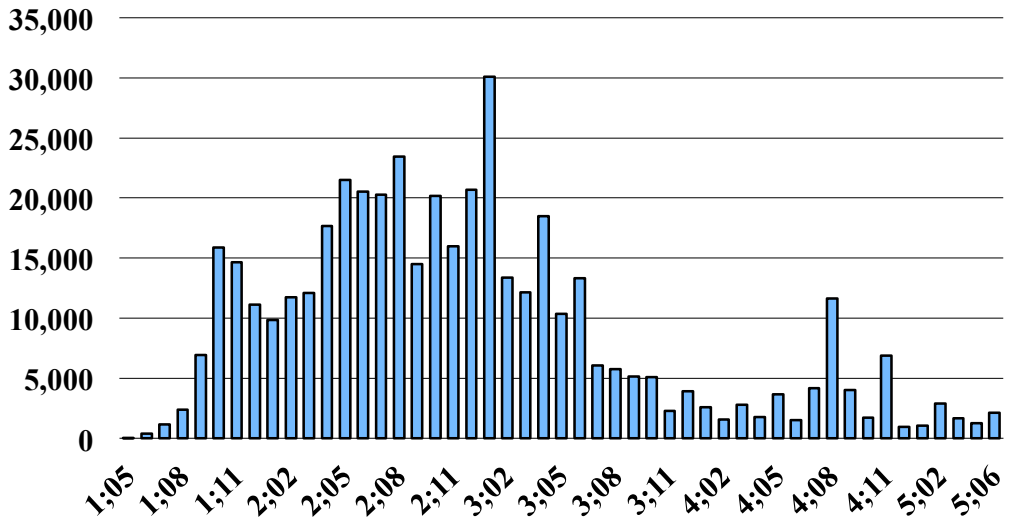

Figure1: Distribution of the number of word tokens per age (years;months) in the Dutch CHILDES corpora

2. CHAT: transcription and coding

The conventions for transcription and coding used in the CHILDES corpora are called CHAT, an abbreviation of Codes for the Human Analysis of Transcripts. CHAT provides a standardized way to organize transcripts. For instance, the example in (1) shows the first lines of a transcript taken from the CHILDES' MACWHINNEY corpus:

(1) Example transcript

```
@Begin
@Languages:  eng
@Participants: CHI Ross Target_Child , MOT Mary Mother
@ID:   eng|macwhinney|CHI|2;7.10|male|normal||Target_Child|||
@ID:   eng|macwhinney|MOT|||||Mother|||
@Media:      21a1, audio
@Date:       04-AUG-1980
*MOT:        okay you talk to Daddy .
*CHI:        <is Daddy with you> [/] is Daddy with you ?
*MOT:        you talk to Daddy on the phone .
*MOT:        no that phone .
@End
```

The example shows main organizational principles of a CHAT transcript. Metadata and data are differentially marked: metadata by a header starting with the symbol "@". These include a header signaling the beginning and the end of the transcript (@Begin, @End), the language(s) spoken by the participants (@Languages), the sound file (@Media), the date of the recording (@Date), etc. The actual data are on lines with the symbol "*" followed by a three letter abbreviation of the speaker's name (*MOT, *CHI). The latter are called the *main tiers* of the transcript, which contain a standardized orthographic transcription of the recorded speech. These

conventions and abbreviations are all clearly presented and discussed in the manuals that are available on the CHILDES website. All corpora in the CHILDES database provide at least information regarding the language, the participants, and @ID information (including the target children's age).

The second aim of CHAT is to provide transcribers with clear conventions about how to transcribe speech recordings. Spoken language use exhibits phenomena such as uncompleted words or utterances, exclamations (e.g., "haha!"), dialectal forms and other non-standard or shortened forms (e.g., *bout* instead of *about*, *gonna* instead of *going to*), onomatopoeias (such as animal sounds), and even unidentifiable material. In order to transcribe these in a consistent way, CHAT conventions were developed. Conventions were also developed for the interactional dynamics such as the consistent marking of pauses, the marking of phenomena such as (self-)interruptions, retracings, overlaps and repetitions. For instance, in example (1) a repetition is marked using the symbol [/] and the repeated part of the utterance is marked with brackets.

In addition to the main tier, the CHAT format permits the insertion of *dependent tiers*. Essentially these are meant for the coding of the material on the main tier. Codings can relate to linguistic enrichment or the explicit marking of extra linguistic information (such as actions, gestures, events). The linguistic enrichment of the transcribed speech can contain many types of information, depending on the aims of the researcher: e.g., a phonemic transcription, morphosyntactic coding, coding of the grammatical relations or speech acts. As an example, a morphosyntactic coding of one utterance of (1) is presented in (2).

(2) Example coding

*CHI: is Daddy with you ?
%mor: cop|be&3S n:prop|Daddy prep|with pro|you ?

The coding exemplified in (2) shows how on a dependent tier, viz. %mor, the wordforms in the main tier are analyzed in three respects: their part-of-speech is coded ("cop" = copula, "n:prop" – proper noun, "prep" = preposition, "pro" = pronoun), the wordforms are lemmatized (e.g., the wordform "is" is a form of the lemma "be"), and the wordforms are morphologically decomposed (e.g., "is" represents the lemma "be" in its third person singular form -- "3S").

In the CHAT manual a great many levels of coding are formally defined, however the format is flexible enough so that researchers can define their own coding categories. The main point about CHAT is that a standard is established, the formal requirements for transcription and coding are defined, so that the CHILDES programs (discussed in the next paragraph) can be easily used with transcripts that comply with the CHAT format. In addition, CHAT is XML compliant, which means that CHAT files can easily be converted to the XML format, which is nowadays the standard for the representation of documents.

3. CLAN: software tools for corpus annotation and corpus analysis

CLAN stands for Computerized Language Analysis. It is a collection of more than 30 integrated software tools with diverse functionality: some tools aid the researcher in the transcription process, other tools perform (semi-) automatic coding of transcripts, CLAN provides basic analytic tools for searching particular strings (e.g., words on

main tiers, particular codes on dependent tiers) and delivers indices, such as a frequency list of the different words on the main tiers of transcript, the co-occurrence patterns of words, or a frequency table of parts of speech and affixes given a morphosyntactic coding on a dependent %mor tier. Other tools output often used indices such as type-token ratios, MLU (Mean Length of Utterance) or VOCD, a measure of lexical diversity.

The CLAN software comes with an editor that can be used to transcribe recordings. In Figure 2 the editor is pictured. The editor window contains a fragment of a transcript in the left part. At the bottom of that window, the sound wave is displayed, and by selecting part of the sound wave, the transcriber can (re-)play the relevant part of the sound file. In the right part of the window, the corresponding video fragment is shown. Also for video, the transcript can be linked with the relevant part of the video recording, and by a simple mouse click, the transcriber can immediately view the part of the video recording he/she is transcribing.

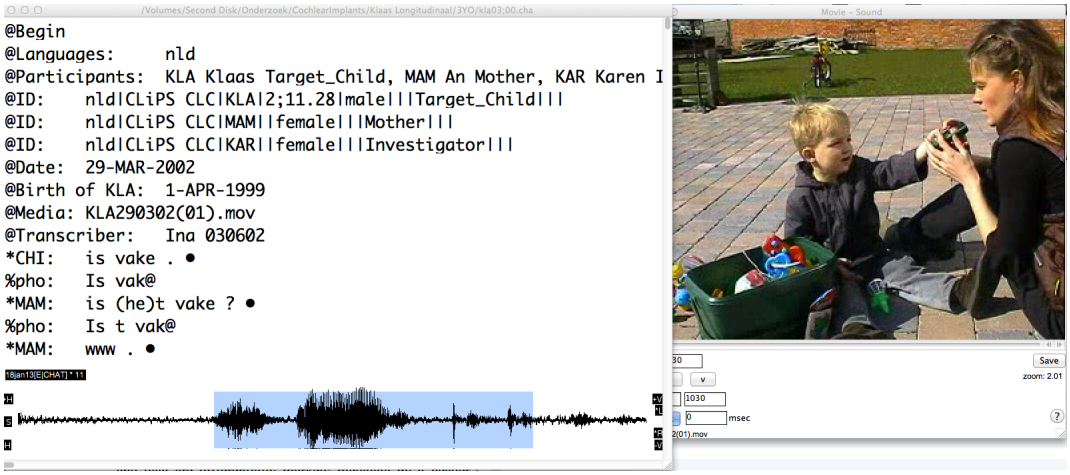

Figure 2: The CLAN editor with a transcript, sound wave, and video fragment.

Of particular interest are a number of tools that assist the researcher by automatically producing a coding tier. The MOR tool automatically inserts a dependent %mor tier (see (2)) in a transcript. A recent addition to the CLAN software involves the automatic insertion of a dependent tier on which dependency grammar tags indicate basic syntactic relations.

PHON is a recent addition to the CHILDES instrumentarium designed and developed by Yvan Rose and Greg Hedlund. Phon is especially geared towards the analysis of phonological data transcribed in CHAT. The program and the associated database supports linkage to the media files, and offers automated services like automatic segmental feature labeling, automatic syllabification, automatic phone alignment, and systematic comparisons between target (model) and actual (produced) phonological forms.

4. The impact of CHILDES on the field

The CHILDES database has had an enormous impact on language acquisition research in the last decades, and specifically on longitudinal studies of observational data. Various reasons can be identified for this crucial role. Going back to the corpora collected in the 1970ies, like Roger Brown's seminal corpus of Adam, Eve and Sarah,

children were recorded and these recordings were subsequently transcribed. In the best case, transcriptions were made on a typewriter. This greatly hampered the large-scale distribution of transcripts that can now be realized via the internet. Moreover, the format of those transcripts differed from researcher to researcher. In other words, there was no standardized way to transcribe children's speech, and no standardized way to code transcripts. The CHAT transcription and coding format has had a significant impact on the transcription and coding practices of child language researchers.

Moreover, collecting and transcribing a (longitudinal or cross-sectional) corpus is a time-consuming and costly enterprise. For instance, for the Spoken Dutch Corpus, a corpus of approximately 10 million spoken words collected between 2000 and 2005, the total budget was 5.5 million € (ca. 0.5 € per word). The required time investment for an orthographic transcription was estimated at 1:8 to 1:40, meaning that one minute of recorded speech needed from 8 to 40 minutes of actual transcription time. Thus for collecting a corpus of a reasonable size, the investment of time and resources is important, and largely transcends the possibilities of individual researchers or research groups. This implies that a maximal reuse of corpora is called for, that the public availability of the CHILDES corpora greatly extends the amount of data available to the individual researcher, and hence that the potential empirical coverage of theories of language acquisition is vastly increased.

CHILDES brought together several technological evolutions in the 1980ies and 1990ies. First of all the introduction of the personal computer, which allowed researchers to have the data they wanted to use "on their desktop", "at their fingertips". In addition, the electronic availability of corpora permitted the fast and accurate analysis of growing amounts of data in shrinking amounts of time. Secondly, the use of the internet made it possible to make data access almost instant: the CHILDES corpora and media can be downloaded, even browsed and analyzed directly over the internet.

Notwithstanding these enormous technological evolutions, one of the main problems of child language corpus studies remains the tedious and time consuming transcription stage. In this respect a further breakthrough is required so that the spoken language recordings of multiple speakers, can be automatically transcribed in a correct and reliable way. Only then the amount of data provided by CHILDES can increase exponentially, and follow the evolution witnessed in area of written language corpora.

Steven Gillis
University of Antwerp

See Also: Roger Brown, Corpus Based Methods

Further Readings

MacWhinney, Brian. (2000). *The CHILDES Project: Tools for analyzing talk. 3rd Edition*. Mahwah: Lawrence Erlbaum.

MacWhinney, Brian. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Corpora in language acquisition research* (pp. 165-197). Amsterdam/Philadelphia: Benjamins.

Rose, Yvan, MacWhinney, Brian, Byrne, Rodrigue, Hedlund, Greg, Maddocks, Keith,

O'Brien, Philip, & Wareham, Todd. (2006). Introducing Phon: A software solution for the study of phonological acquisition. In D. Bamman, T. Magnitskaia & C. Zaller (Eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development* (pp. 489-500). Somerville: Cascadilla Press.

Rose, Yvan, & MacWhinney, Brian. (2013). The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *Handbook of Corpus Phonology*. Oxford: Oxford University Press.

Rose, Yvan. (2013). Corpus-based investigations of child phonological development: Formal and practical considerations. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *Handbook of Corpus Phonology*. Oxford: Oxford University Press.

The Child Language Data Exchange System. http://childes.psy.cmu.edu/