



## Annotation of Child Language Corpora: A comparison of two methods with special emphasis on bimodal bilingual data

Diane Lillo-Martin & Debbie Chen Pichler  
Sign Linguistics Corpora Network  
Workshop 3: Annotation  
Stockholm, Sweden 14-16 June 2010

A handout to accompany this talk is available at:  
<http://web.me.com/dianelillomartin/DLM/Presentations.html>

1



## Acknowledgments

- Collaborators: Ronice Müller de Quadros and Julie Hochgesang
- Warm thanks to:
  - bimodal bilingual children and their families
  - research assistants
- Financial support from:
  - Award Number R01DC009263 from the National Institute on Deafness and Other Communication Disorders. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCD or the NIH.
  - The Gallaudet Research Institute.
  - CNPq (Brazilian National Council of Technological and Scientific Development) Grant #200031/2009-0 and #470111/2007-0.

2



## Longitudinal studies of child language (child language corpora)

- Address a wide variety of research questions
- Each dataset can be mined in many ways
- Complements experimental/cross-sectional study nicely

3



## Challenges of conducting child longitudinal studies

- Balance child's comfort zone and need for a representative sample of language
- Requires real creativity to coax a rich and varied sample out of child
  - Invest in time, get to know child and family, learn what gets them talking/signing
  - Thinking on your feet to follow the child's lead and expand on what the child says

4



## Collaborative story-telling

- Ben 051 (2;07)

*The movie has been deleted from the distribution file.*

5



## Challenges of conducting child longitudinal studies

- Let child do what she wants, yet make sure that conditions are maximized for later transcribability
  - Monitor ambient lighting and sound
  - Film child in rooms without places to hide or too much off-camera space

6

 **Data collection in the dark**

- SAL 002 (1;08)

*The movie has been deleted from the distribution file.*


7

 **Technological tools**


- JIL 019 (2;02)

*The movie has been deleted from the distribution file.*


8

 **Drawbacks of longitudinal spontaneous corpora**


- MacWhinney's (2001) three-headed monster of corpus transcription:
  - Lack of standard format + rapid proliferation of alternative formats
  - Indeterminacy
    - Difficult to determine what was really said/signed
  - Tedium
    - Highly labor-intensive, continually subject to revision and expansion




9

 **CHILDES: Child Language Data Exchange System**



- Started in the early 1980's by Brian MacWhinney and Catherine Snow (with others)
- Goal: to share child language data
- Method:
  - Develop computer software for storing and searching
  - Design conventions compatible with the software and teach these conventions
  - Convince researchers (over 100) to donate their data
  - Make the data freely available on the internet



10

 **CHILDES – Main Points**

- Three main components:
  - CLAN – Computerized Language Analysis
  - CHAT – Codes for the Human Analysis of Transcripts
  - Database (33 languages)
- Additional components
  - Ground rules
  - Guidelines for contributors
  - ...

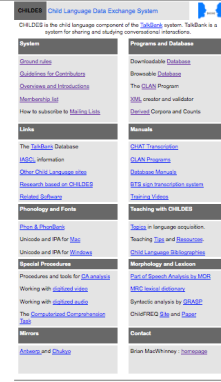



11

**CHILDES**

<http://childes.psy.cmu.edu/>

- System
- Programs and Database
- Links
- Manuals
- Phonology and Fonts
- Teaching with CHILDES
- Special Populations
- Morphology and Lexicon
- Mirrors
- Contact



This page has been accessed 194333 times since Oct 18, 2003

CHILDES is supported by grant R01-HD20868 from NIH/NIDA-D. [Donate](#)

German: [Language](#)

12



## CHILDES - Outcomes

- Major change in many areas of language acquisition research
  - Quantitative, systematic, wider range
- Over 3000 articles published based on CHILDES data (as of 2008)
- Over 1 million hits to website (early 2010)
- Continuing addition of data, increasing types

13



## Sample CHAT transcript

@Situation: CHI is looking at a picture book with MOT

\*CHI: xxx four .

\*CHI: I see four [/] I see four yyy .

%pho: ki:kæ

\*CHI: one # two # three +...

%act: pointing to picture book

\*MOT: those are bunnies .

\*MOT: what is [/] what are the bunnies doing ?

\*CHI: sleeping [?] .

%pho: sipi

%com: tilts head to one side, could be gesture for sleeping

\*CHI: it's dark out [=? darker]

%pho: da:kou

\*MOT: yes, it's time for bed .

\*CHI: goodnight bunnies !



14



## Systems for notation of child sign

- Most child sign researchers use variants of systems developed for adult signing
  - Baker, van den Bogaerde and Woll (2005) discuss many important general considerations
  - Morgan (2005) – Dynamic Space Transcription
  - Takkinen (2005) – HamNoSys
  - Slobin et al. (2001) – BTS (Berkeley Transcription System)

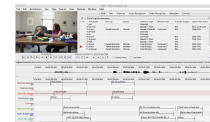


15



## Our goals

- Adoption of ID glossing (Johnston 1991)
  - Transcription focus is on annotating sign lemmas
- Transcription in ELAN
  - Transcript provides consistent information, sufficient for computerized searching
  - Basic transcription avoids analysis as much as possible
  - Analysis by researchers later, using transcript and video (unlike CHILDES, where analysis almost always based on transcript alone)



16



## MLSSA: Multi-Language Sign and Speech Annotation

- Conventionalized notation and procedures crucial for making tri-university collaboration possible (Bibibi project – Binational Bimodal Bilingual study of language acquisition)
- Specifically designed to accommodate bimodal data
  - speech and sign are annotated independently
  - bimodalism is identified at analysis level

17



## MLSSA Procedural conventions

- Lab manager trains transcribers and assigns and tracks progressive additions to transcripts, recorded in online logs accessible to all project members
- We transcribe speech first, as it often helps us identify accompanying signs
- Proofing
- Coding/Analysis

18


 **MLSSA Notational conventions: Comparison with CHILDES**

- MELISSA adopts many CHILDES conventions, but with slight modifications due to:
  - Requirements or capabilities of ELAN
  - Conventions specific to sign language glossing

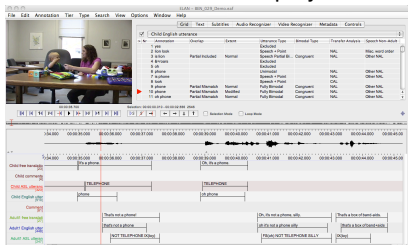
ASL utterance: g(hey) SEE FOUR[/] SEE FOUR YYY  
 Free translation: 'hey...I see four, I see four [garbled]'  
 IX(book)# FOUR DOG[?]  
 'There are four dogs there'  
 DOG[+] DV(sit-in-a-line)  
 'The dogs are lying in a line'




19

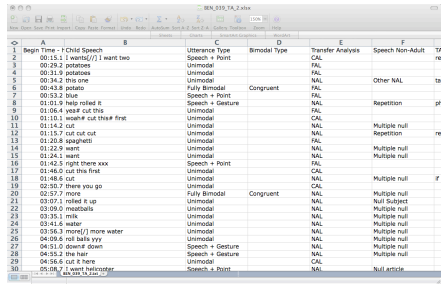
 **Use of MLSSA for research**

- Sample: BEN\_029 (2;01), 00:00:25 – 00:00:59
- Transcribed and coded for two projects.



20

 **Exportation of analysis tiers to Excel**



21

 **Current and future research**

Chen Pichler, Deborah, Quadros, Ronice Müller de, & Lillo-Martin, Diane (2009). Effects of Bimodal Production on Multi-Cyclicity in Early ASL and Libras. Boston University Conference on Language Development. Boston, MA; November 2009. In Jane Chandlee, Katie Franich, Kate Ierman, & Lauren Keil (Eds.), *A Supplement to the Proceedings of the 34th Boston University Conference on Language Development*, April 2010. <http://www.bu.edu/linguistics/BUCLD/supp34.html>.

Chen Pichler, Deborah, Hochgesang, Julie, Lillo-Martin, Diane & Quadros, Ronice (2010). Conventions for Sign and Speech Transcription in Child Bimodal Bilingual Corpora. *Language, Interaction and Acquisition* 1.1, 11-40. Special issue guest edited by Marie-Anne Salandre and Marion Blondel.


Lillo-Martin, Diane, Chen Pichler, Deborah, & Quadros, Ronice Müller de (2009). Best Practices for Building a Bimodal Bi-Lingual Bi-National Corpus of Child Language. Workshop on Sign Language Corpora: Linguistic Issues; London, UK; July 2009.

Lillo-Martin, Diane, Quadros, Ronice Müller de, Koulidobrova, Helen & Chen Pichler, Deborah (2009). Bimodal Bilingual Cross-Language Influence in Unexpected Domains. Generative Approaches to Language Acquisition; Lisbon, Portugal; September 2009. [Proceedings to appear; Cambridge Scholars Press]

Quadros, Ronice Müller de, Lillo-Martin, Diane, & Chen Pichler, Deborah (2010). Two Languages But One Computation: Code-Blending in Bimodal Bilingual Development. To be presented at the conference on Theoretical Issues in Sign Language Research; West Lafayette, IN; October 2010.

Quadros, Ronice Müller de, Lillo-Martin, Diane, Koulidobrova, Helen, & Chen Pichler, Deborah (in progress). Constraints on Cross-Language Influence, Code-Switching, and Code-Blending.

22

 **Works cited**

Baker, Anne, van den Bogaerde, Bepie & Woll, Bencie (2005). Methods and procedures in sign language acquisition studies. *Sign Language & Linguistics* 8:1/2, 7–58.

Johnston, T. (1991). Transcription and Glossing of Sign Language Texts: Examples from Australian Sign Language. *International Journal of Sign Linguistics* 2:1, 3–28.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3<sup>rd</sup> Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, Brian (2001). From CHILDES to TalkBank. In Almgren, M., Barreña, A., Ezeizabarrena, M., Idiazabal, I., & MacWhinney, B. (Eds.), *Research on Child Language Acquisition*, 17-34. Somerville, MA: Cascadia Press.

Morgan, Gary (2005). Transcription of child sign language: A focus on narrative. *Sign Language & Linguistics* 8:1/2, 117–128.

Slobin, D. I., Hoiting, N., Anthony, M., Biederman, Y., Kuntze, M., Undert, R., Pyers, J., Thumann, H., Weinberg, A. (2001). Sign Language Transcription at the Level of Meaning Components: The Berkeley Transcription System (BTS). *Sign Language & Linguistics* 4, 63–96.

Takkinen, Ritva (2005). Some observations on the use of HamNoSys (Hamburg Notation System for Sign Languages) in the context of the phonetic transcription of children's signing. *Sign Language & Linguistics* 8:1/2, 97–116.

23



24