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

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.
  – CNPq (Brazilian National Council of Technological and Scientific Development) Grant #200031/2009-0 and #470111/2007-0.

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

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

Collaborative story-telling
• Ben 051 (2;07)

The movie has been deleted from the distribution file.

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
Data collection in the dark

- SAL 002 (1;08)

The movie has been deleted from the distribution file.

Technological tools

- JIL 019 (2;02)

The movie has been deleted from the distribution file.

Drawbacks of longitudinal spontaneous corpora

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

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

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
  - ...
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

Sample CHAT transcript

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

*CH: **I see four**.  
*Sigh: kik-sector

*CH: **one** # **two** # **three** + ...  
*Act: pointing to picture book

*MOT: those are bunnies.  
*MOT: what is that what are the bunnies doing?  
*CH: sleeping [?].  
*Sigh: sips

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

*CH: it's dark out [+? darker].  
*Sigh: da-kou

*MOT: yes, it's time for bed.  
*CH: goodnight bunnies!

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)

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)

MLSSA: Multi-Language Sign and Speech Annotation

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

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
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/I SEE FOUR YYYY
Free translation: ‘hey…I see four, I see four’ [garbled]

IX(BOOK) FOUR DOG(?)
‘there are four dogs there’
DOG(+) DV(fit-in-a-line)
‘The dogs are lying in a line’

Use of MLSSA for research

- Sample: BEN_029 (2;01), 00:00:25 – 00:00:59
- Transcribed and coded for two projects.

Exportation of analysis tiers to Excel

Current and future research


Works cited


