Corpus based work with CHAT/CLAN: A small catalog

Annick De Houwer
Universität Erfurt

Kate corpus (bilingual Dutch-English child, audio-recorded between 2;7 and 3;4): first steps

- 1979-1980: typewriter transcription (no computer)
- intuitive transcription procedure
- line by line
- semi-orthographic but very close to oral features, including false starts, hesitations
- no codes other than speaker identification
- use of ! & ? to indicate intonation; . for end of utterance; , for pauses
- plus occasional comments
A: Daa! Waa's da?
K: Nog een kaaft.
A: Ja? Ne ziekwaggen haa? Waa doet die? Is't nie schoon?
K: Nee
A: Dat is schoon, haa?
K: Ja!
A: Ja! Welke kleur is da?
K: Moving
A: Movie
E: Moving? Ja (K makes crazy sounds)
K: She’s talking about the museum yesterday.
A: Ja?
E: (to K) Tell her about the things at the museum!
K: What am I doing? The red things or the black things?
A: Directing?
K: No, the black thing and the red thing.
A: The black thing and the red thing? (K confirms with a nod)
E: (talks about the things K saw at the museum and indeed, strange sculpture-like black and red things waving around)
P: O, that’s what she means!
Manual extraction of utterances of interest (typewriter)

Extracted utterances hand counted to yield analysis tables like this (no spreadsheets yet)
As of 1984: No more typewriters!

1985-86: transcript typed (!) into computer; codes; automatic analysis (D. Clynes)

```
ALL UTTERANCES WITH INAPPROPRIATE ELLIPSIS OF AN ARTICLE - FILE ellart.LST

Type 2

151. Emmertie van de seaside.
   !expanding A's previous utt.)
   < 6 MDD Oart NDIM prep DE NE AND NMF
   > UA CS NE ED XS XV XG OR P P P

125. Iceshow I see.
   < 4 EID Oart NMF Ipseg UFPrstem
   > UA CS NY NE EG XV XD D S UT
   178. I think, I'm think that iceshow's.
   < 9 EDD corr Ipseg UFPrstem Ipseg UFPraux UNFstem OR UNFinf demproi NMF VFP
   > UA C11 NY NE EG XG S V S V S SS GS VS

212. Ik heb iceshow gezien.
   ![back in room after being gone for a while]
   < 4 MDE Ipseg UFPraux Oart NE AND NMF UNFdx
   > UA CS NE NF EG XD S V D VRT

243. This is also green one.
   < 5 EDD demproi UFPrCop adv Adjcol ONE
   > UA CS NE EG XG S V A G G

271. Ik heb een nieuwe place to sit en also zetel.
   < 0 MID Ipseg UFPrstem EEN Adjdecl NMF prep inf conj adv Oart NMF
   > UA C2 NE EG XV XD S VT D D D D D LL A2 D2
```
1985
Sent by Brian to Belgium

1989-90: transformation into CHAT with later donation to CHILDES

Many conversations with Brian about how to go about it
Separate Development Hypothesis

• In production, preschool aged children with bilingual input from birth in each language separately fundamentally approach each morphosyntactic system as a separate set (counterevidence so far unknown)

• → this is only possible if children pay very close attention to the input (DH 1987, 1990)

• Using the CHAT version of the Kate corpus and CLAN I looked at Kate’s expression of past verb forms and compared her forms to those in the input
Following the input in BFLA (De Houwer 1997): category frequency

<table>
<thead>
<tr>
<th>Language</th>
<th>Perfect Tense</th>
<th>Simple Past</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate Dutch</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Adult Dutch</td>
<td>200</td>
<td>10</td>
</tr>
<tr>
<td>Kate English</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Adult English</td>
<td>130</td>
<td>20</td>
</tr>
</tbody>
</table>

Study of preschoolers growing up in a bidialectal environment (Antwerp, Belgium) [1994-1998]

Subjects
- N=11 (7 girls)
- age range at start of study: 4;3 - 4;11
- SES: 2 lower; 5 lower middle; 4 middle
- all children in 2nd year preschool

Data collection
- interviews with teachers and mothers
- 1 unstructured home visit by investigators yielding informal notes
- 3 participant observation sessions at school
- pre-structured diary logs kept by mothers
- 3 one-hour unstructured home audio-recordings
Data for 4 children in interaction with family members fully transcribed: CHAT --> on CHILDES > DutchAfrikaans
Broad phonetic transcription, by 2 transcribers (doubled up)
With supporting audio

What is the input to 4-year-olds like in family discourse in a bidialectal setting? (De Houwer, A., 2003. Language variation and local elements in family discourse. Language Variation and Change, 15: 327-347)

• adults use far more utterances with specifically local (dialectal) elements to each other than they do to young children
• specifically non-local elements are used almost exclusively in adult-CHI or CHI-CHI interaction
• 4-year-olds have linguistically different positions from the rest of the family: they use less of the local variety and are addressed less in local elements
The longitudinal project “Language Game” (with Marc H. Bornstein, NIH)

- studies firstborn children and their middle class families (all living in Belgium)
- 4 data collection points: 5, 13, 20, 53 months
- 30 monolingual (NL) families
- 31 bilingual (NL + FR) families: input in 2 languages from birth
- different types of data:
  - all ages: several questionnaires
  - 13 + 20 M: word lists (CDI)
  - all ages: home video- and audio-recordings by silent filer
  - 5 to 20 m: continuous log of CHI social contacts and which languages are spoken to the CHI
  - 53 m: 3 day log of CHI activities

Close transcription (CHILDES; MacWhinney 2000) of video-recordings

- CHI age 5m: part of the data transcribed (N = 15, monolingual NL families; no coding; 3x5 minutes of 60 min. video)
- CHI ages 13m + 20m: videotaped & transcribed MOT-CHI interaction in the home (N = 60/61)
  - 13m: 5/15 mins. joint play + 5/15 mins. eating transcribed
  - 20m: 5/10 mins. joint play + 5/10 mins. eating transcribed + 10 mins. solo play (not transcribed)
  - play sessions: same set of toys/books for all children at both ages
  - independent transcription and coding by 2 coders
- CHI age 53m: part of the data transcribed & coded (short narratives based on picture book done for all CHI)
Substudy focused on the maternal input in Dutch

- 16 MOT-CHI pairs in bilingual families where mothers addressed children in Dutch and 30 matched MOT-CHI pairs in monolingual families
- CHI ages 13 + 20 months; total of 20 mins. interaction
- 13 measures of input frequency:
  - each transcript separately: # turns, utterances, word types, word tokens, morphemes, syllables
  - calculations per transcript: average # utterances/turn, average # words/utterance, highest # words/utterance, average # syllables/utterance, highest # syllables/utterance
Some findings (1): expected discontinuity but stability in the amount of child directed speech over time (De Houwer 2014)

“Although all mothers talked more when children were older, mothers who were more talkative at the younger age were also more talkative at the older age” (p. 53)

<table>
<thead>
<tr>
<th>Bilingual families</th>
<th>13 months compared to 20 months</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word tokens</strong></td>
<td></td>
</tr>
<tr>
<td>highly significant correlation</td>
<td>( p = .87, p \leq .01, \text{two-tailed} )</td>
</tr>
<tr>
<td>N greater at 20 than 13 months</td>
<td>( t(15) = 8.12, p \leq .01 )</td>
</tr>
<tr>
<td><strong>Number of utterances</strong></td>
<td></td>
</tr>
<tr>
<td>highly significant correlation</td>
<td>( p = .77, p \leq .01, \text{two-tailed} )</td>
</tr>
<tr>
<td>N greater at 20 than 13 months</td>
<td>( t(15) = 7.40, p \leq .01 )</td>
</tr>
</tbody>
</table>

Some findings (2): extensive variation amongst mothers in the amount of child directed speech (De Houwer 2014)

<table>
<thead>
<tr>
<th>Extrapolated number of words per hour</th>
<th>13 &amp; 20 months combined</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bilingual families</strong></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>2,715</td>
</tr>
<tr>
<td>range</td>
<td>1,314 - 5,640</td>
</tr>
<tr>
<td><strong>Monolingual families</strong></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>2,673</td>
</tr>
<tr>
<td>range</td>
<td>723 – 4,407</td>
</tr>
</tbody>
</table>
Some findings (3): no differences between mothers in bilingual vs. monolingual families in the amount of child directed speech (De Houwer 2014)

<table>
<thead>
<tr>
<th>Sample result: average # of Dutch utterances/turn (20 mins.)</th>
<th>bilingual families</th>
<th>monolingual families</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>1.43</td>
<td>1.41</td>
</tr>
<tr>
<td>range</td>
<td>1.18 - 2.52</td>
<td>1.19 - 2.24</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.32</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Difference n.s., t(44) = -.25, p = .81

---

**A GIFT**

from Marc H. Bornstein & myself
Giittus
Merci!
deziękę
Danku!

CMU, June 7, 2019