Child language documentation

The sketch acquisition project

Birgit Hellwig (bhellwig@uni-koeln.de)
“7,099 known living languages”

Linguistic diversity

47% speaks 11 languages

96% speaks 250 languages

4% speaks the remaining ~ 7000 languages

Mandarin
Spanisch
Englisch
Hindi
Arabisch
Portugiesisch
Bengali
Russisch
Japanisch
Javanisch
Deutsch

53%
14%
7%
5%
4%
4%
3%
3%
3%
2%
1%
1%
Challenge

“to show how the child’s mind can learn and the adult’s mind can use, with approximately equal ease, any one of this vast range of alternative systems.” (447)
LINGUISTIC DIVERSITY in LANGUAGE EXPERIMENTS and CORPORA

PSYCHOLINGUISTICS CONFERENCES AND JOURNALS

CORPORA

English

German

proportion of unique mentions among 4007 abstracts

representation in Linguistic Data Consortium
22 languages in 550 corpora; weighted by number of corpora

Indo-European

Germanic (3 psycholinguistics; 3 corpora); Romance (5; 3); Slavic (4; 3); Indo-Iranian (1; 3)

Isolates

Korean, Basque

American Sign Language

Altaiic

Austro-Asiatic

Dravidian

Niger-Congo

Tai-Kadai

Uralic

(Anand, Chung & Wagers 2011: 3)
“If we take all the acquisition studies together (experiments and longitudinal studies), we know something about the acquisition of approximately 70 to 80 languages (i.e., approximately 1% of all the languages spoken today). This 1% of languages also includes languages for which only one acquisition study of a single feature exists […]” (p. 144)
Psycholinguistics in the field?

- “These conditions often make it difficult to follow the best-practice approaches to data collection which are commonly assumed in lab-based FLA research.” (Kelly et al. 2015: 287)

- “Still, specialized studies are perhaps best done in larger, less endangered language communities, especially given that many larger, unendangered language communities are also understudied.” (Whalen & McDonough 2015: 3)
Sketch Acquisition Project

• Language documentation
  – plus language acquisition & socialisation
• Project:
  – child corpus of manageable size
  – plus acquisition sketch
• Core group:
  Rebecca Defina, Birgit Hellwig, Shanley Allen, Lucy Davidson, Barb Kelly, Evan Kidd
Sketch Acquisition Project

Workshop series

Workshop on the acquisition of lesser-recorded languages
Cologne 25-26 January 2019
Programme

Workshop on the Acquisition of Lesser-studied Languages
Melbourne 9 August 2019

Participants: Lucy Davidson, Rebecca Defina, Maria Graziano, Birgit Hellwig, Caroline Jones, Dagmar Jung, Barb Kelly, Carmel O'Shannessy, Penelope Schmidt, Gianna Urbanczik, Gillian Wigglesworth, Wanyima Wighton

Archiving & Publishing

Piloting under way: Qaqt, Totoli, Pitjantjatjara, Inuktitut, Dëne Súłiné, Eegima, German, …
Acquisition Corpus

• Severe constraints on:
  – selection of participants
  – sampling intervals
  – amounts of data
  (e.g. Behrens 2008; Demuth 1996; 2008; Eisenbeiß 2006; 2010; Parisse 2019; Tomasello & Stahl 2004)

≈ 1 hour/week recording
  x 2 years (2;0-4;0)
  x 2 children = 208 hours

“[…] the majority of existing child speech samples […] represent only a very small proportion of all the language the child produces and hears – on average around 1%. […] and in some cases 1% sampling is not adequate to answer the question at hand.” (Tomasello & Stahl 2004: 118)
Sketch Corpus

Sketch corpus (ideal scenario): Longitudinal

<table>
<thead>
<tr>
<th>Age (±2 months)</th>
<th>2;0</th>
<th>2;6</th>
<th>3;0</th>
<th>3;6</th>
<th>4;0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child A</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
</tr>
<tr>
<td>Child B</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
</tr>
<tr>
<td>Total</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
</tr>
</tbody>
</table>

- 5 hrs analyzed (10 hrs recorded)
- Scenarious
  - longitudinal (2 children at 5 time points)
  - cross-lagged (2+ children at 2-3 time points each)
  - cross-sectional (10 children at 1 time point each)
### Sketch Corpus

**Sketch corpus (ideal scenario): Longitudinal**

<table>
<thead>
<tr>
<th>Age (±2 months)</th>
<th>2;0</th>
<th>2;6</th>
<th>3;0</th>
<th>3;6</th>
<th>4;0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child A</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
</tr>
<tr>
<td>Child B</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
<td>30(60)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
</tr>
</tbody>
</table>

**More feasible example (Qaqt): Cross-lagged**

<table>
<thead>
<tr>
<th>Age (±2 months)</th>
<th>2;0</th>
<th>2;6</th>
<th>3;0</th>
<th>3;6</th>
<th>4;0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZDL</td>
<td>2;0</td>
<td>2;8</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>YJL</td>
<td>–</td>
<td>–</td>
<td>3;1</td>
<td>3;7</td>
<td>4;0</td>
</tr>
<tr>
<td>YDS</td>
<td>2;0</td>
<td>2;6</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>YRA</td>
<td>–</td>
<td>–</td>
<td>3;2</td>
<td>3;7</td>
<td>4;0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
<td>60(120)</td>
</tr>
</tbody>
</table>
Example

**Qaqet**

Papuan (Baining)

PNG, East New Britain

~ 15,000 speakers

http://qaqet.phil-fak.uni-koeln.de/
- 2014-2022 (VolkswagenStiftung)
- Longitudinal study

<table>
<thead>
<tr>
<th>Child</th>
<th>Age range</th>
<th>Recorded</th>
<th>Annotated</th>
<th>Partly annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZJS</td>
<td>4;3-7;10</td>
<td>112 hrs</td>
<td>37 hrs</td>
<td>38 hrs</td>
</tr>
<tr>
<td>YJL</td>
<td>2;8-6;3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZDL</td>
<td>0;7-4;3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YRA</td>
<td>3;2-4;7</td>
<td>120 hrs</td>
<td>37 hrs</td>
<td>45 hrs</td>
</tr>
<tr>
<td>YDS</td>
<td>1;11-4;10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>232 hrs</td>
<td>74 hrs</td>
<td>83 hrs</td>
</tr>
</tbody>
</table>

+ 5 families dropped out (59 hrs); + 2 families from multilingual subcorpus (151 hrs)
## Sketch corpus

More feasible example (Qaqt): Cross-lagged

<table>
<thead>
<tr>
<th></th>
<th>ZDL</th>
<th>YDS</th>
<th>YJL</th>
<th>Minutes</th>
<th>IUs (Total)</th>
<th>IUs (Adults)</th>
<th>IUs (Children)</th>
<th>IUs (Focal child)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2;0</td>
<td>2;8</td>
<td>–</td>
<td>60</td>
<td>1704</td>
<td>569</td>
<td>1135</td>
<td>452</td>
</tr>
<tr>
<td></td>
<td>2;0</td>
<td>2;6</td>
<td>–</td>
<td>60</td>
<td>1454</td>
<td>228</td>
<td>1226</td>
<td>510</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>3;1</td>
<td>60</td>
<td>2405</td>
<td>585</td>
<td>1820</td>
<td>952</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>3;7</td>
<td>60</td>
<td>1145</td>
<td>136</td>
<td>1009</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>4;0</td>
<td>60</td>
<td>1462</td>
<td>139</td>
<td>1323</td>
<td>711</td>
</tr>
</tbody>
</table>
Tasks

Village
- in/around the house
- + many different interlocutors (adults, children)

Garden
- in garden or garden hut
- + few interlocutors (a parent, a sibling)

Missing setting
- children alone in the bush
Case study: Child-directed speech

- **CDS register:**
  - short, correct & complete; few hesitations & errors
  - exaggerated pitch, high F0, long duration, more pauses
  - restricted vocabulary, here & now
  - nursery vocabulary
  - many questions & imperatives
  - repetitions & variations

Universality?
Father: They will have to find money for a chicken, a small chicken.
Boy [tries to scare his baby brother away]
Boy [to uncle]: Look over there, at the baby’s head.
Father: Like this.
Uncle: The chickens used to be here.
Mother: Those ones there.
Boy [to uncle]: Uncle, he fell down.
Father: Is that our custom, or what?
Uncle: It isn’t.
Father: It’s really not our custom.
Uncle: The custom won’t kill you. Last time I simply gave away a chicken.
“Speech directed to children, and not overheard speech, predicts both children’s later vocabulary […].” (Shneidman & Woodward 2016: 3)

But (Shneidman et al. 2009: 276):

[Diagram showing the process:
A child is exposed to overheard speech in her everyday life → The child forms attention strategies that are adaptive to her role as an observer → The child employs these attention strategies when overhearing conversations → Child is able to learn novel words through overhearing]
CDS in Qaqet

<table>
<thead>
<tr>
<th></th>
<th>Sketch corpus (5 hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child-directed IUs</td>
<td>3.662</td>
</tr>
<tr>
<td>Speaker: Adult</td>
<td>1.323 36%</td>
</tr>
<tr>
<td>Speaker: Older child</td>
<td>2.339 64%</td>
</tr>
</tbody>
</table>

- Controlled study (The Qaqet Pear Story Corpus)
  - longer pauses, higher pitch & greater frequency range
  - fewer disfluencies & hesitations
  - short, less complex, lower MLU
  - (mostly) correct and complete
  - more imperatives & questions

Sketch corpus: Prosody

*ip ngulidresiit nanget nana? paqani nana?*  
‘when do they tell stories? at what time?’

*kua nyitlamaivip? amaivivim?*  
‘do you see the snakes? the two snakes?’
## Sketch corpus: Morphology

<table>
<thead>
<tr>
<th>ADS</th>
<th>CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART=noun</td>
<td>ART=noun ~ noun</td>
</tr>
<tr>
<td>SBJ=verb</td>
<td>SBJ=verb ~ verb</td>
</tr>
<tr>
<td></td>
<td>+ morphophonology</td>
</tr>
</tbody>
</table>

**aɣuukuka** ‘sweet potato’

**a=kuukuk-ka** ‘NM=sweet.potato-SG.M’

---

**YDS**
- *kaakak*
- ee, *kuukuka*

**AMT (mother)**
- *kaakak*
- *akuukuka*
- ee

**YRA**
- *kaakak*
- ussh!

**YDS**
- sweet potato
- yes, sweet potato

**AMT**
- sweet potato
- yes

**YRA**
- sweet potato
- yes

**YDS**
- ussh!

---

**YRA**
- ussh!
Correct, Imitation, Laughter
<table>
<thead>
<tr>
<th>YRA (3;2) to YDS (2;0)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>nyikut</strong></td>
<td>you dig</td>
</tr>
<tr>
<td><strong>unekut</strong></td>
<td>we two dig</td>
</tr>
<tr>
<td><strong>nyikut</strong></td>
<td>you dig</td>
</tr>
<tr>
<td><strong>nyikut</strong></td>
<td>you dig</td>
</tr>
<tr>
<td><strong>nyikut</strong></td>
<td>you dig</td>
</tr>
<tr>
<td><strong>YDS, nyikut</strong></td>
<td>YDS, you dig</td>
</tr>
<tr>
<td><strong>nyikut tamasinep</strong></td>
<td>you dig for spiders</td>
</tr>
<tr>
<td><strong>sinep</strong></td>
<td>spiders</td>
</tr>
<tr>
<td><strong>nyikut iara</strong></td>
<td>you dig here</td>
</tr>
<tr>
<td><strong>nyikura</strong></td>
<td>you dig now</td>
</tr>
</tbody>
</table>
“partial repetitions […], with changes in lexical items, grammatical morphology, and/or word order, maintaining a constant communicative intent” (Küntay & Slobin 1996: 267).

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Verb</th>
<th>Object</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td></td>
<td>you dig</td>
</tr>
<tr>
<td><em>une</em></td>
<td><em>kut</em></td>
<td></td>
<td>we two dig</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td></td>
<td>you dig</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td></td>
<td>you dig</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td></td>
<td>you dig</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td><em>tamasinep</em></td>
<td>you dig for spiders</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>sinep</em></td>
<td>spiders</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kut</em></td>
<td></td>
<td>you dig here</td>
</tr>
<tr>
<td><em>nyi</em></td>
<td><em>kur</em></td>
<td><em>iara</em></td>
<td>you dig now</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>a</em></td>
<td></td>
</tr>
</tbody>
</table>
# Repetition & Variation

**Sketch Corpus (5 hrs)**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Speaker: Adult</th>
<th>Speaker: Older child</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS</td>
<td>3.662</td>
<td>1.323</td>
<td>2.339</td>
</tr>
<tr>
<td>Varied repetitions</td>
<td>895</td>
<td>503</td>
<td>39%</td>
</tr>
<tr>
<td>Exact repetitions</td>
<td>352</td>
<td>98</td>
<td>7%</td>
</tr>
</tbody>
</table>

**Longitudinal Corpus (23 hrs)**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Speaker: Adult</th>
<th>Speaker: Older child</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS</td>
<td>12.968</td>
<td>4.142</td>
<td>8.826</td>
</tr>
<tr>
<td>Varied repetitions</td>
<td>2.724</td>
<td>1.291</td>
<td>31%</td>
</tr>
<tr>
<td>Exact repetitions</td>
<td>1.315</td>
<td>227</td>
<td>5%</td>
</tr>
</tbody>
</table>

Literature: ages 2-4: ~25-30% variations, <1% exact repetitions
Repetition & Variation

AMT (mother) to YDS & YJS

- **utit** we go
- **utiravit** we go up
- **utiravit, serama akun** we go up, into the corn
- **YDS, utit maavit** YDS, we go up here
- **utit** we go
- **YJS, utiravit** YJS, we go up

YRA (3;3) to YDS & YJS

- **guaki, nyan** my friend, come
- **nyan** come
- **YJS, nyan** YJS, come
- **nyan** come
- **nyan** come
- **guaki, nyan** my friend, come
+ many more topics

- Conversational routines
- Lexicon: early words, word classes (nouns/verbs)
- Phonology: first sounds, realization of target sounds, phonological processes
- Morphology: noun/verb morphology, non-target-like forms, MLU
- Combinations: early combinations of words, word order, argument realization
- ...
A note of caution

• Limited database – impacts on the kinds of statements we can make:
  – descriptions of what the children hear & do & say – not of what they know
  – clear focus on qualitative analyses – not quantitative
  – raise topics/questions, to be explored and tested in a larger data set

• Major contribution: to broaden our understanding of the problem space
However, despite the promise of this earlier typologically diverse work, when we fast-forward 30 years we do not see a boom in the study of FLA in typologically diverse languages and culturally different communities [...].” (Kelly & Nordlinger 2014: 180)
Danke – Amatlungena – Tenkyu tru