Acquisition Sketch Project Meeting 2: The corpus

19th July 2023

Sketch Corpus

https://acquisition-sketch.phil-fak.uni-koeln.de/sketch-corpus

Key parameters

Number of children	minimally 2
Ages of children	5 ages - approx. 2, 2½, 3, 3½, 4
Gender of children	attempt a mix of genders
Type of data	preferably natural(istic) interaction with other children and adults
Recording schedule	longitudinal, cross-lagged or cross-sectional
Amount of data to be recorded	minimally 10 hours (= 60 minutes per child and age)
Amount of data to be processed	minimally 5 hours (= 30 minutes per child and age)

Sketch Corpus – Motivations for this setup

- 2;0-4;0: time when large parts of a language are being acquired
  - at 2;0: small vocabulary of first words used in utterances of two or more morphemes
  - at 4;0: major nuts and bolts will be in place (large vocabulary, multi-word utterances, productively using major parts of morphology)
- individual differences:
  - attempting a mix of genders
  - approximating a longitudinal setup as closely as possible (as it allows us to look at a child's development at different points in time)
- natural(istic) interaction:
  - insights into socialization practices
  - access to the ambient language

Sketch scenarios: ideal longitudinal study

Table 1. Sketch corpus: Longitudinal scenario.

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
Child A	30(60)	30(60)	30(60)	30(60)	30(60)
Child B	30(60)	30(60)	30(60)	30(60)	30(60)
Total	60(120)	60(120)	60(120)	60(120)	60(120)

Note. Minutes of transcribed language (suggested recording length in brackets).

Sketch scenarios: Cross-sectional study

Table 2. Sketch corpus: Cross-sectional scenario.

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
Child A	30(60)				
Child B	30(60)				
Child C		30(60)			
Child D		30(60)			
Child E			30(60)		
Child F			30(60)		
Child G				30(60)	
Child H				30(60)	
Child I					30(60)
Child J					30(60)
Total	60(120)	60(120)	60(120)	60(120)	60(120)

Note. Minutes of transcribed language (suggested recording length in brackets).

An intermediate solution is to adopt a ***cross-lagged approach.***

Sketch scenarios: realities

- Many reasons for corpus structure.
  - Existing data.
  - Availability of children.
  - Intermittent access to field site.
  - Length of data collection phase of project.

- We suggest (again) a golden guiding principle of writing a sketch – *any data is better than no data.*


Pitjantjatjara

Pitjantjatjara Sketch Corpus **minutes annotated (minutes recorded)**

Age (±2 months)	2;0	2;6	2;9	3;0	3;6	4;0
Anne	30(117)					
Andrew	30(72)	30(111)	30(58)	30(80)		
Frank		30(182)				60(162)
Rachel				30(42)	30(74)	
Isy					30(58)	60(162)
Total	60(189)	60(293)		60(122)	60(132)	60(162)


Qaqet

Qaqet Sketch Corpus. **minutes annotated (minutes recorded)**

Age (±2 months)	2;0	2;6	3;0	3;1	3;2	3;6	4;0
ZDL (male)	30(245)	30(134)					
YDS (female)	35(303)	28(152)					
YJL (female)			30(362)			30(301)	30(100)
YRA (male)				34(95)		28(360)	28(357)
Total	65(548)	58(286)	64(457)			58(661)	58(457)

Extracted from larger corpus – but limited by what was already transcribed

Considerations:

- match ages as closely as possible
- each child for at least 2 consecutive age points
- mix of genders

Qaqet Sketch Corpus. other children, adults, teens

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
ZDL (male)	1 + 2	1 + 1			
YDS (female)	1 + 2	2 + 2			
YJL (female)			2 + 2	2 + 2 + 1	2 + 1
YRA (male)			1 + 3 + 1	1 + 1	2 + 1

Further considerations:

- further participants

Qaqet Sketch Corpus.

Age (±2 months)	2;0	2;6	3;0	3;6	4;0
ZDL (male)	**village**	**village**			
YDS (female)	**garden**	**garden**			
YJL (female)			**village**	**village**	**village**
YRA (male)			**village**	**garden**	**garden**

Further consideration:

- mix of settings

Village
in/around the house

Bush (missing setting)
children alone in the bush

Garden
in family garden or garden hut


Murrinhpatha

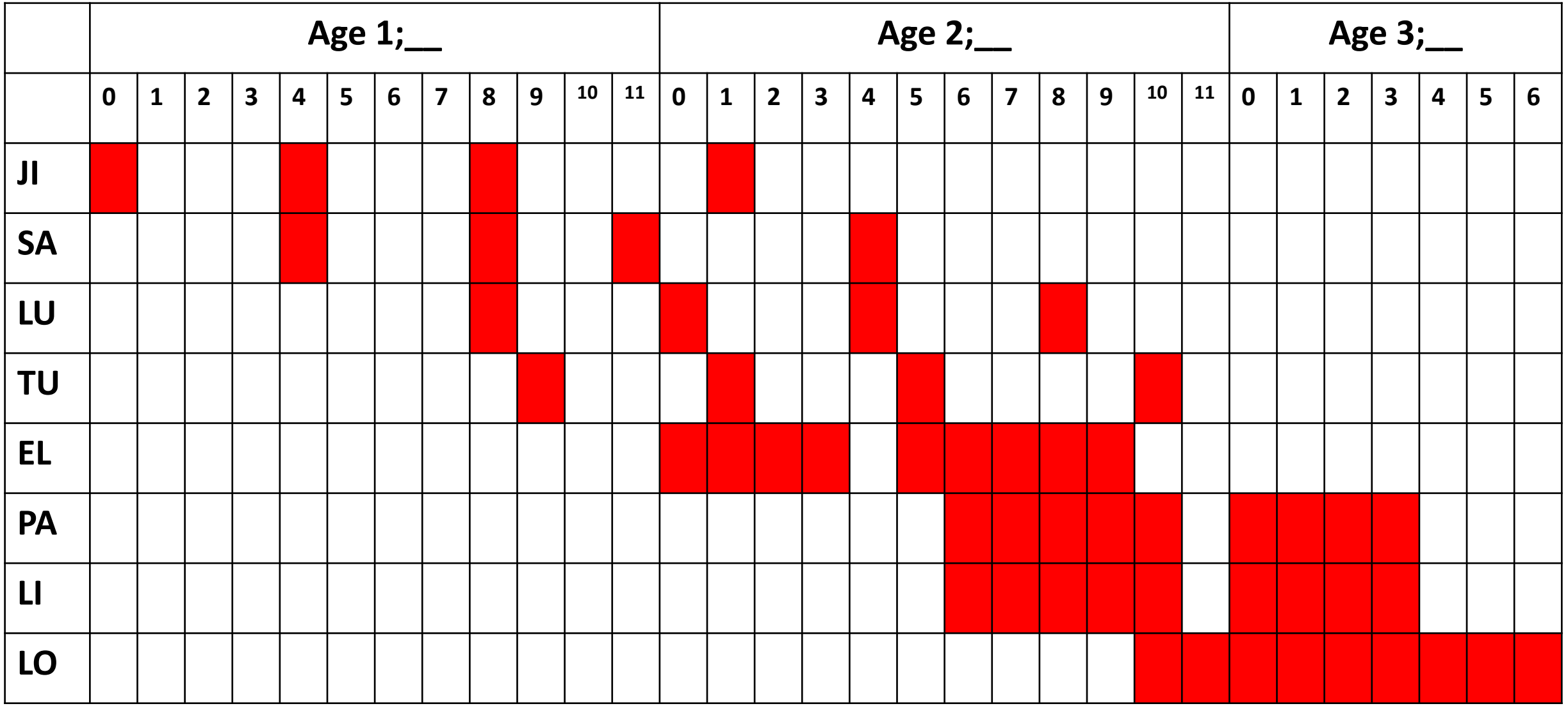
Murrinhpatha Sketch Corpus

minutes annotated (minutes recorded)

Age	**2;6**	**3;0**	**3;6**	**4;0**	**4;6**
(±2 months)	2;7	3;2	3;8	4;1	4;3
Emily (F)	30(54)	30(86)	30(104)	30(54)	30(80)
Acacia (F)	30(86)	30(59)	30(58)	30(44)	30(33)
TOTAL	60(140)	60(145)	60(162)	60(98)	60(113)

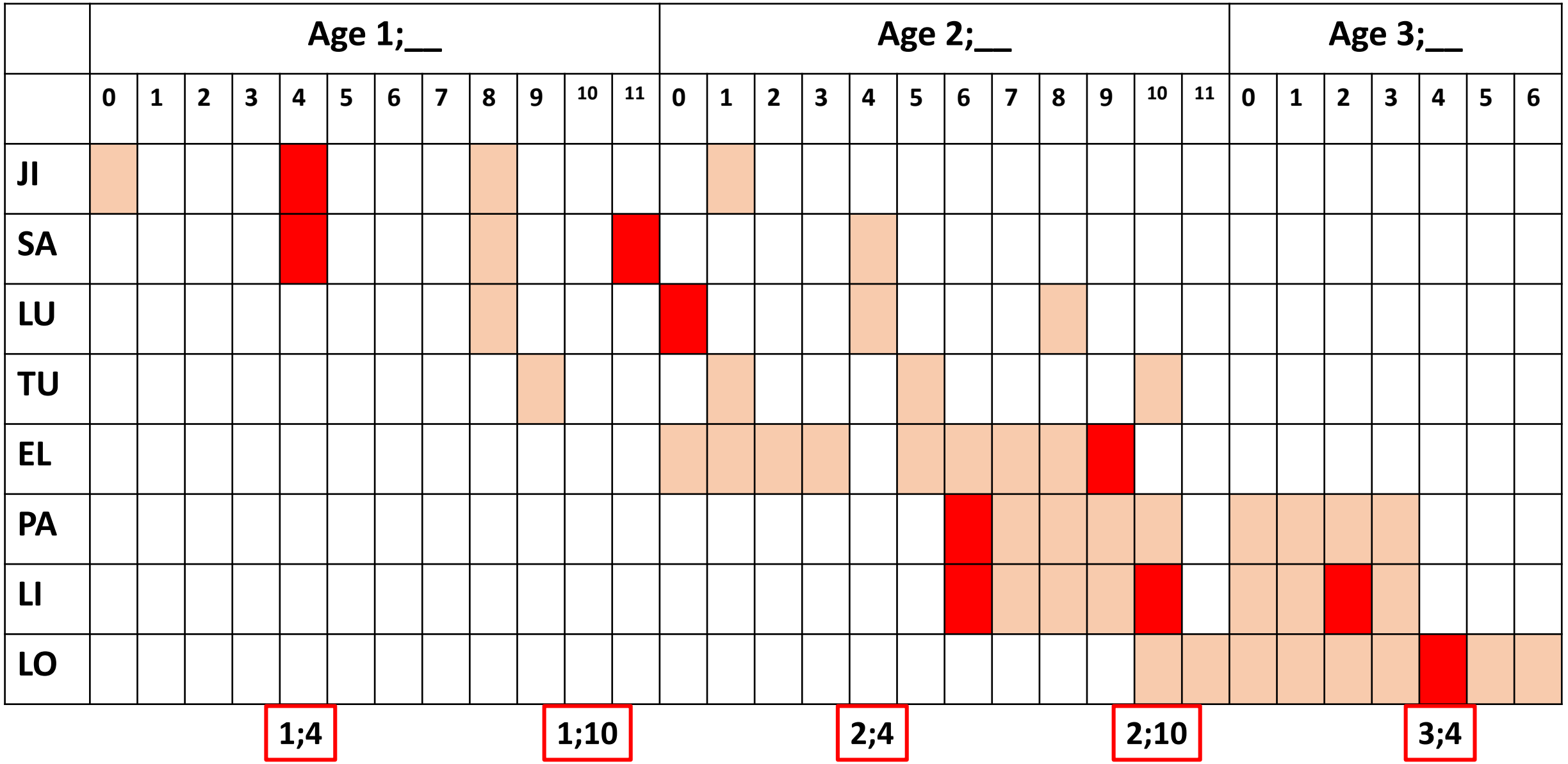
Inuktitut

Inuktitut Full Data Set (recorded 1986-1990)



Recorded ± 4 hours of data per child per age, transcribed ± 2 hours of data per child per age

Inuktitut Sketch Data Set (30 mins from each session marked in red)



Inuktitut Sketch Corpus

Age [± 2 mos]	**1;4** [1;4 / 1;4]	**1;10** [1;11 / 2;0]	**2;4** [2;6 / 2;6]	**2;10** [2;9 / 2;10]	**3;4** [3;2 / 3;3]
Jini (F)	30 (240)				
Sarah (F)	30 (240)	30 (240)			
Lucasi (M)		30 (240)			
Paul (M)			30 (343)		
Lizzie (F)			30 (246)	30 (193)	30 (286)
Elijah (M)				30 (241)	
Louisa (F)					30 (227)
Total	60 (480)	60 (480)	60 (589)	60 (434)	60 (513)

Considerations:

- Only one child available older than 3;3 → limited upper age
- Younger data collected only every 4 months → limited possibilities for younger ages
- Only considered sessions including caregiver speech (many sessions were peers only)
- Prioritized sessions with lots of utterances
- All recordings are indoors (outdoors too windy / children too dispersed)

Website (under construction)

https://acquisition-sketch.phil-fak.uni-koeln.de/sketch-corpus

Information about sketch corpora – including examples and archiving options

Sketch Corpus

The sketch corpus consists of minimally five hours of annotated and archived data. The first part of our manual provides step-by-step suggestions for collecting, archiving and processing the sketch data.

Key parameters

Number of children	minimally 2
Ages of children	5 ages - approx. 2, 2½, 3, 3½, 4
Gender of children	attempt a mix of genders
Type of data	preferably natural(istic) interaction with other children and adults
Recording schedule	longitudinal, cross-lagged or cross-sectional
Amount of data to be recorded	minimally 10 hours (= 60 minutes per child and age)
Amount of data to be processed	minimally 5 hours (= 30 minutes per child and age)

Example sketch corpora

Archiving