

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
<u>Recording schedule</u>	longitudinal, cross-lagged or cross-sectional
<u>Amount of data</u> to be recorded	minimally 10 hours (= 60 minutes per child and age)
<u>Amount of data</u> 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.*

A wide, flat, reddish-brown landscape, likely a dry lake bed or salt flat, dominates the foreground and middle ground. In the background, there are several hills with sparse, dry vegetation in shades of brown and green. The sky is a clear, pale blue. Two people are walking away from the camera in the lower center of the frame, their shadows cast long and dark on the ground. The overall scene is arid and open.

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)

A young child with short hair, wearing a blue sleeveless shirt and dark shorts, is sitting on a light-colored mat on the ground. The child is positioned in the lower-left quadrant of the frame, facing right. The background is a dense, lush green environment with various plants and trees. In the distance, a body of water is visible, partially obscured by the foliage. The overall scene is bright and natural. The word "Qaqet" is overlaid in white text in the center of the image.

Qaqet

Qaqet Sketch Corpus.

minutes annotated (**minutes recorded**)

Age (± 2 months)	2;0	2;6	3;0	3;6	4;0
			3;1	3;2	
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



Murrinpatha Sketch Corpus

minutes annotated (**minutes recorded**)

Age (±2 months)	2;6 2;7	3;0 3;2	3;6 3;8	4;0 4;1	4;6 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)

	Age 1;__												Age 2;__											Age 3;__							
	0	1	2	3	4	5	6	7	8	9	10	11	0	1	2	3	4	5	6	7	8	9	10	11	0	1	2	3	4	5	6
JI	█				█				█					█																	
SA					█				█			█					█														
LU									█			█	█				█				█										
TU										█				█				█					█								
EL													█	█	█	█		█	█	█	█										
PA																					█	█	█		█	█	█	█			
LI																					█	█	█	█	█	█	█	█			
LO																							█	█	█	█	█	█	█	█	█

Recorded \pm 4 hours of data per child per age, transcribed \pm 2 hours of data per child per age

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

- 1 Introduction
- 2 Corpus construction
 - 2.1 Structure of the sketch corpus
 - 2.1.1 Ages and number of children
 - 2.1.2 Amount of data
 - 2.1.3 Participants and content
 - 2.1.4 Rationale for the setup and further reading
 - 2.2 Practical considerations of corpus construction
 - 2.2.1 Getting started: Identifying children and contexts
 - 2.2.2 Recording setup
 - Recording audio data
 - Recording video data
 - Notes on long recordings
 - 2.2.3 Archiving and metadata
 - 2.2.4 Ethical considerations
- 3 Data processing
 - 3.1 Preparing files and transcription sessions
 - 3.1.1 Transcribers
 - 3.1.2 Segmentation
 - 3.1.3 Tiers
 - 3.2 Transcription and translation
 - 3.2.1 Deciphering utterances
 - 3.2.2 Transcription and the adult interpretation
 - 3.2.3 Transcription as data collection
 - 3.3 Beyond transcription
- 4 Summary
- References

Table of contents