Acquisition Sketch Project Meeting 7: Archiving

8th May 2024

Outputs

- Sketch Corpus
- Acquisition Sketch
- Community Materials

Dual role:

- Primary data for the other two
- Output in its own right

→ Archiving

Why archive the sketch corpus?

- Rich resource – usable for multiple purposes
- Academic purposes, e.g.:
    - selected phenomena within a corpus
    - micro-variation across corpora of related/neighbouring languages
    - typological variation across corpora
    - …
- As well as applied and community-oriented purposes, e.g.
    - identifying vocabulary and structures for children's books
    - insights into culturally appropriate learning environments
    - …

Why archive the sketch corpus?

- Rich resource – usable for multiple purposes

  → Preserve sketch data in the face of language endangerment

- Scientific practice – transparency of data and analysis

  → Numerous data repositories

FAIR principles:

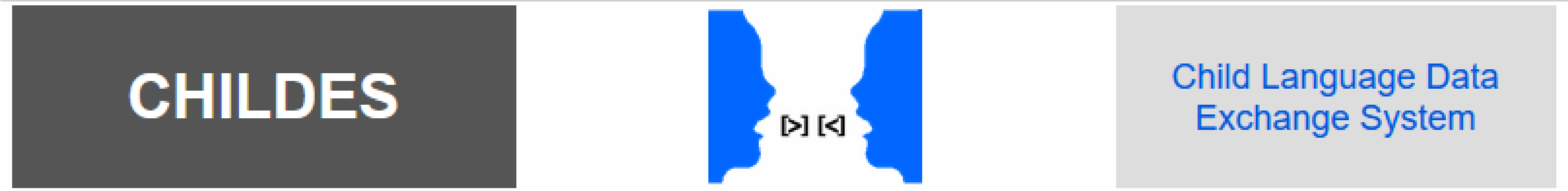
- Findability
- Accessibility
- Interoperability
- Reusability

CARE principles:

- Collective Benefit
- Authority to Control
- Responsibility
- Ethics

Major repositories – language acquisition

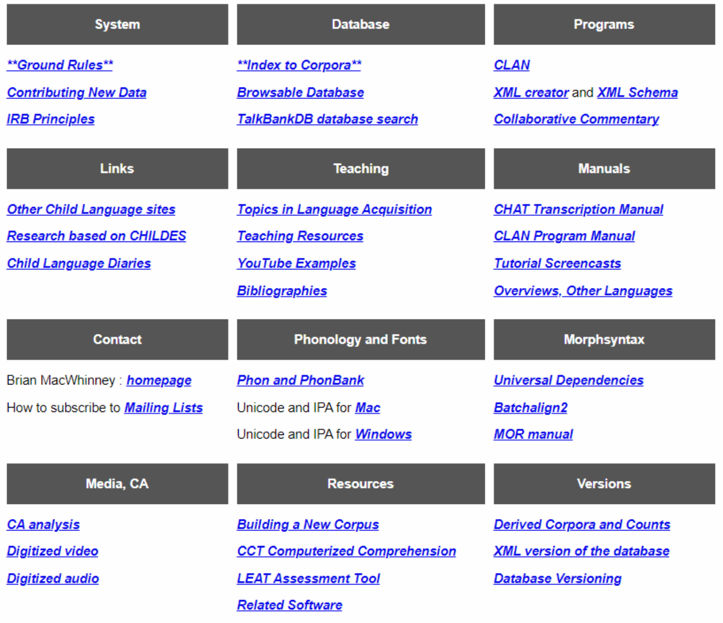
https://childes.talkbank.org/



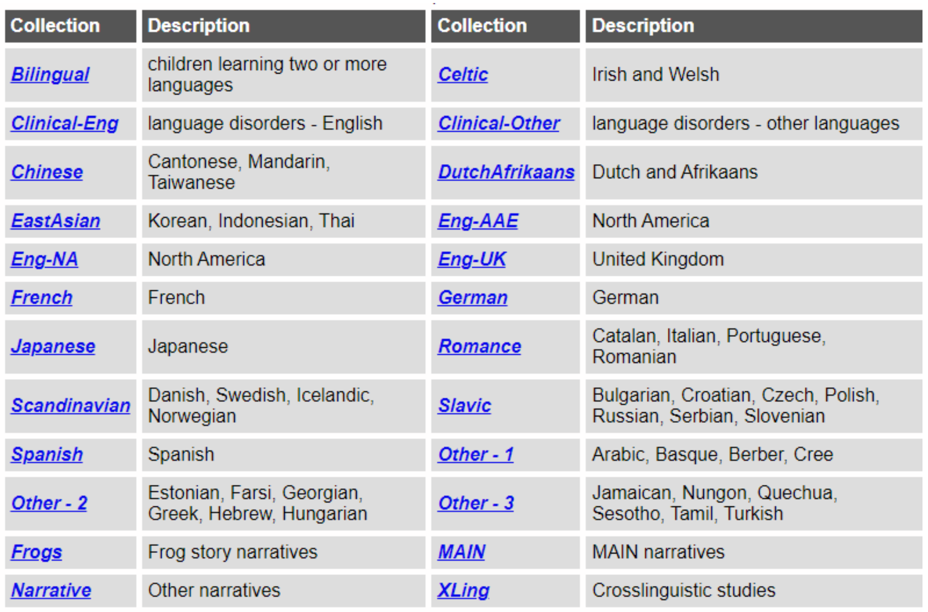
CHILDES

Child Language Data Exchange System

MacWhinney, Brian. 1991. *The CHILDES Project: Tools for Analyzing Talk (1st edition)*. Hillsdale, NJ: Lawrence Erlbaum.

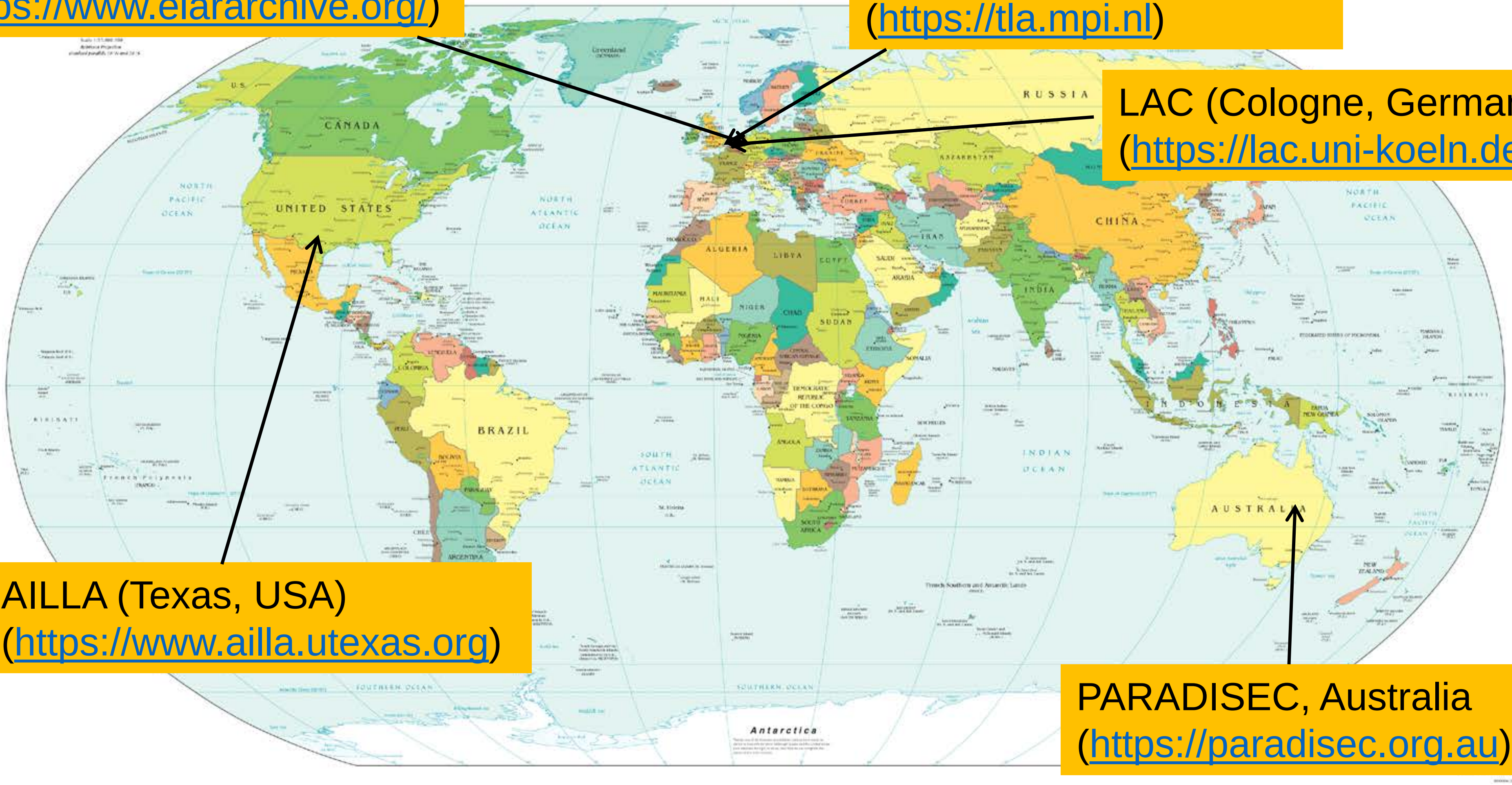


System	Database	Programs
Ground Rules	**Index to Corpora**	CLAN
Contributing New Data	Browsable Database	XML creator and XML Schema
IRB Principles	TalkBankDB database search	Collaborative Commentary
Links	**Teaching**	**Manuals**
Other Child Language sites	Topics in Language Acquisition	CHAT Transcription Manual
Research based on CHILDES	Teaching Resources	CLAN Program Manual
Child Language Diaries	YouTube Examples	Tutorial Screencasts
	Bibliographies	Overviews, Other Languages
Contact	**Phonology and Fonts**	**Morphsyntax**
Brian MacWhinney : homepage	Phon and PhonBank	Universal Dependencies
How to subscribe to Mailing Lists	Unicode and IPA for Mac	Batchalign2
	Unicode and IPA for Windows	MOR manual
Media, CA	**Resources**	**Versions**
CA analysis	Building a New Corpus	Derived Corpora and Counts
Digitized video	CCT Computerized Comprehension	XML version of the database
Digitized audio	LEAT Assessment Tool	Database Versioning
	Related Software	



Collection	Description	Collection	Description
Bilingual	children learning two or more languages	Celtic	Irish and Welsh
Clinical-Eng	language disorders - English	Clinical-Other	language disorders - other languages
Chinese	Cantonese, Mandarin, Taiwanese	DutchAfrikaans	Dutch and Afrikaans
EastAsian	Korean, Indonesian, Thai	Eng-AAE	North America
Eng-NA	North America	Eng-UK	United Kingdom
French	French	German	German
Japanese	Japanese	Romance	Catalan, Italian, Portuguese, Romanian
Scandinavian	Danish, Swedish, Icelandic, Norwegian	Slavic	Bulgarian, Croatian, Czech, Polish, Russian, Serbian, Slovenian
Spanish	Spanish	Other - 1	Arabic, Basque, Berber, Cree
Other - 2	Estonian, Farsi, Georgian, Greek, Hebrew, Hungarian	Other - 3	Jamaican, Nungon, Quechua, Sesotho, Tamil, Turkish
Frogs	Frog story narratives	MAIN	MAIN narratives
Narrative	Other narratives	XLing	Crosslinguistic studies

Major repositories – language documentation



ELAR (Berlin, Germany)
(https://www.elararchive.org/)

TLA (MPI, Netherlands)
(https://tla.mpi.nl)

LAC (Cologne, Germany)
(https://lac.uni-koeln.de)

AILLA (Texas, USA)
(https://www.ailla.utexas.org)

PARADISEC, Australia
(https://paradisec.org.au)

Which archive?

- Good news: you can’t go wrong – all the above language acquisition/ documentation repositories are suitable choices
- Recommendation:
  - Your prior experience and familiarity – with a particular archive, data processing setup, research community/network etc.
- Institutional cooperation with the *Language Archive Cologne* (LAC)
  - Archives the sketch corpus (if you cannot or don’t want to archive elsewhere)
  - Compiles information on all sketch corpora archived at the LAC and elsewhere

Data Center for the Humanities + 〉 *Bestände* + 〉 *Forschungsdatenbestände* + 〉 *Language Data*

Language Data

↓ Language corpora and datasets
↓ External language corpora and datasets
↓ Language corpora and datasets in development

Which archive?

- Archives have their own standard setups and procedures wrt:
  - file and data formats
  - implementing access rights and restrictions
  - providing and structuring metadata information
  - …

  → Discuss the requirements and possibilities early on in your project
- Setups and procedures are mostly applicable to child language data, too

  … but be aware of some special requirements for child data

Child language data

1) Sensitivity of data

2) More extensive metadata

1) Sensitivity of data

- Data often cannot be made publically available:
  - data from minors
  - recordings capture unguarded informal day-to-day interaction
  - metadata captures information on the linguistic and non-linguistic development of the children and on their social networks

→ Think carefully about access rights & restrictions, and about anonymization or pseudonymization

1) Sensitivity of data (ctd.)

- Who has access to what?
- Consent to archive the data
  - access for specific scientific and/or community-related purposes only?
  - access to parts of the data (e.g. only the transcript)?
  - anonymization or pseudonymization?
  - re-negotiating informed consent with the children once they come of age?

NB: Explore the tools that are being developed within the ViCom (Visual Communication) network for anonymizing faces and voices
https://vicom.info/vicom-data-network/

1) Sensitivity of data (ctd.)

- Who has access to what?
- Consent to archive the data
- Beyond archives
    - families (and other participants) – for their own record, but in which form?
    - other community members – e.g. to show videos when recruiting new participants? or to re-use data for community materials?
    - transcribers and translators?

2) Extensive metadata (= data about data)

- Structured metadata recommended by the archive
  - e.g. who are the participants, where and when did the recording take place, which language, which topic? etc.
- Need for more extensive metadata than is typical for adult corpora
  - because this information directly impacts the interpretability of the child data
- Metadata collected during:
  - preparatory stages of corpus construction (→ The SAM, Part II, Section 4; Meeting 7, June 12)
  - transcription process (→ The SAM, Part I, Section 3.2.3; Meeting 6, April 17)

2) Extensive metadata: Dossier of child

(i) Assign an ID or pseudonym, and make sure to use this in publications to protect privacy. The participants may find it fun to suggest their own pseudonyms.

(ii) Name, gender, age (as precisely as possible).

(iii) Any information that you have collected on their linguistic and non-linguistic development, for example their talkativity, their first words, their longest utterances, at which ages they mastered which skills, etc.

(iv) Any information that you have collected on their typical daily routines.

2) Extensive metadata: Dossier of child (ctd.)

Their main interlocutors (even if they do not participate in any of the recordings). This list should minimally include the immediate family (parents, siblings), but it is likely to contain others as well (e.g. grandparents, more distant relatives, neighbors). The goal is to identify and characterize the main interlocutors of the focus children.

(v) ID/pseudonym, name, gender, age.

(vi) Type of relationship to focus child.

(vii) Typical contexts of interaction with focus child.

(viii) Language(s) known, and language(s) typically used with focus child.

2) Extensive metadata: Recording

(i) The structured metadata recommended by the archive for each participant in a session (e.g. ID/pseudonym, name, gender, age, role in the recording) and the session (e.g. date, location, topic).

(ii) Record the ages of all participating children as precisely as possible and calculate them for each session (if possible in the format YY;MM.DD).

(iii) For each participant, record the type of relationship with the focus child.

(iv) A descriptive account of the context of the recording: setting/location (e.g. “in the kitchen hut, next to the fire”), participants and their contributions (e.g. “the adults talk amongst themselves and only rarely interact with the children, while the children play with each other”), main activities (e.g. “the children play with sticks”), and main topics (e.g. “the children talk about building a house”).

Give an overview of data

- As an introduction to the archival collection *and* for the Acquisition Sketch:
  - A table listing the focus children and their ages.
  - Introduction to each of the focus children: Are they older, younger, middle siblings or only children? Which family members do they live with? Which other languages (if any) are used in the home? Is there anything else of note? For example, do they attend formal education?

cf. Meeting 2 (July 19, 2023)

https://qaqet.phil-fak.uni-koeln.de/corpus

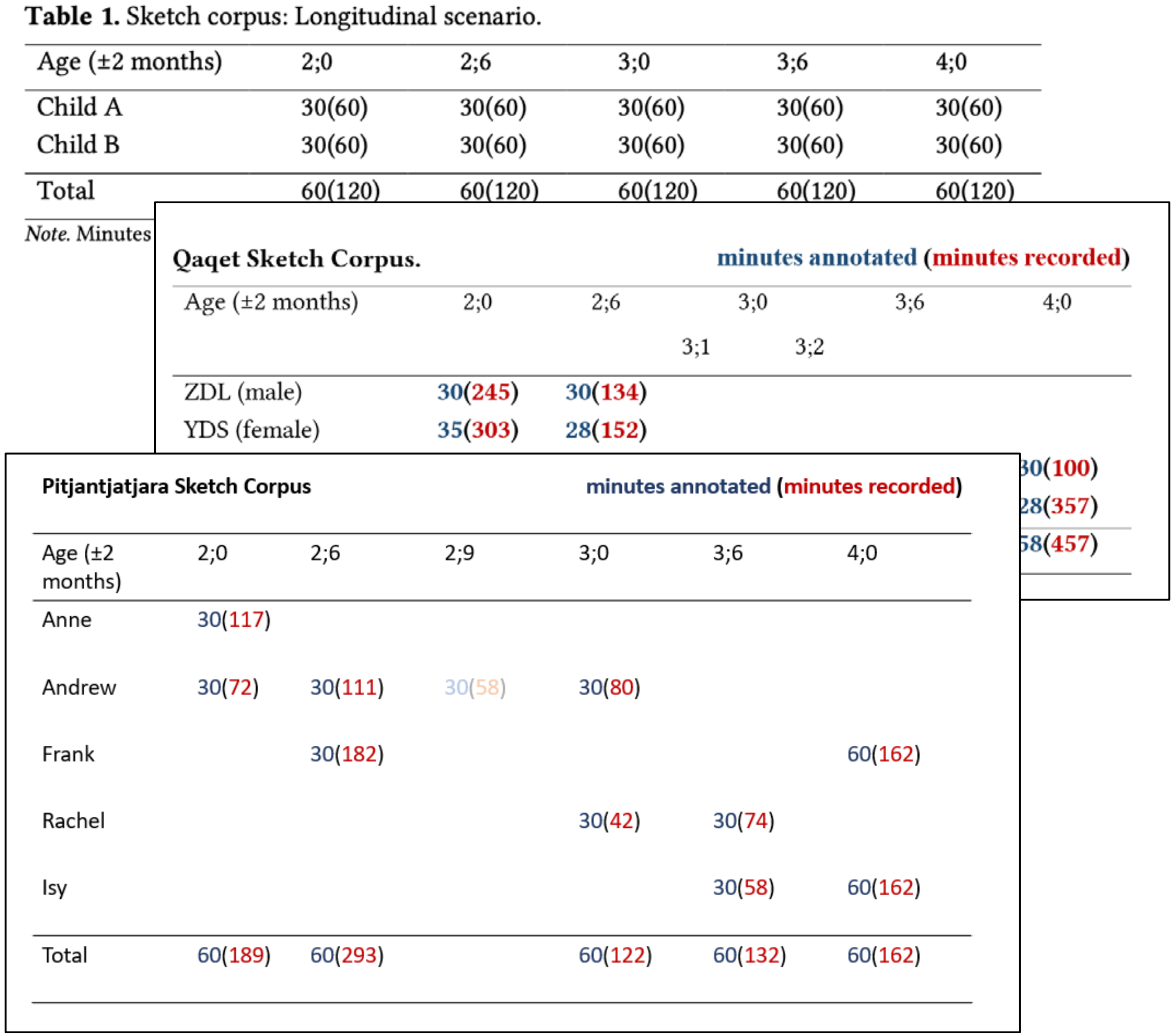


Table 1. Sketch corpus: Longitudinal scenario.

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)

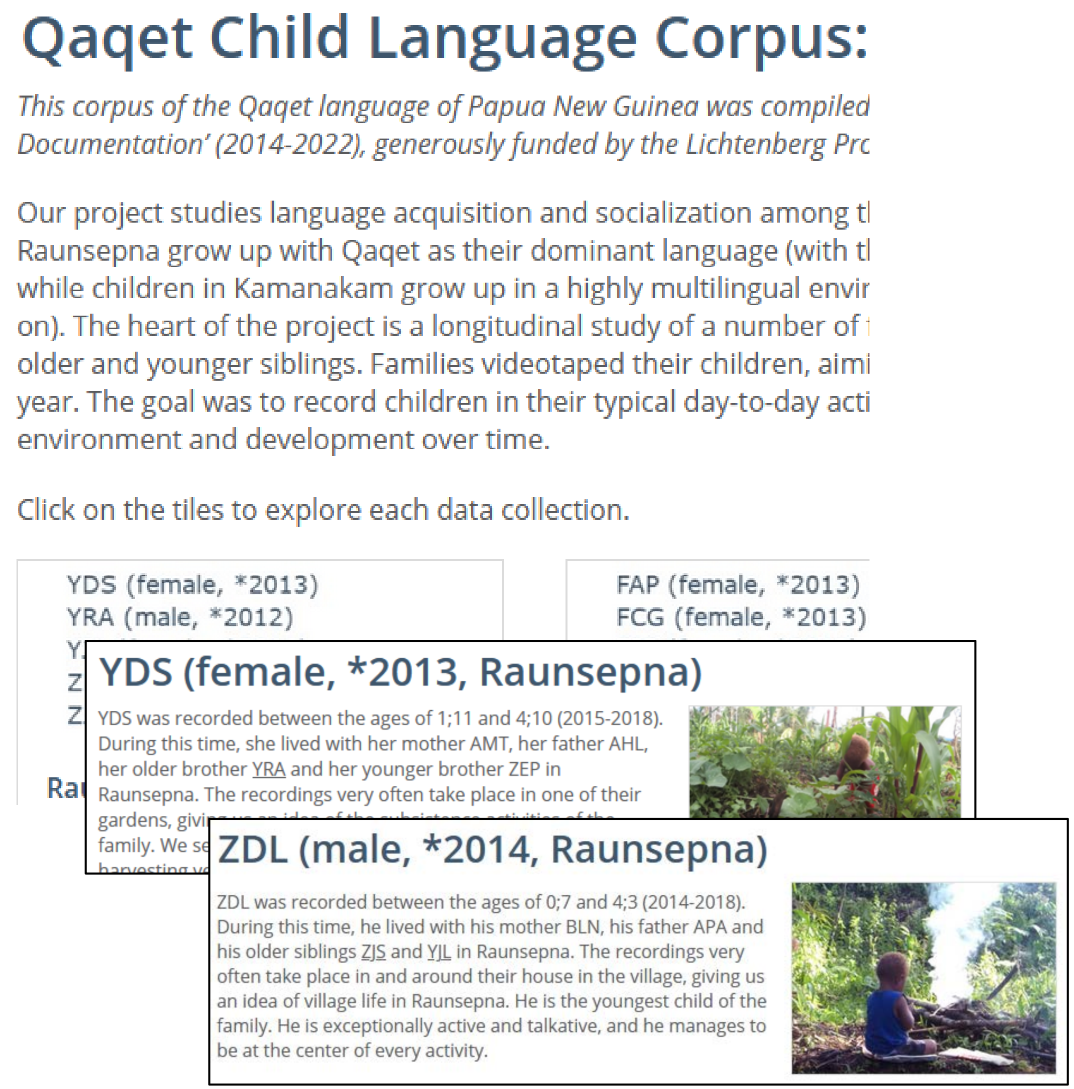
Note. Minutes

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)			
					30(100)
					28(357)
					58(457)

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 Child Language Corpus:

This corpus of the Qaqet language of Papua New Guinea was compiled
Documentation' (2014-2022), generously funded by the Lichtenberg Prc

Our project studies language acquisition and socialization among tl
Raunsepna grow up with Qaqet as their dominant language (with tl
while children in Kamanakam grow up in a highly multilingual envir
on). The heart of the project is a longitudinal study of a number of
older and younger siblings. Families videotaped their children, aimi
year. The goal was to record children in their typical day-to-day acti
environment and development over time.

Click on the tiles to explore each data collection.

YDS (female, *2013)
YRA (male, *2012)
Y
Z
Z

Ra

FAP (female, *2013)
FCG (female, *2013)

YDS (female, *2013, Raunsepna)

YDS was recorded between the ages of 1;11 and 4;10 (2015-2018). During this time, she lived with her mother AMT, her father AHL, her older brother YRA and her younger brother ZEP in Raunsepna. The recordings very often take place in one of their gardens, givi
family. We se

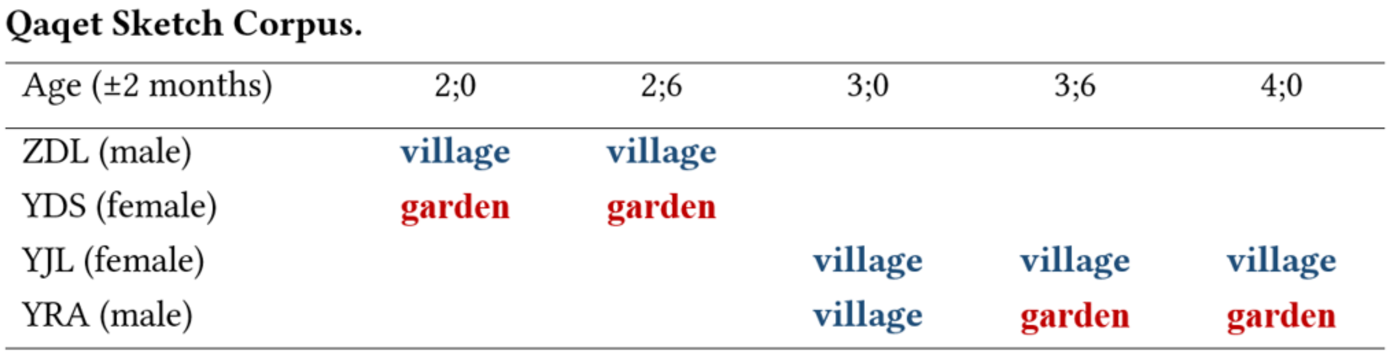
ZDL (male, *2014, Raunsepna)

ZDL was recorded between the ages of 0;7 and 4;3 (2014-2018). During this time, he lived with his mother BLN, his father APA and his older siblings ZJS and YJL in Raunsepna. The recordings very often take place in and around their house in the village, giving us an idea of village life in Raunsepna. He is the youngest child of the family. He is exceptionally active and talkative, and he manages to be at the center of every activity.

Give an overview of data (ctd.)

- As an introduction to the archival collection *and* for the Acquisition Sketch:
  - A brief overview of each of the recordings used for the sketch. In what setting were they recorded (e.g. in the home, in the bush)? In what sorts of activities were the children principally engaged (e.g. playing outside with peers, reading with a caregiver, gathering food, painting, eating dinner).
  - A list of other participants appearing in the recordings. If known, also list the relationship to the focus child. For children, provide (approximate) age.

cf. Meeting 2 (July 19, 2023)

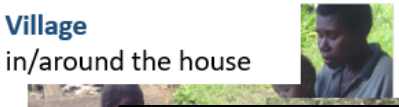


Qaqet Sketch Corpus.

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

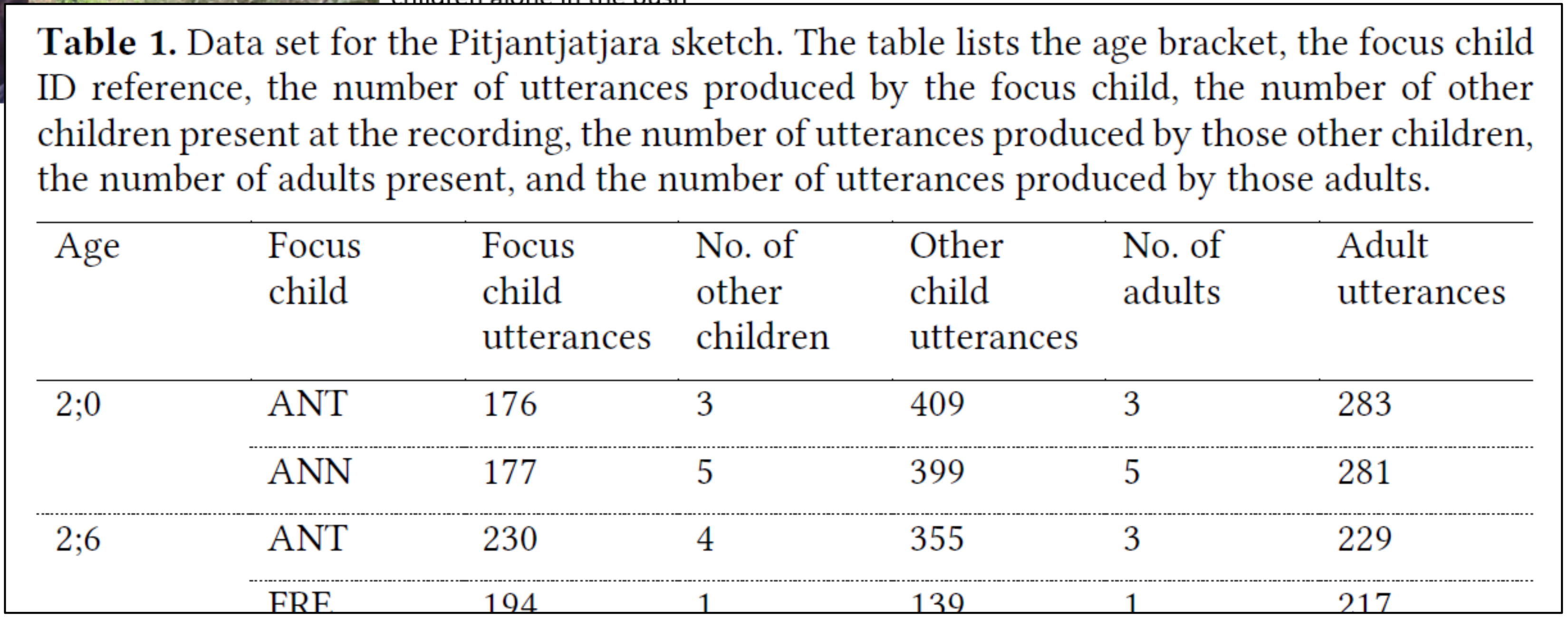


Table 1. Data set for the Pitjantjatjara sketch. The table lists the age bracket, the focus child ID reference, the number of utterances produced by the focus child, the number of other children present at the recording, the number of utterances produced by those other children, the number of adults present, and the number of utterances produced by those adults.

Age	Focus child	Focus child utterances	No. of other children	Other child utterances	No. of adults	Adult utterances
2;0	ANT	176	3	409	3	283
	ANN	177	5	399	5	281
2;6	ANT	230	4	355	3	229
	ERE	194	1	139	1	217

Discussion points

- Which archive?
- Child language data
    - Sensitivity of data
    - More extensive metadata
- Overview of data