Acquisition Sketch Project
Meeting 5:

Data Processing Part 1 – Setup, Transcribers, Segmenting, Tiers

6th March 2024

Agenda:

- Selecting a setup
- Selecting and training transcribers
- Segmenting the data
- Tiers for transcribing and annotation

Selecting a setup

Two most common options

ELAN plus Toolbox or FieldWorks

- Typically used in language documentation
- Shows sound waves and video on screen
- Tiers can be hierarchically connected
- Best option for annotating visual information – e.g. gesture
- https://archive.mpi.nl/tla/elan

CHAT transcription system and CLAN analysis tools in CHILDES

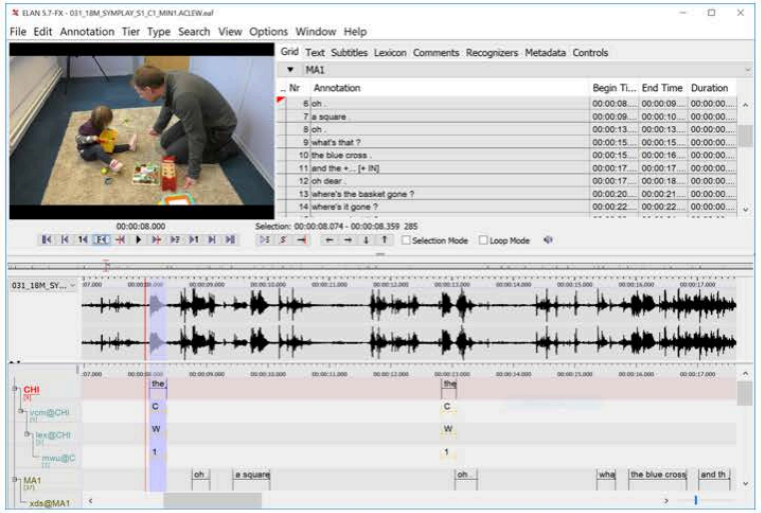
- Typically used in child language acquisition
- Shows sound waves and video on screen
- Tiers not hierarchically connected
- Relatively easy to learn and use
- Dedicated set of analysis tools specific to child language research (CLAN)
- https://childes.talkbank.org

ELAN

Intro | Release notes | Documentation | Download | Citing | Support | Resources

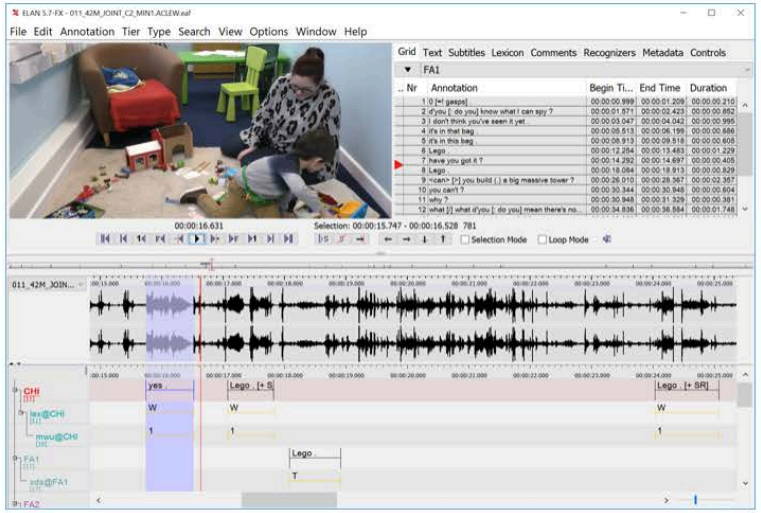
ELAN is an annotation tool for audio and video recordings.

Screenshot 1



A sample from the ACLEW project.

Screenshot 2



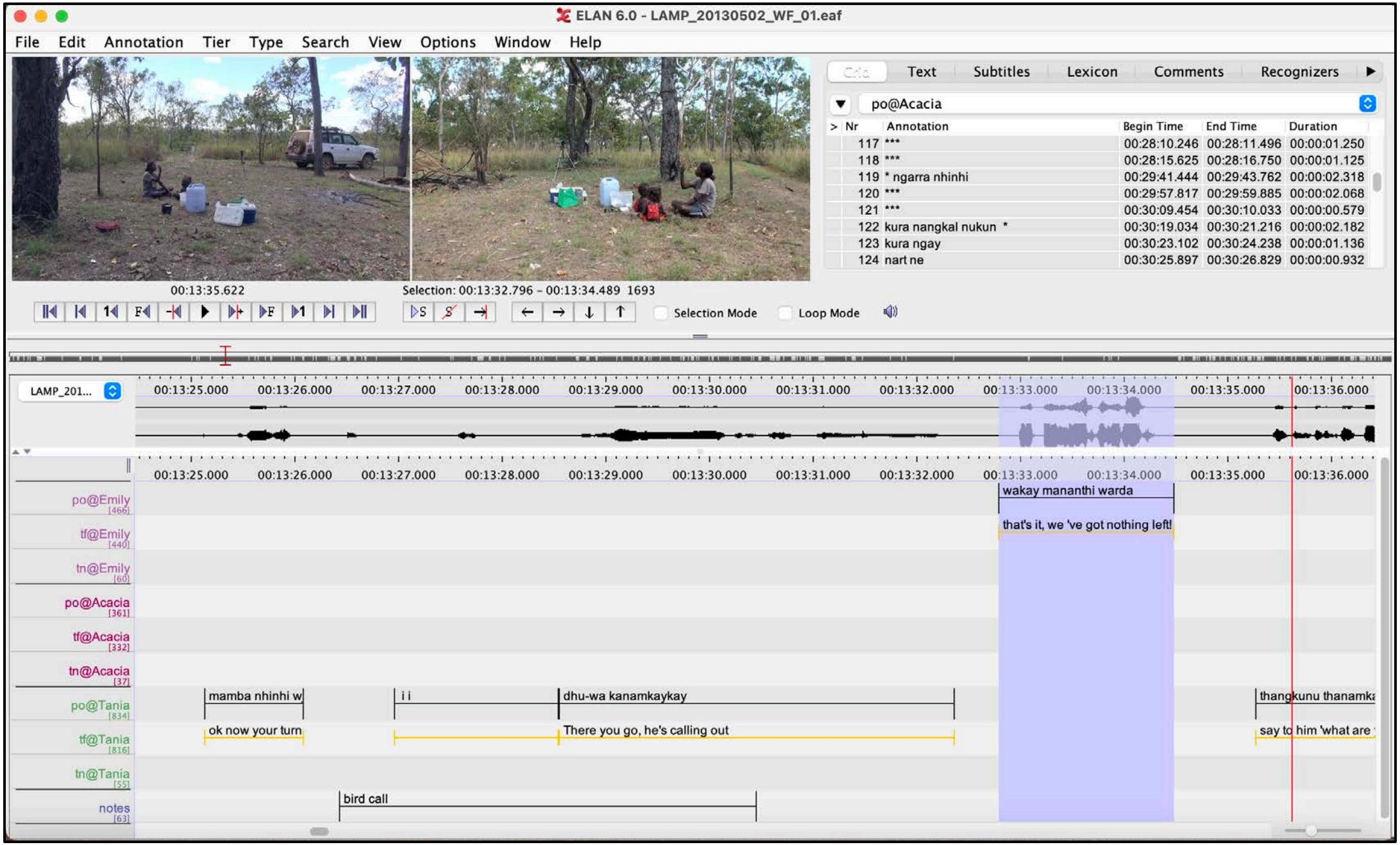
A sample from the ACLEW project.

Description: With **ELAN** a user can add an unlimited number of textual annotations to audio and/or video recordings. An annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media. Annotations can be created on multiple layers, called *tiers*. Tiers can be hierarchically interconnected. An annotation can either be time-aligned to the media or it can refer to other existing annotations. The content of annotations consists of Unicode text and annotation documents are stored in an XML format (EAF).

Main features and characteristics of ELAN:

- provides several ways to view the annotations, each view is connected to and synchronized with the media timeline
- supports creation of multiple tiers and tier hierarchies
- supports Controlled Vocabularies
- allows linking of up to 4 video files with an annotation document
- media support
  - builds on existing, native media frameworks, like Windows Media Player, QuickTime or VLC
  - support for audio and video formats depends on operating system, high performance media playback can usually be achieved
- technical
  - written in the Java programming language
  - distributions available for Windows, macOS and Linux
  - open source, the sources are available under a GPL 3 license

ELAN



Murrinhpatha, data from LAMP project

CHAT

CHILDES

Child Language Data Exchange System

CHILDES is the child language component of the TalkBank system.

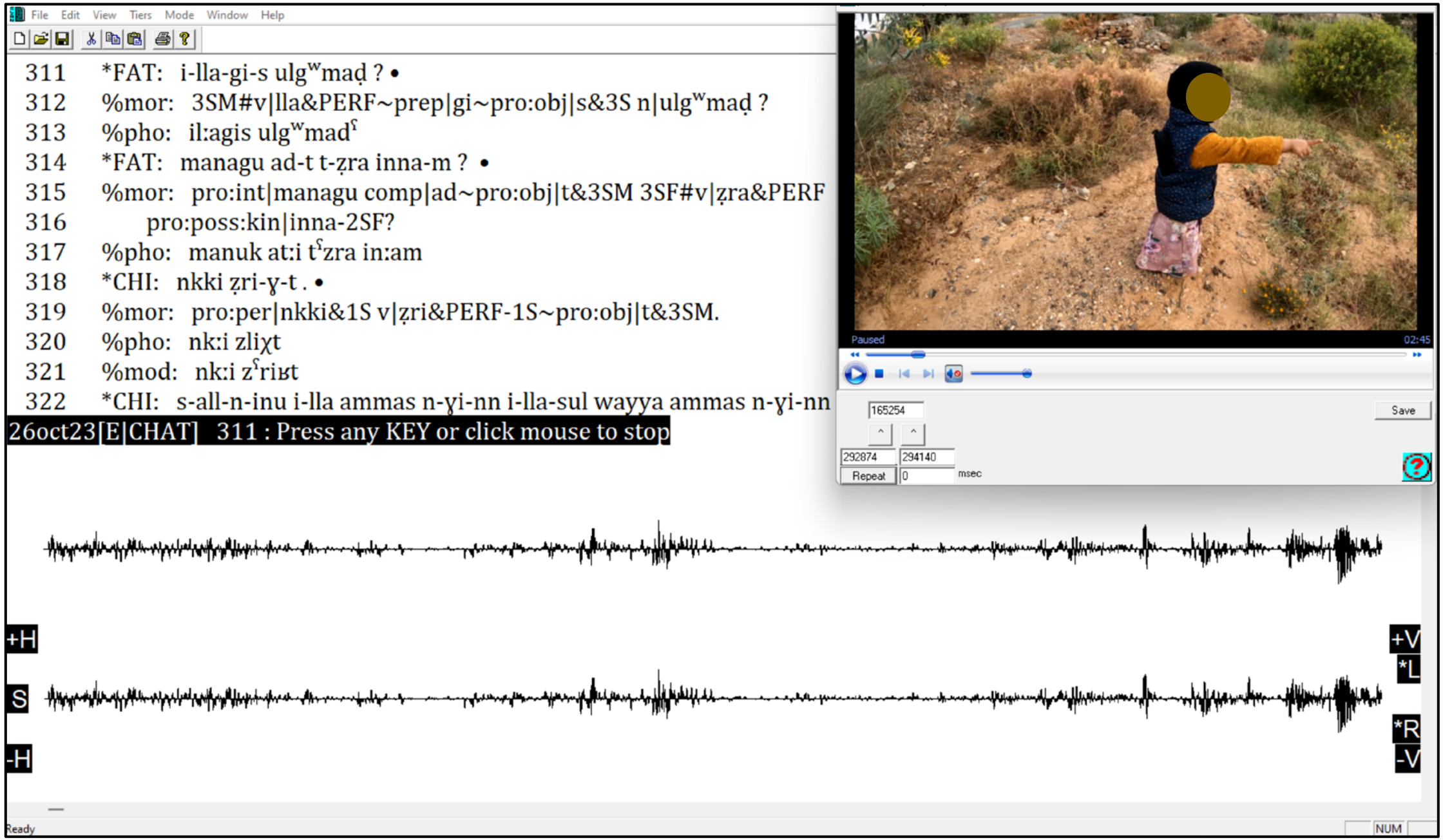
System	Database	Programs
Ground Rules	**Index to Corpora**	CLAN
Contributing New Data	Browsable Database	XML creator and XML Schema
IRB Principles	TalkBankDB database search	Related Software
	Hints on Downloading	

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 Data	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	MOR manual
	Unicode and IPA for Windows	

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

CHAT



Berber, data from Abdellah Elouatiq

Helpful questions to ask

What setup do you already know?

What setup do you have support for (e.g. colleague who can help you)?

What setup is common in your subfield?

Is annotating video important for you (e.g. gesture)?

Is annotating phonetics important for you?

Do you need/want hierarchically connected tiers?

What programs/procedures will you use to analyze the data?

Where do you want to archive the data and what requirements do they have?

Selecting and training transcribers

Selecting transcribers

Typical choices

- You
- Caregivers or family members (alone or with you)
- Other community members (alone or with you)
- Maybe get help from older children and/or target child

Typical criteria

- Fluent speaker of the language
- Able to understand the child's speech
- Knowledge of the child's daily context
- Able to write the language – ideally using a computer
- Sufficient time and patience to complete the job (could be divided in parts)
- Acceptable to the family (ethical considerations)

Examples of transcribers

Inuktitut

- Students from the community studying in Montreal, working at researcher's office

Cree

- Mother of one target child who served as interlocutor for all recordings
- IPA transcription – students at the university trained in Cree phonology

Mongolian

- Researcher and Master's student in psychology who wanted to gain experience

Sesotho

- Mother of each child working together with the researcher

Chintang

- Community members who had studied linguistics

Supporting transcribers

Provide guidelines

- What to transcribe – child speech, caregiver speech, others?
- How to divide intonation units, utterances, etc.
- What to do with repetition, uninterpretable speech, etc.
- Transcribe what child *actually* says, not what you think they intended (that goes on another tier)
- Emphasize need for detail, quality, consistency, etc. as much as is reasonable

Amount of support

- Range from sitting together to fully independent work – depends on your situation
- Brainstorm ideas for motivation

Building in “quality control” if reasonable

Why?

- Increase transcriber confidence and knowledge
- Journals may request evidence of transcriber “reliability”

Hold training sessions

- All transcribers transcribe same short portion and compare (e.g. 2 minutes)
- Repeat a few times till cross-transcriber consistency is high
- Discuss and find solutions for differences
- Add solutions to guidelines document

Make plan to calculate transcriber agreement

- Pairs of transcribers blind-transcribe 10% of data
- Calculate Cohen’s D or % agreement

Segmenting the data

Select portions to transcribe

Scan each recording to select the portion to transcribe

Continuous chunk of 30 minutes

Maximum amount of child and caregiver speech

Avoid excessive crying, adult-only talk, distracting background noise, etc.

Need to listen to / watch recording – finding dense parts in sound file can be misleading

Prepare selected portion for transcription

Prepare audio (or video) files for your selected setup

Integrate audio (or video) file into the file for transcription

Prepare templates of tiers or have workflow in place to automatically generate them

If the transcriber is computer-savvy and has lots of time:

- Start transcribing, segmenting into units as you go

If the transcriber is less computer-savvy and/or has limited time:

- Segment small portions of speech within your setup so the transcriber can just go to each portion directly and not spend time finding the parts to transcribe

Figure 2a. Initial segmentation in the Qaqet corpus (for first-pass transcription).

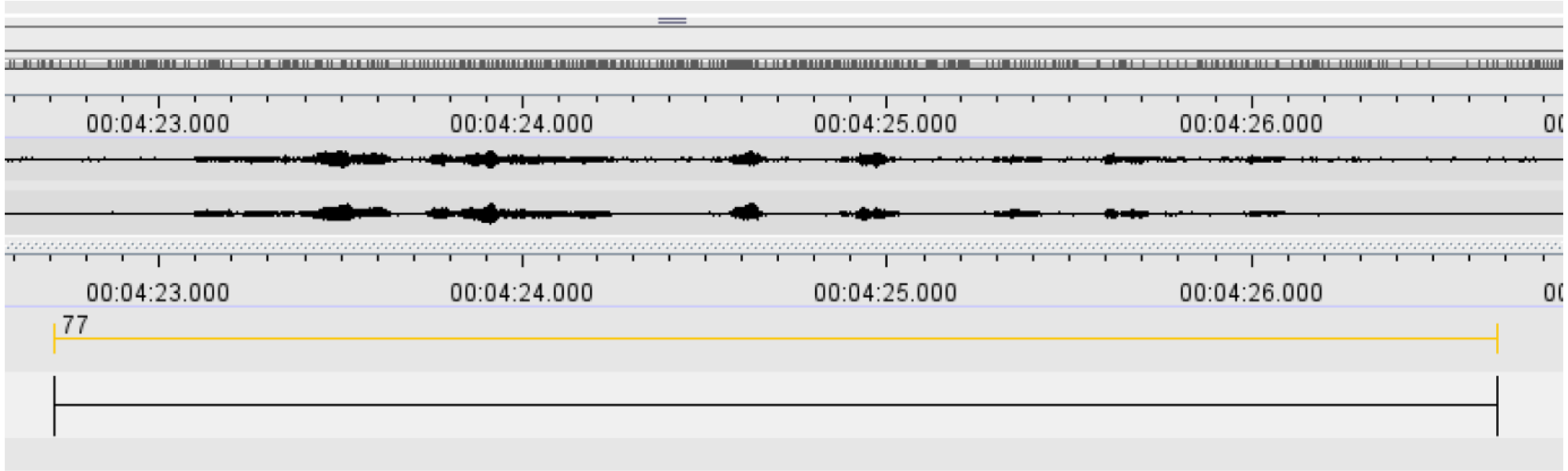
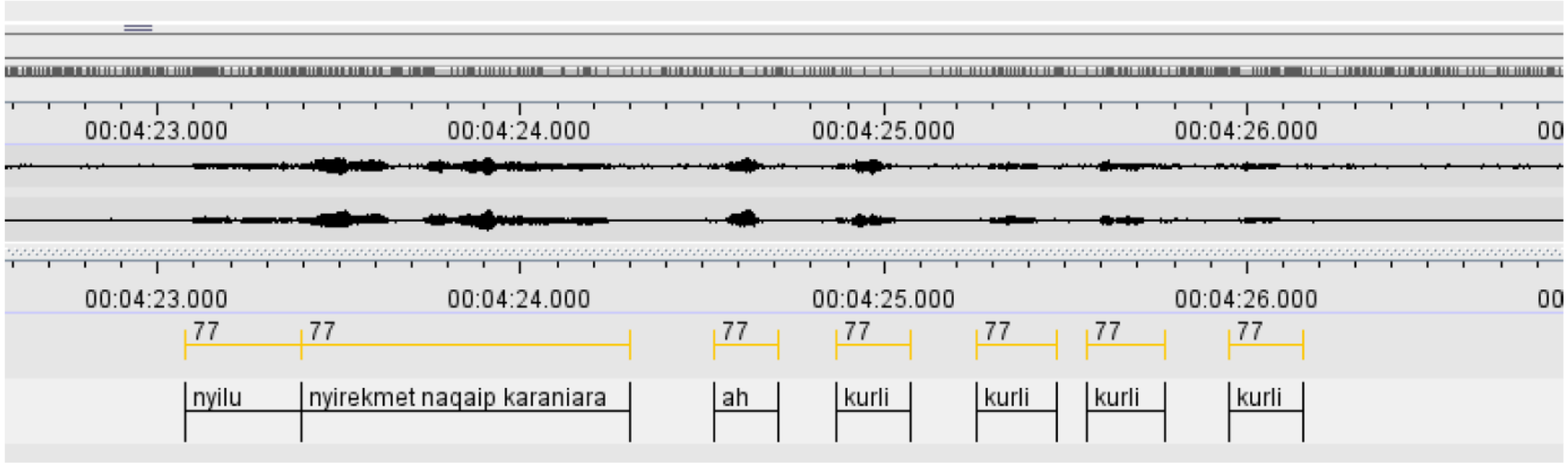


Figure 2b. Final segmentation in the Qaqet corpus (intonation units).



Sketch Acquisition Manual, p. 26

Select unit of segmentation to facilitate analysis

Intonation unit (typical in language documentation – Himmelmann 2006b)

- Segment of speech that occurs with a single prosodic contour
- Likely to begin with a brief pause and to end in a clause-final intonation contour
- Not necessarily identical with a syntactic unit, but often corresponds to a clause or a group of closely related clauses
- Can be as small as a single word or as long as a full sentence or more

Utterance (typical in child language – MacWhinney 2021)

- Continuous piece of speech, by one person, before or after which there is silence on the part of the person
- Typically a syntactic unit
- Can be as small as a single word or as long as a full sentence
- See CHAT Manual, chapter 9

Helpful tips

Child language is often repetitive

- Plan how to deal with repetitions – e.g. transcribe as one unit, decide depending on intonation

Child language is often hard to interpret

- Use video context for clues
- Listen to preceding and following utterances for clues
- Plan strategies to reduce frustration – e.g. maximum number of listens or amount of time before moving on, symbol/code to mark uninterpretable utterance or utterance to come back to later

If transcription resources are limited

- Think carefully about ways to reduce time and cognitive load – e.g. pre-segmentation, ”removing” repetitions from first-pass transcription, providing motivation in procedure / setting / etc.

Tiers for transcribing and annotation

Potential tiers

Minimum

- Transcription in orthography
- Translation into national language, English, etc.
- Morphemic breakdown, gloss, part of speech

Ideas for optional tiers

- Interpretation of child utterance / likely target utterance
- Errors / child-like productions / differences from likely target
- Addressee
- Notes / comments / situational context
- Phonetic transcription
- Annotation for structures of interest

Helpful tips

Create a tier “template” for each utterance (cut/paste or automatically generated)

First pass = just the orthographic transcription and translation

- this will likely take at least 1 hour per 5 minutes of recording
- maybe also include interpretation of child utterance, addressee, and notes

Gradually add information for other tiers as you need them

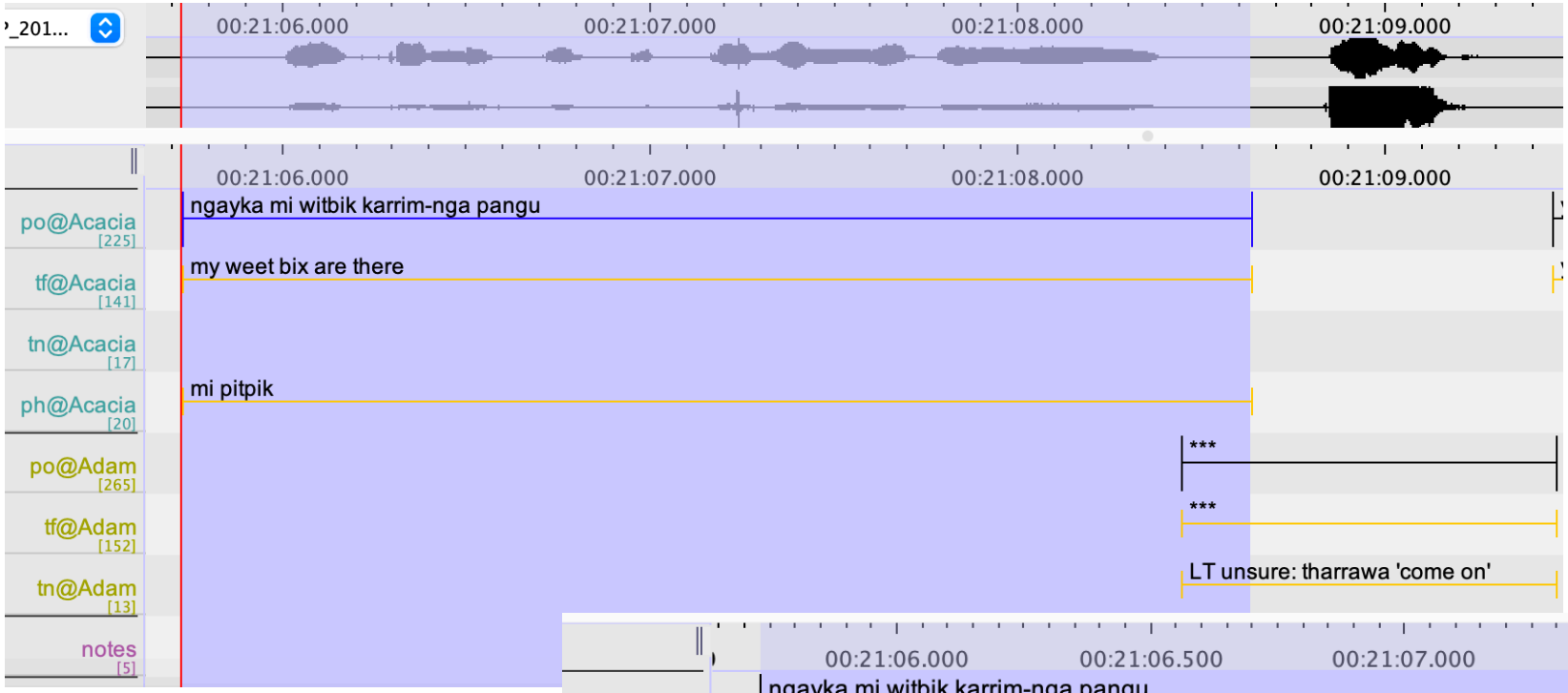
Only use the tiers/information you need for your own questions and analyses

Do minimal annotations for the entire five hours with more detailed annotations for parts of the five hours and/or specific phenomena

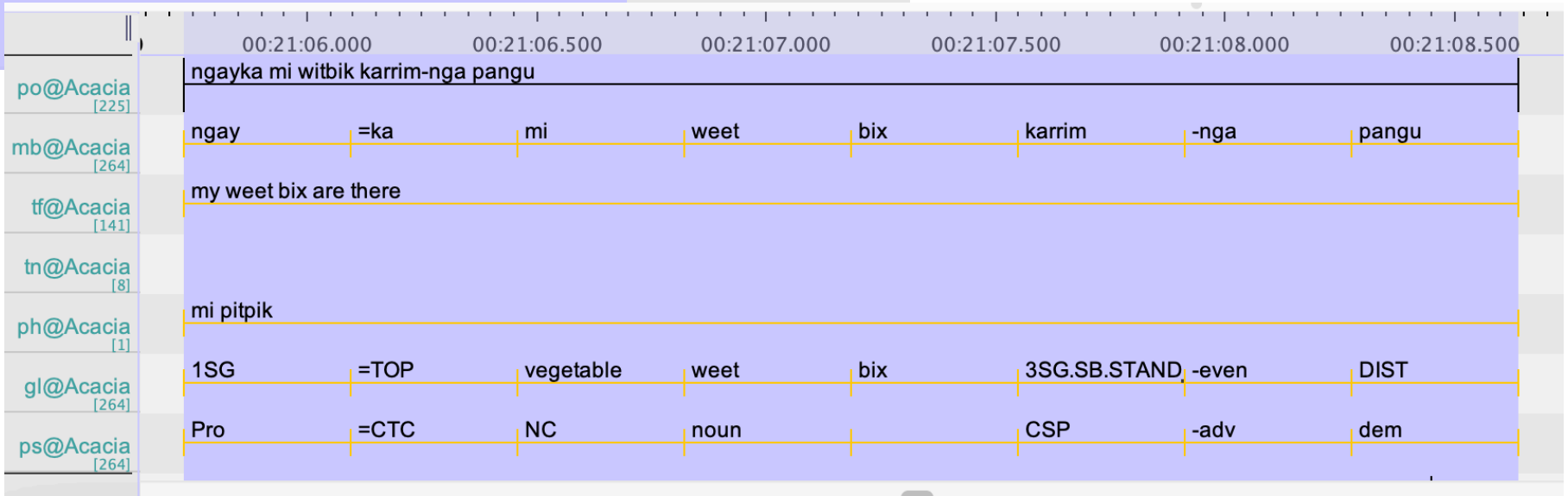
Different people can do different tiers depending on their expertise

- e.g. mother for transcription, RA for morphemic breakdown and gloss

Murrinhpatha in ELAN + Toolbox



po	practical orthography
tf	free translation (English)
tn	transcription notes
ph	phonological notes (focus child)
addr	addressee (adult speech)
mb	morphemic breakdown
gl	gloss
ps	part of speech
notes	notes
***	unclear speech



Murrinhpatha, data from LAMP project

Murrinhpatha in ELAN + Fieldworks

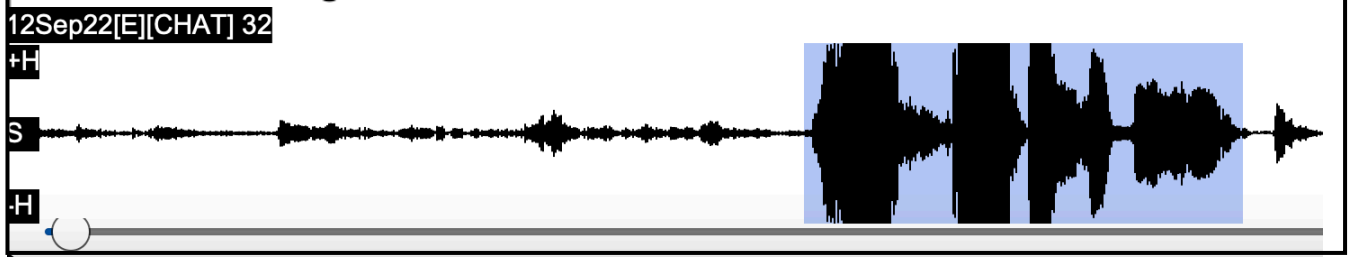
po@Emily [466]		thangkunu *-aykay-yu
tf@Emily [440]		what (?are you) calling out for?
tn@Emily [63]		PROMPTED, similar number of syllables & melody as adult's
ph@Emily [37]		thangkunu yayaykayyo
po@Tania [834]	thangkunu thanamkaykay-ya nana	
tf@Tania [816]	say to him 'what are you calling out for?'	
tn@Tania [59]	'him' is the bird	
addr@Tania [233]	CDS: Emily	
notes [67]	They are talking to a bird (pulupulu) in the tree. Pulupulu is also the nickname and totem of Emily's older brother	

[605]						
Tania_morph-cf-m [3572]	thangkunu	thanam	kay	-ya	na	-na
Tania_morph-gls-e [3573]	why	2SG.BE(4).NFUT	call.out	INTJ	2SG.SAY/DO(8).FUT	3SG.M.IO
Tania_morph-msa- [3571]	INT	CSP	Cov	CTC	CSP	IO
Tania_morph-txt-m [3627]	thangkunu	thanam	kaykay	-ya	na	-na
Tania_morph-type [3571]	stem	stem	stem	suffix	stem	suffix
Tania_phrase-gls-e [836]	say to him 'what are you calling out for?'					
Tania_phrase-segn	232					

po	practical orthography
tf	free translation (English)
tn	transcription notes
ph	phonological notes (child)
addr	addressee (adults)
notes	notes
*	unclear speech

Mongolian in CHAT

```
@Begin
@Languages: mon
@Participants:      MOT Mother, CHI Target_Child
@ID:      mon|Sketch-Mongolian|MOT|25;05|female|||Mother|||
@ID:      mon|Sketch-Mongolian|CHI|3;0.|female|||Target_Child|||
@Media: SAM001, audio
*CHI:     Ügüi. •187_5157•
*CHI:     Zovoogoogüi baina. •4075_5296•
*MOT:     Naadkhyg chini khen gedeg yum.  •8963_9704•
*CHI:     Melkhii gedeg. •10445_10944•
*MOT:     Melkhii gedeg yum uu? •11416_12184•
*CHI:     Mmkhn melkhii gedeg. •12192_14424•
*CHI:     Yoooyo. •18203_18440•
*MOT:     Öör bambaruush baigaa yum uu? •18461_19603•
*CHI:     Baikhgüi. •19488_20624•
```



Initial headers

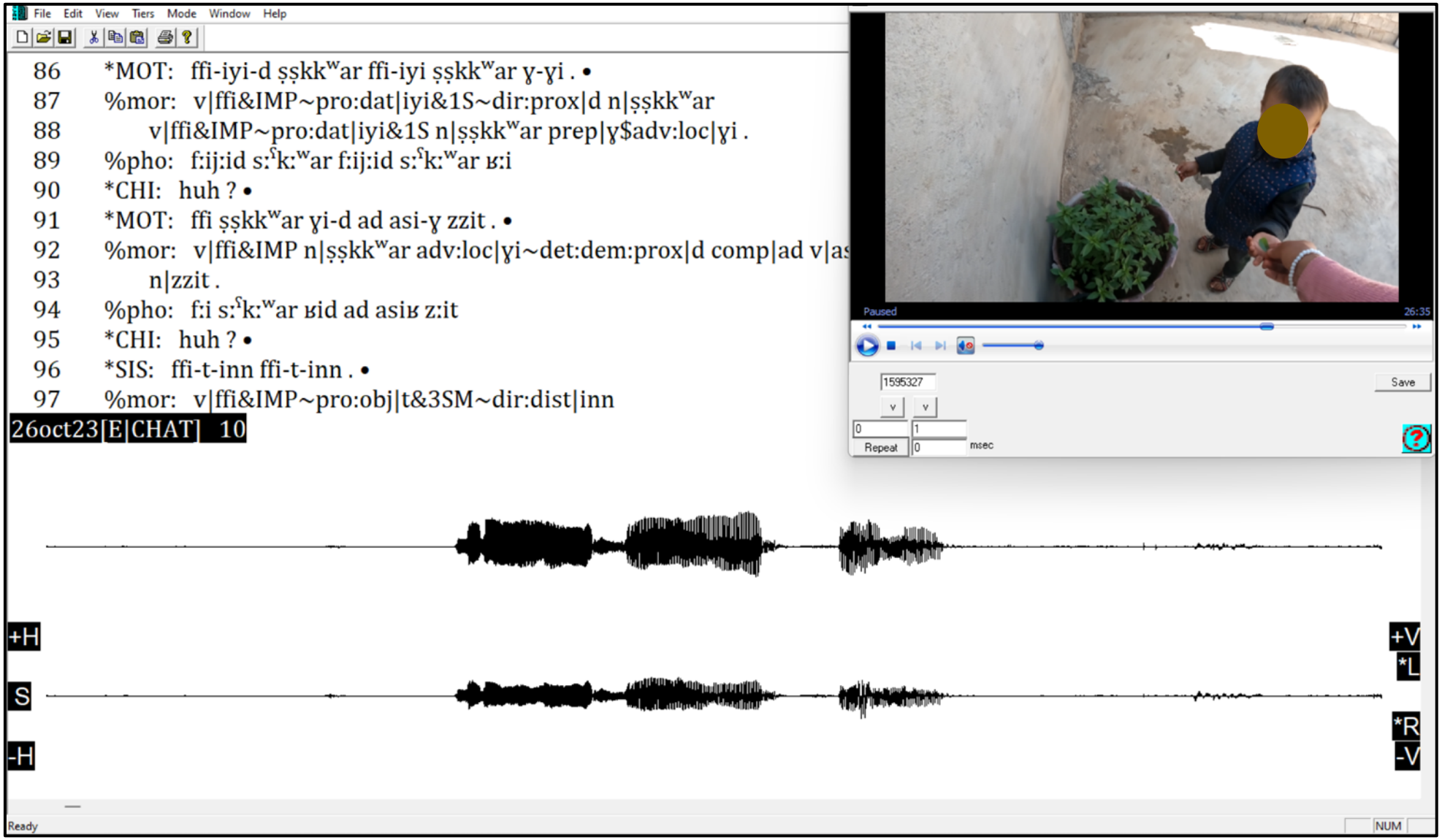
→ Audio and Video Time Marks: •4075_5296• time alignment

→ Transcribing directly from an audio file using sonic mode in CLAN

→ Translation, morphemic breakdown, gloss and other tiers will be done later

Mongolian, data from Dorjderem Byambasuren

Berber in CHAT



*MOT	mother
*CHI	child
*SIS	sister
%mor	morphemic breakdown and gloss
%pho	phonetic transcription

Berber, data from Abdellah Elouatiq

Inuktitut in CHAT

```
*LIZ:    anisaarta.
%eng:    Let's go out.
%mor:    VR|ani^go_out+VV|ADV|saaq^quickly+VI|ta^IMP_1pS .
%cod:    $VER $ITR:vr
%arg:    $YRF $STR $INT $SUB:nul:1:hum:crx:lcn:qun:cxp:ugi:rxx:gsn:prn $WOR:1 $MOR:3
%err:
%tim:    01:02:07
%snd:    "mae14.sd" 3734737 3736477
%add:    DOM; JIN
%sit:    they try to go out the window
%com:
```

%eng	English translation
%mor	morpheme breakdown and gloss
%cod	codes for verb types
%arg	codes for argument structure
%err	errors

%tim	time on tape
%snd	link to audio sound
%add	addressee
%sit	situational context
%com	comments

Inuktitut, Lizzie 2;6.3
Data from Shanley Allen

Cree in Phon extension of CHAT including Praat

Session Editor : B1.01-Daisy-03_08_10*

File Edit Workspace Session View Media Query Analysis Tools Window Help

Play segment | Record: 95 of 886 | Search

Media Player

Speech Analysis

Play segment | Export segment... | Reset | Zoom in | Zoom out | Spectrum | TextGrid

003:41.984 003:42.482 003:42.980

003:41.724

500

No TextGrid found

Click here to generate TextGrid tiers from Phon tiers.

Session Information | Tier Management | Speech Analysis | CCLAS Auto Parser

Record Data

95 Speaker: Daisy | Exclude from searches

Orthography	mâu=tâh âhtut=ikiniwi=ch-h
Morpheme Meaning	this=loc do=passive.3=0.pl
Morpheme Type	p,dem+G.pxl=loc IC.initial=passive=CIN
Translation	We do it like this.
Segment	003:41.724-003:43.347 Play segment
IPA Actual	ondædədɪgɪnʊtʃʰ
IPA Target	maʊdatotɪgənotʃʰ
Actual Morphology	on=dæ dəd=ɪgɪnʊ=tʃʰ
Target Morphology	ˈmaʊ=da tot=ɪgəˈno=tʃʰ
Passives Coding	B; nSAP; Act; T I; Conj; do; ihtutim
Passives Commentary	She is talking about scooping them with the spoon. I don't think it is a peeler but a spoon. scooping or dishing: â ûtih=î=kiniwi=ch=h 'to dish or scoop'
Notes	She is talking about how to peel potatoes. From previous record.

Id: 42fca48c-5a31-4506-ad26-dca4bdc8fa62 Tier: Passives Coding Group: 1 Character: 35

B1/01-Daisy-03_08_10 | 86 warnings

Cree, data from CCLAS project

Cree – data from Phon section of Talkbank

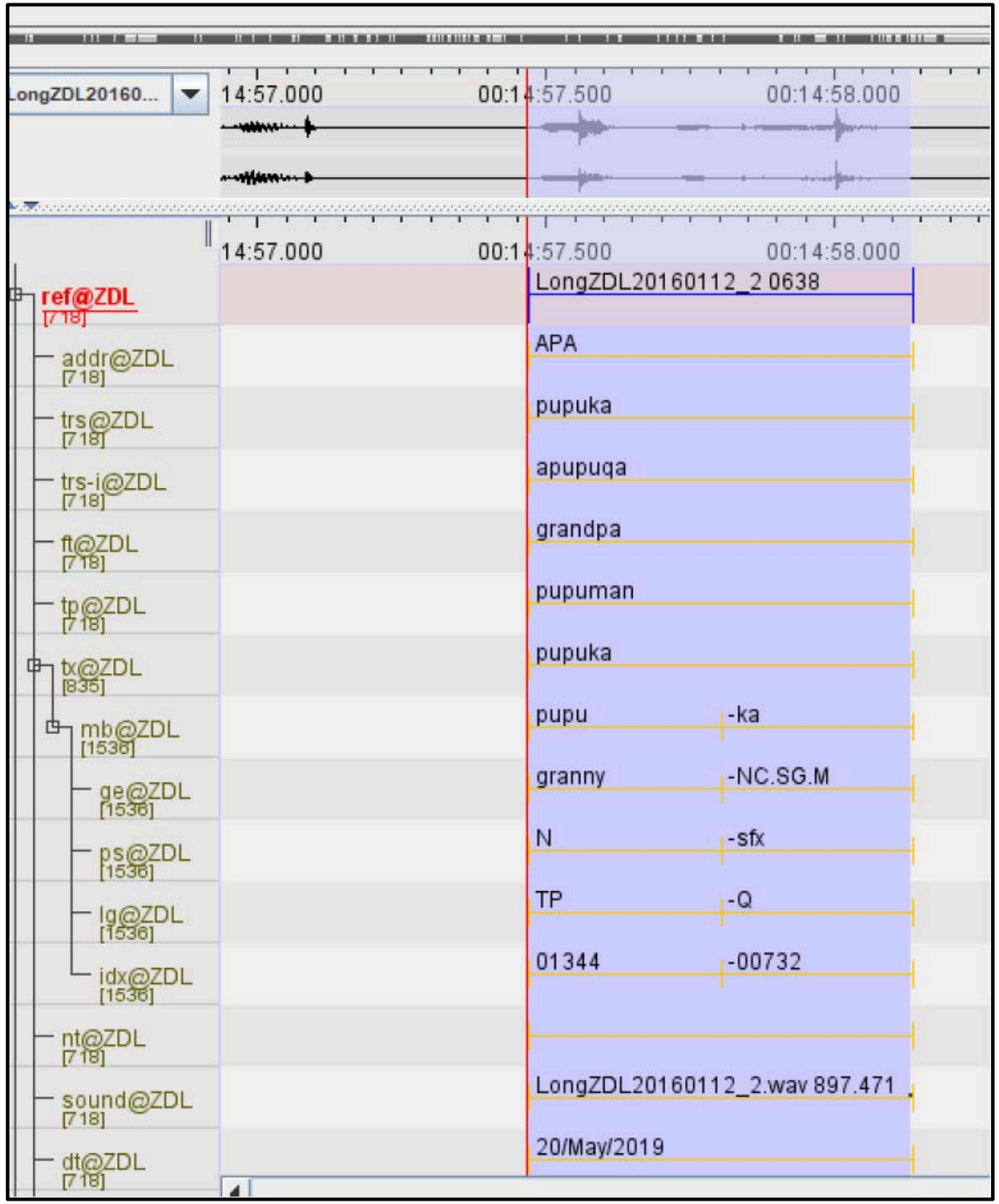
```
CHI: îtuht~â~u . ▶
%pho: tʰəhesaw
%mod: itʃˈdaw
%xactmor: tʰəhes~a~w
%xmodmor: itʃˈd~a~w
%xmormea: go~vai.fin~3.sg
%xmortyp: initial~vai.fin~IIN
%xtrans: is going
```

*CHI	child; orthographic transcription
%pho	IPA transcription - actual
%mod	IPA transcription - target
%xactmor	morphemic breakdown - actual
%xmodmor	morphemic breakdown - target
%xmormea	gloss of morpheme meaning
%xmortyp	gloss of morpheme function
%xtrans	English translation

Ani 3;01.18, CCLAS database
https://phon.talkbank.org/access/Other/Cree/CCLAS.html
https://sla.talkbank.org/TBB/phon/Other/Cree/CCLAS/Ani

Qaqet in ELAN

ref	reference for data file
addr	addressee
trs	transcription
trs-i	interpretation of child utterance
ft	free translation (English)
tp	free translation (Tok Pisin)
tx	text
mb	morphemic breakdown
ge	gloss
ps	part of speech
lg	language
idx	index
nt	notes
sound	link to sound file
dt	date



SAM, p. 27

Qaqet in Toolbox

LongZDL20160112_2.txt

\ref reference	LongZDL20160112_2 0638
\ELANBegin ELAN Begin	897.471
\ELANEnd ELAN End	898.139
\ELANParticipant Participant	ZDL
\addr addressee	APA
\trs transcript (actual utterance)	**pupuka**
\trs-i transcript (interpretation)	**apupuqa**
\ft free translation (English)	grandpa
\tp free translation (Tok Pisin)	*pupuman*
\tx text	**pupuka**
\mb morpheme breaks	**pupu -ka**
\ge gloss (English)	granny -NC.SG.M
\ps part of speech	N -sfx
\lg language	TP -Q
\idx index	01344 -00732
\nt notes	
\sound sound	LongZDL20160112_2.wav 897.471 898.139
\dt date	20/May/2019

Qaqet in CHAT

```
*ZDL:     pupuka
%int:     apupuqa
%eng:     grandpa
%tkp:     pupuman
%mor:     n|pupu=grandparent&TP+nc|ka=sg&m&Q
%pos:     n+sfx
%add:     APA
%not:
%snd:     LongZDL20160112_2.wav 897.471 898.139
```

%int	interpretation
%eng	English translation
%tkp	Tok Pisin translation
%mor	morphemic breakdown and gloss

%pos	part of speech
%add	addressee
%not	notes
%snd	link to audio file

Questions and Discussion

Good luck in data processing!