Conventions for annotation and transcription of the MINT-project

Modulating child language acquisition through parent-child interaction, MAW:2011.007

Funded by Marcus and Amalia Wallenbergs foundation

Tove Gerholm

Department of Linguistics
Research group: Language development
2013-2018

Stockholms universitet

Contents

Introduction

The MINT-project set out to uncover the mechanisms of multimodal behavior in child language acquisition. The project was funded by Marcus and Amalia Wallenberg for the years 2013-2018. The researchers involved, and responsible for the structuring and implementation of the project from the start, were Tove Gerholm (PI), Iris-Corinna Schwarz, Lisa Gustavsson and Eva Klintfors (subsequently left the project).

In October 2013 a letter of invitation was sent to 1000 randomly chosen parents in the Stockholm area who had children born between the 1st of August and the 30th of September 2013. The invitation informed the parents of the aim of the MINT-project and asked whether they were interested in participating in the study. If interesed, they should come to Stockholm University every third month, to be recorded together with their child. They were also informed of their right to withdraw from the study at any point, and that the data gathered was to be treated according to the ethical regulations of Stockholm University and the Swedish Research Council.

Of the 1000 invited parents, 85 replied and were willing to participate. In the 4 ½ years the study has been running since then, 14 families have withdrawn from the study and 71 remain. The recordings were conducted 4 times a year for the first 3 years and twice a year during the fourth year. When the children were 4 ½ years old, a subset of the group (32 children) were randomly allotted to participation. The same procedure will be performed as they turn 5 years old, Aug-Sept 2018.

The recordings are conducted in a studio equipped with a few pillows and some age-appropriate toys. Three stuffed animals have been present throughout the project, named Mo, Na, and Li. The studio has three stationary cameras (Canon HDMI model X A10) on the walls, and one InAction Camera (Go-Pro Hero 3) attached to the parent's chest. Apart from the camera microphones, both child and parent have microphones (Sennheiser model eW 100 G2) pinned to their clothes.

The first 10 minutes are "free interaction" where the parent and child do whatever they want to, given the limitations of the room and setting. Then follows, in most recordings, a set of semi-structured tests, lead by a researcher or research assistant. These tests target skills like gaze following/joint attention, awareness of symbolic play and routines such as Play House, perceptive language tasks, working memory tasks, imitation tasks, etc. As the children grew older, executive functions were assessed as well as understanding of quantifiers, prepositions, etc. From 9 months and onwards, the parents filled in SCDI in connection to each visit. SCDI is the Swedish version of the MacArthur Communicative Development Inventories (CDI). The SCDI instrument assesses communicative and language abilities in children aged 8-48 months by means of parental reports (Berglund & Eriksson, 2000a, b; Eriksson, 2017).

In order to transform the video recordings to written form, a transcription and annotation process commenced directly after the first recording sessions. The annotations and transcriptions are done in the ELAN-software (Sloetjes and Wittenburg: 2008; https://tla.mpi.nl/tools/tla-tools/elan/download/). The conventions used and developed throughout the project stem from various sources (including Gerholm, 2007; Nilsson Björkenstam & Wirén, 2013). Research assistants Stina Andersson, Freya Eriksson, Fatima Guseinova, David Pagmar, Linnea Rask, Hanna Rönnqvist, Johanna Schelhaas & Sofia Tahbaz) have been employed in the project and have aided in the fine-tuning of the conventions.

The following pages are meant to work as a guide to others who are in the process of annotation and transcription work. Any questions in relation to the conventions or the project can be sent to tove.gerholm@ling.su.se

Aims and goals

The aims of the project are:

- To reach a better understanding of the role played by different modalities in language development, and to understand the relation between modalities
- To build a multimodal database for Swedish children in the ages 0;3-5;0
- To use the knowledge gained to model the language acquisition process
- To identify factors beneficial for language outcome

Our quest was to identify and describe how different interactional means work together during the first four to five years and if/how they change in prominence. Questions guiding us in the set-up of the project were: 1) How does the child's vocal development relate to modalities other than the vocal/verbal? 2) How does parental multimodal behavior correspond to child multimodal behavior? 3) Are there any behaviors in particular (child or parent) that correlate to a child's later language proficiency? 4) What differences and similarities are there in the behavioral repertoire of different child-parent dyads? and finally 5) Are there particular interactional patterns that, more than others, appear to aid the language acquisition process? These questions were then elaborated into an overall structure as indicated in Figure 1.

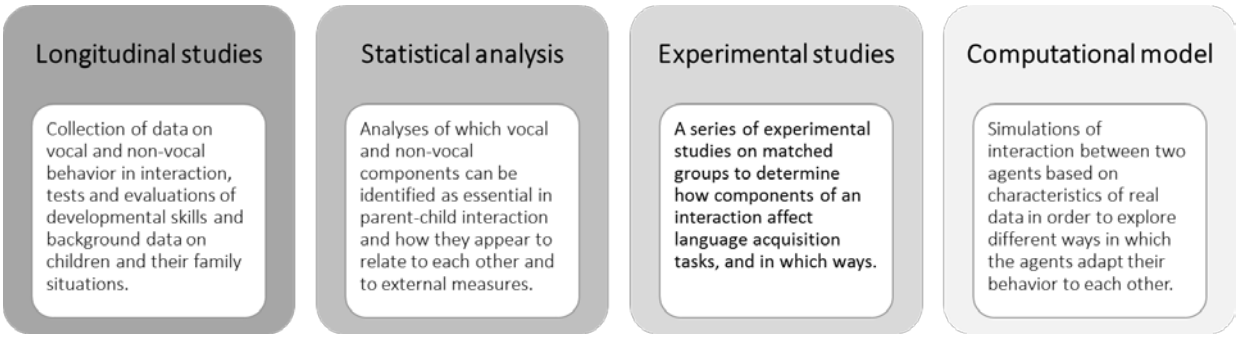


Figure 1.

Through the longitudinal study, we gathered three different types of data: i) vocal and non-vocal behavior from child and parent as they interacted in the studio, ii) test-results on language production/perception, working memory, executive functions (like inhibition, selective attention, cognitive flexibility) and phonological awareness; and iii) background data on family income, education, health etc.

To find patterns in this data, different statistical methodologies will be used (e.g., logistic regression analysis and factor analysis). To test the reliability of these results, matched groups will be used to experimentally test findings from the statistical analyses. In parallel to this, work on modeling the acquisition process will be done (Marklund et al., 2017). At present, no experimental studies have been conducted and we are awaiting the transcription of a larger number of files in order to proceed with statistical analyses on larger parts of the data (see Publications and Presentations for the studies done on smaller parts of data).

Participants and test data

1000 invitations were sent out to parents who had children born in August and September of 2013. The addresses were bought randomized from the Swedish Tax Authority. Of these 1000 invitations, 85 families contacted us and enrolled in the study. There was no payment for participation; the only benefit for the parents was copies of the video recordings of their children and themselves, and a copy of the vocabulary growth reports they themselves provided for their children in connection to each visit. We had 49 boys and 36 girls starting the study and the sample of 71 children remaining at the age of four-and-a-half consisted of 40 boys and 31 girls. Of these, 23 have one language in addition to Swedish and 6 have yet more additional languages. The languages represented in the group are Spanish, Catalan, Persian, Serbian, Thai, Armenian, Norwegian, English, Russian, Polish, Greek, Finnish, Portuguese, Danish and Dutch. Information from the follow-up questionnaire indicates that of the 71 remaining children, 14 had contact with health care (psychologists, physiotherapists, etc.) in relation to developmental issues of a more general kind and three children had visited or continued contact with a speech therapist. In total, six children had been diagnosed with some kind of delayed development.
During the 4 ½ years of recordings, the following data has been collected in addition to the "free interaction":

Table 1.

Test	Description	Age(s)
Imitation	Imitation games (tongue protrusion; waving; knocking; jumping with blocks; etc.).	0;3, 0;6, 0;9; 1;0, 1;3; 1;6, 1;9; 2;0
Deferred imitation	Games where activities from previous sessions (3 months earlier) are re-introduced.	1;3, 1;6, 1;9; 2,0
Joint attention	Testing the child's ability to follow gaze, gesture and vocalizations. Research-led.	1;0, 1;3, 1;6, 1;9
Routines	Testing the child's understanding of routines such as Bedtime/Coffee break/ Play Doctor/Play House.	1;9, 2;0, 2;3, 2;6, 3;6
Perceptive language	Testing the child's understanding of prepositions, animate-inanimate, and quantifiers.	2;6, 3;6, 4;0; 4;6
Working memory	Forward and backward digit span.	2;9, 3;0, 4;0
Language comprehension	Peabody Picture Vocabulary Test.	2;9, 4;0
Executive Functions	Dimensional Change Card Sorting Task; Flanker Task; Inhibition task (tapping pen).	3;0, 4;0; 4;6
Phonological identification	Digital tablet test of phoneme recognition and discrimination.	3;0
Socio-emotional interaction	Interaction analyses from Free Play with parent and researcher.	All recordings from 0;3 to 4;6
Narrative skills	The Renfrew Bus Story.	4;6
Language production and comprehension	Parent questionnaire (SCDI).	All recordings from 0;9 to 4;6
Behavioral data	Strengths and Difficulties Questionnaire, and Epistemic Curiosity Questionnaire. Parental questionnaires.	3;6

Transcriptions and annotations

Transcribing speech, and annotating behavior in general, includes an aspect of interpretation of the data. Our ambition was to make this interpretation as transparent as possible. Even though speech as such has been investigated at length throughout history, we still struggle with some basic definitions of how to go about transcribing it. Typically, the research question sets the stage for how much detail is needed but creating a corpus possible to use for many different purposes makes this a tricky aspect as well. We chose to follow orthography for the most part, in this making the data searchable. However, we did conform to spelling closer to pronunciation in cases where the established pronunciation has moved further from traditional spelling, for example the word “mig” (“me”) which is always pronounced “mej” [mɛj] and thus spelled that way in the transcript. Also some words are never actually pronounced in colloquial speech but contracted with the preceding word, and we therefor did not write them out either but marked them in the transcripts using ‘, for example “där är” /dæ:r æ:r/ (“there are” becoming “där’e” [dæ:re]). Second, how to define an utterance? To be able to make searches where the different tiers are related to one another, we needed to keep utterances short. To have long utterances would mean that too many instances of gestures, gaze shifting, changes in touch behavior would co-occur with the utterance. Thus, we settled on breaking up continuous speech either when there was a 0.5 sec pause or when the utterance had continued for approximately 5 sec. In these latter cases we used semantic and/or syntactic cues to break the utterance into two.

When an annotator has finished a file, s/he filles in a document named “Information from annotator on ready file”. This document is saved together with the transcriptions and holds information on any peculiarities in/with the transcription and annotation, e.g., information on whether the child cried throughout, whether other people entered the room, whether a mobile phone disturbed the session, whether the technology failed in any way, etc.

The different levels of interpretation applied and the modalities to which we applied them are listed below. This is followed by the complete annotation key.

Vocal/verbal – parent: Orthographic transcription of parental utterances. In addition, we added labels for specific types of utterances, e.g., #FS for Formulaic Speech, #VR for adult-directed.

Vocal/verbal – child: During the first 3 recordings (0;3, 0;6, and 0;9) we used Controlled Vocabulary[1] (CV) for child vocalizations. This was to avoid the huge inconsistency in spelling which would have been the case if 10 people had tried to make out and create a spelling for crying sounds, whines, grunts etc. Basically, the categories were divided into consonant sounds, vowel sounds and sounds of joy/distress/anger.

Gesture – parent: Gestures were divided into categories based on how the annotator interpreted them in the context. They were also described in words, e.g., “right index finger to toy, then back”. This gives us the opportunity to look at i) how frequent different categories of gestures are; but also, ii) whether some movements are regarded as different categories depending on context.

Gesture – child: During the first two recordings (0;3 and 0;6) we used CV for child gestures. Most movements were made with the child’s whole body and were too time consuming to describe in words. From 9 months onward, we used the labels of the CV but wrote them by hand, giving us the possibility to also add gesture categories from the adult set. This was called for since some children started to use gestures more similar to the adult ones, e.g., pointing.

[1] *Controlled Vocabulary (CV) is an opportunity to lock the options in a tier so that only a few pre-chosen tags appear as you mark a time-unit. This speeds-up the process of annotation considerably.*

Gaze – parent + child: We can only make qualified guesses as to absolute focus of gaze, since we do not use eye-tracking devices. However, based on the four cameras and the possibility to zoom in, we have a relatively good notion of when parent and/or child alternate gaze. This has been used to classify the most important changes using a CV containing meeting of gaze, looking at researcher, hands, toys and other.

Touch – parent + child: Being as touch is the most recent addition to the multimodal community of research in regard to language acquisition, we did not have much prior work upon which to base our decisions. We used two tiers per individual, the first one for the dominant hand and for when both hands were used in the same movement/grip, and the second tier for touch behavior that was carried out with the second (often left) hand while the first hand was doing something else. We also indicated which body part was touching/being touched and in what manner. If many parts were in connection with each other simultaneously, the hierarchy in which we annotated them was head to feet, where touching the head or face (in the general case) was regarded as more important to tag than a simultaneously touching of a leg or foot. Skin touch was marked with an S, as it has been shown to differ from cloth touch (Ackerley et al., 2014).

Facial expression – parent: Facial expressions can be measured with great accuracy using FACS (Ekman & Friesen, 1978). We did not do this since i) the subjects moved in such a way that their expressions could not easily be classified with FACS, but also since it was regarded as too time consuming. Advised by a trained FACS-coder, we choose a few labels for what we thought would be easily recognizable and possibly frequent facial expressions. These were used for parents and children starting from 1;3 years of age.

Mood – child: During the first year (0;3, 0;6; 0;9, 1;0) the children were not annotated for facial expression since we had difficulty making out expressions in often roundish and sometimes sleeping children. However, mood was more easily detectable since children tend to use their whole body in expressing frustration, anticipation, joy, etc. From 15 months, we started to use facial expressions for the children as well, leaving out the mood tier.

Context – general: A context tier was used to keep track of researchers entering and leaving the room, and whether or not the child and parent were engaged in a common activity.

Transcription and annotation key

Tiers

1 TIER per person (child and adult)
Vocal: orthographic transcription but close to the actual sounds for *ja* (jag), *de* (det), *va* (var/vad), *å* (att/och) and *dom* (de/dem). Otherwise, keep to orthography in order to make it searchable. If an adult has deviant speech in any way, such as creaky voice, do not mark this in the tier but add information as a comment to the whole file in "Information from annotator on ready file".

IMPORTANT: caps are only used for loud speech (see below). Names of persons, places, Mo/Na/Li are written with lower-case letters.

Vocal/verbal tier

Label	Explanation
[??]	A message from the annotator to have the tag checked, e.g. word [??].
?	Question intonation, placed after the intended word/phrase.
(?)	Unsure transcription.
( )	Swallowed/omitted sounds, e.g. ”den h(är) (het)te Mo”; ”vi(lk)en fin” [Eng: "loo(k) a(t) (th)at"; "i do(n't) wan(t) it"].
&	Interrupted (word), e.g.: ”och &ko kolla här då” [Eng: ”and &lo look here”].
&(phrase)	Interruption (phrases), e.g.: ”det har vi sagt & (för att han hela tiden) eftersom han alltid dreglar på fjärrkontrollen” [Eng: "we said that &(because the entire time) because he was always drooling on the remote control"].
xxx	One or more unknown/inaudible words, e.g.: ”ta den då ta xxx bollen” [Eng: ”take that take xxx the ball”].
a-z	Nonwords with communicative function, e.g. “huh?”, imitation or vocal illustration (“nam nam nam” to eat), or sound effects (“hå!”), e.g.: *nam nam nam*.
:	Extended sounds are marked with colon, e.g.: “hå:”.
Ex *ee*	Filled pauses are transcribed as the sound, e.g.: ”Ee mm hm aa öh”.
-	Continuation intonation, e.g.: ”Ja ska gå å-”.
_	Disfluency due to hesitation, e.g.: “jjjja det tror jag” [Eng: "yyyyes, i think so"], “neeej” are marked as “j_a det tror jag”, “ne_j” [Eng: "nooo" are marked as "y_es", "n_"].
‘	Marks typical (or atypical) reductions, e.g: “har’u”, “är’e”, “var’e” [Eng: "y'ave", "there're", "would’nt’ve"].

CAPITAL	Speech clearly louder than the surrounding speech.
.word.	Speech clearly softer than the surrounding speech. Not whispering.
#VI	Whispering, e.g.: Den e [#VI jättefin] [Eng: it's [#VI really nice].
{word}	A child's deficit pronunciation for a word or when the child imitates the intonation of a word, e.g.: 'appa' but it is evident in the context that the target word is "lampa" → {lampa}.
"word/sound"	Irritation/anger.
%word/sound%	Distorted speech (changed intonation pattern, sudden use of (other) dialect, cartoon figure speech, etc.).
~	Creaking voice, marked on both sides of the word/phrase, e.g.: hon kom ~så~ långt [Eng: she came ~so~ far].
+	Paus <0,5 sec.
(h:)	Audible in or out breathing.
¤word/sound¤	Crying/whining voice. For children 3 months CV is applied instead. Crying/whining adults (imitation?), use approximate spelling, e.g.: ¤uhhhhhh¤
!word/sound!	Excited speech, screaming etc.
/approximation of sound/	Used for coughs, hummings, panting, whissling, kissing sounds, etc. /grunting x 2/ = repeated sound. Also used for other sources of sounds than vocal, e.g. /clapping sound/. Also used for /sound/ for 12-month olds' (and older) sounds in between babbling and words proper.
#word/sound#	Laughter. Put # around the utterance produced while laughing, e.g.: #ja det va de värsta# ja varit med om. When only laughter: ###.
¨word/sound¨	Singing, humming.
#VR	Adult directed speech (usually in interaction with experimental leader); child directed speech is unmarked/standard. Place [ ] around the utterance, e.g.: [#VR ska jag läsa i boken?] ja ska vi läsa? [Eng: [#VR shall I read the book?] yes shall we read?].
#IN	Ingressive speech. Place [ ] around the utterance if it is part of a longer sequence: [#IN ja de vore ju] en nåd att stilla bedja om [Eng: #IN yeah that would sure be] a consummation devoutly to be wished].
#LA	Word or phrase in other language than Swedish. [#LA the thing] du vet [Eng/Spa: [#LA la cosa] you know]. If you can't identify the language [#LA xxx].

#FS	Formulaic Speech (frozen phrases). A FS could be an idiom like ”better late than never”, but also (in CDS) expressions that reoccur among many parents and that you recognize, e.g. ”titta lampa”, ”kossan säger… [#FS_å hur låter kossan/vad säger kossan?]. [Eng: ”look lamp”, ”the cow says"… [#FS_oh what sound does the cow make/what does the cow say?].
#UG	Ungrammatical or unsemantic utterances: [#UG_utterance]. Example: [#UG_den va in hons hand][Eng: [#UG_is was in she's hand]; [#UG_ni har två många][Eng: [#UG_you have three many]]; [#UG_nelly hon ramla bakom på stolen][Eng: [#UG_nelly she fall behind on the chair]]; [#UG_fast ja har långt hår för ja e ju blond][Eng: [#UG:_well I have long hair ’cause I’m blond]]; [#UG_ja kunde skriva de själv utan å stava][Eng:[#UG_I could write it myself without spelling]]; etc.

Controlled Vocabularies

CV VOCAL TIER (children 3, 6 and 9 months)	Grunting; Panting; Cooing; Babbling; Laughing; Whining/Crying; Screaming/Shrieking; Other.
CV GESTURE-BABY TIER (children 3 and 6 months; from 9 monhts use lables but without CV)	Wiggle whole body; Wiggle with arms; Wiggle with legs; Grab r-hand; Grab l-hand; Grab both hands; Object in mouth; Other. OBS! The hierarchy for Gesture Baby is: mouth →hands/upper body → feet/lower body.
CV GESTURE-CHILD 1 and GESTURE CHILD 2 (children from 9 monhts) OBS! TWO TIERS FROM 9 MONTHS! OBS! FROM 15 MONTHS, USE ADULT TAGS!	Tier 1 is used for the most prominent movement (hands → head → feet). If both hands are used for different purposes, use Tier 1 for the right hand and Tier 2 for the left hand. Use a combination of terms from CV GESTURE BABY and GESTURE ADULT (see below).
CV MOOD TIER (children 3-12 monhts). From 15 monhts, use FACIAL	Content-Alert; Excited; Frustrated; Anticipating; Sad/angry; Other; Out of frame.

EXPRESSION for the child.	
CV GAZE-CHILD TIER (children all ages).	qp-gaze (looks at parent); qr-gaze (looks at researcher); o-gaze (looks at our chosen objects); qa-gaze (looks at others stuff or unclear); h-gaze (looks at own or parents hands); out of frame.
CV GAZE-PARENT TIER	qc-gaze (looks at child); qr-gaze (looks at researcher); o-gaze (looks at our chosen objects); qa-gaze (looks at others stuff or unclear); h-gaze (looks at own or parents hands); out of frame.
CV FACIAL EXPRESSION-PARENT/CHILD (from 15 months)	Fear; Joy; Neutral; Sad; Angry; Irritated; Surprise; Interest; Anticipation; Out of frame; Concern; Drama.
CV DISTANCE (one per recording)	Body contact; Within reach; Out of reach.
CV CONTEXT TIER (one per recording)	ENTER_researcher; EXIT_researcher; Instructions (talk between researcher and parent; instructions); Conversing_adult (small talk researcher-parent); CONVERSING_child (talking to child, when none of the following tags apply); PLAY_mo/na/li (play with Mo, Na or Li); PLAY_peekaboo; PLAY_singing (singing, rhymes, rhythmical sounds); PLAY_book (play with books); PLAY_object (play with any of the other toys available); PLAY_non-toy (parents, own clothes, etc.); Other (when nothing above fits the situation). Always judge context based on the adult behavior.

Gesture tier

One per child (from 9 months, use two per child, and from 12 months, use the annotation tags from adult AND child). Mark beginning and end time for gesture. Right and Left hand respectively. The dominant hand is used for GESTURE 1. Describe the gesture in words (e.g. both hands out from body, palms facing up, "it's all gone"). Annotate FUNCTION whenever possible (**DEICTIC, ICONIC, EMBLEM, EMPHATIC, GROOMING, EMOTIVE, OTHER**). Example: DEICTIC_points with right hand index towars the bunny on the floor. One gesture can have many functions like ACTION/EMBLEM_makes peek-a-boo 4x.

DEICTIC = pointing gesture (whole hand, index finger, middle finger etc. Inlcude pointing with gaze and with object – mark which kind of pointing it is).

ICONIC = decriptive gestures symbolising e.g. shape, distance, height etc. Visual element of actions could also appear, e.g. "pull back" together with a movement of pulling something.

EMBLEM = conventional gestures like waving, clap hands, put index-finger in front of lips for "husch", etc.

EMPHATIC = gestures/movements/actions marking rhythm, e.g. drumming with fingers on table; other gestures used for marking time or to emphasize something said.

EMOTIVE = gestures/movements appearing together with emotional utterances of some sort (vocal, verbal, physical, etc). Positive and negative; could sometimes be "regular" gestures (like the ones above) but used with a different force. In these cases, mark as both EMOTIVE and, for example, EMBLEM.

- Some facial expressions are marked as gestures, e.g. /smiles/: EMOTIVE_smile.

- Some actions are marked as gestures, e.g. /hugs/: EMOTIVE_hugs child; EMOTIVE_/kissing sound 3x).

GROOMING = gestures/movements like adjusting hair, clothes; scratching, pick the nose, etc.

SHOW/OFFER = a child/adult holds out an object in order to show it (often gaze alternates between object and person). Sometimes show and offer are similar. Use SHOW/OFFER if uncertain, otherwise pick one of them.

ACTION = movements where you are uncertain if it is a conventional gesture or something else, e.g. movements/actions with toys. Decribe the activity in words. Example: rhymes and clapping games ("itsy bitsy…" etc).

OTHER = gestures/movements not fitting in any description above.

Touch tier

TOUCH TIER 1 and 2: (two per child/parent). 1 is the dominant, use it for the right hand/foot and when both hands/feet are used. 2 is used for the left hand/foot or when they do different things with their right and left hand/foot. All kinds of touch should be included, disregard aspects like perceived intentionality. Mark the parts and forms of movements in the following order:

R/L_Bodypart toucher-Bodypart touched_Type of touch _(S).

E.g. R_hand-foot_stroke_S (= right hand touches foot, skin-to-skin touch). If it is on CHILD TOUCH TIER, the child is touching the parent, and vice verse for PARENT TOUCH TIER. When both hands/feet are used, write the body part in plural and do not use R/L, e.g. hands-leg stroke. If there are more touch than there are tiers, use the following hierarchy and disregard the least important: head → hands/arms → leg/feet.

R/L	Right and Left
Body parts	
Hair	Hair, not if you touch the head.
Head	
Face	
Upper (half)	Use when many parts of the upper body are touched or when the parent lifts up or supports the child.
Chest	
Back	
Abdomen	
Arm(s)	
Hand(s)	
Finger(s)	Use instead of "hand" if fingers are used more specifically.
Lower (half)	Lower body, see description for Upper body.

Leg(s)	
Foot/Feet	
Bottom	
Side(s)	
Manner of touch	
Stroke	Stroke/caress, the same direction or back and forth, often quite slowly.
Groom	Touch with a practical purpose, adjust clothes, wipe away something, adjusting the microphone etc.
Hold	Hold a body part or support the child. Takes more “power” than Rest (below).
Pull	Pull towards you.
Push	Push away from you.
Lift	Lift, and only the lifting, not the holding child in air.
Rub	Rub back and forth, often quite fast.
Rest	Rest, touch without movement or pressure. Not a hold-movement.
Play	Playful touch, often with feet/hands. Looks very different from case to case.
(Tickle)	Tickle. Now included in Play.
Press	Press, no grip and not with the purpose to move the child or body part (Push).
Poke	Poke, similar to Press, but mostly with a finger or sometimes the toes (not the whole foot).
Kiss	
Suck	
Kick	Kick, if the child for example lies on his/her back and wiggles and kicks on the parent’s legs.
Pat	
Scratch	
Hit	When child/parent hits with hand/s.
Hug	Can be used when both hug each other (then tag on both TIERs) but also when one hugs the other.
Other	Use when there is some other kind of manner, e.g. brush.
Skin-to-skin	
S	When they touch cloth, do not use the S.

Comment tier

Use this tier for all kinds of uncertainties or comments connected to the annotation on other tiers. Mark with initials and date. Also mark which tier the comment concerns. If needed, have two comment tiers.

Additional markings

Repetitions

Applies to VOCAL & GESTURE TIERS.

After a repetition (defined as: identical or close to identical repetition of utterance, word, sound, gesture, action) add on the same tier [#RE_name1/copy-name2/source]. Name 1 is the person repeating and Name 2 is the source. Self-repetitions are marked [#RE_name-name]. Ett exempel:

Ex. Parent: are we to read the book?

Researcher: yes, read the book [#RE_researcher-parent]

Child: [#DEICTIC_points with r-index to teddybear]

Parent: [#DEICTIC_points with r-index to teddybear][#RE_parent-child]

Ex. Parent: do you wanna read the book

Parent: come let's read the book [#RE_parent-parent]

"Close to identical" has to be consider on a case by case basis. We want to catch expansions, for example.

Ex. Child: train

Parent: yes look, it's a train [#RE_parent-child]

It is better to include more than less in the #RE-tag. As for time-sequences, the repetition does not need to follow the source directly. Do not mark the source, only the repetition of it. As for gestures, a tighter time-frame is needed, the gesture should be repeated in close proximity to the source-gesture. However, if a parent makes "itsy bitsy spider"-movements (or other heavily conventionalized gesture sequences) and the child (without prompting) makes the same movements, mark it as a repetition even if the movements lag behind in time.

Changes as children grow older

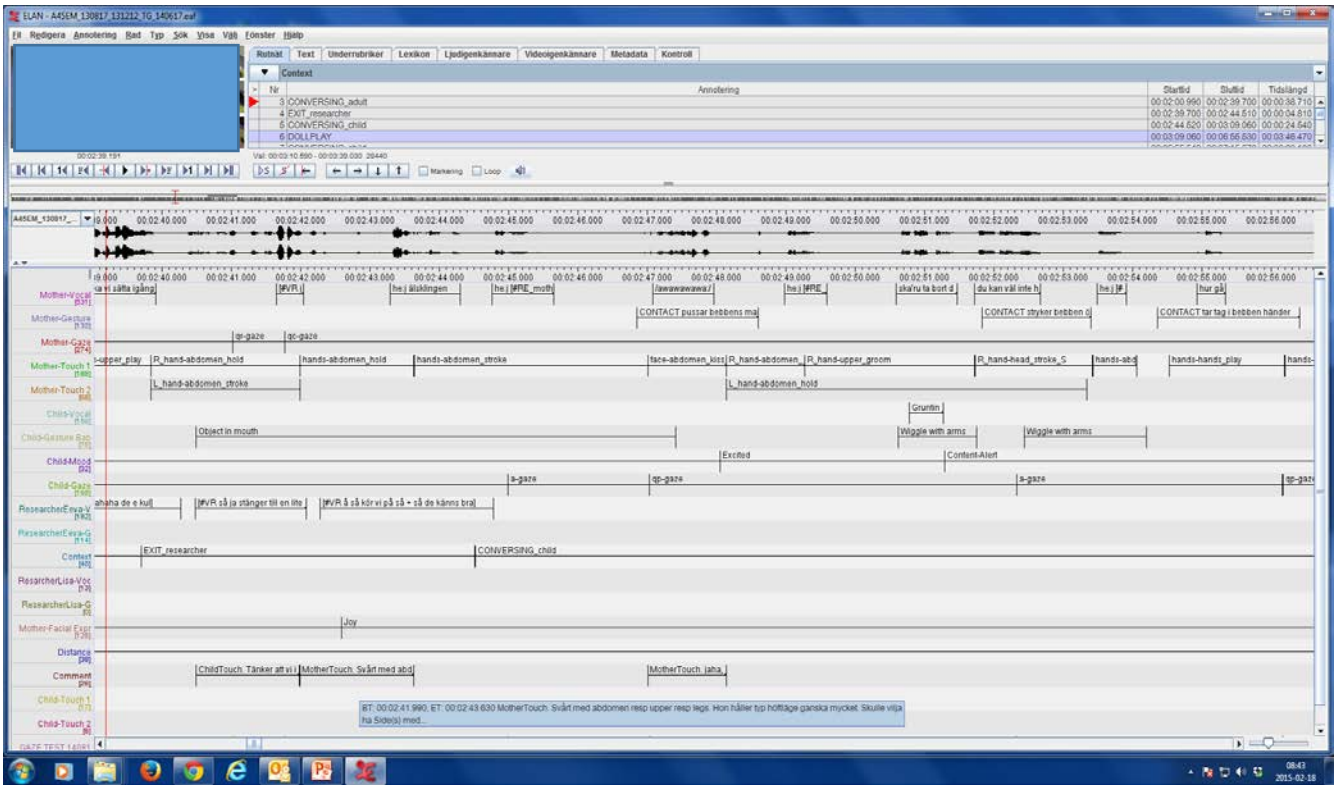
Age (months)	Notes	CHANGES TO KEEP IN MIND (*marks a change).
3	CV for CHILD VOCAL, BABY GESTURE, CHILD MOOD	

6	CV for CHILD VOCAL, BABY GESTURE, CHILD MOOD	
9	CV for CHILD VOCAL, CHILD MOOD	*CHILD GESTURE: no CV any longer. *CHILD GESTURE 2 tiers: CHILD GESTURE 1 for primary gestures, CHILD GESTURE 2 for secondary gestures OBS! Combine adult- and child tag in the Gesture-TIERS.
12	CV för CHILD MOOD	*No CV for CHILD VOCAL any longer.
15	CV för CHILD FACIAL EXPRESSION	*No CHILD MOOD any longer, CHILD FACIAL EXPRESSION using the same CV as PARENT FACIAL EXPRESSION. *Adult gesture tags are preferred.

Annotations in ELAN

Below is an illustration of how a transcribed and annotated file looks. ELAN allows for all 4 cameras being included and you can easily swap between them to get the best angle of a participant.

Figure 2.



Discussion

Transcription and annotation work. It is often stated that the research question determines the level of detail needed in a transcription of vocal/verbal interaction. In this case we had some clear ideas of the specific studies we wanted to perform on the data, but we also wanted to create a corpus that could be of use to others. One available option is to go with phonetics completely and use the IPA throughout. The other end of the spectrum is to cling to spelling rules and orthography, and sort of "clean up" the vernacular and write down what you believe the interlocutors opted for. Our choice ended up in between these options. To use IPA would probably have given us better consistency (given the transcribers are skilled in using IPA) but would have made large-scale searches on a lexical level very difficult. Using strict orthographic rules of Swedish would have speeded up the transcription process, and increased the consistency, but would make all idiosyncratic aspects of the data – and the individuals within it – undetectable. It would also have proved difficult with the children's language as the first phases of language production relies on the interlocutor's interpretation ability. Using a transcription policy where we relied mostly on orthographic rules but made adjustments to these when we saw it fit, had the advantage to capture early stages of language production and also to capture the actual input children receive in terms of repetitions, self-corrected utterances, ungrammatical language and pronunciation idiosyncrasies among both parents and children. The Swedish orthography is fairly opaque, making the task manageable. English would be difficult and Spanish easier. Probably, each language will have to find its own way to handle this complicated task of transforming wavelengths into letters. The main gain with keeping the orthographic base was to enable searches in the corpus. By adding symbols for different kinds of phonological or semantic aspects, we also manage to search for frequent trends in the language or for developmental traits.

Regardless of the choice of detail, transcribing free interaction is tedious work. No matter how detailed the instructions are, there will be differences in consistency in terms of spelling used, accuracy in pauses, length of utterances etc. And this is if you have managed to capture what is actually said. All researchers who have ever transcribed will recognize the experience of going back to a transcribed file, listening to it again, and realizing the person in fact said something else than what you were sure you had heard and thus transcribed. The MINT-database will have faults of this kind, and the plan for handling this is that future users supply information as they discover faulty transcripts or annotations in their research process. In time, this will improve the data and make it as consistent as possible.

Vocal/verbal behaviour. The tier for verbalizations is named vocal/verbal and this is to capture our interest in vocal behavior preceding and accompanying language. Parents make all kind of noises as they interact with their children, especially when the children are infants. They also tend to imitate the noises made by the children. Adding to this is all the different forms of laughter, giggling, sighs and grunts that accompany the interaction. It is – as far as we are aware – not known if these metalinguistic traits are culture specific and/or if they relate to language development in any manner. By transcribing them along with the strictly verbal code, we get an opportunity to test this further on.

The annotations made on vocal/verbal behavior relate to areas that were currently of interest to the researchers involved in the project. We therefore have a symbol for Formulaic Speech (#FS), ungrammatical/unsemantic language (#UG), repetitions (#RE)[2], adult-directed speech (#VR)[3],

[2] *Repetitions refer both to when the speaker repeats his/her own utterance and to instances when they repeat the interlocutors' utterance or gesture. In each case the #RE is followed by the information of who repeats whom, e.g. #RE_mother-child indicating that it is the mother repeating the child's previous utterance.*

[3] *Child-directed speech was considered the norm, and everything uttered in the room while the parent and child were alone was considered child-directed. The marking of adult-directed enable us to study pronunciation differences and other differences that might adhere to speech performed in these two different conditions.*

ingressive speech (#IN)[4] and code switches between Swedish and other languages (#LA). In studies on other matters of interest, researchers will in the future have to add more and other labels to the data. Further, phonetic and detailed acoustic analyses of the whole material can be performed on information from the audio files. This information will be added to the annotations as searchable tags. For example, fundamental frequency can be analyzed and marked-up with information regarding pitch; and formant frequencies of target words (*Mo*, *Na* and *Li*) can be marked-up with information about the first four formant's frequencies throughout the vowel. Additionally, a continuous representation of the vocal intensity across whole utterances can be added and made searchable in the corpus.

Gaze behaviour. Gaze made use of CV, which speeded up the annotation procedure. The drawback is that details are missing. The gaze behavior we decided was most interesting was gaze at the interlocutor's face/eyes, her/his hands, or at the objects they had to play with. All other gaze behavior we called Other. This "other" could be the cameras, the ceiling, out through the window, etc. We also made changes in the guidelines along the way – and then went back and corrected earlier annotations – as we realized something to be important. This was the case for gaze at hands, where we found evidence in recent research (Yu & Ballard, 2002; Yu & Smith, 2017) for this kind of gaze being of particular interest in word learning.

Gestural behaviour. As for gestural behavior, we had difficulty in knowing what could potentially be of interest at a later stage. The compromise reached was to use labels for some form-related gesture behaviors (e.g. iconic gestures, emblems, and deictic pointing). However, the co-speech gestures most frequent in speech are completely context-dependent, as a gesture could be used iconic, deictic, emblematic, emphatic, or have a number of different functions. The choice settled on, was to use a label for all gestures, even if the label was Other (i.e. not iconic/emblem/emphatic/deictic) and then add a basic description of what was taking place, e.g. right hand to stomach, or both hands forming a circle in front of chest. In studies on gestures we thus have the opportunity to search for what has, by the transcriber, been interpreted in a particular way, but also search for all instances with gestures that contain right hand, left hand, both hands, and different shapes of the hands. However, this meant that we could not use CV, but had to write down even the labels. Hopefully this will turn out to have been worthwhile.

Touch behaviour. Touch behavior was the behavior of which we had least prior knowledge. By engaging a researcher who had studied mother-child touch (Agrawal, 2010) we managed to settle on a list of behaviors that could prove important and still be limited enough to make the annotation plausible in terms of time and effort. No CV could be used but we limited the body parts and manners of touch that was allowed to choose between. We also made a specific symbol for when the touch included skin-to-skin, since this particular behavior has turned out important in other studies (Ackerley et al., 2014). Body parts were further ranked in a hierarchy where we deemed movements of the upper body (head, hands, arms) as more important than those of the lower body (legs, feet). If we had to make choices for what to include in an annotation (although we had two tiers for touch), the choice would be to include the touch of the upper body prior to lower body.

Mood and facial expression. It has been suggested that infants living with depressed mothers are delayed in their language acquisition (Murray et al., 1993; Sohr-Preston & Scaramella, 2006). We had no means to check for parental depressive symptoms, but even without a depression, mood and facial expression affect the quality of an interaction as well as the activities taking place in the interaction lab. By adding a tier for mood (for the infant) and facial expressions for the parents and the child (when s/he turned 15 months) we can find the recordings that stick out due to a large proportion of distressed children or angry/irritated parents. We were also curious as to if the parents' facial expression changes as the children get older or if the facial expressions depend on the researcher being

[4] *Swedish is known to use ingressive speech, at least when it comes to some well-known phrases and words such as affirmative "yes" in reply to questions. In-breathing in itself can, in the right context, mark a yes or agreement. By adding the tag #IN we were looking for other cases of ingressive speech, something which would have been very difficult to find in searches on a lexical level at a later stage.*

in the room or not. Cultural aspects are difficult to investigate in a Swedish sample, but the possibility to compare our findings with those from other cultures is available.

Context, activity and distance. The importance of these tiers relates to meta-information. We will need to know if specific gestures, utterances and behaviors are restricted to specific contexts and activities, or part of the "general behavior" if there is such a thing. Ritual-like behaviors are frequent in child-adult interactions and it could be valuable to be able to sort out and disentangle rituals from other kinds of interactions. A ritual like "itsy bitsy spider" and putting dolls to bed, to take two examples, would include use of particular gestural behaviors and frozen phrases like "night night" and so on. Some parts of the transcriptions also include the parents or children singing and the activity-tier makes it easy to see when a singing correlates with "nursery rhyme" and when it appears within "typical" interaction, where it is more sparse and unexpected.

References

Ackerley, R., Backlund Wasling, H., Liljencrantz, J., Olausson, H., Johnson, R. D., & Wessberg, J. (2014). Human C-Tactile Afferents Are Tuned to the Temperature of a Skin-Stroking Caress. *The Journal of Neuroscience, 34*(8), 2879-2883

Agrawal, P. (2010). *Development of a coding system for touch: An analysis of early mother-child interaction*. (Unpublished Doctoral Dissertation) Indian Institute of Technology, Delhi University, India.

Berglund, E. & Eriksson, M. 2000a. Communicative Development in Swedish Children 16-28 months old: The Swedish Early Communicative Development Inventory--Words and Sentences. *Scandinavian Journal of Psychology 41*, 133-144.

Berglund, E. & Eriksson, M. 2000b. Reliability and content validity of a new instrument for assessment of communicative skills and language abilities in young Swedish children. Logopedics Phoniatrics Vocology 25, 176-185.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System*, Palo Alto, California: Consulting Psychologists Press. Retrieved from http://www.paulekman.com/product-category/facs/

Eriksson, M. (2017). The Swedish Communicative Development Inventory III: Parent reports on language in preschool children. *International Journal of Behavioral Development*, *41*(5), 647. doi:10.1177/0165025416644078

Gerholm, T. (2007). *Socialization of verbal and nonverbal emotive expressions in young children*, Dept. of Linguistics, Stockholm University, 2007. Doctoral dissertation.

Marklund, E., Pagmar, D., Gerholm, T. & Gustavsson, L. (2017). Computational simulations of temporal vocalization behavior in adult-child interaction. Proceedings of Interspeech 2017, pp. 2208-2212. Interspeech 2017: Situationed Interaction, Stockholm, Sweden, Augut 20-24. DOI: 10.21437/Interspeech.2017-1289

Murray, L., Kempton, C., Woolgar, M., & Hooper, R. (1993). Depressed mothers' speech to their infants and its relation to infant gender and cognitive development. *Journal of Child Psychology and Psychiatry*, *34*(7), 1083-1101.

Nilsson Björkenstam, K. & Wirén, M. (2013). Multimodal annotation of parent-child interaction in a free-play setting; In: Multimodal Corpora 2013: Beyond Audio and Video / [ed] J. Edlund, D. Heylen, P. Paggio, 2013; Conference contribution

Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by Category: ELAN and ISO DCR. In LREC.

Sohr-Preston, S.L. & Scaramella, L.V. (2006). Implications of Timing of Maternal Depressive Symptoms for Early Cognitive and Language Development. *Clinical Child and Family Psychology Review*, *9*(1), 65–83. doi:10.1007/s10567-006-0004-2

Yu, C., & Ballard, D. H. (2002). Understanding Human Behaviors Based on Eye-Head-Hand Coordination

Yu, C., & Smith, L. (2017). Hand-Eye Coordination Predicts Joint Attention, Child Development, DOI: 10.1111/cdev.12730

Stockholms universitet/Stockholm University
Department of Linguistics
SE-106 91 Stockholm
www.su.se

Stockholms universitet