[PREPRINT]
Text accepted (18.07.2013) for publication in "Interaction Studies".
The final publication is available at https://benjamins.com/#catalog/journals/is/main

Tutoring in adult-child interaction. On the loop of the tutor's action modification and the recipient's gaze

Karola Pitsch[1], Anna-Lisa Vollmer[2,5], Katharina Rohlfing[3], Jannik Fritsch[4], Britta Wrede[5]

[1]Interactional Linguistics & Human-Robot-Interaction Group, CITEC, Bielefeld University, Germany
[2]Centre of Robotics and Neural Systems, Plymouth University, UK
[3]Emergentist Semantics Group, CITEC, Bielefeld University, Germany
[4]Honda Research Institute Europe GmbH, Germany
[5]Applied Informatics, Bielefeld University, Germany

Abstract

Research of tutoring in parent-infant interaction has shown that tutors – when presenting some action – modify both their verbal and manual performance for the learner ('motherese', 'motionese'). Investigating the sources and effects of the tutors' action modifications, we suggest an interactional account of 'motionese'. Using video-data from a semi-experimental study in which parents taught their 8 to 11 month old infants how to nest a set of differently sized cups, we found that the tutors' action modifications (in particular: high arches) functioned as an orienting device to guide the infant's visual attention (gaze). Action modification and the recipient's gaze can be seen to have a reciprocal sequential relationship and to constitute a constant loop of mutual adjustments. Implications are discussed for developmental research and for robotic 'Social Learning'. We argue that a robot system could use on-line feedback strategies (e.g. gaze) to pro-actively shape a tutor's action presentation as it emerges.

Keywords: adult-child-interaction, motionese, gaze, interactional coordination, feedback, social learning, tutoring, conversation analysis, quantification

1. Introduction

Recent years have shown a rise in the development of technical systems that can adapt to changing environments, and which can be controlled using naturalistic forms of communication. In this context, researchers strive to endow robotic systems with mechanisms that make it possible for lay users to teach a system new behaviors by way of ordinary language and interaction. Within this 'Social Learning' paradigm, tutoring and imitation scenarios play an important role: a human tutor presents and explains a task to a robot, who observes the human, presumably understands the action, and in turn, attempts to imitate it (Breazeal & Scasselati, 2002; Steels & Kaplan, 2001; Wrede et al., 2008; Cangelosi et al., 2010). This framework creates a set of challenges: Not only are sophisticated learning algorithms required, but also the technical system has to understand what and when to imitate (Breazeal & Scassellati, 2002; Nagai & Rohlfing, 2009). It has to analyze the human's multimodal conduct, and to understand how the tutor structures actions, renders certain aspects salient, distinguishes between task-related and social actions, and so forth. However, robotic systems have only limited perceptual and cognitive abilities. This opens up interesting parallels to very young infants, such that tutoring in adult-child-interaction has been considered as an empirical model for robotic learning (Rohlfing et al., 2006; Zukow-Goldring, 2006; Zukow-Goldring & Arbib, 2007). When presenting new tasks to their infants, parents carefully modify both their speech ('motherese') and their actions ('motionese') to render specific aspects of the action more salient (Fernald & Mazzie, 1991; Brand et al., 2002, 2007; Brand & Shallcross, 2008; Rohlfing et al., 2006). Since this type of tutor conduct could provide input also highly suitable to a robot's observational capabilities, a systematic description of 'motionese' features has been realized (Rohlfing et al., 2006; Vollmer et al., 2009 a,b). Focusing on the tutor's actions, the *sources* and *effects* of these action modifications in the concrete interaction between tutor and learner have been disregarded so far.

In this paper, we adopt an *interactional* perspective and suggest that 'motionese' behavior in adult-child-interaction constitutes a phenomenon of interactional coordination. Starting with the basic Conversation Analytic assumption of the participants' constant 'mutual monitoring' and 'online analysis' (Mondada, 2006), we explore how the tutor's action presentation is co-constructed by both tutor and learner. They are thus considered as one interactional learning system (Bruner, 1985).

Investigating video recordings from a setting in which parents show to their infants how to stack differently sized cups we address the following questions: How is the tutor's (manual) action presentation instantiated and shaped moment-by-moment with regard to the infant's conduct? What function(s) do the tutor's actions fulfill?
In what follows, we first provide a brief overview of tutoring in adult-child interaction (section 2), we then present the corpus (section 3) and the conceptual and methodological groundwork for the analysis (section 4) and lastly, the variability in the tutors' hand motions (section 5). We propose a novel analytical research chain, that links qualitative research informed by Conversation Analysis (section 6) with subsequent systematization of interactional paths and quantification across a larger corpus (section 7). Results are summarized (section 8) and discussed with regard to their implications for tutoring in adult-child interaction and robotic 'Social Learning' (section 9).

2. Tutoring in adult-child interaction

According to the socio-constructionist approach, learning is a social endeavor rooted in the situated and communicational practices of collaborating co-participants (Wertsch et al., 1980; Fogel 1993). In these interactions, the activity of tutoring plays an important role: An expert/tutor helps the novice/learner to understand new actions and attempts to provide support tailored to the learner's specific needs (Gergely & Csibra, 2005; Zukow-Goldring & Arbib, 2007; Zukow-Goldring, 2012). In doing so, the tutor adjusts her presentation to the learner's displayed abilities and state of understanding (cf. 'scaffolding', Bruner, 1985; Vygotsky, 1978) and e.g. gradually reduces support as the learner's ability to perform a given task increases (Pea, 2004). Thus, tutor and learner can be considered an interactional learning system (Bruner 1985) and the activity of tutoring as the participants' joint production (Zukow-Goldring, 2006; Zukow-Goldring & Arbib, 2007). In line with this social perspective on learning, Conversation Analysis considers human communication as the collaborative achievement of interacting co-participants. Thus, we suggest that the tutor's 'motionese' conduct is also best understood as an interactional co-construction. Investigating video recordings of parent-infant tutoring enables us to reveal communicational structures and understand how participants create suitable learning conditions, but does not allow to make any claims about potential changes to the cognitive system (Mondada & Pekarek-Döhler, 2000; Dausendschön-Gay, 2003; Pitsch, 2006). As such, our data does not provide empirical grounds to investigate whether the child undergoes a learning process, so that we conceptually conceive of the child as 'recipient' (as opposed to a 'learner').

When tutoring in adult-child interaction has been investigated from developmental perspectives, the tutor's and learner's actions have mostly been considered individually. Focusing on the *tutor's action presentations*, studies have shown that when parents present new actions/objects to their infants (versus to other adults), they carefully adapt both their speech ('motherese') and their actions ('motionese') to the child rendering specific aspects salient (Fernald & Mazzie, 1991; Brand et al., 2002, 2007; Gogate et al., 2000). Recent research on 'motionese' has shown that parents perform shorter motions with more pauses for their infants (Rohlfing et al., 2006). In terms of measurable action parameters, such as roundness, pauses and pace, the tutors' hand motions differed significantly between adult-adult- and adult-child interaction, and most prominently for very young infants aged 8 to 11 months (Vollmer et al., 2009a, b). While these studies describe and compare features of the tutors' performance precisely, they cannot explain the *sources* and interactional *effects* of the tutor's action modifications.

The *impact of a tutor's actions on the recipient* has been investigated in experimental settings in which infants were shown video clips (as opposed to an actual real-life performance) of action presentations. The study revealed that infants aged 6 to 13 months preferred to look at action presentations using 'motionese' features (Brand & Shallcross, 2008). Indeed, there is evidence that 7-month-olds benefited from action modifications and temporal synchrony between a shaking/leaping gesture and verbalizations when learning syllables (Gogate et al., 2000).[1] Eye-tracking studies suggest that the infants' visual focus of attention can be interpreted as an indicator of their understanding of an action presentation. By 6 months of age, when infants were shown a video clip in which a human hand repeatedly moved objects from a defined starting point to a target object, they were able to follow the hand movement with a short delay. By 12 months, similarly to adults, they were able to anticipate the presenter's hand movements (Falck-Ytter et al., 2006; Gredebäck et al., 2009). Thus, there is evidence that the tutor's action

[1] For older infants (30 to 46 months), different tutor strategies ('communicational styles') have been attested to enhance a child's understanding of a word, such as verbally focusing the child's attention (Reese et al., 1993).

modifications are beneficial to the infant's perception of an action, and that young infants seem to display their understanding of actions through different forms of gaze behavior. However, these studies used video clips of action presentations as stimuli and thus isolated the infant's conduct from the social situation. No conclusion can be drawn about how the tutor's action presentation might impact the recipient during the unfolding course of interaction. For the phenomenon of 'motherese' (i.e. a tutor's verbal action modifications), Smith & Trainor (2008) suggest that the *mothers' acoustic cues change with regard to the infant's feedback*. In an experimental study, during which the mother could see/hear her child on a video screen, they found that mothers raised their pitch significantly as a reaction to their infant positively engaged (by an experimenter) after her attempt to make him happy through their vocalization.

When investigating the real-time *interaction between parent and infant*, some studies point to the central role of the adult's guidance in helping the infant orient to relevant features of an action. Estigarribia & Clark (2007) showed a pattern of subsequent interactional steps for toddlers aged 18 months to 3 years (and thus considerably older than the pre-lexical group considered in this study): (i) the adult attempts to direct the infant's gaze to a relevant object, (ii) the child orients to that object, (iii) the adult introduces new information about the object and (iv) attempts to maintain the infant's attention on the object. In particular, parents relied on the child's first gaze to an object as an indicator of attention, and verbal attention-management was used more often with younger children than with older ones. While these studies conceive of the infant primarily as an observer of the tutor's action presentation, Zukow-Goldring (2006, 2007, 2012) suggests that the caregiver's and the infant's perceiving and acting may be dynamically coupled in 'assisted imitation': When a child attempts (and partially fails) to reproduce an observed action, the tutor helps the child by guiding her body, and eventually reproduces her action presentation resulting in a set of repeated, slightly modified versions (see also De Léon, 2008). As a result, such an interactional perspective points to the relevance of guiding the infant's attention (Zukow-Goldring & Ferko, 1994; Zukow-Goldring, 1996, 1997, 2001; Rader & Zukow-Goldring, 2010), how the task can be integrated into the sequential structure of explaining actions and that tutors modify their actions in repeated presentations. However, they do not address the variability in the tutors' 'motionese' behavior as it emerges on-line.

In summary, there is evidence (i) that tutors modify their action presentation with regard to the infant's age/cognitive abilities; (ii) that the infant's gaze behavior might serve as an indicator of their understanding of repeated actions; and (iii) that the learner's focus of attention plays a central role in the interplay with the tutor. This raises questions about the extent to which an interactional perspective on 'motionese' might be able to shed some light on the variability of the tutor's action presentation: Specifically how are the tutor's manual actions instantiated and shaped moment-by-moment with regard to the infant's conduct? Which function(s) do the tutor's actions fulfill? Could there be a systematic relationship between the tutor's manual action modifications and the learner's focus of attention?

3. Data: Nesting cups

The investigation is based on a corpus of videorecorded data from a semi-experimental study in which 67 pairs of parents were asked to present a set of 10 manipulative tasks to their infants (8 to 30 months) and to another adult (conditions: adult-child- and adult-adult-interaction (ACI, AAI)). The tutor and her co-participant were seated across a table facing each other and used both talk and manual actions for demonstration, while being videotaped with two cameras (Rohlfing et al., 2006). The analysis presented here focuses on the task of 'nesting differently sized cups' as it encompasses a set of simple, repeated manual sub-actions during which the participants' conduct can be compared. Also, it requires the recipient to attentively observe the different parts of the action, i.e. to look at the right place at the right moment in time. The study focuses on the tutors' presentations to pre-lexical infants aged 8 to 11 months (N=18 infants, 10 male, 8 female) as they have been found to prompt the most significant 'motionese' conduct from their tutors (Vollmer et al., 2009b). Also, pre-lexical infants tend to be able to only observe this particular action presentation without attempting to reproduce the action afterwards (Lock & Zukow-Goldring 2010; Zukow 1990). In this way, the analysis can concentrate on the presentation phase while ignoring the child's subsequent attempts of action reproduction.

At the beginning of the 'nesting cups' task, an experimenter places a tray with a set of differently sized and colored cups on the table in front of the tutor and instructs her to demonstrate to the infant how to nest the cups into each other. The adult is asked to start the action with the largest cup, i.e. to first place the green cup in the blue cup (a1), then the yellow into the green cup nested in the blue one (a2), and finally the red into the three already nested cups (a3) (procedure 1, Fig. 1a). However, some parents reversed the order, placing the red cup into the yellow one (*a1), then the yellow cup containing the red one into the green cup (*a2) and finally nested the green cup (containing the red and yellow ones) into the blue cup (*a3) (procedure 2, Fig. 1b).

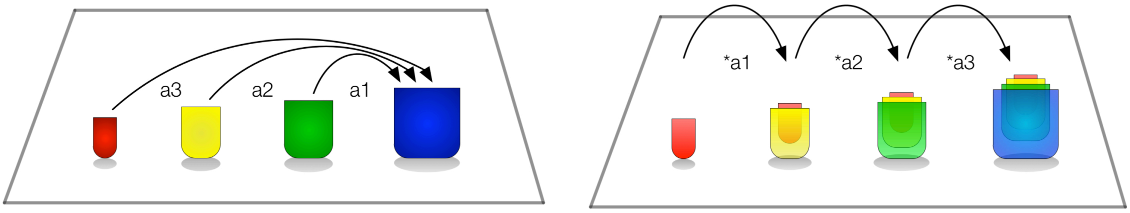


Fig. 1: Nesting cups. **(a)** Procedure 1: a1 – a2 – a3 (considered for quantitative analysis), and **(b)** Procedure 2: *a1 – *a2 – *a3.

For the qualitative analysis (section 6), we initially considered both procedures performing the nesting action. For the subsequent quantitative analysis (section 7), the corpus was narrowed down to include only those cases which used procedure 1 so that a data set was obtained in which the tutor's hand trajectories were comparable. In this way, the quantification was based on a set of 17 parents (9m, 8f), which produced a total of 51 sub-actions for the nesting cups task.

4. Methods: From Conversation Analysis to Formalized Systematization

Our goal was to understand tutoring interactions in humans to motivate robotic social learning, making our task two-fold: Firstly, the participants' multimodal conduct needs to be reconstructed as methodological solutions to re-occuring interactional problems, and systematic communicative procedures and their interactional functions need to be explored. Secondly, these results are to be formalized as descriptions of action sequences and quantified as interactional paths across the corpus. This requires a novel interdisciplinary analytic chain, that links qualitative micro-analysis of videotaped data, modeling and quantification, as well as manual and computational investigation of interaction.

4.1 Conversation Analysis: Basic Assumptions and Analytic Methodology

Exploring the interactional function of 'motionese' requires an analytical framework that enables us to understand the tutor's actions as a sequentially organized interactional achievement. In this line, our analysis is informed by Conversation Analysis (CA) which offers both a specific theoretical perspective on social interaction and a methodology for micro-analysis of social interaction (ten Have, 1999; Sidnell & Stivers, 2013). While CA has originally been developed on the basis of audio data, in recent years a growing corpus of studies has emerged which take into consideration spoken, bodily and material resources and which share the same analytic commitments and concerns (e.g. Heath & Luff, 2013). Thus, our research begins with a set of basic assumptions about human communication:

- *Task orientation*. The participants' actions can be reconstructed as methodological solutions, deployed to solve recurring interactional tasks.
- *Interactivity & Co-construction*. Actions within interaction are considered as a 'joint accomplishment' of collaborating co-participants (as opposed to an individual's act).
- '*Mutual monitoring' & 'online analysis'*. Participants constantly monitor each other, interpret the co-participants' conduct and display their 'online analysis' publicly through their conduct, which, in turn, shapes the coparticipants' actions as they emerge (Goodwin, 1981; Mondada, 2006; Pitsch, 2006; Schmitt & Deppermann, 2007).
- *Sequentiality*. Actions within interaction are built so that they display their relationship to the immediately preceding action and that they make subsequent actions relevant. In this way, the structural organization of a sequence of actions can be revealed and the absence of an otherwise expectable action (i.e. a verbal turn or other 'interactional moves') can be accounted for.
- *Multimodality*. During face-to-face interaction, participants make use of the full range of communicational resources at hand, i.e. speech, prosody, facial expressions, gaze, gestures, bodily conduct, spatial behavior, structures in the environment etc., and deploy them as complex communicative 'gestalts' (Goodwin, 2000).

In this view, the tutor's hand motions previously described as 'motionese' conduct are conceived of as a co-production of tutor and learner, and as a methodological solution to a practical problem that the researcher has to reconstruct in her analysis. The investigation proceeds through a set of manual case analyses. These involve repeated inspection of videorecorded data, and detailed transcription of talk and embodied actions to uncover the precise timing and relationship of the co-participants' actions.

4.2 Integrating Conversation Analysis into Interdisciplinary Research

Aiming at contributing to robotic research, our investigation constitutes a case of ‘applied’ – though basic – research (Richards & Seedhouse, 2005). Some methodological issues need consideration.

(1) *Semi-experimental setting*. Typically, CA explores audiovisual recordings of *naturalistic* everyday interactions, i.e. not specifically arranged for the purpose of analysis. The data investigated here has been elicited in a semi-experimental setting. The recording has been arranged by the researcher for a specific purpose which involved asking the participants to carry out defined tasks, e.g. showing how to nest differently sized cups. In this way a data set of comparable situations is created where participants interact using their own authentic choice of communicational resources and procedures to solve locally occurring practical (interactional) problems. In this sense, the ensuing interations are considered as social events and the data as our ‘field’ to be explored ethnographically (compare e.g. ‘interviews’ as data, ten Have, 1999, 162-181).

(2) *Pre-existing research question and ‘unmotivated examination’*. The aim of CA- and CA-informed analysis is not to test pre-defined hypotheses, but that the researcher should undertake an open, ‘unmotivated examination’ to derive the participants’ procedures and conceptual issues ‘from the data themselves’ (Sacks, 1984, 27; Sacks & Garfinkel, 1986; Lynch, 1993). In the case of ‘motionese’, the phenomenon has been shown to be a prominent feature of tutor conduct. To further explore its interactional function, we thus pursued a pre-existing research question. Yet, the participants’ methods deployed to solve this practical problem need to be reconstructed from the data themselves (ten Have, 1999; Mondada & Pekarek-Döhler, 2000; Richards & Seedhouse, 2005).

(3) *Focus on particularly salient features of the multimodal complexity*. The context of robotics research requires the reduction of multimodal complexity to a small set of the most relevant dimensions. Technical systems have limited perceptual capabilities, and detection and classification of conduct needs to be automated. Here, a particular focus is placed on the visual dimension of the humans’ conduct, as it is particularly relevant for the participants themselves (see section 6.1) and previous research points to its saliency (section 2). While this might seem contrary to the basic assumption of ‘multimodality’, it has been applied broadly at the auditory level of conversations.

(4) *Linking qualitative research with formalization and quantification*. As a qualitative

approach, CA informed research is built around a collection of case analyses, that reconstruct interactional procedures by contrasting different versions of a (similar) practice as well as deviant cases. In the context of robotics research, however, it is relevant to determine whether some interactional procedure is likely to occur frequently, and to detail expectable courses of actions including an estimation of the likelihood of prospective actions. Such quantification is challenging as the goal is not to count isolated events, but rather to consider courses of inter-individual actions, the conditions under which they are valid and the functions they assume (Schegloff, 1993; Heritage & Robinson, 2006, Robinson & Heritage, 2006). To do so, the following elements need to be defined: (a) the 'denominator', i.e. the "environments of possible *relevant* occurrence" of a phenomenon; (b) the 'numerator', i.e. a "set of types of occurrences whose presence should count as events […] and whose non-occurrence should count as absence"; and (c) the "domain or universe being characterized" (Schegloff, 1993). In our case, we propose to formally identify and quantify sequences of action and alternative trajectories (requirements a and b) in a specified context (requirement c) and, as far as data allow, to test whether observable differences are significant on the corpus level.

4.3 Semi-automatic motion tracking and manual annotation

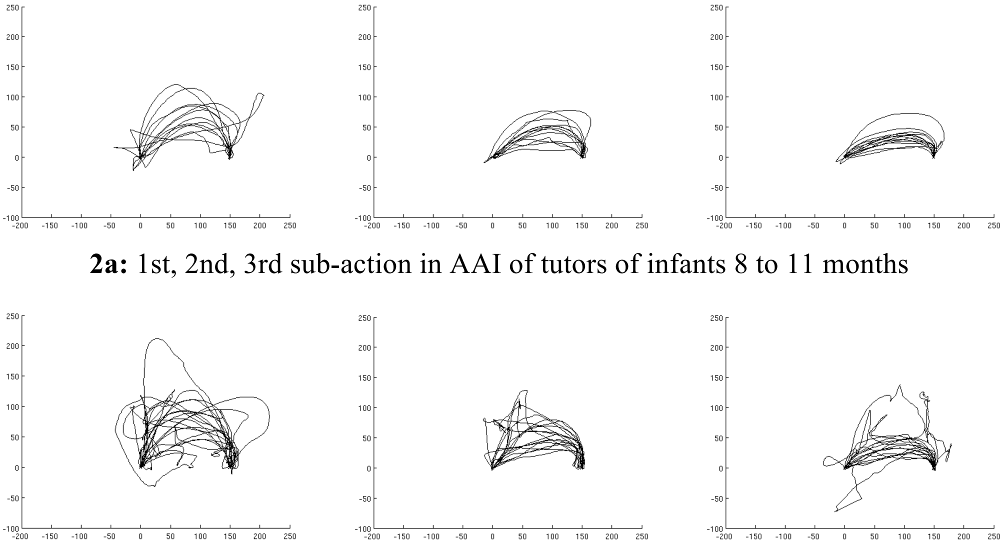
Based on the qualitative analyses, the tutoring interactions were annotated using a set of reduced features containing the most salient information: (i) the structure of the presenter's actions (sub-actions a1, a2, a3, see Fig. 1), (ii) the tutor's hand motions, (iii) the tutor's and (iv) the recipient's gaze direction. While (i), (iii) and (iv) were manually annotated using the timeline-based annotation software ELAN (http://tla.mpi.nl/tools/tla-tools/elan/), the tutors' hand motions (ii) were captured with a semi-automatic 2D motion tracker (Vollmer et al., 2009a). By registering the x and y coordinates of the tutor's hand in the video frame at any point in time, the hand motions can be visualized and measured. Importing these novel methods from computer science enabled us to overcome the challenge of capturing ephemeral visual phenomena such as gestures or body movements, and to describe their actual performance (e.g. trajectories, velocity, acceleration, distance etc).

4.4 Formalization and quantification

To formalize and systematize the interactional procedures, measures and algorithms need to be found to assess the qualitative analyses with computational methods. Thereby an alternative way of building collections of cases is explored that might be able to support CA informed research when assessing larger corpora. Quantitative analyses are undertaken on the basis of the annotations and the motion tracking data. The timestamps and annotation values are parsed and loaded into MATLAB for further processing, i.e. for visualization, algorithmic systematization and quantification (see section 7).

5. Starting Point: Variability of hand trajectories

In following with the observations in the literature of the tutor's actions modifications (Brandt et al., 2002, 2007; Rohlfing et al., 2006; Vollmer et al., 2009a,b), analysis begins with the actual individual hand motions. To gain a better understanding of the tutors' 'motionese' conduct, the tracked hand trajectories have been visualized by plotting them both in comparison to each other and over the corresponding video-clips. Across the data set, this shows that the tutors' actions have a relatively homogenous parabolic shape in the AAI-condition (Fig.2a). The trajectories in the ACI-condition, in contrast, show more variation (Fig.2b). i.e. higher arches and extensive modulations, particularly in the 1st sub-action.

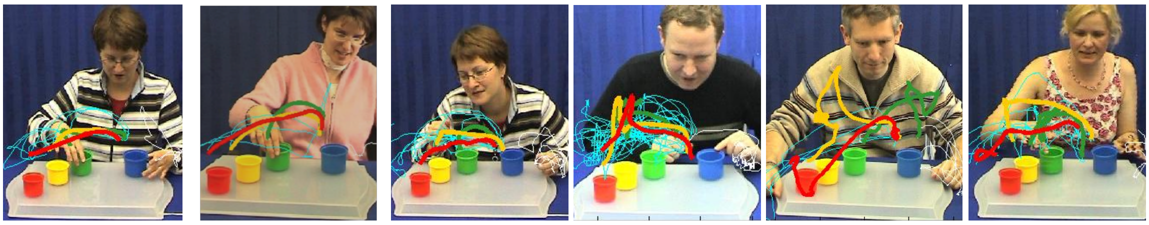


2a: 1st, 2nd, 3rd sub-action in AAI of tutors of infants 8 to 11 months

2b: 1st, 2nd, 3rd sub-action in ACI of tutors of infants 8 to 11 months

Fig. 2: Normalized hand trajectories of groups of tutors in Adult-Child (ACI) and Adult-Adult interaction (AAI)

Considering the participants' individual cases, i.e. the tutors' three consecutive hand trajectories, reveals a set of patterns more differentiated than those suggested for AAI (round arc) and ACI (squared arc) on the same corpus in Rohlfing et al. (2006). We find (i) cases, in which the presenters' hand trajectories are flat without being particularly marked points (Fig. 3a, 3c); (ii) cases, in which the trajectories are more pronounced with a small peak towards the end (Fig. 3b); (iii) cases, in which the presenter's hands perform a peak or modulation at the onset (Fig. 3d, 3e); (iv) combinations of these trajectory types, in particular those in which the first two nesting actions (green, yellow) show a high/pronounced shape, while the third action (red) is performed in a rather flat manner (Fig. 3e, 3f). From an interactional perspective, this raises questions about the functions of the tutors' action modifications and the interactional task they are designed to solve.



3a (AA) 3b (AA) 3c (AC) 3d (AC) 3e (AC) 3f (AC)

Fig. 3: Individual hand trajectories of the tutor in Adult-Adult Interaction (AAI) and Adult-child interaction (ACI). Green/yellow/red trajectories mark the actions of stacking the cup of the corresponding color into the blue one; thin lines represent movements without a cup.

6. Interactional Loop: Orienting attention to aspects of an action

As a first step, the analysis aims to investigate the functions of the tutors' hand motions and therefore explores the interaction between tutor and child with regard to the sequential structure of actions. In the analysis, a special focus will be placed on the role of the tutor's action modifications and the recipient's foci of attention (section 2 and 4).

6.1 Interplay between the tutor's hand motions and the infant's gaze

For the infant to understand the 'nesting cups' task, she must pay careful attention to the tutor's actions so that she can grasp the role of the differently sized objects, and the ordering of apparently repetitive actions. However, since young infants' cognitive abilities are not fully developed, an important task for the tutor consists in establishing and maintaining the infant's attention to relevant features of the action.

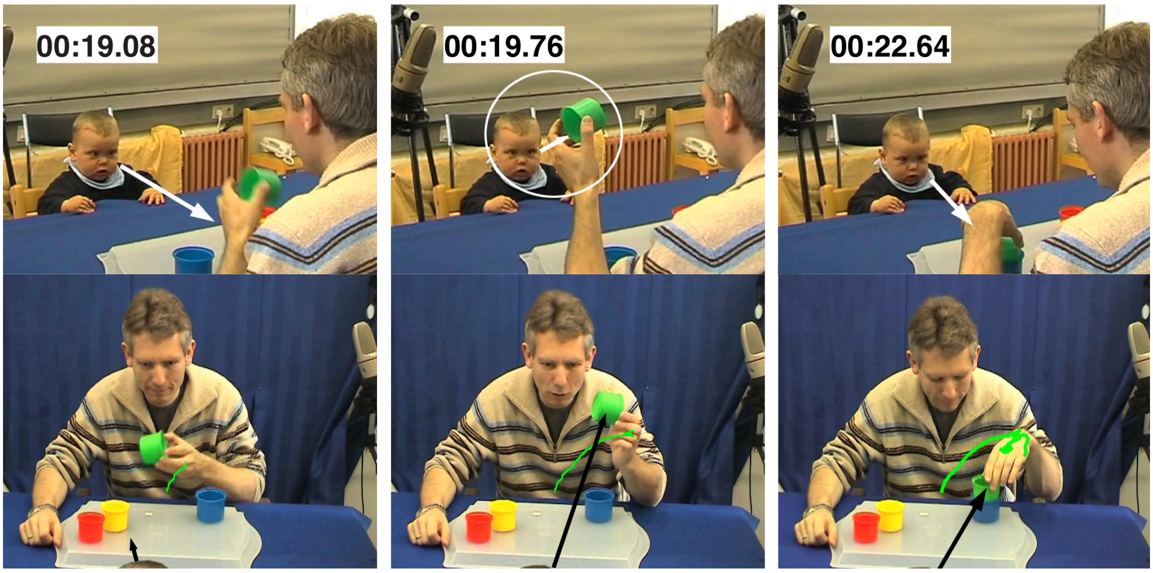
6.1.1 Tutor's hand motions as procedures for orienting attention

Consider the following fragment 1a where a father presents the 'nesting cups' task to his 8-month-old son. After the experimenter has placed the tray with the cups on the table and briefly instructed the father (T1), he picks up the green cup while observing the child (00.19.08). The infant's gaze is initially oriented to the cups/tray, but now his orientation shifts: Within a 0.3 second delay, he begins to follow the green cup. Observing the infant's reactions, the father stops the upward motion of his hand in mid-air. Once the infant's gaze has reached the green cup (00.19.76), he begins to also verbally invite the infant to look (01: "HAVE A LOOK") and to explain: "FIRST of all we take the GREEN (one)" (01). During the explanation, he shakes the green cup in the infant's visual field and the infant indeed maintains his attention on the object. At the end of his explanation (i.e. on "GREEN"), the tutor lowers his hand to drop the green cup into the blue one. Again, the infant's gaze follows within a short delay (00.22.64).

Fragment 1-a (VP001_1_FC): Green cup (a1)[2]

```
01 T1:                      |↑KUCK mal; ERST nehmen wir den GRÜ:|Nen; (1.0) |
                              LOOK;     FIRST take  we  the GREEN one;
   T1-act: |g grab|g lift |hold & shake                         |g place   |
   R1-gaz: @Ø   |@g    |>>> |@g                                             |
                     *19.08    *19.76                                   *22.64
```

[2] For the transcription, each line represents the conduct of the tutor (T) or the recipient (R). Their synchronization is inspired by music partitions via vertical bars ("|"). For the verbal level, we use the conventions described in GAT (Selting et al., 2010). As a basic rule, all words are written in lower-case and without punctuation, so that the latter can be used to describe prosodic phenomena (capital letters = emphasis, "." = elongation of the preceding sound, ";" = falling intonation, "," or "↑" rising intonation"). The tutor's actions are annotated in the line 'T-act' (e.g. "g/y/r grab" = grabs the green/yellow/red cup). Underlining in the line "T" represents the tutor's gaze being directed to the recipient. The recipient's gaze behavior is annotated in the line 'R-gaz' making a distinction between static (e.g. @g) and moving (>>>) gaze.



First, this short fragment shows that the tutor's presentation is a multimodal action consisting of both a manipulative action and an accompanying verbal explanation. At times, talk and action are organized synchronously and thereby provide motion/speech 'packages' separated by combined talk/action pauses (Schillingmann et al., 2009). At the same time, when the tutor starts the action, he is silent and only begins to talk once the cup is already in the air, such that, at least for this initial stage, we can observe the impact of the tutor's visual actions without interference from verbal input. Second, the sequential ordering of the tutor's and learner's actions, and the small delay in the recipient's gaze with regard to T1's hand motion suggest that the tutor's action constitutes a first interactional move that elicits the learner's gaze to follow as a second move. The learner thus appears to use the tutor's hand motion as an orienting device to indicate where to look.

This fragment provides a very clear example of the role of the tutor's hand motion. Since the tutor is silent at the onset when moving his hand and inducing the learner's gaze shift, this suggests a sequential relationship between hand motion and gaze. In other cases, the tutor will often also talk while moving a cup from place to place, making it difficult to disambiguate whether the learner is reacting to the entire multimodal gestalt or to a specific feature of it. Nevertheless, we will focus on the visual dimension, since (i) 'nesting cups' is, in the first place, an embodied action, (ii) we cannot assume that very young infants understand the talk (though prosody and rhythm are relevant), (iii) the tutor's hand motion is the most salient and consistent activity throughout our cases, and

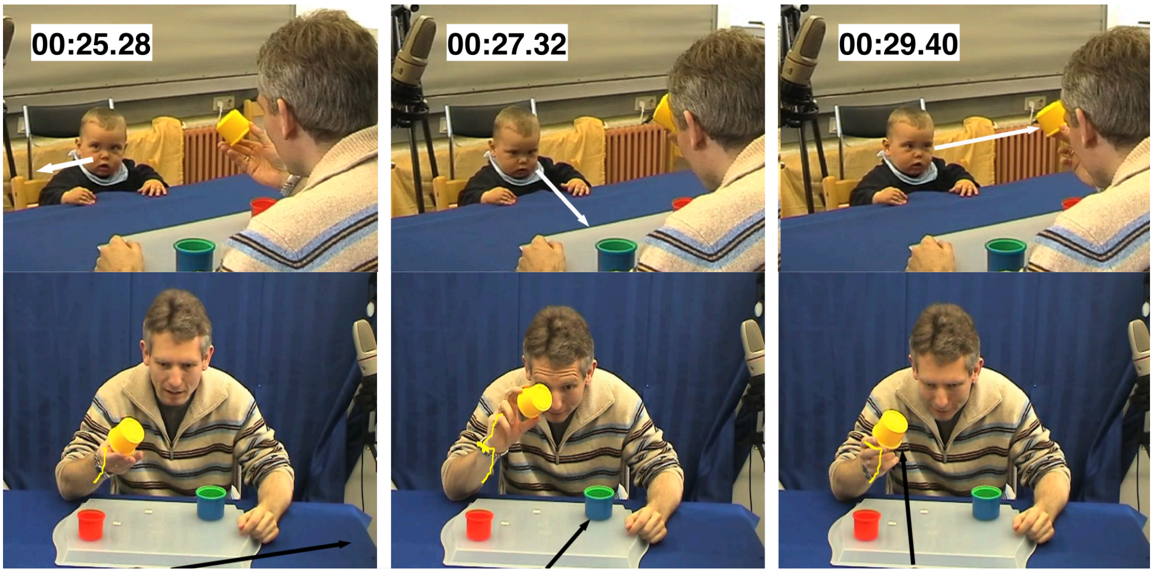
(iv) the child can be seen to systematically orient toward it.

6.1.2 “Interactional loop” between tutor’s hand motions and learner’s gaze

Resuming the analysis after the first nesting action, the tutor’s hand pauses for a second while the child’s gaze turns from the blue/green cup to the side where the experimenter has been hiding. On the opposite side of the tray, the father lifts the next cup (yellow) (00.25.28) and begins to verbally prepare for the next action with the temporal marker “THE:N” (02). The child, however, does not react and shifts his gaze entirely to the opposite side. The tutor reacts by initiating a repair. First he interrupts both his hand motion and verbal utterance. Then he begins to shake the cup while verbally calling for the child’s attention “HELLO RAINER; LOOK here”. This procedure, in turn, induces a shift in the child’s orientation, who then turns towards the yellow cup (00.27.32). In contrast to the first nesting action, this procedure constitutes an upgrade: an attention getter in which the tutor verbally addresses the child with marked prosody while syncronizing a shaking object (e.g. Zukow-Goldring, 1997; Zukow & Ferko, 1994). In this case, it proves successful.

Fragment 1-b (VP001_1_FC). Green cup (a2)

```
02 T1:     (0.5) |DA:NN,|(0.8) ↑HA:LLO ↑RAI|NER;|HIERher kucken;|(0.2) DANN |den |GELBEN
                  THE:N,       HE:LLO  <name;>  HERE    look;          THEN  the  YELLOW
   T1-act: |y grab|y lift|y shake                                |y lift     |   |y place
   R1-gaz:  >>>>>>|@Ø                        |>>>>|@y                     |@b
                        *25.28                      *27.32                     *29.40
```



Once the child’s gaze reaches the hand/cup, the tutor resumes his action showing that, to his understanding, the appropriate conditions for tutoring have been reinstated, in this case, the recipient paying visual attention to the action. The tutor again lifts the cup and restarts his verbal presentation: “THEN the yellow one” (02). The infant, however, does not follow this upward motion with his eyes. Instead, his gaze shifts to the blue cup, i.e. the action’s goal position (00.29.40). The tutor treats this orientation shift as a relevant next move, as he does not attempt to alter it. Rather, he follows the infant’s orientation and places the yellow cup into the blue one.

This second fragment extends the initial analysis on two levels. (i) The repair action reveals the tutor’s *explicit* orientation to the child’s visual attention, and shows the relevance of his gaze trajectory. At the same time, variants of the first orientation procedure emerge as consecutive upgrades, in which the hand motion appears as the constant.

(a1): [lifting hand/cup]

(a2): [lifting hand/cup] + [verbal marker (then)]

(a2): [shaking hand/cup] + [address terms (hello; name), imperative (look)]

(ii) With regard to the interplay of the tutor’s and recipient’s actions, this fragment reveals that not only is the child’s gaze orientation influenced by the tutor’s hand motion, but also that the recipient shapes the tutor’s manual action in a complementary fashion. The tutor adjusts the trajectory of his hand motion and its concrete timing to the recipient’s shifting focus of attention to attract his gaze or waits for it to arrive. In this sense, the tutor’s action presentation is not only constituted by the task of moving the object from place A to B, but becomes ostensive at specific moments and assumes the role of a communicative gesture. Conceptualizing this behavior as an *interactional loop* between the tutor’s hand motions and the recipient’s gaze opens up a new perspective on ’motionese’. It explains how a tutor’s manual action is shaped locally in the actual interaction with the learner, and posits an *interactional account* for the previously observed variability in action presentations.

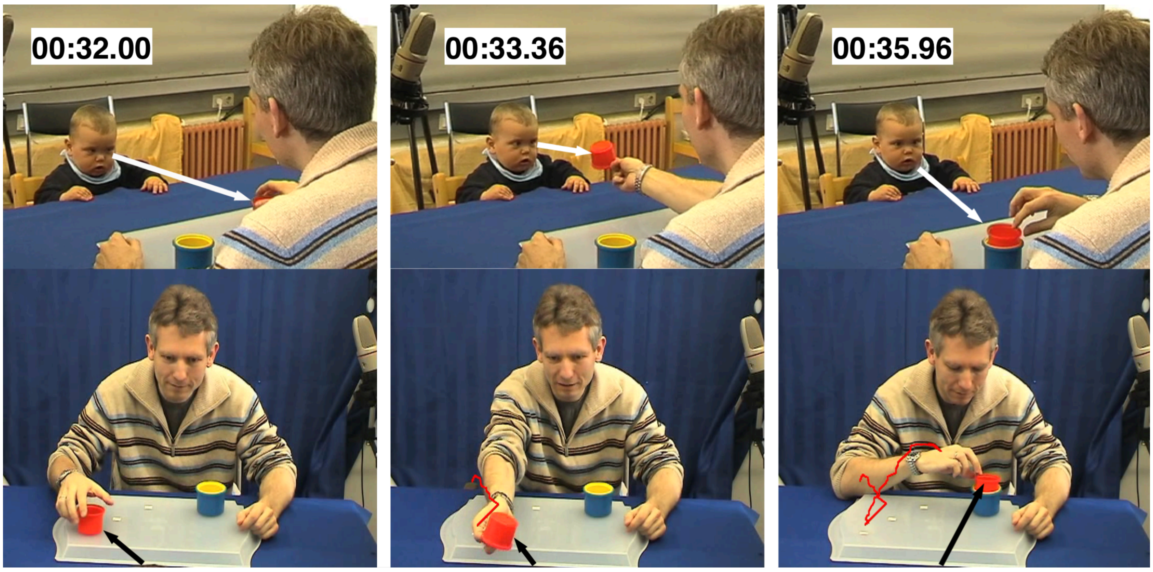
6.1.3 Comparing the trajectories of the different nesting actions

To initiate the third nesting action, the father moves his right hand over to the remaining (red) cup. As the tutor takes the cup, the child’s gaze again begins to follow the hand

(00.32.00). When the child’s line of sight reaches the red cup, the father moves it closer to the infant (00.33.36) and also verbally prefaces the next action (“A:ND THEN,”).[3] The father then moves the red cup over to the blue one, which again induces the child’s gaze to follow after a short delay.

Fragment 1-c (VP001_1_FC). Red cup (a3)

```
03 T1:      (1.0) |(0.2)    |U:ND DANN, (0.2) |den ROten;   |
                             A:ND THEN          the RED one;
   T1-act:              |r grab  |r forward          |r place
   R1-gaz:  .               |>>|@r (follow)                    |
                              *32.00    *33.36         *35.96
```



Of particular interest here is the trajectory covered by the tutor’s hand. While, in the first and second nesting actions, his hand motion is characterized by *high* trajectories attracting the child’s gaze, this third transport action is performed with a particularly *flat* trajectory (00.35.96). This corresponds to how the infant displayed at the end of the second nesting action that he was able to correctly project the tutor’s next action (looking to the goal position even before the tutor’s hand had moved to it) and, importantly, that the tutor had treated this conduct as relevant (following the infant’s gaze with the cup without attempting to repair it). Thus, it appears that under certain conditions (i.e. the child being oriented to the action and having displayed the correct expectations about the

[3] Given the 2D-representation of the tutor’s hand motions, this axial movement cannot be rendered in the visualizations and/or the quantification.

relevant next steps) the tutor re-adjusts his previously *pronounced* hand motions to perform *flat* trajectories, similar to those found in adult-adult-interaction. Thus, not only do the tutor's hand motions function as gestures to orient the recipient's attention at each single step, but, in contrast to repetitive actions, they also display the tutor's interpretation of the recipient's current/changing state of knowledge about the task.

6.2 Anticipating next actions and its impact on the tutor's presentation

The observations made above raise questions about whether tutors might express their understanding of the recipient's current state of knowledge in their hand motions (and possibly also other features) across the different versions of a repetitive action.

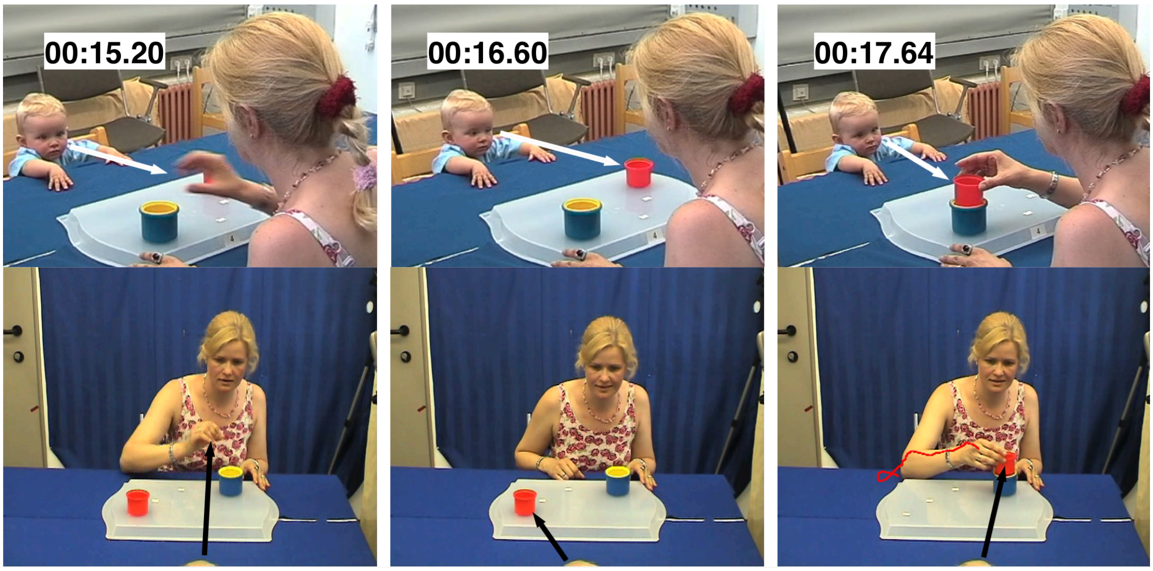
Consider a second interaction fragment that contains the pattern of hand trajectories shown in Fig. 2.c: high arch (1st action), high arch (2nd action), flat motion (3rd action). When the mother transports the first two cups (green, yellow), the infant's gaze follows her presentation and hand motions, similar to fragment 1. However, after the mother has dropped the second cup into the blue one (00.15.20), the infant quickly gazes to the left side, where the remaining (red) cup is placed (00.16.00). The video shows the presenter's right hand still located next to the blue cup, while the infant's gaze is already focused on the red cup. The infant thus initiates the next action with her gaze, and the adult follows. In this reversed dynamic, the infant performs the first move by gazing to the next object, and the adult responds with a second move by following with her hand. The infant anticipates the next relevant step in the presentation and thereby shows her understanding of the ongoing action.

Fragment 2 (VP040_3_MC). Green, yellow and red cup (a1, a2, a3)

```
01 T2:      |↑SPATZ KUCK |mal; |(0.5) |kuck=mal den GRÜ:Nen,|stecken wir HIE:R rein, (0.2)
             DARLING LOOK                look      the GREEN one place   we  HERE   into
   T2-act:  |g grab            |g lift||                      |g>b                        .
   R2-gaz:   @T2           |@g     (follow)                                               .

02 T2:      |(0.2) .hhh |und den  |GEL|BEN |(0.2) in |den GRÜNen, (0.2)|WEG ist=er; (0.5) |
                         and the   YELLOW ONE      in  the GREEN one     GONE it=is
   VP-act:  |           |y lift             |        |y>g              |
   R2-gaz:   .                  |>>>|@y (follow)                                          |
                                                                                     *15.20

03 T2:      |(0.1)|(0.2)|(0.4)|(0.2) und |der RO::TE, (0.4)|(0.1) ↑DSCHUBB;
                                     and  the RE::D ONE              (exclamation)
   T2-act:              |to r |r lift   |r>g                |
   R2-gaz:  |>>   |@r          (follow)
                  *16.60                        *17.64
```



When the tutor then takes the red cup and moves it over to the blue one, the transporting action is different from the first two movements. While the first two actions are performed with a high trajectory, the third one is produced with a particularly flat trajectory (00.17.64). The tutor thus *treats* the infant's gaze behavior as anticipating the next action and as revealing her knowledge and understanding about the ongoing action. This example reveals that the tutor treats the recipient's gaze behavior as an online indicator of his understanding and that a tutor's manipulative action is sensitive to the recipient's display of knowledge.

However, at the same time, not all tutors treat the infants' anticipatory gaze behavior as displaying their state of action understanding. Some parents treat the infants' anticipatory gaze as displaying a lack of attention to the ongoing presentation by attempting to re-orient or repair the infants' attention through a modified, higher hand motion (during the action presentation, see section 6.1). Yet other tutors do not show any reaction to the infants' anticipatory gaze at all. Thus, tutors exhibit different systematic interpretations of the recipient's conduct resulting into different forms of consecutive action paths.

6.3 Losing the recipient's attention and problematic understanding

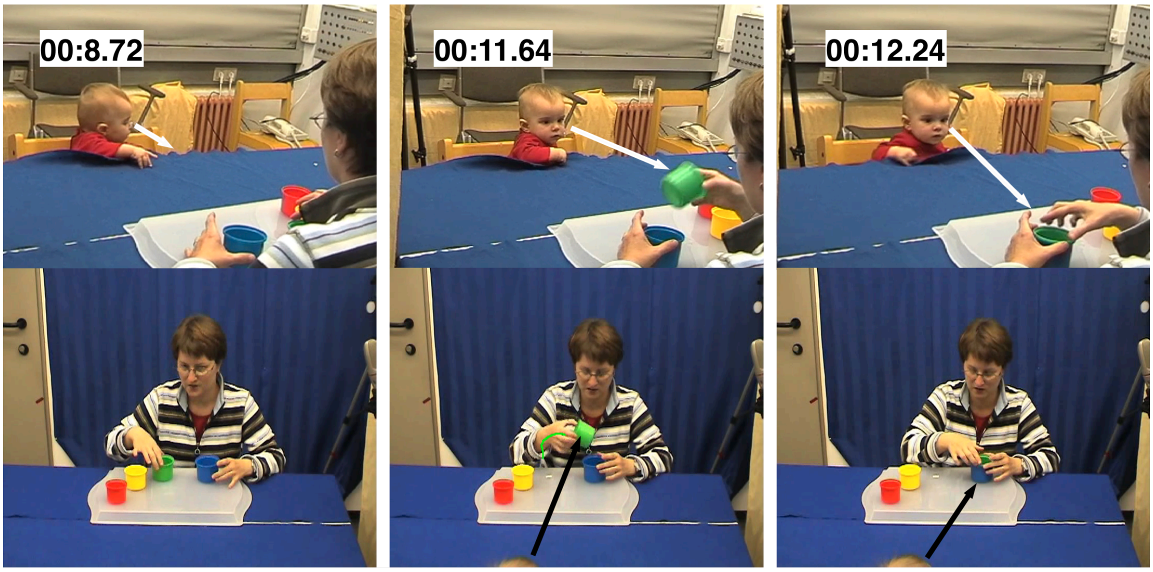
To test the validity of the interplay between the tutor's action modification and the recipient's gaze, these observations can be contrasted with additional cases where parents perform similar motions under different conditions. For example, if *pronounced* hand motions indeed serve as orienting devices for the child, and *flat* hand trajectories are produced upon a display of understanding (likely to occuring in the 3rd nesting action), we could hypothesize that *flat* hand motions are not designed to engage the infant. In the video data, this can be seen either as the child not following the action presentation or by disattending. As a counter example, consider a third fragment where the mother uses particularly flat hand motions (Fig. 2c). When the cups are placed on the table, the child is gazing towards the floor. The mother then verbally calls for the child's attention ("INA; HAVE a look"), which induces the child to gaze to the cups, although she quickly re-orients to the floor when the mother takes the cups (00.08.72).

Fragment 3 (VP052_3_MC). Green, yellow and red cup (a1, a2, a3)

```
01 T3:    ↑INA; KUCK mal; (0.2) ↑I|na; (0.5) |DA; die BEcher kennst du auch; |ne,
           <name> LOOK;           INa           THERE; the cups know  you also; right,
   R3-gaz: @Ø                        |>>>>>>>>>>|@b                              |>>>
```

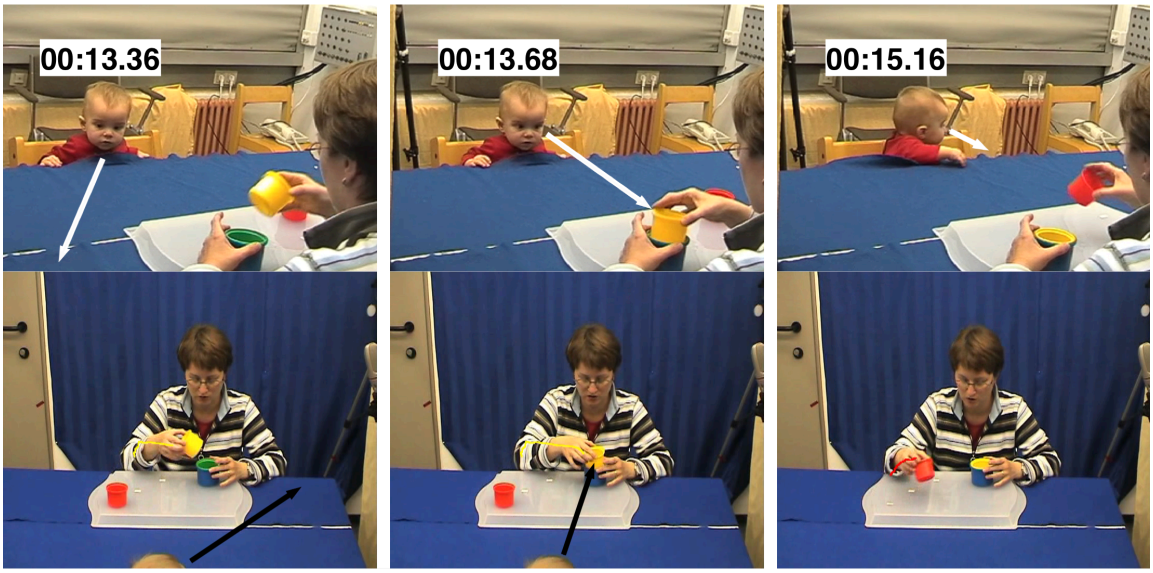
Again, the mother calls the child ("Ina"), the child then re-orients to her, she notices that the infant is paying attention and quickly takes the green cup. She lifts it slightly (00.11.64) and moves it over to the blue cup. After a small delay, the child's gaze follows to the blue cup (00.12.24).

```
02 T3:      |(0.5)|↑Ina; (0.5)|kuck mal  |↑HI:E|R; (0.2) den kann man da REIN |stel|len;
                    Ina          look       HERE           the can  one there INTO place
   T3-act:  |to g |                |grab g |g>bl                         |
   R3-gaz:  |@Ø                 |>>|@g (following)                                |>>>>
                                     *08.72                    *11.64                *12.24
```



Then, while the mother's right hand moves to the yellow cup, the child's gaze remains in the opposite direction. The mother briefly gazes at the child, sees her lack of interest, and regardless, reorients to the cup and moves it straight to the blue one (00.13.36). Right at the moment when the yellow cup is being dropped into the blue one (00.13.68), the child reorients toward it. Thus, although the child has not attended to the actual action, it appears to the mother, who only looks at her recipient at the end of the action, as if the child is gazing correctly and attending to the action. The mother, then proceeds to take the red cup and move it into the blue one with another flat motion. The child turns away immediately (00.15.16).

```
03 T3:     (0.2)|(0.2) und |DIEsen kann man da |rein |stellen; |(0.4) und den RO|ten;
                       and  THIS   can  one there into place          and the RED one;
   T3-act:      |y>b                           |               |r>b             |
   R3-gaz: @ø              |@bl/y                    |>>>>>>>>>|@ø
                   *13.36   *13.68                              *15.16
```



This conduct supports our hypothesis and suggests that presentational actions *without* modification are not appropriate to helping a child attend to the relevant events of the presentation, and thus to follow and comprehend the action.

6.4 Dual orientation between task and co-participant

The interactional outcome of a tutor's non-motionese conduct, as seen in fragment 3 (section 6.3), raises questions about (i) what the recipient is actually able to grasp from the action presentation and (ii) what the tutor can know about her current understanding. In the case of fragment 3, the infant's gaze behavior entails that she is only able to witness a fraction of the tutor's actual presentation. The infant sees the complete action once (a1), then watches as the second cup is placed in the blue one (a2) followed by the tutor taking the third cup (a3).

(i) a1: first cup (green): grab – lift – transport – drop.
(ii) a2: second cup (yellow): – drop.
(iii) a3: third cup (red): grab

Fig. 4a. Fragment 3 – Infant's perception of the tutor's presentation.

The mother orients to the child only very briefly at the end of a nesting action or during a pause in the action when she moves her hand back to the next cup in the sequence (underlined). Thus, her gaze toward the infant coincides with the rare moments when the infant is paying attention to the action.

(i) a1: first cup (green): grab – lift – transport – drop.
(ii) a2: second cup (yellow): – drop.
(iii) a3: third cup (red): grab

Fig. 4b. Fragment 3 – The infant's perception of the tutor's presentation and instants when the tutor's gaze is directed at the infant (underlined).

These short instances of parallel joint focus provide solid ground for the mother to incorrectly assume that the infant is attending the entire action presentation:

(i) a1: first cup (green): grab – lift – transport – drop.
(ii) a2: second cup (yellow): grab – lift – transport – drop.
(iii) a3: third cup (red): grab – lift – transport – ??

Fig. 4c. Fragment 3 – The tutor's likely impression of the infant's perception of the presentation.

This discrepancy between the child's actual witnessing of the presentation and the tutor's awareness of it yields an important issue. In order to be able to micro-coordinate with the recipient, the tutor needs to visually orient toward him. However, tasks like ‘nesting cups’ do not allow the tutor to monitor the learner exclusively as she must also look at the objects involved in the task. Thus, a form of ‘dual orientation’ is required to allow the tutor to notice where the recipient is orienting so as to judge her current state of knowledge while being able to manipulate objects as necessary. While the tutor in fragment 3 uses a pattern of dual orientation where she only very briefly glances at the child at the end of each nesting action, the tutors in fragments 1 and 2 solve this problem in a different way. They look at the child while transporting the cup and only very briefly glance towards the objects before dropping the smaller cup into the larger one. This allows them to orient to the recipient and to micro-coordinate their actions with them. We call these different forms of handling the task of dual orientation ‘task-oriented’ vs. ‘recipient-oriented’.

7. From empirical observation to quantification of interactional paths

To make our findings from the CA-informed analysis usable for robotics research, the interactional procedures and patterns need to be formalized and systematized. Artificial systems need to know which actions are likely to occur after a certain event, how they should be classified, and what this might mean in terms of the interactional organization. Also, since it is unknown how some interactions will proceed, computational modeling requires an understanding of relevant variants in action sequences and a probabilistic estimation of the resulting different interactional paths. Therefore, in what follows, task-oriented and recipient-oriented tutor actions will be initially distinguished before aspects of the interactional loop are analyzed quantitatively for the tutors’ recipient-oriented sub-actions.

7.1 Motionese behavior: Task-oriented vs. recipient-oriented

The qualitative analysis has revealed a difference in the ways in which tutors handle the ‘dual orientation’ between the recipient and the objects involved in the task (recipient-oriented vs. task-oriented). This difference has an impact on the tutor’s ability to micro-coordinate her actions with those of the recipient. The examples analyzed in section 6 suggest that a task-oriented tutor produces rather flat motion trajectories when she is only marginally aware of the recipient's actions and thus less likely to organize her attention. In contrast, a recipient-oriented tutor appears to produce more pronounced hand motions. By taking this into account, it can be hypothesized that tutors who orient themselves to the recipient exhibit more 'motionese' features in their action demonstrations than those who orient to the task.

A formal description of the phenomenon involves the following two steps. (i) Depending on the tutor's gaze behavior (toward objects for task-oriented, and toward recipient for recipient-oriented) the data is separated into two classes. We define a sub-action (a1, a2, a3) as belonging to the task-oriented category if the tutor gazes at the recipient for a maximum of 25% of the time while moving the cup. This 25% threshold is based on the adult gazing patterns, and is a simplification of the phenomenon described in section 6.4. All other sub-actions where the tutor gazes to the recipient longer than 25% of the time it takes for the cup to be transported, fall in the category of recipient-oriented sub-actions. Using this criteria, 51 sub-actions derived from the nesting actions of 17 parents (9m, 8f), were automatically divided into task-oriented (20 sub-actions by 11 parents) and recipient-oriented (31 sub-actions by 13 parents). The intra-subject variability is detailed as follows (Fig. 5):

- 6 parents exhibited only recipient-oriented sub-actions (a1, a2, a3)
- 4 parents showed only task-oriented sub-actions (a1, a2, a3),
- 5 parents produced recipient-oriented sub-actions for a1 and a2, and a task-oriented for a3.
- 1 parent produced a task-oriented sub-action for a1 and a2 and recipient-oriented for a3.
- 1 parent used a task-oriented sub-action for a1 and recipient-oriented for a2 and a3.

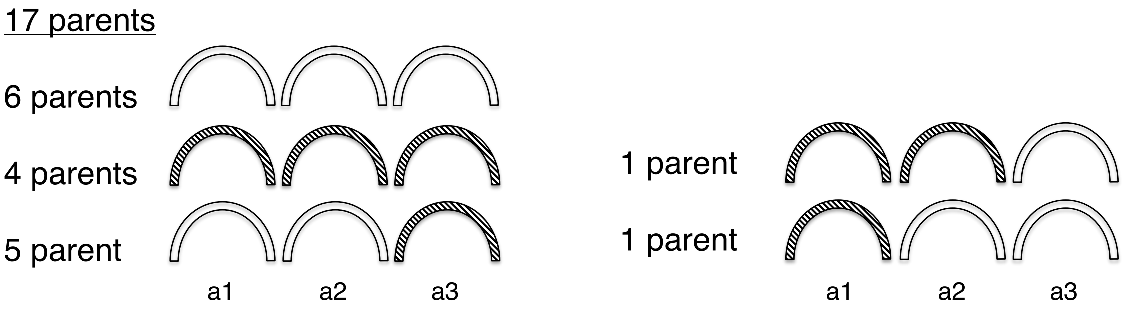


Fig. 5. Intra-subject variability. Number of parents showing recipient-oriented (white arcs) and task-oriented (hatched arcs) sub-actions for the three cup-transports, a1, a2, a3.

(ii) For these two classes of sub-actions, the tutors' 'motionese' features are calculated. For this, the annotations of the action structure intervals (a1, a2, a3) are combined with the hand trajectory values (Fig. 6) and analyzed for the set of measures introduced in Vollmer et al. (2009a), namely, action length, velocity, acceleration, range, total/average length of motion pause, and pace.

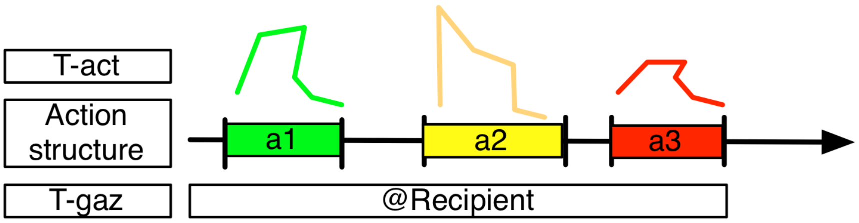


Fig. 6. Information used to classify sub-actions and calculate 'motionese' measures. Tutor's hand trajectories for the three sub-actions a1, a2, a3 (T-act), annotations of action structure intervals with the three cup transports a1, a2, and a3 and the action pauses in between the sub-actions p1 and p2 (action structure), and annotations of the tutor's gaze intervals (T-gaz).

Independent samples *t*-tests were conducted to compare 'motionese' features in the tutors' action demonstrations in task- and recipient-oriented sub-action conditions (within the data set of ACI with infants aged 8 to 11 months). The analysis revealed a significantly stronger index of 'motionese' behavior in the recipient-oriented (r-o) compared to the task-oriented (t-o) sub-actions. The recipient-oriented sub-actions were found to:

- be longer (action length [*s*], r-o. *M* = 3.25, *SD* = 2.06, t-o. *M* = 1.13, *SD* = 0.35, *t*(33) = -5.62***, *p* < 0.001),
- be performed at a lower speed (velocity [*100 pixels/s*], r-o. *M* = 0.09, *SD* = 0.05, t-o. *M* = 0.15, *SD* = 0.07, *t*(49) = 3.63***, *p* < 0.001), acceleration [*100 pixels/s*^{2}], r-o. *M* = 1.08, *SD* = 0.78, t-o. *M* = 2.14, *SD* = 1.33, *t*(27) = 3.2**, *p* < 0.01), and pace (the

duration of each motion divided by the duration of the preceding pause, r-o. *M* = 8.77, *SD* = 11.18, t-o. *M* = 19.16, *SD* = 12.91, *t*(43) = 2.87**, *p* < 0.01),

- exhibit more range (the covered motion path divided by the distance between subaction on- and offset, r-o. *M* = 3.21, *SD* = 1.72, t-o. *M* = 2.15, *SD* = 0.91, *t*(48) = -2.85**, *p* < 0.01),
- exhibit more motion pauses [*%*] (total (r-o. *M* = 6.03, *SD* = 10, t-o. *M* = 0.13, *SD* = 0.57, *t*(26) = -3.1**, *p* < 0.01),
- exhibit greater average length of motion pauses [*frames (with 25fps)*] (r-o. *M* = 6.27, *SD* = 8.83, t-o. *M* = 0.11, *SD* = 0.47, *t*(26) = 3.62***, *p* < 0.001).

To ensure that the average findings were not an artifact of individual-specific styles of presentation (Reese et al., 1993), we manually compared values of the task- and recipient-oriented sub-actions of the 7 parents with sub-actions in different categories (see Fig.5). The comparison of values in one parent reflected the overall findings of the comparison of sub-actions. Thus, the analysis supports the hypothesis that the tutor's 'motionese' conduct is linked to the concept of recipient design (Sacks et al., 1974) in the actual interaction. Not only does the mere physical presence of an infant (as opposed to an adult) play a role in the tutor's 'motionese' behavior (compare Herberg et al., 2008), but it appears that the tutor's stepwise, local monitoring and 'online analysis' of the recipient's actions are the basic condition for the observed conduct.

7.2 Tutor's hand motions as an orienting device

In the case of recipient-oriented tutoring, the qualitative analysis has revealed an interactional loop between the tutor's hand motions and the recipient's gaze. In particular, it has been shown that the tutor's high, upward hand motions function as an orienting device to attract and guide the recipient's attention (section 6.1). In a next step, we aim at formally investigating this 'interactional loop' across the corpus. Describing the tutor's orienting devices and their function algorithmically requires investigating time-based sequential structures of interactional coordination (in contrast to the simultaneous occurrence of one participant's actions, as in section 7.1) and leads to identifying trajectories of action.

A formal description of the tutor's orienting device and its impact on the recipient encompasses the following set of operations:[4]

(i) Identify the beginning of a sub-action (Fig. 7: onset of 'a1' in the annotation line 'action structure').

(ii) Identify whether the adult, at that moment, gazes at the infant's actions (Fig. 7: '@Recipient' in the annotation line 'T-gaze').

(iii) Identify the infant's orientation, i.e. gaze direction, at the beginning of the sub-action. Orientation can be classified as two sub-groups: (iii-a) the infant is attentive and gazes at the cups or the tutor vs. (iii-b) the infant is non-attentive and gazes elsewhere (Fig. 7:. '@Object/Tutor (+)' vs. '@Elsewhere (-)' in the annotation line 'I-gaz').

(iv) Identify whether the tutor's hand motion performs a high trajectory during the nesting action (Fig. 7:. 'high' in the annotation line 'T-act'). For this analysis, 'high trajectories' are defined by calculating the height of a trajectory peak in relation to the height of the remaining motion trajectory.[5] On this basis, they are defined as lying above a threshold calculated by adding the standard deviation of the trajectory height of the three sub-actions (a1, a2, a3) to their mean trajectory height.

(v) Analyze the infant's reaction once the tutor's hand motion has reached the defined threshold (Fig. 7:. '@Object/Tutor (+)' vs. '@Elsewhere (-)' in the annotation line 'R-gaz'), i.e. Does the infant's gaze follow the tutor's hand or not?

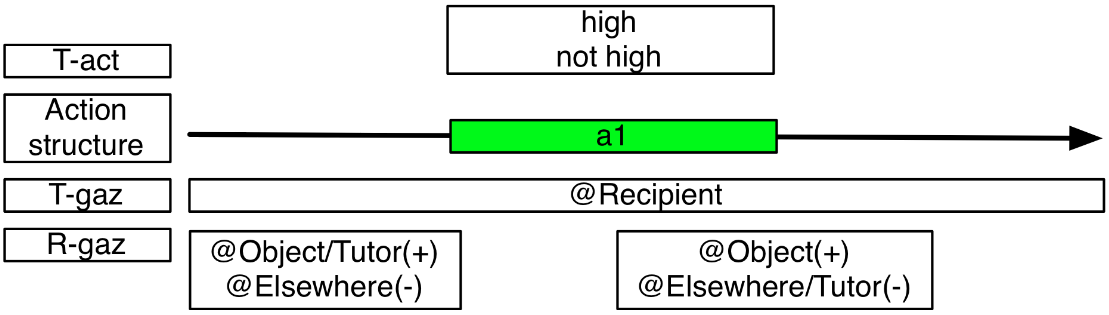


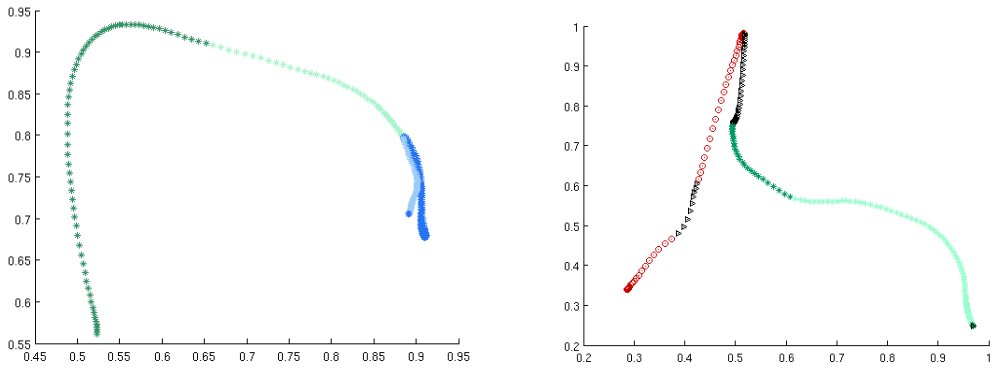
Fig. 7. Information used to find orienting devices: Tutor's hand motions as either high or low arcs (iii) (T-act), annotations of action structure intervals with the three cup transports a1, a2, and a3 and the action pauses between sub-actions p1 and p2 (i) (action structure), annotations of the

[4] This analysis was performed offline on the existing corpus although the sequence of actions is described in procedural terms so that its principles can also be valid for the online analysis in real-time interaction.

[5] This definition enables us to define a high trajectory for every tutor individually, as some tutors might generally tend to demonstrate with rather flat hand motions and also use rather flat motions as orienting devices compared to other tutors.

tutor's gaze intervals (ii) (T-gaz), and the infant's attention at the beginning of a sub-action (iv) and her reaction to the high arc in the tutor's demonstration (v) (R-gaz).

Applying this formal description as a classifier to the set of recipient-oriented sub-actions (as the task-oriented sub-actions have been shown to not exhibit any 'motionese' conduct) allows us to identify both types of orienting devices revealed in the qualitative analysis (see section 6.1): (i) Figure 8a depicts the prototypical case of the tutor pro-actively engaging the infant's attention. At the onset of the action, the child's gaze is already oriented toward the cup/tutor's hand (marked as green asterisks), and the tutor's high onset incites the child's gaze to follow the cup's trajectory. (ii) Figure 8b represents the case in which the infant's visual attention is not oriented toward the cups/tutor's hands at the onset (red circled line) and the tutor attempts to repair it by lifting her hand/the object. After the peak in the tutor's hand trajectory, the infant's gaze shifts to the moving cup and is thus re-oriented to the action.



a) Tutor's hand motion pro-actively engages infant's attention (case iii-a).

b) Tutor's hand motion re-orients and thus repairs the infant's gaze (iii-b)

the demonststrator is aware that	the demonstrator is **not** aware that	the recipient's gaze is
✳	✳	directed toward the hand/object
◈	◈	directed at the tutor
⊙	⊙	directed elsewhere
▷	▷	shifting
✳	✳	anticipating
⊡	⊡	directed toward the goal in-between sub-actions

c) Legend

Fig. 8. Tutor's high hand motion trajectories used as orienting devices to focus the infant's attention. The hand trajectories of individual subjects' actions were normalized ([0,1] on x- and

y-axis corresponding to a maximum movement extension during the nesting cups action) and are shown in relation to the infant's gaze.

The automated corpus analysis provides a systematic overview of the courses of action. Figure 9 shows the different interactional paths extracted by the classifier, detailing at each stage the relevant option that follows, and how often they occur in our corpus. It should be noted that the small size of the sub-samples prevents statistical relationships from being reasonably calculated for the classifier output.[6]

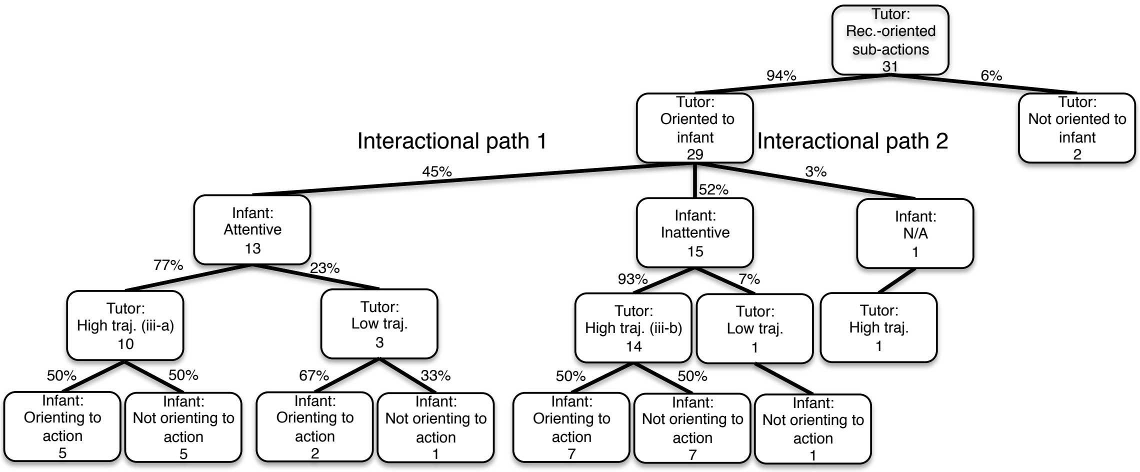

Tutor: Rec.-oriented sub-actions 31
94%
6%
Tutor: Oriented to infant 29
Tutor: Not oriented to infant 2
Interactional path 1
Interactional path 2
45%
52%
3%
Infant: Attentive 13
Infant: Inattentive 15
Infant: N/A 1
77%
23%
93%
7%
Tutor: High traj. (iii-a) 10
Tutor: Low traj. 3
Tutor: High traj. (iii-b) 14
Tutor: Low traj. 1
Tutor: High traj. 1
50%
50%
67%
33%
50%
50%
Infant: Orienting to action 5
Infant: Not orienting to action 5
Infant: Orienting to action 2
Infant: Not orienting to action 1
Infant: Orienting to action 7
Infant: Not orienting to action 7
Infant: Not orienting to action 1

Fig. 9. Courses of action and their frequency of occurrence. Numbers in boxes depict numbers of sub-actions in each class for each step. Percentages on the arcs indicate the percentage of each child node (class) relative to the respective parent.

For the 31 sub-actions of the recipient-oriented group, the tutor monitors the infant right at the beginning of the sub-action in 29 cases.

Interactional path 1 (infant attentive at the beginning). If the infant is attentive to the tutor's actions, the tutor performs the action presentation (1-a) with high trajectories in 77% of the cases (i.e. 10 cases, iii-a), and (1-b) with low trajectories in 23% (i.e. 3 cases). Thus, extending our initial qualitative single case analysis to the corpus level, we find evidence that tutors act predominantly following a model of providing visual guidance to

[6] To verify the choice of features and our definition of an orienting device/repair activity, the results obtained by the classifier were compared to an independent qualitative analysis of the same sub-actions informed by CA methodology as presented in section 6. We found that the results were consistent. For iii-b, the manual analysis reported only one additional sub-action with orienting device/repair activity that the computational analysis did not find. In this case, the tutor treated as appropriate an instance where the recipient reoriented her gaze direction to the tutor's face instead of to the cup.

the infant's attention, and thereby attempt to *pre-empt* potential problems that could interfere with the tutoring. At the same time, the infants react to the highly articulated action presentations in 50% of the cases (i.e. 5 cases) by sustaining their attention to the action; in the other 50%, they re-orient either to the tutor's face (20%, i.e. 2 cases; these could still be considered relevant to the action) or elsewhere (30%, i.e. 3 cases). Thus, the interactional procedure turns out to be successful in 50% of the cases (considering also the child's gaze to the tutor's face as relevant, resp. 70%) of the cases. – In case '1-b' the infant's initial orientation to the action remains focused also when the tutor uses low trajectories (2 cases, i.e. 66%) or shifts to the tutor's face (1 case, 33%).

Interactional path 2 (infant non-attentive at the beginning). If the infant is not attentive to the tutor's actions, the tutor performs the presentation (2-a) with a high trajectory in 93% of the cases (14 cases) and thus presumably attempts to *attract* the infant's gaze to the action. The infant reacts to this tutoring behavior by re-orienting to the tutor 50% of the time (7 cases) and by *not* orienting the other half of the time. The success rate of the tutor's 'attention repair activity' (iii-b) being about 50%. The corpus shows only one case where an inattentive infant is presented with low action trajectories where the child responds with persistent inattention to the action.

The analysis reveals that the two cases of the interactional procedure initially identified in the qualitative analysis as (iii-a) a strategy to continuously secure and guide the infant's attention (section 6.1.1) or (iii-b) to repair the infant's lack of attention (section 6.1.2), can be formally described and algorithmically detected. Further, they were found to be deployed frequently by the tutor. Their interactional success in terms of inciting a particular reaction from the learner ranges at about 50% (for iii-a. resp. 70%). This suggests, on the one hand, that interactional regularities can be detected formally, and that typical courses of action can be described. On the other hand, the recipient's reactions are not easily predictable, which is especially the case for very young infants just becoming familiarized with social routines. This points to the contingent nature of social interaction and reminds us that empirically observed regularities should not be conceived of as interactional 'rules' to be directly transferred into an artificial system (see Button 1990). Rather, that revealing interactional procedures, and considering them in the empirical context of *alternative* courses of action should be the basis for modeling interaction in technical systems.

Additionally, information about the timing of actions can be retained. We found that the tutor’s orienting device needed between 0.52 to 1.25 seconds (*M* = 0.83, *SD* = 0.28) to induce a change in the child’s focus of attention (case (iii-b). The repair action was measured from the delay at the start of the sub-action until the child reached a relevant point with their gaze).[7] If the design of a robot system were to be motivated by human conduct, it should also attempt to adjust its focus of attention and follow the tutor’s action in a comparable time frame.

7.3 Anticipating next actions

Our qualitative analysis suggests that some infants anticipate the next relevant action during a tutor’s demonstration, as displayed in their gaze behavior. Tutors respond to this in several ways (section 6.1 and 6.2) and treat the behaviors as

- displays of understanding, in which case they respond with a flat 3rd nesting action.
- displaying a lack of attention to the ongoing action and thus in need of repair. They respond with either an online action modulation or with a higher subsequent (3rd) nesting action;
- not requiring any specific reaction or change in activity.

Locating these interactional patterns computationally in the corpus of recipient-oriented tutoring requires the following formalization steps (in which each classification is related to an observable action by the tutor):

(i) Formalize and identify the moments at which the infant’s gaze anticipates the tutor’s next action (see the appendix for algorithm details).

(ii) Investigate the tutor’s reaction to the infant’s anticipatory gaze, and classify it as (a) understanding of action, (b) repair, or (c) indifference.

Applying these classifications to the corpus reveals an astonishingly high total number of events, i.e. 23 in the set of 17 interactions, in which the infants anticipate during the sub-actions and/or the nesting pauses. In 20 cases the tutor is oriented toward the infant such that she is indeed aware of the infant’s anticipatory action. However, in comparison to the case described in section 6.2 the formalization reveals also a set of structurally

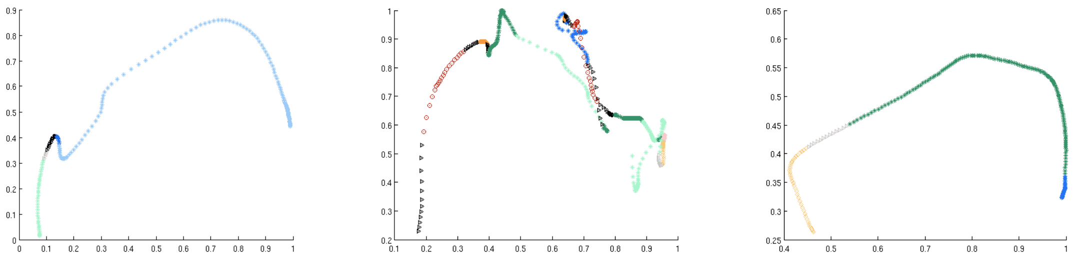
[7] For group iii-a where the infant’s gaze is directed to the relevant object at the beginning of the action, the annotations don’t allow measuring a change in the infant’s orientation. This would only be possible using eye-tracking technology.

different forms of anticipation, and thus provides a more detailed view of the phenomenon.

(a) **Anticipation during nesting pauses.** This type of anticipation occurs when the tutor's empty hand travels back to pick up the next cup, similarly to the case presented in section 6.2. In 2 cases, the infant anticipates during the nesting pause, for which we find a flat nesting sub-action (defined as described in section 7.2) potentially used by the tutor to indicate her interpretation of the infant's gaze behavior as displaying her understanding of the current action.

(b) **Anticipation during sub-action** (Fig. 10a and b). We find 7 cases (7 children, 4 in a1, 2 in a2, 1 in a3), in which the infant anticipates the goal in an ongoing sub-action (Fig. 10a). In two of these cases, the tutor reacts with a *flat* trajectory in the subsequent sub-action, thus possibly displaying her interpretation of the infant's gaze as *understanding* of the action. In four cases, the tutor reacts with a *higher* trajectory in the next sub-action, thus possibly showing that she interprets the infant's conduct as a *lack of attention*. Similarly, in one case (a3) the tutor reacts immediately to the infant's anticipation and makes an online adjustment to her nesting movement with an elevated hand motion to re-orient the infant's attention to the object (Fig. 10b).

(c) **Action-final anticipation** (Fig. 10c). One type of infant anticipation revealed by the algorithm (and ignored in the qualitative analysis) occurs at the moment when the adult's hand hovers right above the big blue cup just before she drops the small cup into the bigger one (and analogously, in the pauses just before grasping the next cup). This form of anticipation is very short, barely visible in the video-data and completes only the almost finished action. It is not clear whether tutors acknowledge these forms of anticipation, as there is no real opportunity to react to them while performing nesting actions. In our corpus, 11 of the 20 cases belong to this type.



a) Sub-action trajectory with anticipation during

b) Sub-action trajectory with anticipation during sub-

c) Sub-action trajectory with action-final anticipation.

sub-action.	action treated as being in need of immediate repair.

Fig. 10. Patterns of infant's anticipatory gaze behavior. The infant's anticipatory gaze is marked as a blue-asterisk trajectory during the tutor's nesting action (for the legend see Fig. 8c).

The automated corpus analysis provides a systematic overview of the interactional paths extracted by the classifier as shown in Figure 11. It details the relevant subsequent interactional option at each stage, and how often it occurs in our corpus.

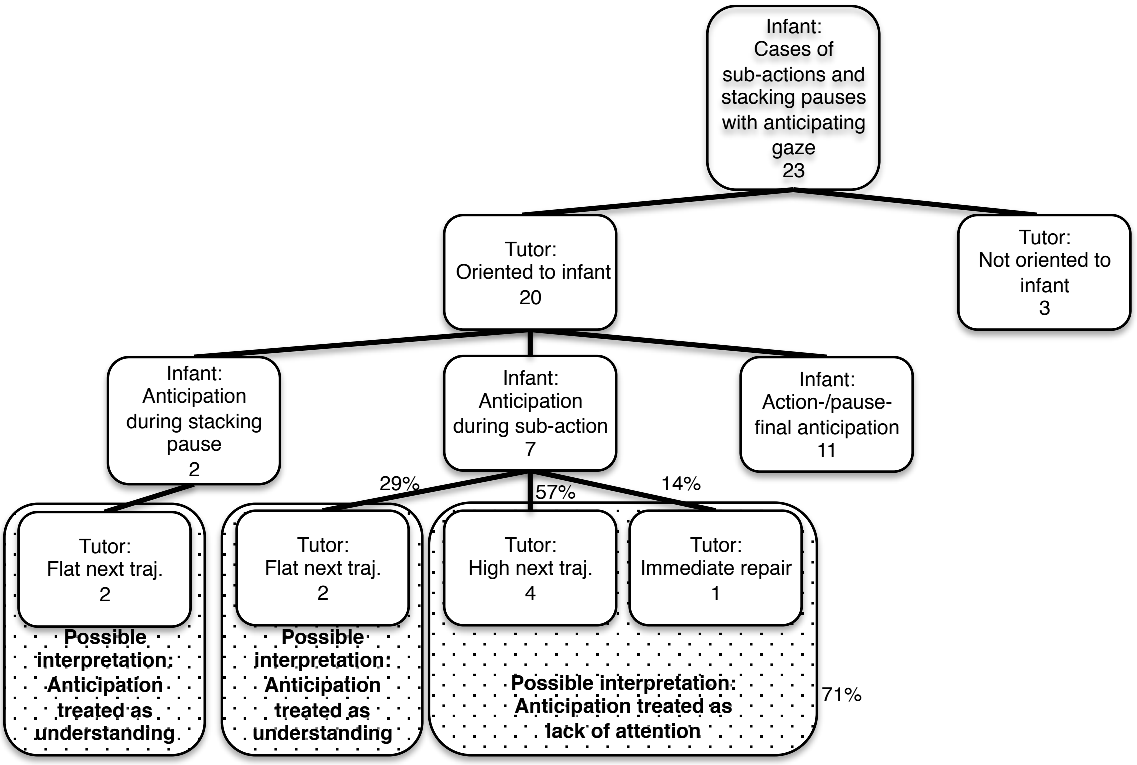

Infant: Cases of sub-actions and stacking pauses with anticipating gaze 23
Tutor: Oriented to infant 20
Tutor: Not oriented to infant 3
Infant: Anticipation during stacking pause 2
Infant: Anticipation during sub-action 7
Infant: Action-/pause-final anticipation 11
29%
57%
14%
Tutor: Flat next traj. 2
Tutor: Flat next traj. 2
Tutor: High next traj. 4
Tutor: Immediate repair 1
Possible interpretation: Anticipation treated as understanding
Possible interpretation: Anticipation treated as understanding
Possible interpretation: Anticipation treated as lack of attention
71%

Fig. 11. Classification of cases with anticipatory gaze. Numbers in boxes depict the quantity of sub-actions in each class for each step. Percentages on arcs indicate the percentage of the child node (class) relative to the respective parent.

This corpus analysis both supports the initial qualitative results and reveals further details. Across the corpus, most tutors are indeed sensitive to the infants' anticipatory gaze in that they adjust the trajectory of their nesting action. While the investigation began with a comfortably sized corpus (N=18 infants of pre-lexical age, with 36 parents presenting the nesting cup action), the variability of the task performance (nesting procedure 1 vs. 2), the difference in the tutor's orientation to the recipient vs. to the task, and the different options of subsequent reactions towards the infant's conduct result in a data sample that is too small to conduct valid statistical comparisons. Thus, only tendencies can be reported and interpreted according to the hypotheses of the qualitative analysis. These are as follows.

1. The infant's anticipatory gaze *during* the nesting *pauses* seems to be interpreted by the tutor, as shown by a flat, non-ostensive hand motion during the subsequent nesting action, as the infant displaying some understanding of the next relevant action (in 100% of the cases (N=2)).

2. However, the infant's anticipatory gaze *during* the nesting *sub-action* is only rarely considered as a display of knowledge (29% of the cases (N=2)). Instead, it tends to be considered as a lack of attention (71% of the cases (N=5)) requiring repair either immediately (N=1) or during the next nesting action (N=4) with a higher or more modulated hand trajectory.

8. Summary of the main results

Motivated by the phenomenon of 'motionese' (Brand et al., 2002; Brand et al., 2007; Rohlfing et al., 2006) whereby tutors modify their actions when presenting a manipulative task to young infants, and its relevance for robotic social learning, we sought to investigate its functions, sources and effects in social interaction. We used video-recordings from a semi-experimental tutoring setting (see section 3), in which the tutor presents the activity of nesting differently sized cups to a pre-lexical infant (8 to 11 months) while they were seated face-to-face across a table. We applied a combined qualitative and quantitative approach for initial case-analysis to discover the phenomenon (section 6) and subsequent formalization and quantification across the corpus (section 7). Starting from the Conversation Analytic ideas of co-construction and online-monitoring, our investigation has revealed how the tutor's action presentation is interleaved with the learner's conduct (Zukow-Goldring, 2006; Zukow-Goldring & Arbib, 2007) on a micro-level suggesting an *interactional* account of 'motionese'. The analytic results can be summarized as follows:

(1) Tutors handle the task-inherent 'dual orientation' between the recipient and the objects involved differently. A tutor who closely monitors the recipient's actions during the presentation (recipient-oriented) is able to micro-coordinate her actions with those of the recipient and thus to adjust her actions with regard to the observed conduct. A tutor whose gaze remains on the objects (object-oriented) is not able to do so. Quantified

across the corpus, this difference is measurable in that recipient-oriented sub-actions were significantly longer, performed with lower speed, more range and longer motion pauses. Thus, the mere physical presence of an infant was insufficient to affect the tutor's 'motionese' behavior (compare Herberg et al., 2008; Knoll & Scharrer, 2007). Rather the mutual adjustments between tutor and learner in a social situation and the learner's feedback appear to be the basic condition for the observed 'motionese' conduct.

(2) In the case of recipient-oriented tutoring, the shape of a tutor's hand trajectory during the action presentation and the learner's gaze are closely linked. The tutor's action modification and the recipient's gaze can be seen to have a reciprocal sequential relationship and constitute a constant loop of mutual adjustments. The tutors' hand motions (and their variability) can be construed as an interactional procedure for securing and organizing the infant's visual attention to the ongoing task.

(3) In this loop, the tutor's hand motions (in particular: upward motions resulting in high trajectories, 25 out of 29 cases) function as orienting devices to attract and guide the infant's attention. Applying the formal description of time-based sequential structures as a classifier to the corpus, the automated analysis provided a systematic overview of the different interactional paths: (i) If at the onset of the action, the infant is attentive to the tutor's actions, the tutor performs the action presentation with high trajectories 77% of the time and with low trajectories in 23% of the cases, and thus tries to *pre-empt* potential problems which could disturb the tutoring. These orienting devices are successful in 50% of the cases. (ii) If the infant is not attentive to the tutor's actions at the onset, the tutor performs the presentation with a high trajectory in 93% of the cases. This *repair-action* induces a shift in the infant's gaze and re-orients her attention to the action in about 50% of the cases. The success rates of the tutors' orienting devices of about 50% suggests that – while interactional procedures can be formally described – the infant's reactions are not easily predictable. Due to this contingency of social interaction, empirically observed regularities should not be conceived of as interactional 'rules', but rather considered in the form of alternative courses of action to be the basis of modeling in technical systems.

(4) Some infants anticipate the next relevant action during the tutor's demonstration with their gaze. The following types of anticipation were found: Action-final anticipation, anticipation during sub-action, and anticipation during a nesting pause. Tutor's treat this anticipatory gaze behavior differently resulting in different hand trajectories: The infant's

anticipation *during pauses* was likely to be treated as a display of action understanding, whereas anticipation *during sub-actions* was interpreted as 'lack of attention' slightly more often.

9. Discussion & Implications

The analysis carried out in this paper shows how a change in research paradigm can lead to new insights about a specific phenomenon. While the phenomenon of 'motionese' had been revealed in developmental studies using an individualistic approach, its investigation from a socio-constructionist and interactionist-sequential perspective has brought to light its sources and effects as they are rooted in the interaction between tutor and recipient. In doing so, the methodological approach of combining qualitative and quantitative analysis has led us from revealing a phenomenon 'from the data themselves' to its systematic and formalized description on the corpus-level. These analyses and results have implications for both tutoring in adult-child interaction and robotic social learning.

(1) *Tutoring in adult-child interaction*. (i) The difference between investigating tutoring interactions in a semi-experimental setting and under controlled laboratory conditions becomes visible with regard to results on the infant's understanding of a presented action. A range of laboratory studies take the infant's gaze as an indicator of novelty of some information and/or hypotheses about some action. In this line, eye-tracking studies have shown that infants were able to anticipate the goal of a presenter's reaching actions at the age of 14 months, but they were not able to do so at 10 months (Gredebäck et al., 2009; Falck-Ytter et al., 2006). In contrast, our data suggest that infants begin to anticipate next actions at 8 months which might be explained by the different resources available to the infant: In the eye-tracking studies infants were shown a systematically moving hand, abstracted from the experimenter's head/body, and without any verbal explanation. In our study infants were immersed in a dynamic and multimodal interaction process with the tutor. The infant had access to the full range of the tutor's communicational resources, i.e. talk, gaze, head orientation etc. (see e.g. Streeck, 1993 on hand motions being made relevant by gaze). And the tutor was able to adjust her conduct to the participant's actions and provide an explanation tailored to the recipient's current state of participation. While our semi-experimental setting was designed to

provide good conditions to investigate the phenomenon of ‘motionese’, further research is required to understand how it is organized under naturalistic interactional conditions.

(ii) The analysis presented in this paper has shown the relevance of understanding actions as emergent interactional products. Previous literature has found interactional patterns in the form of a tutor guiding an infant’s focus of attention, which consists of the following succession of actions: (i) the adult tries to direct the infant’s gaze toward a relevant object, (ii) the child orients to that object, (iii) the adult then introduces new information about the object and (iv) attempts to maintain the infant’s attention on the object (Estigarribia & Clark, 2007). The infant’s gaze was coded into fixed categories (looking at tutor, object, elsewhere) and the tutor’s action in terms of static types of gestures and verbal attention getters. However, our approach has started from the idea that actions are not only sequentially organized, but also occur simultaneously and are based on principles of ‘mutual monitoring’ and ‘online analysis’. In this way, the task to ‘orient the co-participant’ becomes part and parcel of the tutor’s action presentation itself.

In sum, tutoring and learning interactions are best understood as a genuine multimodal and dynamic interactional process embedded in a rich environment.

(2) *Robotic Social Learning*. (i) Robotic learning approaches have a longstanding tradition in conceiving of learning as a process in which the human tutor’s conduct is understood as providing training data for the algorithms to be trained *offline*. More recently, learning approaches have begun to consider the social dimension and that an autonomous system could learn from directly interacting with the ‘human in the loop’ (Steels & Kaplan 2001, Breazeal 2002, Rothwell et al., 2011, Lyon et al., 2012). However, as in the developmental sciences, such tutoring situations have predominantly been regarded as one-way communication, in which the robot *passively* observes the tutor’s actions without contributing to the social situation. For example, Herberg et al. (2008) presented their subjects with *static pictures* of the assumed recipient and investigated whether tutors modified their actions for different types of learners (non-anthropomorphic computer, adult, infant). Despite the asocial nature of the situation, they found that tutors produced simple perceptual modifications for the computer, i.e. motions that were more punctuated and with a wider range. In contrast, speaking to an imaginary infant did not produce any features characteristic of motherese (Knoll & Scharrer, 2007). In those cases where a dialogic perspective was taken and the robot system programmed to give feedback to the tutor’s presentation, the robot’s feedback consists generally of a

positive/negative verbal comment *after* the tutor had finished her presentation (e.g. Alissandrakis et al., 2011). In contrast, our analysis of tutoring in adult-child interaction reveals the impact the learner has on the tutor's actual emerging presentation. We suggest that, despite what is implied by existing models, for the learner it is insufficient to simply *observe* the tutor's actions in order to build a representation. We propose an alternative conceptualization: A robotic system that is supposed to learn by interacting with a social partner, is immersed in a situated interaction with the tutor and could through its own multimodal conduct influence the tutor's ongoing presentation. The robot could signal through its gaze and other features (see also Vollmer et al., 2010), information about its current state of participation, focus of attention, state of cognitive development, and which parts of the presented actions it knows/understands already vs. which appear to be new. We hypothesize that in doing so, the system could influence the tutor's action presentation with regard to aspects such as speed, shape of trajectories, etc. Knowing about the interactional consequences of its own conduct would give a robot system a powerful instrument with which to pro-actively shape the tutor's presentation for its own benefits. Bringing human-robot-interaction to this level of micro-coordination between participants could pave the ways to a genuine *interactionist* version of social robotics.

(ii) For such an interactionist view, the robot system would need to be able to react to the tutor's conduct at a fine-grained level and therefore process information incrementally. As the description of the orienting procedures in adult-child interaction show, human reactions are challenging to predict, albeit systematic. Due to the unpredictable and contingent nature of social interaction, empirically observed regularities should not be conceived of as interactional 'rules', but rather considered in the form of alternative courses of action when used as the basis of modeling in technical systems. The description of alternate interactional paths detailing the expectable frequency of occurrence of a relevant next action might help with probabilistic estimation in computational modeling.

Based on these observations, future work should address the following issues. Firstly, with regard to the analysis of tutoring in adult-child interaction, the observations on the tutor's (manual) action modifications and the recipient's gaze should be re-integrated with verbal actions and other forms of conduct. Secondly, the concept of the 'interactional loop' could be implemented in a robot system to test to what extent such an interactional account and, in particular, which cues would be functional in human-robot-interaction (see Vollmer, 2011; Pitsch et al., 2012, 2013). Additional questions include the following: Could a robot assume a pro-active role as learner and provide its teacher with systematic cues about its current state of participation and of its emergent understanding of the presented action? Which forms of conduct could be used as interactional cues in human-robot interaction? How would they relate to those observed in human-human interaction?

Appendix

We formalize ‘anticipation’ by first establishing a rule. We define an infant’s gaze interval as an ‘anticipatory gaze’, if the child gazes to the object which will be transported next; this includes that she has previously looked at a relevant (as opposed to a random) position. The gazing directions were taken from the annotations. As a first step, we used the original notations to group gazes according to their positions related to the tutoring situation and task. We used the term *relevant* to describe those positions of the action demonstration which are involved in the transport of the cups at a certain point in time. Accordingly, ‘relevant gaze’ describes gaze to a relevant position. For the first sub-action, when the green cup is transported, the relevant positions are the green cup and the tutor’s hand that transports it. The child cannot anticipate at the beginning of the sub-action because she has not yet seen the tutor perform the action. However, when the child follows the transport of the cup and anticipates its goal position, this extension of the infant’s gaze is considered to be anticipatory. We refer to the goal position as *anticipating* position. Depending on the sub-part of the action (see Fig. 1(a)), the definition of relevant and anticipating positions was operationalized using the following rules.

relevant a1 = {green cup, parent's transporting hand a1}
anticipating a1 = {};
= {blue cup}, if child gaze was relevant before in a1.

In the subsequent pause (p1), the parent’s hand does not transport a cup, but travels to and grasps the next cup to be stacked. The hand grasping the second cup is the same hand that will transport it during the second sub-action (a2). Only this hand is considered to be a relevant gaze target. Anticipatory gaze can only take place when the child is gazing toward a cup which could be relevant next (i.e. a cup which could be transported in the subsequent sub-action (a2)), which in this case could be the yellow or red cup.

relevant p1 = {parent's transporting hand a2}
anticipating p1 = {yellow cup, red cup}

For the transport of the second cup (sub-action a2), the relevant position is again the transported cup (here: yellow) and the hand transporting it. Anticipation can occur as in sub-action a1 – when the child has gazed to a relevant position, before gazing to the goal position - but also, when the child has anticipated a cup which could be relevant next in the pause beforehand, i.e. in p1.

relevant a2 = {yellow cup, parent's transporting hand a2}
anticipating a2 = {},
= {blue cup}, if child gaze was relevant in a2 before,
= {blue cup}, if child gaze was anticipating in p1.

In the second pause (p2) and the third cup transport (a3), the classes are defined analogously.

relevant p2 = {parent's transporting hand a3}
anticipating p2 = {red cup}

relevant a3 = {red cup, parent's transporting hand a3}
anticipating a3 = {}
= {blue cup}, if child gaze was relevant in a3 before,
= {blue cup}, if child gaze was anticipating in p2.

Acknowledgement

Karola Pitsch gratefully acknowledges financial support from the FP7 European Project

ITALK (ICT-214668) and from the Volkswagen Foundation (Dilthey Fellowship). Anna-Lisa Vollmer acknowledges financial support from Honda Research Institute Europe for the project 'Acquiring and Utilizing Correlation Patterns across Multiple Input Modalities for Developmental Learning' and from the EU project RobotDoc (235065) from the FP7 Marie Curie Actions ITN. An early version of the analysis can be found in Pitsch et al. (2009) and in presentations at the Research Institute for Cognition and Robotics, Bielefeld.

Bibliography

Alissandrakis, A., Syrdal, D.S., & Miyake, Y. (2011). Helping robots imitate: Acknowledgement of, and adaptation to, the robot's feedback to a human task demonstration. In K. Dautenhahn, & J. Saunders (Eds.), *New frontiers in human-robot interaction* (pp. 9-33). Amsterdam: John Benjamins.

Brand, R. J., & Shallcross, W. L. (2008). Infants prefer motionese to adult-directed action. *Developmental Science*, *11*(6), 853-861.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': Modifications in mother's infant-directed actions. *Developmental Science*, *5*(1), 72-83.

Brand, R. J., Shallcross, W. L., Sabatos, M. G., & Massie, K. P. (2007). Fine-Grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action. *Infancy*, *11*(2), 203-214.

Breazeal, C., & Scassellati, B. (2002). Challenges in building robots that imitate people. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in animals and artifacts* (pp. 363-389). Cambridge, MA: MIT Press.

Breazeal, C. (2002): *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Bruner, J. S. (1985). The role of interaction formats in language acquisition. In J.P. Forgas (Ed.), *Language and social situations* (pp. 31-46). New York, NY: Springer.

Button, G. (1990). Going up a blind alley. Conflating conversation analysis and computational modelling. In P. Luff, N. Gilbert, & D. M. Frohlich (Eds.), *Computers and conversation* (pp. 67-90). San Diego: Academic Press.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J, Belpame, T., Sandini, G., Nori, F., Fadiga, L., Wrede, B., Rohlfing, K., Tuci, E., Dautenhahn, K., Saunders, J. & Zeschel, A. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development, 2*(3), 167-195.

Dausendschön-Gay, U. (2003). Producing and learning to produce utterances in social interaction. *Eurosla Yearbook*, 3, 207-228.

De León, L. (2008). The emergent participant: Interactive patterns in the socialization of Tzotzil (Mayan) infants. *Journal of Linguistic Anthropology,* 8, 131-161.

Estigarribia, B., & Clark, E. V. (2007). Getting and maintaining attention in talk to young children. *Journal of Child Language*, *34*(4), 799-814.

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878-879.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209-21.

Fogel, A. (1993). *Developing through relationships: Origins of communication, self, and culture*. Chicago, IL: University of Chicago Press.

Gergely, G., & Csibra, G. (2005). The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy. *Interaction Studies*, *6*(3), 463-481.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development, 71*(4), 878-894.

Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York, NY: Academic Press.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, *32*, 1489-1522.

Gredebäck, G., Stasiewicz, D., Falck-Ytter, T., Rosander, K., & von Hofsten, C. (2009). Action type and goal type modulate goal-directed gaze shifts in 14-month-old infants. *Developmental Psychology*, *45*(4), 1190-1194.

Heath, C. & Luff, P. (2013). Embodied Action and Organizational Activity. In J. Sidnell & T. Stivers (Ed.), *The Handbook of Conversation Analysis* (pp. 283-307). Chichester, West Sussex: Wiley-Blackwell.

Herberg, J. S., Saylor, M. M., Ratanaswasd, P., Levin, D. T., & Wilkes, D. M. (2008). Audience-Contingent variation in action demonstrations for humans and computers. *Cognitive Science*, *32*(6), 1003-1020.

Heritage, J. & Robinson, J. D. (2006). The Structure of Patients' Presenting Concerns:

Physicians' Opening Questions. *Health Communication, 19*(2), 89-102.

Knoll, M., & Scharrer, L. (2007). Acoustic and affective comparisons of natural and imaginary infant-, foreigner-and adult-directed speech. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, 1414-1417.

Lock, A. & Zukow-Goldring, P. (2010). Preverbal Communication. In J.G. Bremner & T. Wachs (Eds.), *The Wiley-Blackwell Handbook of Infant Development*, *Vol. 1, Basic Research* (pp. 395-425). Chichester, West Sussex: Wiley-Blackwell.

Lyon, C., Nehaniv, C.L. & Saunders, J. (2012). Interactive Language Learning by Robots. The Transition form Babbling to Word Forms. *PLoS ONE* 7(6):e38236.

Lynch, M. (1993). *Scientific practice and ordinary action: Ethnomethodology and social studies of science*. Cambridge: Cambridge University Press.

Mondada, L. (2006). Participants' online analysis and multimodal practices: Projecting the end of the turn and closing of the sequence. *Discourse Studies (Special Issue: Discourse, Interaction and Cognition)*, *8*(1), 117-129.

Mondada, L., & Pekarek-Döhler, S. (2000). Interaction sociale et cognition située. Quels modèles pour la recherche sur l'acquisition des langues? *Acquistion et Interaction en Langue Etrangère (AILE), 12*, 147-174.

Nagai, Y., & Rohlfing, K. (2009). Computational analysis of motionese. Toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development, 10(1)*, 44-54.

Pea, R.D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences*, 13(3), 423-451.

Pitsch, K. (2006). *Sprache, Körper, intermediäre Objekte: Zur Multimodalität der Interaktion im bilingualen Geschichtsunterricht*. Doctoral dissertation. Bielefeld University. Faculty of Linguistics and Literary Studies. urn:nbn:de:hbz:361-17464. http://bieson.ub.uni-bielefeld.de/volltexte/2010/1746/

Pitsch, K., Vollmer, A. -L., Fritsch, J., Wrede, B., Rohlfing, K., & Sagerer, G. (2009). On the loop of action modification and the recipient's gaze in adult-child interaction. *Proceedings of the Gestures and Speech in Interaction Conference* (*GESPIN 2009).* Poznan, Poland, 6 pages.

Pitsch, K., Lohan, K. S., Rohlfing, K., Saunders, J., Nehaniv, C. L., & Wrede, B. (2012).

Better be reactive at the beginning. Implications of the first seconds of an encounter for the tutoring style in human-robot-interaction. *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man 2012),* Paris, France, 974-981.

Pitsch, K., Vollmer, A.-L. & Mühlig, M. (2013). Robot feedback shapes the tutor's presentation. How a robot's online gaze strategies lead to micro-adaptation of the human's conduct. *Interaction Studies, 14*(2), 268-296.

Rader, N. V., & Zukow-Goldring, P. (2010). How the hands control attention during early word learning. *Gesture, 10*, *2*(3), 202-221.

Reese, E., Haden, C. A., & Fivush, R. (1993). Mother-child conversations about the past: Relationships of style and memory over time. *Cognitive Development*, *8*(4), 403-430.

Richards, K. & Seedhouse, P. (2005). *Applying conversation analysis*. Macmillan: Palgrave.

Rohlfing, K., Fritsch, J., Wrede, B., & Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, *20*(10), 1183-199.

Robinson, J.D. & Heritage, J. (2006): Physicians' opening questions and patients' satisfaction. *Patient Education and Counseling,* 60, 279-285.

Rothwell, A., Lyon, C., Nehaniv, C.L. & Saunders, J. (2011): From babbling towards first words: the emergenc of speech in a robot in real-time interaction. In: *Proceedings of the IEEE Symposium on Artificial Life*, 86-91.

Sacks, H. (1984). Notes on methodology. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action. Studies in conversation analysis* (pp. 21-27). Cambridge.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696-735.

Sacks & Garfinkel (1986): On formal structures of practical action. In H. Garfinkel (Ed.). *Ethnomethodological Studies of Work* (pp. 160-193). London: Routledge & Paul.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, *26*(1), 88-128.

Schillingmann, L., Wrede, B., & Rohlfing, K. J. (2009). A computational model of acoustic packaging. *IEEE Transactions on Autonomous Mental Development, 1*(4), 226-237.

Schmitt, R., & Deppermann, A. (2007). Monitoring und Koordination als

Voraussetzungen der multimodalen Konstitution von Interaktionsraum. In R. Schmitt (Ed.), *Koordination: Beiträge zur Analyse multimodaler Kommunikation* (pp. 95-128). Tübingen: Narr.

Selting, M., Auer, A., & et al. (2010). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung 10*, 353-402.

Sidnell, J., & Stivers, T. (2013) (Ed.): *The Handbook of Conversation Analysis*. Chichester, West Sussex: Wiley-Blackwell.

Smith, N.A. & Trainor, L.J. (2008). Infant-Directed Speech Is Modulated by Infant Feedback. *Infancy 13*(4), 410-420.

Steels, L. & Kaplan, F. (2001). AIBO's first words. The social learning of language and meaning. *Evolution of Communication,* 4(1), 3-32.

Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs*, *60*(4), 275-299.

ten Have, P. (1999). *Doing conversation analysis*. A practical guide. London: Sage.

Vollmer, A.-L., Lohan, K. S., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., Rohlfing, K. & Wrede, B. (2009a). People modify their tutoring behavior in robot-directed interaction for action learning. *Proceedings of the 8th International Conference on Development and Learning (ICDL 2009)*, Shanghai, China, 1-6.

Vollmer, A. -L., Lohan, K. S., Fritsch, J., Rohlfing, K., & Wrede, B. (2009b). Which 'motionese' parameters change with children's age? Paper presented at the *Cognitive development society's biennial meeting 2009*, San Antonia, Texas.

Vollmer, A.-L., Pitsch, K., Lohan, K., Fritsch, J., Rohlfing, K., & Wrede, B. (2010). Developing feedback: How children of different age contribute to a tutoring interaction with adults. *Proceedings of the 9th International Conference on Development and Learning (ICDL 2010)*, Ann Arbor, Michigan, 6 pages.

Vollmer, A.-L. (2011). *Measurement and analysis of interactive behavior in tutoring action with children and robots*. PhD thesis. Bielefeld University. Faculty of Technology. urn:nbn:de:hbz:361-24251023. http://pub.uni-bielefeld.de/publication/2425102.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Havard University Press.

Wertsch, J.V., McNamee, G.D., McLane, J.B. & Budwig, N. (1980). The adult-child dyad as a problem-solving system, *Child Development, 51,* 1215-1221.

Wrede, B., Rohlfing, K., Hanheide, M., & Sagerer, G. (2008). Towards learning by interacting. *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot-Interaction* (*HRI 2008)*, Amsterdam, NL, 139-150.

Zukow, P. G. (1990). Socio-perceptual bases for the emergence of language: An alternative to innatist approaches. *Developmental Psychobiology, 23*(7), 705-726.

Zukow-Goldring, P. (1996). Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. *Early Development and Parenting, 5*(4), 195-211.

Zukow-Goldring, P. (1997). A social ecological realist approach to the emergence of the lexicon: Educating attention to amodal invariants in gesture and speech. In C. Dent-Read & P. Zukow-Goldring (Eds.), *Evolving Explanations of Development: Ecological Approaches to Organism-Environment Systems* (199-250). Washington, DC: American Psychological Association.

Zukow-Goldring, P. (2001). Perceiving referring actions: Latino and euro-american infants and caregivers comprehending speech. *Children's Language,* 11, 139-165.

Zukow-Goldring, P. (2006). Assisted imitation: Affordances, effectivities, and the mirror system in early language development. In M. A. Arbib (Ed.), *From action to language via the mirror neuron system* (pp. 469-500). New York, NY: Cambridge University Press.

Zukow-Goldring, P. (2012). Assisted imitation: First steps in the seed model of language development. *Language Sciences, 34*(5), 569-582.

Zukow-Goldring, P., & Arbib, M. A. (2007). Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention. *Neurocomputing*, *70*(13-15), 2181-2193.

Zukow-Goldring, P. & Ferko, K. R. (1994). An ecological approach to the emergence of the lexicon. Socializing attention. In V. John-Steiner, C.P. Panofsky & L.W. Smith (Eds.), *Sociocultural Approaches to Language and Literacy: An Interactionist Perspective* (pp. 170-190). New York, NY: Cambridge University Press.