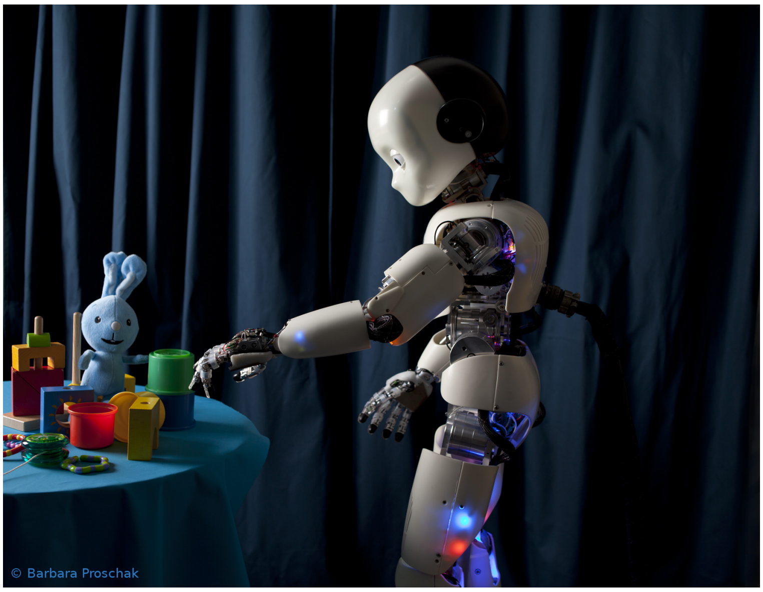
A MODEL OF CONTINGENCY DETECTION TO SPOT TUTORING BEHAVIOR AND RESPOND TO OSTENSIVE CUES IN HUMAN-ROBOT-INTERACTION

KATRIN SOLVEIG LOHAN



Dipl. Inform.

CoR-Lab, AG-AI
Faculty of Technology
University Bielefeld

2011

Katrin Solveig Lohan: *A model of contingency detection to spot tutoring behavior and respond to ostensive cues in Human-Robot-Interaction*, Dipl. Inform., © 2011

"Wenn ich's mir recht überlege, hat das ganze Überlegen keinen Sinn."
*Janosch (*1931)*

ACKNOWLEDGMENTS

Many thanks to everybody who were involved in helping me to create this dissertation. I am very thankful for the support of my supervisors and I am glad that I had the opportunity to write my thesis at the Bielefeld University.
Regarding the typography and other help, many thanks go to Konstantinos Theofilis, Sascha Griffith, Brigitte Lohan, Alessandra Sciutti, Christian Dondrup, Frederic Siepmann and again to my supervisors.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

OS	Ostensive Signals
EDD	Eye Direction Detector
SAM	Shared Attention Mechanism
HRI	Human Robot Interaction
HHI	Human Human Interaction
ACI	Adult Child Interaction
AAI	Adult Adult Interaction
ARI	Adult Robot Interaction
PCA	Principal Component Analysis
NMF	Non-negative Matrix Factorization
CA	Conversation Analysis
RP	Reaction Patterns
MLU	Mean Length of Utterance
SUS	System Usability Scale

1 INTRODUCTION

This work investigates the strategies used in parent-child interactions. The knowledge gained from these interactions (with the parents acting as tutors) was transferred to the iCub humanoid robot platform. The iCub's behavioral model was based on the observed strategies used by children to interact socially with their parents. The main goal was to verify whether these strategies can create a social interaction between a robot and a human tutor. It was also of interest to see if the robot can benefit from the resulting mechanism and induce an appropriate response from the human. The whole work was evaluated by comparing results of tutor-robot interactions with parent-child interactions.

1.1 MOTIVATION

From learning by observation, robotic research has moved towards investigations of learning by interaction [18], [134]. Within this research paradigm, a robot learns how to, for example, label an object, identify its properties or how to perform actions with it, from a tutor with whom it interacts. Thus, instead of learning about the action from observation, the robot gets the information from the interaction with the tutor. This information is depending on what aspects of the action the tutor considers crucial. This research direction is inspired by findings from developmental studies on human children and primates, pointing to the fact that learning takes place in a social environment [25], [118], [119]. Accordingly, instead of just responding to and memorising a signal, a learner receives support from the interactions with its social partners, the resulting situation and its own experience about such interactions [18]. Recently, driven by the idea that learning through observation or imitation is limited because the observed action does not always reveals its meaning, bootstrapping or scaffolding processes have received increased attention for supporting learning. Zukow-Goldring et al. [137] studied how a learner is actually provided with additional social information by a teacher or a tutor that demonstrates where it is important to pay attention to, e.g., the goal, means or constraints of a task[137]. In these interactions, it is essential that the tutor makes sure that the learner is receptive and ready to learn. The reciprocal contribution, i.e., the guidance of attention by the tutor and the manifestation of receptivity by the learner, seems to follow certain interactive irregularities [21], [39], [92].

1.2 STRUCTURE OF THE PRESENTED WORK

The need for evaluating interactions between humans in order to create a robotic system that can benefit from tutoring strategies, is presented in Chapter 2. A first version of the behavioral model, resulting from observations regarding attention mechanisms, was implemented on a simulated robot (Akachan), used as an interaction partner for humans. The use of a simulator gave rise to interesting questions about differences between robotic platforms and how these can affect the interaction. All these are presented in Chapter 3. Insights on an improved model are presented in Chapter 4. The results of the studies conducted using this improved model and further revisions of the model are presented in Chapter 5. Finally, discussion about the results and the future work is presented in Chapter 6.

2 INSIGHTS FROM DEVELOPMENTAL PSYCHOLOGY

In this chapter, related work on developmental psychology, as well as results of studying parent-child interaction, are presented. In research on developmental psychology, tutoring behavior (section 2.1.1) has been identified as scaffolding learning processes of infants. Infants seem sensitive to tutoring situations (section 2.1) and they detect these by ostensive cues [24]. Some social signals such as eye-gaze, child-directed speech, child-directed motion and contingency have been shown to serve as ostensive cues (section 2.2.1). The concept of contingency describes exchanges in which two agents interact with each other reciprocally. Csibra and Gergely argued that contingency is a characteristic of ostensive stimulus in a tutoring situation [24]. For transferring knowledge into this field of developmental robotic,s it is necessary to model a robot in a way that it can be treated similarly as a human, or even better, as an infant. The robot has to be sensitive to the ostensive stimuli but also has to induce tutoring behavior using its feedback capabilities. These questions were targeted by analysing 65 tutoring parent-child pairs, as well as conducting more focused studies with two different robotic platforms.

2.1 RELATED WORK: ATTENTION MECHANISMS AND FEATURES IN A TUTORING SITUATION

The term tutoring situation, in this context, describes a face to face interaction in which one interaction partner, the *tutor*, has more information about a specific object or task than the other interaction partner, the *learner*. Face to face interaction has a developmental function for infants [120]. This function can be understood as a mechanism that is triggering the capability to acquire new knowledge. In particular, the focus is placed on tutoring situations in which the tutor is an adult and the recipient is either an infant, an adult or a robot. Interactive teaching is often using objects as an example based learning [85], [37], [36]. The analysis is concentrated on the information transfer during the task presentation of toys or tools used in everyday life.

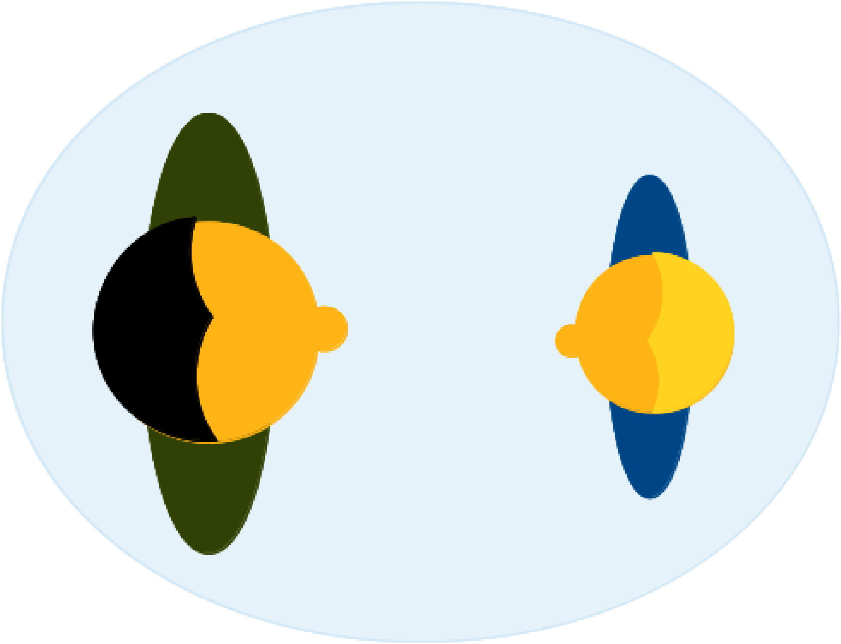


Figure 2.1: Face to face interaction.

2.1.1 *Infants' benefit from tutoring behavior*

In recent research, learning by observation has moved toward the learning by interaction paradigm [134], suggesting that interaction with a caregiver is needed by infants to learn language. There are claims that, already very young infants make use of information arising from interactive situations. The potential of interactive situations for learning, where important parts are highlighted by linguistic and non-linguistic features [48], is interesting. The joint attention of a mother with her child provided by multi-modal cues [48], seems to guide the attention of the child toward those aspects of the sequence which are relevant to the child. The benefit of learning by interaction is not limited to language acquisition, but it has been shown that infants attend more to a visual event which is highlighted by speech [73]. It is envisioned that agents will learn from humans by simply interacting with each other. So far, little is known about interactive processes and the feedback strategies involved. Yet, in order to learn, a learner will typically need to be provided with information given by a teacher who not only gives certain structure to the interaction but also instructs and demonstrates the learning contents. The given information can only be effective if the learner is receptive to the "right" point. There are two supportive aspects in interactive situations, as suggested by Kuhl [59]:

1. in face to face communication, attention and arousal of the infants are higher than in video situations
2. in live interactions, joint attention mechanisms might provide the infants with additional information about relevant phonemic contrasts, but not in video settings.

To assure the receptiveness of the learner, the tutor makes use of interactive regularities by checking the learner's behavior. The term contingency has been suggested to encompass such regularities in interaction. More specifically, it refers to a temporal sequence of behavior and reaction [42], [18]. It has been shown that contingency is an important factor in interactions with infants and contributes to their cognitive development [109]. In social learning, infants benefit from the behavior of their tutors [25]. The regular checking of this behavior, as well as the modification in the contingency used by teachers towards a learner, helps infants to filter the information that is crucial for learning [13]. There have been several modifications in the behavior of a tutor described in developmental research, for example, motionese, motherese and contingency (section 2.2). Overall, there is a polemic behavior of the tutor to highlight new information to the learner. By using this polemic behavior, the tutor seems to build a frame around the important new information which is conducted from explicit meaningful signals (*ostensive signals* (OS)). Around this frame, the signals seem to be overlaid with noise, meaning that they appear more blurry. But one major aspect of the detection of such frames is that a recipient must have the "right" point in his/her focus. It is also important to establish a correlation between the correct multimodal cues. Hence, the attention is influencing all feedback signals.

2.1.2 *Attention is needed to create a social interaction*

William James [51] defined attention as: "...Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought..."

James believed that three physiological processes played a role in the implementation of attention: the accommodation or adjustment of the sensory organs, the anticipatory preparation from within the ideational centers concerned with the object to which attention is paid, and an afflux of blood to the ideational center.

The word attention is describing a process to verify or cheek where a subject should focus on at the current time. This process needs to take care of all sources of information provided from the environment to the subject. These sources can be defined as all the sensors the subject is equipped with and all internal states that the subject can have. The attention process is enriching the most important points or associations between information sources to change the focus of the subject. Concerning the focus the obvious starting point is vision, thus, in robotics, the visual attention mechanism is mainly targeted.

Based on this definition, Tsotsos [121] defined visual attention as: "...Attention is the process by which the brain controls and tunes information processing...."

"...Attention is the set of mechanisms required to tune the search processes in vision to achieve their best performance for a given task (even if the task is free-viewing)...."

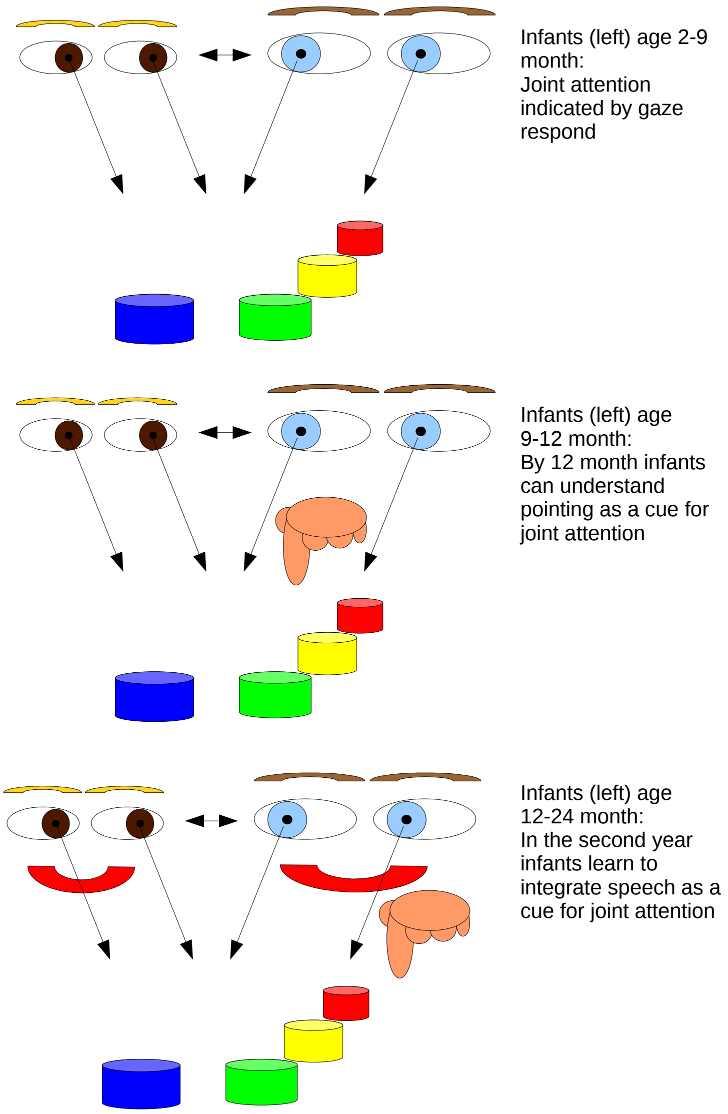
As visual attention is the most targeted part of attention, it is the main focus of this work. It seems that, even if there is still no perfect definition of attention, the subdivision into parts, as in the definition of James, seems plausible. From the perspective of interaction, the concept of attention and mechanisms that are required to tune the search process can be limited. Even though we do not fully understand the concept of attention, we can try to specify the required mechanisms. For an interaction as a tutoring situation, which is the target of this work, the mechanisms will be limited to: joint attention, saliency and anticipation.

In more detail, in a tutoring situation, we could limit the needs of attention to:

- from the learner's perspective, the sensitivity to perceive contingent Ostensive Signals (OS) from the teacher
- from the teacher's perspective, sending contingent signals and making the information salient to the learner
- from the learner's perspective, giving feedback to the teacher by anticipating the object to which attention is paid
- from the teacher's perspective, detecting the feedback and take it into account for the presentation structure to create a joint attention frame.

2.1.2.1 *Joint attention in a tutoring interaction*

The term *joint attention* has been used for quite a while now to define an observed behavior which occurs in a social interaction. Joint attention, also referred to as deictic gaze by Butterworth [17], is a mechanism that can be observed in a triadic interaction, where a triadic interaction is defined as the interaction between two partners and an object that is the target of this interaction. A simpler definition of joint attention could be "looking where someone else is looking" [17]. A broader definition would be that, joint attention is the idea that humans make inferences from observable behaviors of other humans by attending to the objects and events that these other humans attend to. This has been recognized as a critical component in human-robot interactions [136]. A key factor for detecting joint attention is to detect the gazing behavior of the interaction partners. Baron-Cohen [8] describes two neurocognitive mechanisms that have evolved to answer the question if someone is the target of another organism's attention (Eye Direction Detector (EDD)) and the question of how to create a shared focus of attention with another organism (Shared Attention Mechanism (SAM)). The SAM is only used in a triadic interaction. Therefore, it seems to be the mechanism that creates the observed behavior named joint attention. The emergence of joint or shared attention has been identified as playing a number of important roles in the social and cognitive development of an infant. Human social interactions can be very complex and comprise of multiple levels of coordination, from high-level linguistic exchanges, to low-level couplings and decouplings of bodily movements. In particular, the temporal patterns of eye-gaze coordination between interacting humans, including mutual eye fixations as well as following gaze shifts to perceivable objects in the environment, play a critical role in the establishment of mutual rapport and understanding [79]. The joint attention behavior seems to develop during the first year of life in humans. A joint attention paradigm proposed by Scaife and Bruner [102] indicated that 30% of 2 months old infants turned their heads to follow the line of regard from a model. Between 11-14 months, 100% of the infants were capable of turning their heads in the expected direction. In contrast to the study proposed by Woodward [133], results show that infants at the age of 3-6 months are responding on gazing behavior in the appropriate way, but they do not seem to understand gaze as an object-directed action until the age of 9-12 months. When taking not only gaze but also pointing as a joint attention behavior, only infants at the age of 12 months respond appropriately. Tomasello [117] follows up on these different steps of development and goes even further in the direction of infants acquiring linguistic skills. The development of joint attention could thus be extended to the second year of life where the first linguistic conversations take place. For a summary see Figure 2.2.



Infants (left) age 2-9 month:
Joint attention indicated by gaze respond

Infants (left) age 9-12 month:
By 12 month infants can understand pointing as a cue for joint attention

Infants (left) age 12-24 month:
In the second year infants learn to integrate speech as a cue for joint attention

Figure 2.2: Development of infants ability to detect and use joint attention.

2.1.2.2 *Saliency in a tutoring interaction*

The concept of saliency maps, or a mechanism that is detecting salient regions, is based on the need of the individual to accentuate important features occurring in the world. Saliency could thus occur in every sensory stream that is important to an individual. There are two common definitions of saliency depending on the research area:

> "In neuroscience salience (also called saliency) of an item - be it an object, a person, a pixel, etc. - is the state or quality by which it stands out relative to its neighbors. Saliency detection

> is considered to be a key attentional mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data. Saliency typically arises from contrasts between items and their neighborhood, such as a red dot surrounded by white dots, a flickering message indicator of an answering machine, or a loud noise in an otherwise quiet environment. Saliency detection is often studied in the context of the visual system, but similar mechanisms operate in other sensory systems.
> When attention deployment is driven by salient stimuli, it is considered to be bottom-up, memory-free, and reactive. Attention can also be guided by top-down, memory-dependent, or anticipatory mechanisms, such as when looking ahead of moving objects or sideways before crossing streets. Humans and other animals cannot pay attention to more than one or very few items simultaneously, so they are faced with the challenge of continuously integrating and prioritising different bottom-up and top-down influences..."[1].

> "Salience is the state or condition of being prominent. Salience refers to the relative importance or prominence of a part of a sign. The salience of a particular sign, when considered in the context of others, helps an individual to quickly rank large amounts of information by importance and thus give attention to that which is the most important. This process keeps an individual from being overwhelmed with information overload..."[1].

The first computational model of a visual saliency system was proposed by Itti and Koch [50]. The idea behind the model was to decompose an image into its important features, like color, intensity and orientation, to create maps with scaling values for each feature and sum them up in a single saliency map (Fig.2.3). Many researchers have been working on this concept of saliency maps and more and more features have been added. Also, other ways of combining these feature maps have been explored. For more details, see section 3.2.6.

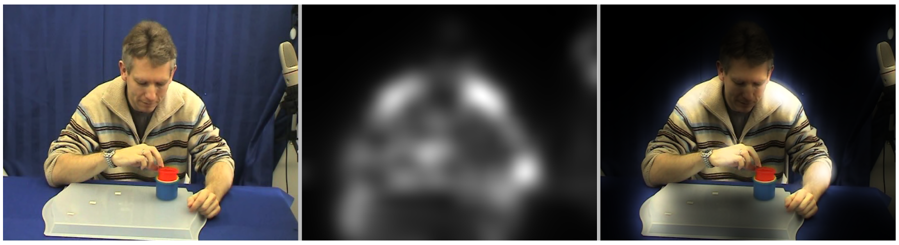


Figure 2.3: Detecting salient regions with the saliency computational model proposed by Itti et al[50]. The left image is the input image, in the middle is the saliency map and on the right is the combination of both of them.

2.1.2.3 *Anticipation in a tutoring interaction*

After getting an idea about what is important in an interaction by detecting a salient region or by following the tutor using a joint attention mechanism, important information can be gathered from a tutoring situation. To give feedback about this knowledge and to predict the following steps, anticipation is needed. The ability to anticipate movements starts to develop in the second half of the first year of human life. The ability to predict the movements is judged by the infants gaze to the object or to the predicted goal of the objects movement [30]. Anticipations are ubiquitous, e.g., each time we switch on our TV, we anticipate that it will start up and deliver some colorful moving pictures [96]. The prediction of movements is a very important factor in action planning and the abilities to predict the actions of someone seems to be directly linked to the speed and accuracy at which a task is performed [52].

2.1.3 *Children prefer contingent actions*

Csibra [23] argues that the prerequisite to identify sequential organization of a co-activity is the mechanism of contingency. The term contingency is considered with regards to a temporal sequence of behavior and reaction [42]. Spatial contingency points to the fact that e.g. when somebody claps its hands, the sound will be perceivable from the same spatial region as the movement of the hands. Contingent intensity, for example, can be seen between the decrease of height and decrease of pitch. Watson [130] assumes that children are born equipped with a contingency detector module that allows them, not only to detect contingency and discern whether they are a part of a true interaction, but also to expect contingent behavior and to try to elicit it. As already mentioned above, most studies focus on contingency as the temporal sequence of behavior. Here, the probability of a temporal related sequence is analysed: A prospective probability refers to the conditional probability of an approaching stimulus as a function of an emitted response while retrospective probability refers to the conditional probability that a stimulus event was presented by an emitted response [41], [71]. A mirroring response can be considered as perfectly contingent [71]. Before the age of 3 months, infants prefer perfect contigencies over intermittent ones and react to a sudden still face of the care-giver with reduced smiling and gazing [81], [109], [71]. Moreover, when different events are shown to infants, they do not learn that they belong together if the time span between the events exceeds 1 second [54]. With reference to the system of child and caregiver, Keller et al. [54] have further shown that "one common feature of verbal, nonverbal and intermodal maternal responses toward infants communicative signals in face-to-face interactions, is a general propensity to react within short intervals of less than 1 second, that is,

the time range during which small infants can detect contingencies...."
As shown in developmental research, there seems to be a developmental shift around the age of 3 months. At that age, infants start to prefer high but intermittent contingency [109], [71]. This development is rooted in contingent interactions with adults.
As already pointed out in section 2.2.5, similar symbiotic behavior can be observed for the temporal contingency: Not only do infants prefer contingent behavior from their caregivers [120], [11] and try to elicit it [53], [87] but there is also evidence that parents intuitively produce contingent actions [61]. With increasing age of their children, mothers have been shown to decrease their level of contingency according to infant's the increase of development for a certain task [55].
Along the lines of Natural Pedagogy [25], it is argued here that, in order to be part of a social interaction, an artificial system needs to be equipped with mechanisms that detect and produce contingent behavior. Interestingly, this hypothesis has been tested in a kindergarten environment in which Tanaka et al. [115] investigated to what extent a robot will be treated as a peer rather than a toy when it behaves in a predictable manner, i.e., displaying contingent behavior. The researchers took a within subject design and analysed several sessions of interactions, for which the robot was displaying different skills. More specifically, in the 11th session, the robot showed simple reflex-like contingent behavior by giggling immediately after being touched on the head. The authors report that after this change, in children, the distribution of the touch behavior towards the robot converged to the distribution of the touch behavior toward the peer [115]. Thus, children treated the robot more like a peer. In sum, Tanaka and his colleagues [115] have shown that when a system produces a contingent behavior, it gains more attention and toddlers socialise with it for a sustained period of time. In addition, in studies where children interacted with a PLEO robot, it was investigated how users attempt to establish coordinated 'sequences of action' with the robot and different strategies were revealed:

1. experimenting and organising their own actions with regard to the contingency exhibited by the system

2. making use of a 'mediator' who observes both the robot's and the infant's actions and helps to 'translate' the robot's actions in terms of what next actions are relevant for the user

3. observation and technical reasoning [90]

This strongly suggests that the capability of producing a contingent behavior will facilitate human robot interaction. Yet, for a system to learn from a human, it is necessary that it can not only produce contingent behavior but also can detect it. This could be realised by gathering features that tutoring behavior exhibits in different modalities. These features should guide the development of a tutoring spotter system. The capability to spot a tutor will enable the system to pay attention to an ostensive action and the crucial parts or circumstances that are helpful in resolving the problem of what and when to imitate [86]. Furthermore, mechanisms that detect and produce contingency can be a precursor of later dialogical competencies as described in the framework of grounding. While contingency mainly describes a temporal pattern where one event occurs as an answer to a previous one, grounding relies on semantic information in the sense that one event, or speech act, needs to be grounded by an interaction partner through a signal of understanding.

2.1.4 *Summary*

In this section, related work has been presented, focusing on the perspective of a child's attention mechanism in an tutoring interaction, with a goal to create a description of an tutoring interaction model. The human attention mechanism includes a saliency mechanism, the capability for joint attention, and is able to anticipate. Fig. 2.4 shows the model of the tutoring scenario used in this work. This scenario was adapted and revised according to the findings of each stage of this work. Finally, the preference for a contingency interaction of children was presented.

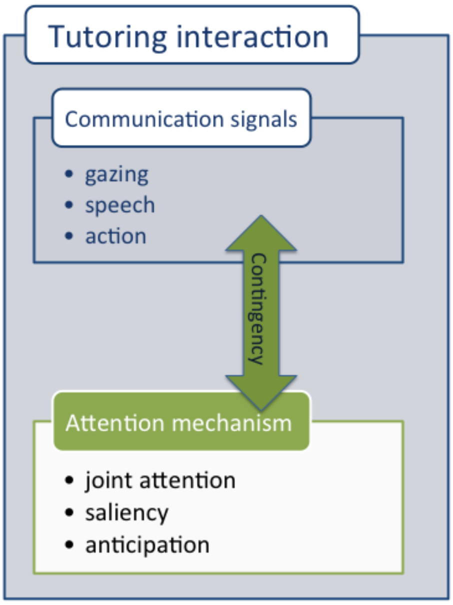


Figure 2.4: A model of a tutoring interaction between adult and child: the attention mechanism of a child.

2.2 RELATED WORK: OSTENSIVE SIGNALS GUIDE ATTENTION

2.2.1 *Ostensive signals highlighting novel information*

Csibra and Gergely [24] highlight the importance of the pedagogic behavior that is crucial to the understanding of some actions: "...pedagogy essentially created a new way of information transfer among individuals through the use of ostensive communication". In their work, they give the example of peeling a hard fruit or carving away pieces of wood with a tool. The movement and the tool in both actions are the same, but the goals and the reasons for the action are very different. While it is easy to infer the goal of the action when peeling a fruit, i.e. accessing the edible parts, it is not obvious what is intended in the case of the wood carving. Therefore, tutoring is crucial for a learner to understand the goal correctly. Ostensive signals (OS) have been measured and modeled in several disciplines by focusing on single sources like motion (*motionese*), speech (*motherese*) and structuring behavior (*contingency*).

2.2.2 *Innate structuring mechanism (contingency)*

The concept of contingency has been identified as a mechanism which has several characteristics. Gergely and Watson [42] presented a conceptual distinction among three occurrences of contingent behavior in humans. They distinguished the occurrence into spatial relation contingency, temporal contingency and sensor related contingency. These three often occur in a hybrid form. Spatial relation contingency describes a relation in space between objects or individuals in terms of a combined occurrence. Gergely and Watson [42] have given a nice example: "You recall being in a room watching a person making an impassioned speech. You recall his facial expressions. Among other events that transpired, you recall three instances in which he pounded his fist on the lectern while at the podium. One blow was hard, one soft, and one slightly softer yet. The order in which these occurred is not clear in your memory, however. Suppose you enter a room. The room is empty except for the presence of three flowers. They differ only in size. One is large, one is smaller, and one is yet slightly smaller (2,6 and 14 inches)... Suppose, however, that we do not remember the variation in the relative size differences among the flowers. Instead, our limited memory provides us only with images of where things happened. We now note that the flowers reside at three places on the lectern. More than that, there is a flower at each place we recall the speaker hitting the lectern - one at the lower left corner, one in the center, and one midline at the top. We do not know the temporal order of the flowers appearance nor do we have any evidence of correlated variation in the sensory quality of the flowers. Yet, despite the lack of temporal or sensory pattern information, it is clear that the pattern of spatial positioning alone provides a powerful implication for the attribution of causal relatedness between podium pounding and the presence of the flowers."

The temporal contingency is often described as a certain type of rhymes or joint attention. It has been described as a mechanism that helps us to structure diadic interaction in time. Gergely and Watson [42] stated that "many studies have shown infants to be sensitive to situations in which their behavior is followed in time by a stimulus event (e.g. a vocalisation is followed by an auditory or visual stimulus, [12], [94], or a leg movement is followed by movement of a mobile [100], [99], [129])."

The concept of sensor related contingency tries to measure the relatedness between human sensory input, e.g., when you move a cup you can feel it in your hand and you can see it moving. Or, in the example of Gergely and Watson [42], you can hear three different sounds and see the person hitting on the lectern. This could help you to distinguish the three episodes of hitting.

The measurement of contingency has several starting points, due to the expanded concept. Most of the measures are concentrating on one of the forms of contingency described above. Some of them try to measure a hybrid occurrence of contingency. E.g., Brand et al. try to measure the temporal contingency of a tutor by calculating the frequency of eye gaze bouts in a tutor [14]. While Movellan [80] is measuring the temporal contingency of the occurrence of a vocal respond to a given vocalization and is taking the interaction loop into account, Brand is focusing exclusively on the tutor.

In summary, contingency is a mechanism responsible for how stimulus and reaction are bound together.

2.2.3 *Motionese: a polemic hand motion to guide infant's attention*

To quantify the polemic motion of parents towards their child, several metrics have been investigated by Rohlfing et al. [98] and Brand et al. [14]. The metrics they found give a particular insight into the hand motion of the tutor. They take into account the enlargement, the speed and the smoothness of the presented motion. By comparing these metrics, a tutoring situation towards an infant can be distinguished from one towards an adult [98], [14]. In Chapter 3, a method for transferring these metrics to a robot directed tutoring behavior is presented.

2.2.4 *Child directed speech scaffolds infants (motherese)*

The term motherese has been defined by Gogate et al. [43] and specifies infant directed speech that is motionese in terms of intensity and amplitude. It has been shown that subjects, when asked to speak even to an imaginary infant, were not able to produce speech that exhibits all the features that are characteristic for motherese as it is produced in real ACI [56]. The use of certain key words is another important factor of the motherese behavior towards infants. The tendency of using motherese towards robots is presented in Chapter 3. Some argue [43], [24] that the activated behavior underlying the tutoring process for motionese and motherese is driven by the same ability.

2.2.5 *Children prefer faces*

Current developmental research on predispositions towards social environment proposes that human children are born with a bias to faces [106]. This results in the behavior that even newborns will orient toward and attend to faces. Currently, it is still under debate whether this preference is innate: Pascalis and Kelly [88] argue that, because face processing is a complex task, different forms of individual capabilities are observed. The authors conclude that such a system is unlikely to be developed at a very young age. In contrast, research in developmental psychology [78] as well as in developmental robotics [33], shows that rapid learning is possible to establish this preference behavior. Such behavior allows for receptivity towards other humans who are the source of social signals. The function of such a receptivity can be viewed in terms of evolutionary establishment of how caring for the infants can be solicited. In addition, the function can also be seen in terms of cultural knowledge transmission. Csibra and Gergely [24] argue that human infants "are adapted to transfer knowledge to, and receive knowledge from, conspecifics through teaching...." Thus, predispositions allow for this adaptation. More specifically, because indefinite information from the environment can confuse the child, a selection of information is necessary. This means that children need to recognize when a situation is communicative and therefore especially suited for knowledge transmission [25]. Interestingly, the means by which such a situation is recognizable (such as ostension, child-directed speech or child-directed action) are simultaneously linked to information reduction. Hence, the benefit of such a situation is twofold: first, children recognise that this is a communicative situation, and second, the input is reduced and tailored for the receivers in this situation.

For example, the bias towards faces is viewed as linked to OS such as making eye contact. Eye contact, or the later ability to establish a joint focus of attention [118], is a powerful indicator communicating that knowledge is shared [119], [24]. Joint attention is necessary for navigating a joint focus on the environment. This means that children can see what object or event is of interest by monitoring the eye gaze of the interaction partner. At the same time, the interaction partner

can draw some inferences about the child's interest from her or his eye-gaze. An object or event can become a matter of joint action only if the interaction partners manage to coordinate and constantly monitor their attention. Such a "co-activity" [39] establishes one aspect of the ability to communicate relevant information [24]. This ability is also acknowledged from the perspective of developmental robotics [6] as a concrete mechanism of a reciprocal interactive contribution making social learning possible.
It is important to emphasise once more that infants are not only biased towards faces but there is actually a system comprising of a child being receptive and its parent complementing this receptivity by a specific form of interaction [39]. This system is kept ongoing when this co-activity is coordinated. The mechanism of contingency provides both sequential and co-occurring coordination.

2.2.6 *Summary*

In this section, related work about tutoring interactions between adults and children has been presented. This section focused on the communicated signals produced by the adults. It was shown that these signals are specialised to target children by using exaggeration. These exaggerated signals (or ostensive signals) have been shown to be supportive to the children. It has been presented that, for describing a tutoring interaction, both interaction partners were to be taken into account. Therefore, both perspectives of the communication signals,

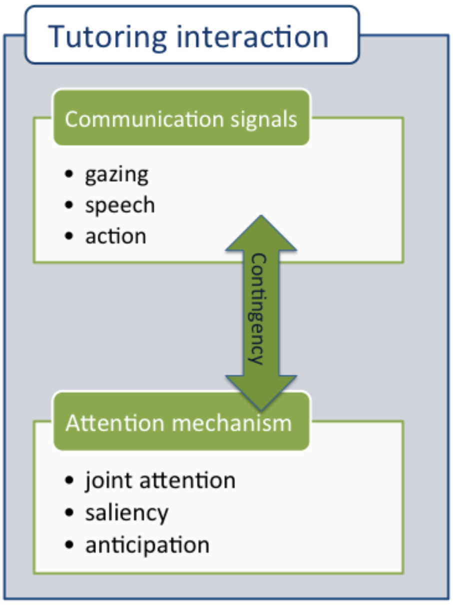


Figure 2.5: A model of a tutoring interaction between adult and child: the communication signals of an child.

as well as the attention mechanism, have to be studied to understand a tutoring interaction (Fig. 2.5).

2.3 ANALYSIS: CRUCIAL FACTORS FOR A SUCCESSFUL INTERACTION

Learning by interaction is a paradigm in which a robot learns to bind a signal, like an object, with some meta information, like a grasping action for an object, from a tutor. This is in contrast to learning by observation, where the robot would try to learn the grasping by observing the grasping task, without getting meta information about it. In the learning by interaction paradigm, the robot will get more information about what the tutor considers to be crucial aspects of the action. Learning that takes place in a social environment is inspired from the observations made in developmental studies on human children and primates. Scaffolding or bootstrapping processes have been identified as supportive for learning because the observed action does not always reveal its meaning. Scaffolding behavior is also increasing attention [137]. In an interaction with a learner, it is essential that the tutor makes sure that the learner is receptive and ready to learn. The reciprocal contribution, i.e. the guidance of attention by a tutor on one hand and the manifestation of receptivity by a learner on the other hand, seems to follow certain interactive regularities [21], [39], [92]. Analysis of parent-child interactions have revealed that:

- depending on their age and linguistic capabilities, as investigated for the different groups, infants provide different kinds of feedback (Table 2.1). More specifically, in group 1 focusing on pre-lexical infants of the age of 8 to 11 months, feedback consists primarily of gazing behavior displaying the infant's state of attention. In group 2, with early lexical children (12 to 23 months of age), children begin to anticipate next actions through the direction of gaze and use more gestures and other modalities with which they provide the parent with information about their actual understanding of the presented action. This becomes more evident in group 3 of lexical children (24 to 30 months of age), in which the infants' feedback pinpoints to the structure of the action more systematically. Thus, the child provides "action guides", i.e. it times its own (verbal and bodily) action in relation to the adult's presentation as the feedback occurs after the first sub-action and/or at the end of the second and third sub-action.

- across the age groups, the infants' feedback seems to operate on two levels: as continuous involvement (e.g. through gaze) and at discrete points within the structure of the interaction (e.g. through pointing gestures at objects).

Regarding the interactional loop between adult and infants, two patterns were found:

- Considering the precise timing of the infant's gaze in relation to the adult's hand movements, the infant's gaze follows the

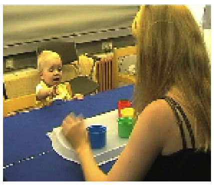
Age	Pre-lexical children 8-11 month	Early lexical children 12-24 month	Lexical children 25-30 month
			
Gazing	State of attention	Anticipating next actions	Structures action, solicit input
Interaction	More gestures and other modalities give information about the understanding of the presented action	Verbal communication negotiating novel aspects of the action	
Action			"Action guides" Gestural anticipating of action; Attempts to handle objects

Table 2.1: Different kinds of feedback as a result of parent-child interactions

current actions or anticipates the next relevant action. The latter is mostly the case for children in the early lexical and lexical groups (2 and 3).

- Considering the precise timing of the infant's gaze in relation to the adult's verbal utterance "look"/"guck mal", its function changes with the infant's age: While it serves to grab the child's attention in group 1, it becomes a structuring signal that marks important points of the demonstration to the children, in group 2 and 3.

From these results, the following implications for the development of robot systems that should learn within and from social interaction can be derived: A robot's feedback, depending on its modality, should be provided continuously or transmitted at specific moments in time, making use of multimodal conduct; in this way it is possible for the robot to influence the presenter's actions.
For the concrete mechanisms of a reciprocal interactive contribution, it seems that there is more than just additional social information that the child can take advantage of. It seems that, from a very early age, children are biased towards such interactional exchanges. The focus here is on two aspects that contribute to such biases: Children's preference to look at faces and their preference for contingent actions. These two aspects are considered as crucial and linked to each other and were chosen for the subsequent model of the tutoring spotter system. In the next sections, these aspects are elaborated and evidence for their importance is provided from developmental psychology.

2.3.1 *Motionese Corpus*

The Motionese Corpus was conducted for a study on the topic of parent-child interaction by Rohlfing [98], with 64 participating parents and their children. The parents and their child were sitting on opposite sides at a table facing each other (Fig. 2.6). Details about the participants can be seen in Table 2.2. The parents were instructed to present several randomised small tasks to the child and then to another adult. As a result, there are interaction recordings captured with two video cameras (Fig. 2.6). The corpus comprises of three age groups: pre-lexical children (8-11 months of age), early lexical children (12-23 months of age) and lexical children (24-30 months of age) (Table 2.2). The second group was originally divided into 2 subgroups, because of the drastic increase in the infant's vocabulary [10], but here, subgroups 2a and 2b will be counted as one group.

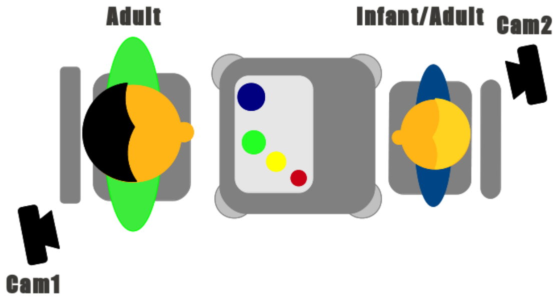


Figure 2.6: Motionese setting: There are two cameras recording the scene. The interaction partners are seated opposite to each other and the object is placed on the table in front of the tutor.

	Group 1	Group 2		Group 3
		Group 2a	Group 2b	
Infants age in months	8-11	12-17	18-24	25-30
Number of infants	18	15	16	18
Gender of infants	10m, 8f	8m, 7f	9m, 7f	7m, 11f

Table 2.2: Infants age, number of infants and gender of infants which participated in the motionese study.

The tasks the parents had to fulfill appeared in randomised order (Fig. 2.7):

- Lamp task: The adult had to show the child how to switch a table lamp on and off by pulling on a cord.
- Cup-stacking task: The adult had to stack a green, a yellow and a red plastic cup one after the other into a larger blue cup.

- Minihausen task: The adult had to restore a certain building block configuration (shown on a small image on the tablet) by adding several blocks to a prearranged building block basis.
- Drawing board task: The adult had to show the infant how to use stamps on a magnetic drawing board.
- Bell task: The adult had to show the child how to use a table bell.
- Shelf task: The adult had to explain to the child how a shelf with sliding doors is opened and closed.
- Book task: The adult had to put 3 small books into a box that has a lid. The box is closed at the beginning of the task.
- Ring task: The adult had to put 3 small plastic rings into a box similar to the box in the book task.
- Bag task: The adult had to show how a bag with a zipper is opened.
- Salt shaker task: The adult had to explain how a salt shaker can be used to pour some salt onto a blue lid.

Only the Cup-stacking task, the Minihausen task and the Salt shaker task are taken into account here as examples.

2.3.2 *Looming: a way of making a movement salient*

As child-directed action was shown to be important in a tutoring situation, Matatyaho and Gogate [72] investigated further the kind of action that is typically used. They found that the looming action, which is the action that describes a movement of a tutor moving an object towards a learner's face, is used more frequently than upward or backward motions in temporal synchrony with the spoken words. This looming motion is likely to highlight a novel word-object relations [44].

According to Regan and Beverley [95], looming is a movement that is oscillating in the size of an object. The movement is a motion away from the body of the presenter. When the distance is getting smaller for the learner the object's appearance to be bigger. Regan and Beverley have shown that the human visual pathway is even more sensitive to looming movement than to a movement that is oscillating sideways. Studying interaction with young children, Matatyaho and Gogate [72] showed that, when mothers introduce a novel word for an object to their infants, they use showing gestures, like looming and shaking motion, in conjunction with temporal synchrony. These looming forms of actions are likely to highlight novel word-object

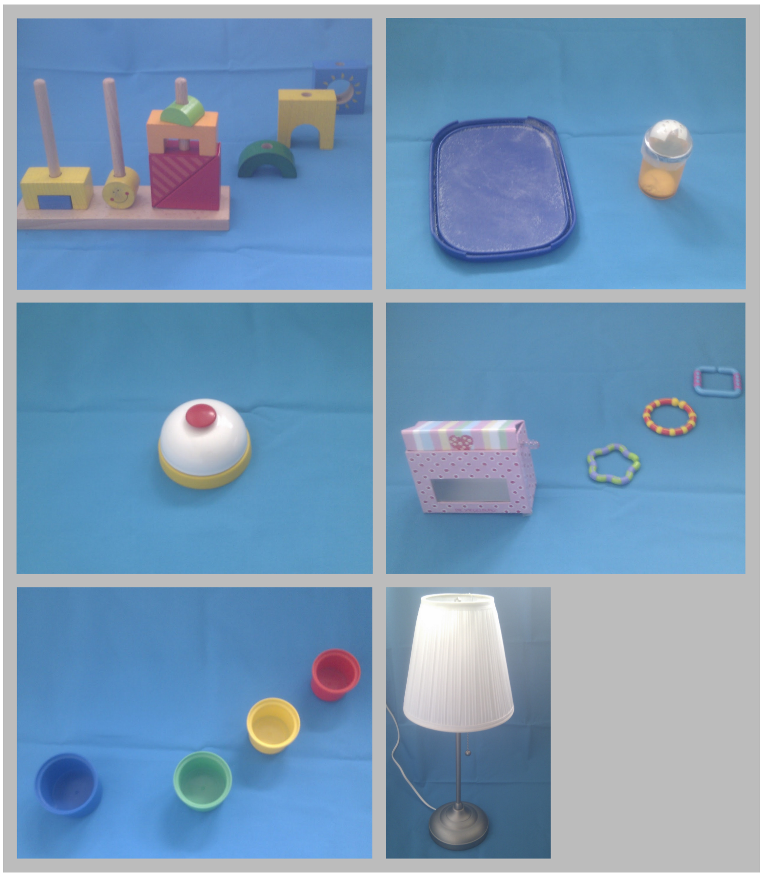


Figure 2.7: Six of the used objects presented in the motionese study. Minihausen, salt shaker, bell, ring, cup-stacking, and lamp task.

relation for young infants. Therefore, looming movement seems to be an important signal in child-directed action. Taking the Motionese corpus into consideration, an investigation, regarding whether parents perform looming action and what they are saying when they perform this kind of movement, was conducted. For the analysis, the focus was on two different items: the salt shaker and the cups (Fig. 2.9). The sub-action structures (Fig. 2.8) were annotated. These sub-actions are based on the *landmarks* that were found during the performance of the task. Landmarks are important points in the task, for example when a movement starts or a sub-tasks ends. Also transcriptions of the parent's speech and the actions adopted to gain children's attention (e.g., looming) were performed.

For each sub-action, the looming periods were calculated and the transcript was cut based on these periods, followed by a counting of the uttered words that were classified as "naming, attention getter and others" words. This categorisation was done based on the theory that there are important words, *keywords* that facilitate the internal structure of an utterance.

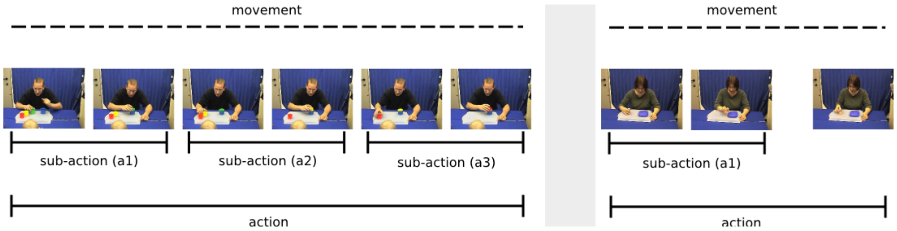


Figure 2.8: This figure presents the enclosed sub-actions of the cup stacking and the salt shaker task.

DEFINITION: *Keywords* are the words that are used to reveal the internal structure of an author's reasoning. While they are used primarily for rhetoric, they are also used in a strictly grammatical sense for structural composition, reasoning, and comprehension. Indeed, they are an essential part of any language [132].
There are many different types of keyword categories including: conclusion, continuation, contrast, emphasis, evidence, illustration, and sequence. Each category serves its own function, as do the keywords inside of a given category.

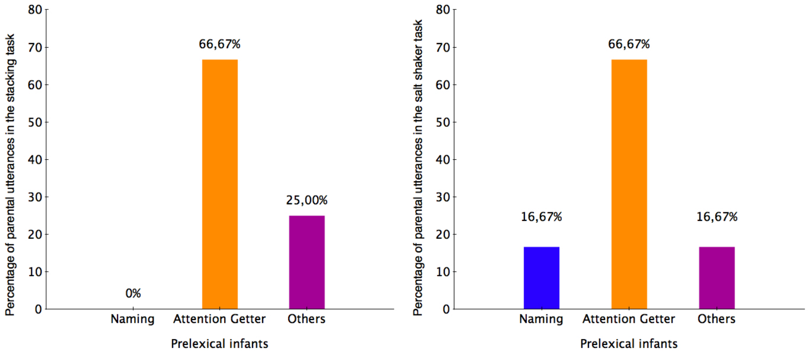


Figure 2.9: The content of parental speech when performing looming movement: Naming (= naming/labeling objects), Attention getter (= trying to gain child's attention by saying "look!"), and Others

For the cups, that are mostly goal-oriented objects, it was found that looming movement was performed by only 40% of the 35 participating parents. In contrast, when demonstrating the salt shaker, that is a mostly manner-oriented object 76% of the 37 participating parents performed looming movement during their interaction with a child. It was also analyzed what the parents were saying while performing the looming movement.
The findings of Matatyaho and Gogate [72] showed that parents perform this kind of movement when they teach novel object labels to their infants. However, in that study, parents were actually asked to teach novel labels to their children. By contrast, in the study presented

here, parents were free to decide whether they wanted to introduce object labels or show actions to the children. Two hypotheses can be formed, based on the findings of Matatyaho and Gogate [72]:

1. Looming action conveys object labels
2. Looming action attracts children's attention and is performed, not only to convey the object label, but most importantly to get the child to attend to the object.

The second hypothesis suggest that, in a setting more natural than the one provided by Matatyaho and Gogate [72], looming action is performed for the purpose of gaining the child's attention, as the parents devoted the most part of the looming movements to attention getting statement like calling the child's name or saying "guck mal! [look here!]". Support for this hypothesis can be found in Fig. 2.9.

2.3.3 *Learners' age affects the tutoring behavior*

The metrics concepts, described in section 2.2, represent the tutoring behavior in terms of motion, language and structuring behavior. Other studies have shown that there is a difference in tutoring behavior between addressing a child or an adult and that young infants aged 6 to 8 months prefer motionese [14]. In a different age span, a difference can also be found [127]. The effects of children's age on motionese, defined as modified action demonstration [13], [98], is investigated. In the study presented here, parents demonstrated a function of an object (cups stacking) towards their infant and towards another adult. Parental behavior in three different age groups was analysed: parents of pre-lexical (8-11 months), early lexical (12-24 months) and advanced lexical (25-30 months) children. In this analysis, objective metrics of hand trajectories, providing data about their shape and time structure, were used. The results suggest that actions, chosen to attract attention by providing more range, can primarily be found in interaction with younger infants, whose attention needs more guidance. Interactions with older children seem to benefit either from the increase of children attention abilities or the use of other means (such as language) by their parents to attract their attention. In contrast, parameters that appear to be more in charge of structuring the action by organising it in motion pauses, seem to persist over the age and verbal capabilities.

2.3.4 *Infants anticipate the tutor's actions*

Current research suggests that there are differences in the anticipation, in relation to the type of a presented task and that anticipation is present in the presence of goal oriented actions, i.e. in *PATH*-directed tasks [30], [97], [38].

According to Talmy [114], there are four components to the semantics of motion events.
These are (i) a figure, which is an object of perception that moves, (ii) the motion, which is the actual description of a motion event, (iii) a path, which consists of a source location, a trajectory and a goal and (iv) a ground, which is a landmark or several landmarks with respect to which figure moves [107]. Each of these components can be emphasised (explicitly encoded, or as Talmy [114] puts it "windowed for attention") or omitted (not explicitly named) in a respective utterance. Different tasks can therefore be distinguished as more *MANNER*- or *PATH*-oriented. More precisely:

MANNER DIRECTED TASK: *MANNER* of motion refers to a type of distinct motion described by a particular verb, e.g., running, tumbling, sliding, walking, crawling, etc. A *MANNER* utterance contains the means, medium or speed information (see also Fig. 2.10) and is mostly encoded in the verb in satellite-framed languages.

PATH DIRECTED TASK: *PATH* of motion refers to the direction of the movement, e.g., movement into, out of, across, etc. A *PATH* utterance contains a source, a trajectory and a goal (see also Fig. 2.10) and is often encoded in the prepositions, but also in adverbs.

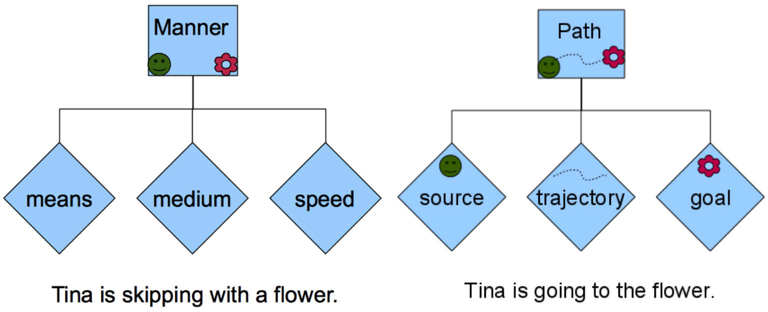


Figure 2.10: *MANNER* and *PATH* directed utterances can be classified in terms of the shown parts.

These two concepts are borrowed from the linguistic theory described in [112]. They can be encoded in the verb as part of its core meaning, or in a separate particle associated with the verb (a satellite, as for instance a preposition). According to Talmy [112], [113], [114], [49] languages can be distinguished in satellite-framed and verb-framed, as a function of the how the motion is encoded. The former expresses the semantic components of the motion event, motion and *MANNER*, conflated in the verb, and the *PATH* in a satellite. The latter conflates motion in the verb, and expresses *MANNER* in a separate

expression. All Germanic languages are satellite-framed languages (e.g., English and German).
An example of the use of *MANNER* and *PATH* utterances in human communication could be the following sentence:

Tina	ran	into	the room.
FIGURE	MOTION+MANNER	PATH	GOAL

This sentence conveys information about the event occurring in the external world. There is a figure (Tina) in motion, moving in a particular *MANNER* (i.e. running, not skipping) forward along a *PATH* that crosses a boundary into a goal location (i.e., the room).
It is suggested here that a fundamental step in the acquisition of these concepts is the parental tutoring, in which the windowing of attention is performed differently for more *PATH*- or more *MANNER*-oriented tasks. These salient parts have been identified, based on the transferred meta information of the performed task. In the Motionese Corpus (section 2.3.1) we selected the cup stacking task as one example of a *PATH*-oriented task (meta-information is the size of the cups) and the salt shaker as an example for a *MANNER*-oriented task (the meta-information is that there is salt in the salt shaker and it can be used for salting by turning it).
In the work of Pruden et al. [93], it was found that children acquire the capability to encode the description of a *PATH*-oriented motion first. Based on this knowledge, an analysis of the Motionese corpus, in which the transfer of the description of motion into the presentation of a task in a way which allows to classify the salient parts of the presentation into more *PATH*- or more *MANNER*-oriented, is done.
By looking back at the occurring utterances during the sub-actions (Fig. 2.8) of the cup stacking and salt shaker task the classification was verified. Finally it was examined if the utterances are *PATH*- or *MANNER*-oriented for the cup stacking task.

Based on these results, it can be confirmed that the cup stacking task is a more *PATH*-oriented task than the salt shaker task. It was also found, that in the salt shaker task, there is a tendency to use more *MANNER*-related utterances towards the lexical children than towards the early lexical and pre-lexical children. Overall, there were only few *MANNER*- and *PATH*-related utterances compared to the overall utterances in the sub-actions (Fig. 2.8) selected. But upon inspection of the utterances in relation to the action, it was found that there are many looming actions (section 2.3.2) in the motion and attention getters in the speech that could be excluded from the analysis. Also, these utterances contained descriptions of static states of affairs, attention getters (see looming above), social interactions with the child, one-word utterances and action-descriptions.

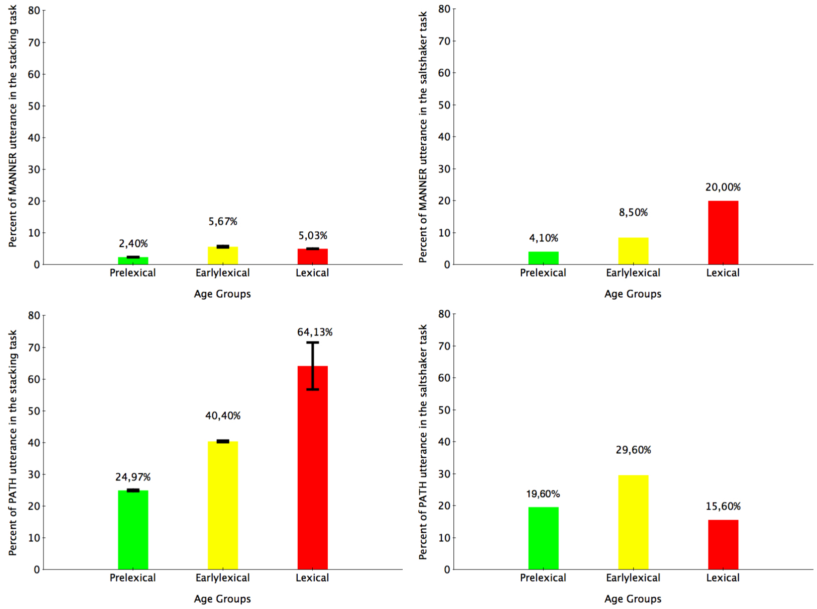


Figure 2.11: On the left, is the percentage of *MANNER*- and *PATH*-oriented utterances by the parents while performing the cup stacking task. On the right, the utterances presented during the saltshaker task. The colors are an encoding of the age groups.

After classifying the cup stacking task as a *PATH*- and the salt shaker task as a *MANNER*-oriented task, the analysis and results for the anticipation behavior shown by the children's eye gazing behavior took place. The Motionese Corpus was recorded with ordinary video cameras. Therefore, the analysis was done manually and frame by frame to get the highest accuracy possible for the eye gaze directions of the children.

Analysis

For the salt shaker task, the data were encoded based on the occurrence of three features:

1. Are there movements to draw the child's attention to the manipulated object? E.g., shaking movements, moving the object closer to the child.

2. Is the child anticipating? A short look at the lid at the beginning of the task or looking at the salt on the lid at the end of the task were not counted as anticipatory gazing behavior.

3. Is child looking at the salt shaker all the time while the adult performs the task?

For the cup stacking task, data based on action segmentated into sub-actions were encoded. The action was segmented based on the transportation of the red, yellow and green cup into the blue one (Fig. 2.8).

Then five features for each sub-action were encoded:

1. Was the child displaying an anticipatory gazing behavior from the active cup to the blue cup?

2. If yes, how long is the time-interval between the cups being lifted off the tablet and the anticipation? This time-interval was measured in frames. A normal video has 25 frames per second. This implies that every frame covers the time-interval of 0.04 seconds. This is the maximum accuracy we could get by using the video data.

3. Is the child anticipating how long the time-interval between the first arrival of the infant's eyes and the cup's arrival at the blue cup is? This time-interval was also measured in frames.

4. Are there any movements to draw the child's attention back to the manipulated object after the child anticipated or right at the beginning of the task? E.g., shaking movements, moving the object closer to the child.

5. If there were any attention drawing movements performed with the object, did they draw the child's attention back to the object?

Condition	Pre-lexical children	Early lexical children	Lexical children
Participant's age in months	8-11	12-24	25-30
Number of participants	24	50	27
Unusable trials for salt shaker	3	5	6
Unusable trials for cups stacking green cup	12	21	12
Unusable trials for cups stacking yellow cup	13	22	11
Unusable trials for cups stacking red cup	14	26	15

Table 2.3: Unusable trials were selected on the basis of the task performance and the video quality.

Results

In the salt shaker task, 97,8% of that age group and nearly 99% of all children did not anticipate during the trial. In 92% of all children, the parents used a looming behavior to shift the attention of the child to the object.

For the cup stacking task, the analysis showed that, the children in group 2 and 3, distinctly anticipated more than the children in group 1 (Table 2.4). The rate of anticipating children in group 2 and 3, in comparison to group 1, was up to two times higher in the green and yellow cup part and up to four times higher in the red cup part (Fig. 2.12). There were significant differences in the percentage of anticipative gazing behavior between the prelexical and the early lexical infants (p=0.04). There were also significant differences in the percentage of anticipative gazing behavior between the prelexical and the lexical infants (p=0.03).

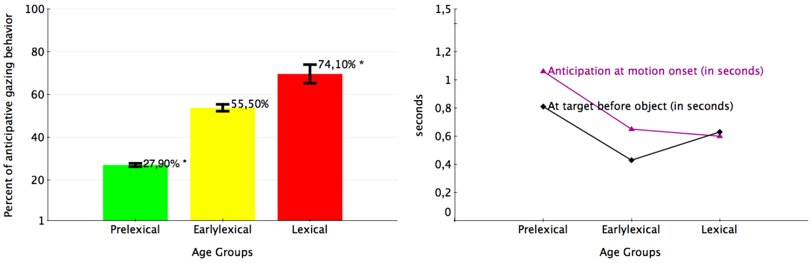


Figure 2.12: On the left the analysis of the percentage anticipative gaze presented by the infants while viewing their parents performing the cup stacking task. On the right, the timing when the infants start anticipating in relation to the motion onset of the cups (purple) and in relation to the arrival of the object at the target position (black).

The average time interval, between the start of the objects movement and anticipation, strongly decreases over all three cup parts in all age groups. The time interval, between anticipation and the cup's arrival at the blue cup increased over time in all three groups, but these data were fluctuating a bit more (Fig. 2.12) .

In summary, only 23% of children at the age of 8-11 months showed an anticipating eye gazing behavior to the goal position of the cup, where at the age of 25-30 months, ca. 70% of children showed an anticipating behavior. Considering speech, it was found that the parents tended to give more *MANNER*-related instructions towards the lexical children than towards the pre-lexical and early lexical children. Also, for the *PATH*-related instruction, the parents gave more utterances like that towards the lexical children than towards the pre-lexical and early lexical children. A correlation was found between anticipative gazing behavior and the occurring of *MANNER*- and *PATH*-related instructions the parents gave over the age groups. The older the children, the more the anticipation that could be found and the older the children, the more the occurring of *MANNER*- and *PATH*-related instructions that the parents gave.

Pre-lexical children	Anticipation after start (in seconds)	At target before object (in seconds)	Attention movement
Green Cup	~1,65 (0,72/2,24)	~0,52 (0,12/0,96)	0/12 (0%)
Yellow Cup	~1,03 (0,56/1,48)	~1,06 (0,52/1,64)	0/11 (0%)
Red Cup	~0,52 (0,44/0,60)	~0,86 (0,44/1,28)	1/10 (10%)
Early lexical children	Anticipation after start (in seconds)	At target before object (in seconds)	Attention movement
Green Cup	~0,90 (0,32/1,88)	~0,40 (0,16/1,20)	2/29 (6,9%)
Yellow Cup	~0,64 (0,20/1,52)	~0,30 (0,08/0,72)	0/28 (0%)
Red Cup	~0,40 (0,08/0,92)	~0,58 (0,12/2,88)	1/24 (4,2%)
Lexical children	Anticipation after start (in seconds)	At target before object (in seconds)	Attention movement
Green Cup	~1,01 (0,32/1,44)	~0,54 (0,40/0,88)	3/15 (15%)
Yellow Cup	~0,44 (0,08/1,16)	~0,56 (0,12/1,08)	0/16 (0%)
Red Cup	~0,35 (0,04/1,00)	~0,78 (0,24/1,24)	1/12 (8,3%)

Table 2.4: The first and second column contain the average time between start of the object movement and anticipation and the average time between anticipation and the arrival of the moving cup at the blue cup in seconds. The brackets contain the lowest and highest amount of time for that feature. The last column contains the number of trials that contained a movement to draw the child's attention back to the object.

2.3.5 *Summary*

In this section, crucial factors for creating and maintaining a successful interaction, were presented. The motionese corpus was introduced and a first analysis of the interaction was presented. In it, the metrics for looming action, anticipatory gazing behavior and language concepts (manner and path) were studied. It was found that looming action, anticipatory gaze and language concepts, as well as the learners age, affect the tutor interaction (Fig. 2.13).

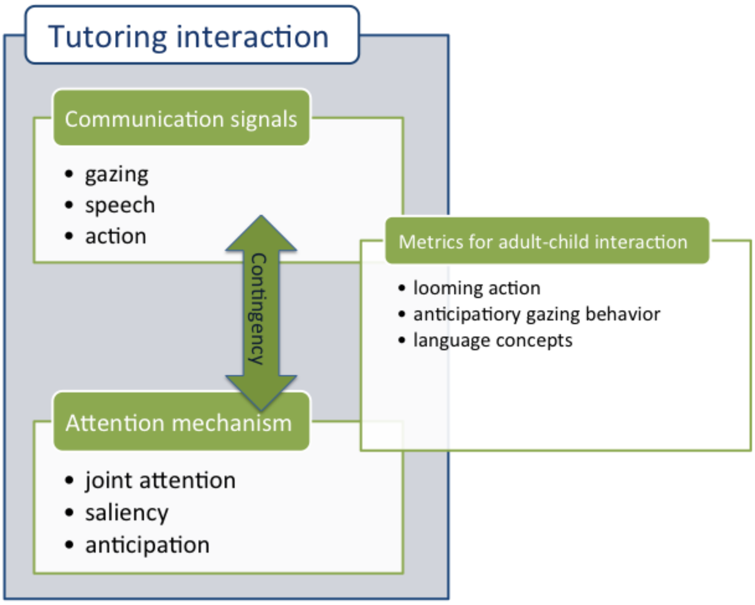


Figure 2.13: A model of a tutoring interaction between adult and child: metrics.

3 INSIGHTS FROM HUMAN ROBOT INTERACTION

In recent years, a lot of work has been done in the field of Human Robot Interaction (HRI). A great deal of this work has focused on the production of robotic behavior as, for example, in infant-like robot behavior. Here, the focus of the research is on studying how the behavior of the robot is interpreted by a human tutor. To produce realistic infant-like robot behavior, it is essential to understand how humans perceive robots' movements. This field of research is very broad and many researchers have studied it starting from very different points of view. Koay et al. [57] have shown that the movements of a mobile robot can induce discomfort when the robot does not respect the social space to a person, i.e., when it is blocking the path of person or coming too close. Perhaps the most interesting part of this field lies in the gazing behavior of both interaction partners. The gazing behavior is very important in Human Human Interaction (HHI) as a great deal of information is transported through the gazing behavior. Gazing behavior has a high importance in HRI, too. Mito et al. [76], [77] found that humans show the same "breaking eye contact" behavior when interacting with humans and robots. Furthermore, they state that this behavior can be an evaluation of an android's human-likeness. Later on, results will be presented in comparison to an infant as well as an adult as learner in a tutoring situation (sections 3.2.2 and 3.2.3). In addition to this, the problem of which robot behavior is adequate for evoking the desired effects, has to be solved. Lopes et al. [69] give a lot of information about neck-eye coordination of infants to create behaviors that are very infant-like. The gained knowledge about tutoring behavior will be transferred to a robotic platform to make the system more likely to be accepted as a social interaction partner. To do so, the difference between the embodiment of a human and a robot has to be considered (section 3.2.5).

3.1 RELATED WORK: USING SALIENCY TO CONTROL ROBOTIC GAZING BEHAVIOR

3.1.1 *Creating attention with a saliency system*

The Ackachan system [64], consist of two parts: the robot simulation and a model of a Saliency-based Visual Attention System (Fig. 3.1).

- The simulation
  The simulated robot Ackachan has four degrees of freedom (DoF). The eyes have two DoF, for the eyelids and the mouth. Only the eyes are controlled by the outcome of the Saliency-based Visual Attention System. The other two DoF follow a random pattern.

- Visual attention system
  The model, inspired by the behavior and the neural mechanism of primates, can detect salient locations in a scene that stand out from the surroundings with respect to color, intensity, orientation, flicker (i.e., change in the brightness) and motion (i.e. optical flow) [84].

The calculation of the most salient point is based on the model proposed by Itti and Koch [50].
In the Ackachan system, the simulation is equipped with a web cam. The saliency system is applying several linear filters (e.g. colors, intensity, orientations) on the video input. Based on the resulting images, there is a normalization and an across-scale combination step. In the end, all resulting images are linearly combined and the point with the highest saliency is calculated on the resulting saliency map. The system used could run in near real time; the computation is done for 9 frames per second. Based on the results of the saliency calculation, the robot simulation is looking to the most salient point in the scene. As it was mentioned in the Chapter 2, an attention system seems to consist of different parts, like joint attention, saliency and anticipation mechanisms. It has been argued [83] that a saliency system can create joint attention (Fig. 3.2). Consequently, a saliency system might be a good starting point for modelling the attention system of a robot.

But the question is whether that system would induce an equal tutoring behavior in a human child. This question is important because only if the behavior towards a robot is similar to the one towards a child, the gained knowledge from the parent child diads, that have previously been studied, can be used. If it is possible to transfer this knowledge, it would improve the ability of a robot to know when a human is trying to transfer important knowledge to the robot and what this knowledge is about.

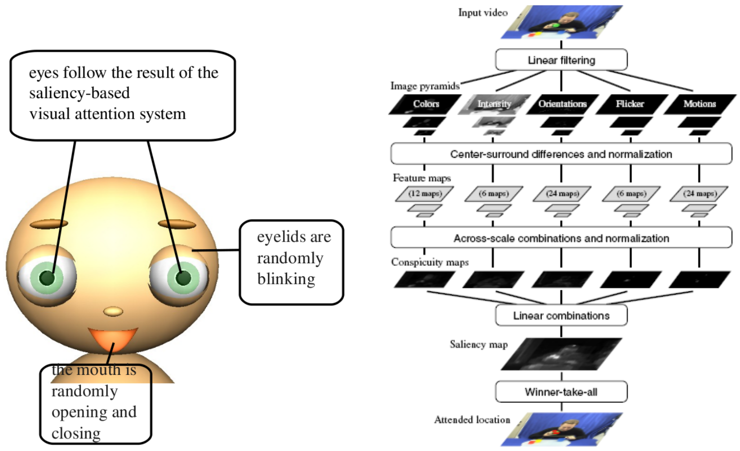

eyes follow the result of the saliency-based visual attention system
eyelids are randomly blinking
the mouth is randomly opening and closing
Input video
Linear filtering
Image pyramids
Colors
Intensity
Orientations
Flicker
Motions
Center-surround differences and normalization
Feature maps
(12 maps)
(6 maps)
(24 maps)
(6 maps)
(24 maps)
Across-scale combinations and normalization
Conspicuity maps
Linear combinations
Saliency map
Winner-take-all
Attended location

Figure 3.1: The left figure shows the degrees of freedom and the behavior of the Ackachan. The right one shows the Saliency-based Visual Attention System. [84]

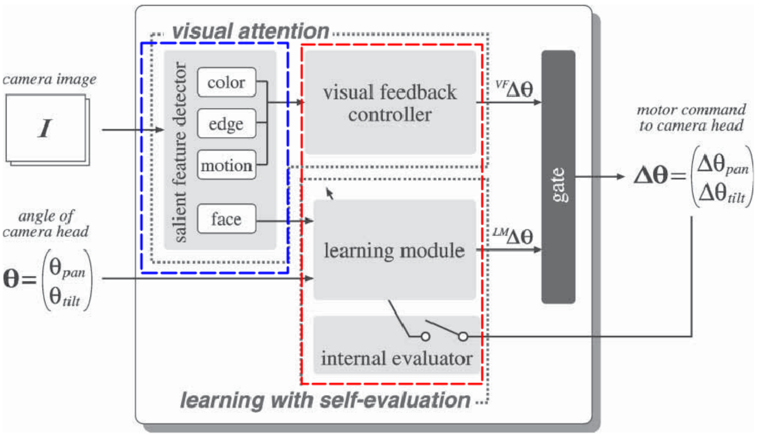

visual attention
camera image
I
salient feature detector
color
edge
motion
face
visual feedback controller
$^{VF}\Delta\theta$
learning module
$^{LM}\Delta\theta$
internal evaluator
gate
motor command to camera head
$\Delta\theta = \begin{pmatrix}\Delta\theta_{pan}\\ \Delta\theta_{tilt}\end{pmatrix}$
angle of camera head
$\theta = \begin{pmatrix}\theta_{pan}\\ \theta_{tilt}\end{pmatrix}$
learning with self-evaluation

Figure 3.2: The blue part describes the linear filters used for the saliency calculation. The red part describes on the one hand the saliency point where to look at and gives on the other hand a special rating onto the face orientation map. The result is that the face gets a better rating as a salient point [83].

3.1.2 *The iCub humanoid robot*

The iCub robot [75] is 104cm tall, weighs approximately 22 kg and has 53 degrees of freedom in total, consisting of 16 controlled degrees of freedom for each arm (with 9 controlled degrees of freedom in each hand), 6 controlled degrees of freedom in each leg, 3 controlled degrees of freedom for the torso and 6 controlled degrees of freedom for the head. It is equipped with a wide variety of sensors, like positional sensors (absolute position encoders) for the joints, gyroscopes, accelerometers and force/torque sensors. It also has two digital cameras,

positioned in its eyes, to give a realistic point-of-view for the user. The iCub robot was chosen because it is a child-like humanoid platform. There is a PC104 card in the head of the robot handling all sensory and motor-state data and controls the communication with external computers that normally are responsible for the more complex and CPU intensive tasks. One of these tasks is, for example, handling the iKinGazeCtrl module (that is used in this thesis), a software module, that uses inverse kinematics to calculate the iCub's head and eye positions according to a given set of 3D coordinates [89]. The connection between the PC104 card and external computers are handled by the YARP midelleware [74] .
This robotic platform was used in two different setups:

1. in the first one, the robot was equipped with the same system as the Ackachan. Even though the iCub has two cameras on board, an external web cam was used, to be more in line with the Ackachan system and to solve the ego-motion problem.

2. in the second one, the robot was equipped with a model that will be presented in Chapter 4. This model was trying to detect a tutoring situation in an interaction.

To transfer the knowledge gained from analysing the interaction with the simulated robot, the question whether the embodiment of the iCub had an effect on the behavior of the tutor had to be asked. This question is targeted in the following section.

3.1.3 *Summary*

In this section, related work has been presented, focusing on the perspective of a child-like humanoid robot in order to create an attention mechanism in an tutoring interaction. As a first step, a saliency mechanism was used to study the focus of attention of such a robot (Fig. 3.3). Finally, the iCub humanoid robot platform was introduced.

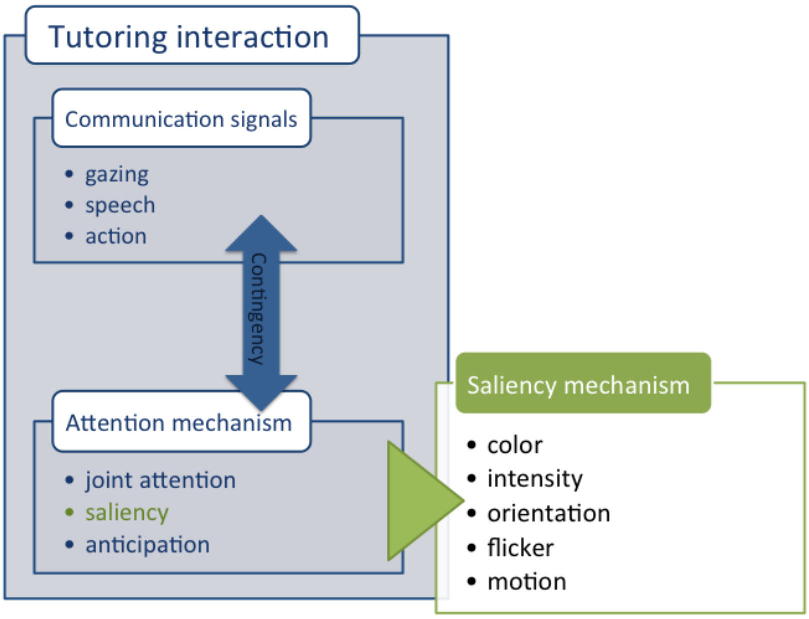


Figure 3.3: A model of a tutoring interaction between adult and child: the saliency mechanism.

3.2 ANALYSIS: ATTENTION IN A TUTORING SITUATION WITH A ROBOT

The outcomes of the analysis of ACI and AAI (section 2.3.1), by comparing those tutoring situations with an ARI, had to be transferred and used. To transfer this knowledge, it is important to, not only take into account the behavior, of the tutor (Chapter 2), but the robot should also produce acceptable feedback (section 2.3). However, there are even more things to be considered, in order to make the robot benefit from the induced tutoring behavior, because there is a difference between the embodiments of a human and a robot.
The results from the experiments with two different robotic platforms are presented here. The first one is using the simulated robot Ackachan [84] (section 3.1.1). The other one is using the iCub embodied robot (section 3.1.2) .

3.2.1 *Robot-Directed Interaction Experiment (RDIE)*

There were 31 adults (14 females and 17 male) participating in this experiment. Out of them, 7 were parents (Table 3.1).

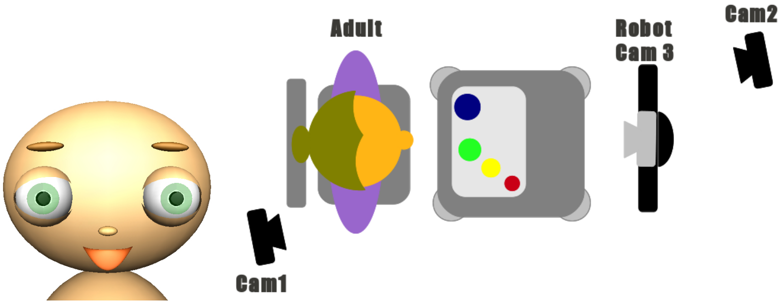


Figure 3.4: The left picture shows the simulated robot. The right picture shows the Ackachan-directed interaction setting. There are four cameras recording the scene. The subject is seated opposite to the robot and the object is laid on the table in front of the tutor.

Participants age	18 - 65 years; median = 29 years
Number of participants with children	7
Number of participants	31
Gender of participants	14 female, 17 males
Estimated age of the robot	0.6 - 10 years; median = 3 years

Table 3.1: Participants age, number of participants with children, number of participants, gender of participants and estimated age of the robot of participants which participated in the Robot-Directed Interaction Experiment.

The participants were instructed to present 6 of the tasks that were used in the Motionese Corpus (section 2.3.1) to a simulated interaction partner (Fig. 3.1). The tasks the participants had to fulfill appeared in randomized order:

- Lamp task: The adult had to show the child how to switch a table lamp on and off by pulling on a cord.
- Cup-stacking task: The adult had to stack a green, a yellow and a red plastic cup one after the other in a larger blue cup.
- Minihausen task: The adult had to restore a certain building block configuration (shown on a small image on the tablet) by adding several blocks to a prearranged building block basis.
- Bell task: The adult had to show the child how to use a table bell.
- Ring task: The adult had to put 3 small plastic rings into a box similar to the box in the book task.
- Salt shaker task: The adult had to explain how a salt shaker could be used to pour some salt onto a blue lid.

The interaction partner was an infant-looking virtual robot with a saliency-based visual attention system (Fig. 3.1.1). The robot's eyes were following the most salient point in the scene, which was computed by color, movement, and other features, in accordanc with [82]. More details about the system and the study can be found in section 3.1.1.

3.2.2 *Analysing the different dimensions of behavior adaptation*

In this section, the main focus of the study with the Ackachan robot system was on the tutoring behavior towards this robot system, in contrast to tutoring behavior towards a child or an adult. This comparison was done by taking into account the tutor's hand movement and eye gazing behavior. The results were published in [126]. First, the question targeted was which end of the scale of tutoring behavior the one towards a robot is, in contrast to towards an adult and towards a child. Researchers often assume that tutoring towards a robot and towards a child is very similar (Fig. 3.5).

The study was similar to Herberg et al. [47], exploring whether people will modify their actions when interacting with a machine was pursued. In contrast to Herberg et al. though who used a computer, the interaction here was investigated with a virtual robot. For this purpose, the real interactions with the artificial system, not just a picture of the partner as in the study presented by Herberg et al., were analysed and the results were compared to the ones obtained from real

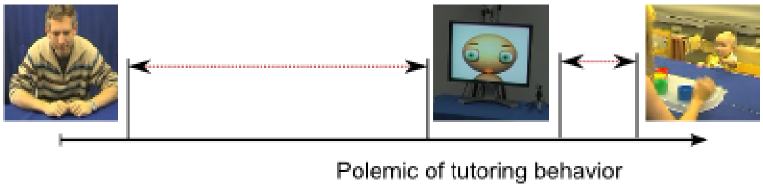


Figure 3.5: The scale of tutoring behavior between towards an adult, towards a child and towards a robot.

interactions with a child and an adult. A range of metrics, allowing a fine-grained analysis of performed motions and their changes in the interaction as it unfolds, were used.

3.2.2.1 *Experiment*

To quantify the behavior of a tutor towards a robot in relation to the behavior toward a child or an adult, the data from the Motionese Corpus (section 2.3.1) and the RDIE (section 3.2.1), were analysed.
From the Motionese Corpus (section 2.3.1), the pre-lexical children group, comprising 12 families of 8 to 11 months old children, was selected. The focus was on the analysis of the cups stacking task, because it offers the best comparability in motion performance. A subgroup of 8 parents (4 fathers and 4 mothers) for the Adult Child Interaction (ACI) and a subgroup of 12 parents (7 fathers and 5 mothers) for the Adult Adult Interaction (AAI) were further selected, because of the quality of the video and the sound and the way in which the action was performed. Only those parents were selected, who started the task by putting the green cup into the blue one (Fig. 3.6a1 and Fig. 2.8).
For the Adult Robot Interaction (ARI) 12 participants (8 female and 4 male) from the Robot-Directed Interaction Experiment (RDIE) (see Section 3.2.1), who performed the task in a comparable manner, were selected.
The participants were instructed to demonstrate several objects to an interaction partner, while explaining to him/her how to do it (Fig. 3.1). Again the stacking-cups task was chosen for the analysis.

3.2.2.2 *Data Analysis*

The goal of the analysis was to investigate the tutoring behavior from two perspectives: motionese and contingency. The videos were coded semi-automatically to obtain data for the 2D hand trajectories and the eye gaze directions.

For all annotations, the video captured by camera 2 (Fig. 2.6 and 3.1) was used. It showed the front view of the demonstrator and was therefore best suited for action, movement, and gaze annotations, which are discussed in detail below.

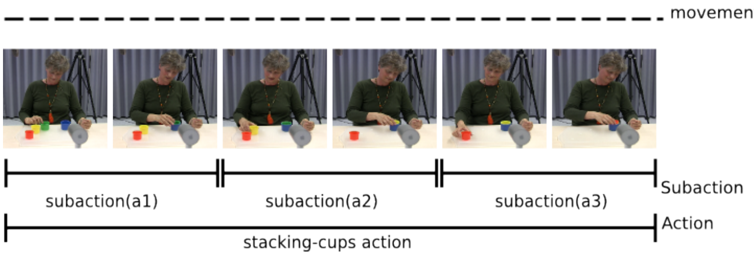


Figure 3.6: This graphic shows an example for the structure of an 'action', 'sub-action', and 'movement'.

Action Segmentation: For analysing the data, the action of the stacking-cups and additionally, the sub-actions (a1-a3) of grasping one cup until releasing it into the end position (Fig. 3.6 and Fig. 2.8) were marked in the video. The above were defined as following:

1. *action*, as the whole process of transporting all objects to their goal positions.

2. *sub-action*, as the process of transporting one object to its goal position.

3. *movement*, as phases where the velocity of the hand was above a certain threshold. All other phases were defined as pauses.

Motionese

Hand Trajectories: The videos of the two experiments were encoded via a semi-automatic hand tracker system (Fig. 3.7). The system was written as a plugin for a graphical shell, iceWing [62], and made it possible to track both hands with an optical flow-based algorithm [70]. The system allowed manual adjustment in case of tracking deviation. This tracking system was used instead of the previously used 3D body model system [98]. Since 3D results in [98] were not significant, 2D analysis was performed to show more consistent results. Additionally, the new system was easily accessible for non-expert users.

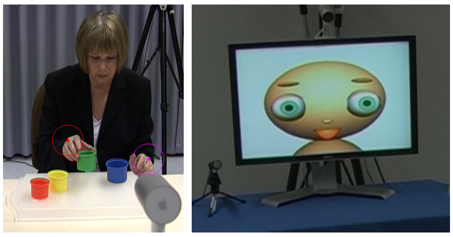
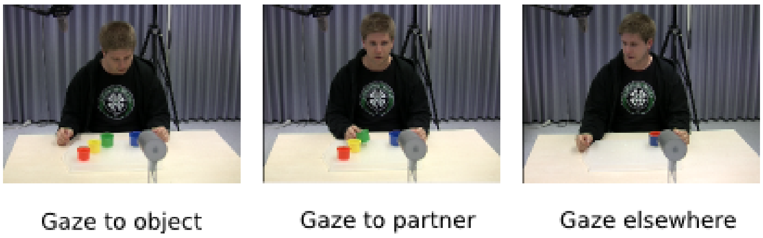


Figure 3.7: In the left picture, the red and violet circles depict the tracking regions which are tracked by the hand tracker system. The points in the middle of the circles are the resulting points for the 2D hand trajectory. In the right picture, the virtual robot used is shown.

Contingency

Eye Gaze: In annotating the eye gaze directions using the software Interact [2], a distinction between looking at the interaction partner and looking at the object was made (see Fig. 3.8).



Gaze to object

Gaze to partner

Gaze elsewhere

Figure 3.8: These three pictures show the difference between looking to the object (left), looking to the interaction partner (middle) and looking somewhere else (right).

3.2.2.3 *Metrics*

Motionese

For quantifying motionese and contingency, seventeen variables were taken into account, related to the 2D hand trajectories that derived from the videos and the eye gaze bout annotations produced with Interact.

Motionese is operationalised in terms of velocity, acceleration, pace, roundness and motion pauses as defined in [98]. Rohlfing et al. [98] automatically segmented the task into movements and pauses based on hand velocity.

Velocity was calculated using the derivative of the 2-dimensional hand coordinates of the hand that performed the action per frame. Rohlfing et al. did not find a significant effect for velocity for the 3D posture tracking data. Their 2D hand tracking data showed a statistically significant trend that hand movement in AAI is faster than in ACI.

Acceleration is thus calculated from the as the hand velocity.

Pace was defined for each movement by dividing the duration of the movement (in ms) by the duration of the preceding pause (in ms). For pace, Rohlfing et al. found almost significant differences comparing ACI and AAI. Their results suggest that pace values in ACI are lower than in AAI.

Roundness of a movement was defined by a covered motion path (in meters) divided by the distance between motion on- and offset (in meters). Thus, a higher value in roundness means rounder movements. Rohlfing et al. found that hand movement is significantly rounder in AAI compared to ACI.

Frequency of motion pauses was defined as the number of motion pauses per minute. Therefore, the number of motion pauses was calculated

automatically using the segmentation mentioned above (Fig. 3.6). Furthermore, the *average length of motion pauses* (in frames) and *total length of motion pauses*, as the percentage of time of the action without movement, were calculated.
Additionally, eye gaze trajectory during the actual transportation of the cups, when performing the task, was taken into account. For each video and setting, the exact video frames of the beginnings and ends of the transportation for each of the three cups were annotated by hand (Fig. 3.6). That way, variables were defined for each individual sub-action (a1, a2, a3) and also changes in the demonstrator's behavior in the course of completing the task, were detected.
Sub-action specific velocity was calculated as the average velocity for sub-actions a1, a2, and a3 each.
Sub-action specific acceleration was calculated as the average acceleration for sub-actions a1, a2, and a3.
Range was defined as the covered motion path divided by the distance between motion, i.e. sub-action, on- and offset.
Action length denoted the overall action length and was measured from the beginning of sub-action a1 to the end of sub-action a3.

Contingency

Watson [130] describes contingency as the human infant's means for detecting socially responsive agents and therefore postulates the existence of an innate contingency detection module as one of the most fundamental innate modules. He formally defines the contingent temporal relation of two events, for example, a response R and a stimulus reward $S*$, as two conditional probabilities. The first, called the sufficiency index, measures the probability of a stimulus reward $S*$, given a span of time t, following a response R, $P(S*|Rt)$. The second, called the necessity index, measures the probability of the response given time t prior to the reward stimulus, $P(R|tS*)$ [130]. "Contingency detection is crucially involved in an infant's progressively developing awareness of his or her internal affective states" [24]. So contingency seems to be an important factor for an infant to learn more about it self and its surrounding. "The discovery that another agent's gaze is a cue worthy of monitoring relies on the infant's ability to detect the contingency structure in interactions with that agent" [32]. The contingency of the interactions was quantified in terms of metrics related to eye gaze, as defined in [14] for measuring interactiveness.
Following the metrics of Brand et al. [14], the metric for contingency was used for our data collection. The *frequency of eye-gaze bouts to interaction partner*, i.e. eye gaze bouts per minute, was calculated from the Interact annotations. Also, the *average length of eye-gaze bout to interaction partner* and the *total length of eye-gaze bouts to interaction partner* as the percentage of time of the action spent gazing at the interaction

partner were calculated. Brand et al. found that infants received significantly more eye-gaze bouts per minute [14], so the frequency of eye-gaze bouts to the interaction partner was significantly higher in ACI than in AAI. The total and average length of eye-gaze bouts to the interaction partner in their study was significantly greater in ACI than in AAI.

Equivalent metrics were calculated for the eye gaze on the demonstrated object. In particular, values were obtained for *frequency of eye-gaze bouts to object*, *average length of eye-gaze bout to object*, and *total length of eye-gaze bouts to object* as the percentage of time of the action spent gazing at the object.

3.2.2.4 *Results*

A short qualitative summary of the results can be found in Table 3.2. For quantitative results and the detailed analysis see [126].

Compared to AAI, ACI shows	Compared to ACI, ARI shows	Compared to AAI, ARI shows
slower hand movement	slower hand movement	slower hand movement
lower hand movement acceleration	lower hand movement acceleration	lower hand movement acceleration
smaller pace	smaller pace	smaller pace
less round movement		less round movement
greater range and therefor more exaggerated movement	greater range and therewith more exaggerated movement in the first sub-action	greater range and therewith more exaggerated movement
higher frequency of motion pauses		higher frequency of motion pauses
greater average length of motion pauses	greater average length of motion pauses	greater average length of motion pauses
greater total length of motion pauses	greater total length of motion pauses	greater total length of motion pauses
longer action	longer action	longer action
more frequent eye-gaze bouts to the interaction partner	less frequent eye-gaze bouts to the interaction partner	
on average longer eye-gaze bouts to the interaction partner	on average shorter eye-gaze bouts to the interaction partner	
more time spent gazing at the interaction partner	less time spent gazing at the interaction partner	
higher frequency of eye-gaze bouts to object	lower frequency of eye-gaze bouts to object	lower frequency of eye-gaze bouts to object
smaller average length of eye-gaze bout to object	greater average length of eye-gaze bout to object	
smaller total length of eye-gaze bouts to object	greater total length of eye-gaze bouts to object	

Table 3.2: A short summary of the results.

3.2.2.5 *Discussion and Conclusion*

In summary, the results showed a differentiated picture for modifications in human-robot interaction. On one hand, it was found that a robot receives even more strongly accentuated input than an infant: almost all hand movement-related variables, when pooled over the whole action sequence, showed a significant difference, or at least a trend, between the three conditions with a clear ordering (AAI $>$ ACI $>$ ARI). ARI movements can therefore be characterised as slower (velocity, acceleration, and pace), more exaggerated (range) and less round (roundness) than AAI movements. In contrast to ACI, where the tutoring behavior seemed to have lots of variability, in the ARI, more stability could be observed. This suggests that ARI allows to control for the parameters of the learner and is therefore a promising method for studying tutoring behavior. On the other hand, the contingency metrics showed less contingent eye gazing behavior in ARI than in ACI (frequency and length of eye-gaze bouts to interaction partner).

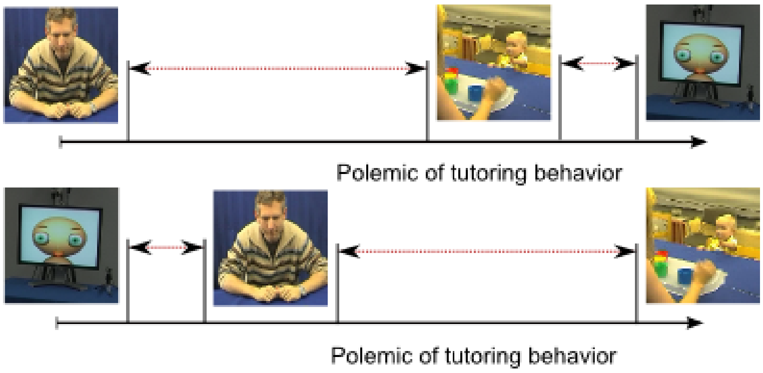


Figure 3.9: In the first scale, it is shown how the motionese features are scaling the ACI,AAI and ARI. In the second scale, it is shown where the results of the Contingency values for ACI, AAI and ARI can be sorted.

These results raise an interesting question: Why is the behavior of the tutors in the ARI condition less contingent than in the ACI condition? As contingency is a bi-directional phenomenon, it is likely to be related to the robot's feedback behavior. Indeed, while the frequency of motion pauses was similar in ARI and ACI, the length of motion pauses was significantly longer in ARI than in AAI and ACI indicating that the tutor was waiting - possibly in vain - for a sign of understanding from the robot. The lower amount of eye-gaze bouts to the interaction partner in ARI as opposed to ACI could be interpreted similarly: as the tutor was not receiving the expected feedback of understanding from the robot, she or he was not searching for eye-contact with the robot. In future research, a closer focus will be on the feedback behavior and identifing the important signals in a bi-directional interaction.

These results have important inplications for human-robot interaction in developmental robotics. They indicate that the behavior of the robot shapes the behavior of the tutor. Although all tutors showed strong modifications in their movement behavior towards a robot, thus stressing important aspects of the demonstrated action, they did not increase their contingency behavior as other tutors would do in interactions with infants. Even though the purely reactive behavior of the robot in our study induced parent-like teaching (as indicated in a qualitative study by Nagai et al. [84]), it did not seem to be sufficient to produce a contingent interaction. As studies show, contingent behavior is an important feature for learning in human development. Thus, in order for robots to be able to learn from a human tutor, they should have the capability to engage in a contingent interaction. Further studies need to be carried out to find out if these metrics are generalising over different taks.

3.2.3 *Do the measures for ostensive signals generalize over different tasks?*

The results presented above, in section 3.2.2, are very promising, but, at that time, only data from one task were considered. The question is whether these results will generalise over different tasks. In [65] the question is posed: "Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations?". Results of a task with a similar structure will be presented based on a more fine-grained analysis of the eye gaze behavior in order to:

- show how the findings by Vollmer et al.[126] hold for a different task
- analyse the structure of eye-gaze behavior over time and
- discuss these results with regards to what extent the observed modifications of behavior can be interpreted as ostensive signals in human-robot interaction.

3.2.3.1 *Experiment*

Again, data were selected from parent-infant and adult-robot interactions (sections 2.3.1 and 3.2.1). From the overall set of items that were presented, the "Minihausen" task was selected. This task is similar to the cups stacking task as it is a rather goal-directed action, with three sub-goals to be reached. Results from analysis of motionese and contingency features in parent-infant and adult-robot interaction have shown that, while motionese features of infant-directed and robot-directed interactions are similar, they diverge for contingency metrics, indicating that contingency is impaired in human-robot-interaction. Here, the question to what extent these results are decisive for the statement that motionese, as well as contingency features, serve the function of OS, will be asked.

Subjects Motionese Corpus

As before, the pre-lexical children group, comprising of 12 families of 8 to 11 months old children from the Motionese Corpus (section 2.3.1), were selected.
The "Minihausen" task, which was to sequentially pick up the blue (a1), the yellow (a2) and the green (a3) block and put them on the wooden base with three poles on the white tray, was selected here (Fig. 3.10 a1).

Subjects Robot-Directed Interaction Experiment (RDIE)

From the RDIE, 12 participants were selected (8 female and 4 male), who performed the task in a comparable manner (section 3.2.1).

3.2.3.2 *Data Analysis*

The data analysis was again structured into two groups, one that measures motionese and another one that is used to measure contingency. The videos were coded semi-automatically to obtain data for the 2D hand trajectories and the eye gaze directions. The semi-automatic video annotation was done with an optical flow tracker, that was manually corrected when the tracker was losing the tracked point (the hands).

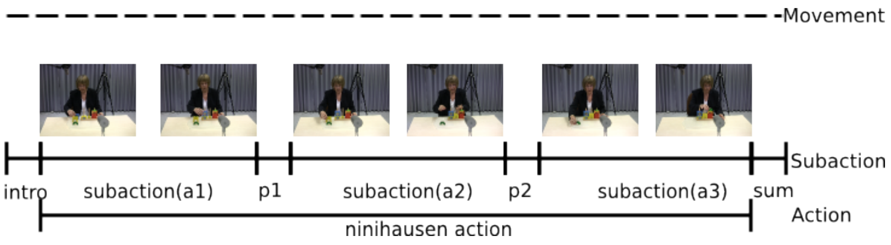


Figure 3.10: The action was divided into movement and pause parts and into sub-actions. This figure shows an example of the structure of an 'Action', 'Sub-action'(intro = introduction and sum = summary), and 'Movement'.

3.2.3.3 *Annotations*

For all annotations, the video captured by camera 2 was used (Fig. 2.6 and 3.4). It showed the front view of the demonstrator and was therefore best suited for action, movement, and gaze annotations, which are discussed in detail below. A similar action segmentation was selected as in section 2.8.

Action Segmentation:

Motionese

For analysing the data, the action of the "Minihausen" task and the sub-actions (a1-a3) of grasping one block until releasing it onto the end position (Fig. 3.10), were marked in the video. The definitions of section 3.2.2.2 were used for action, sub-action and movement definitions. The same system was also used to identify the hand trajectories were used, also.

Contingency

Eye gaze: In annotating the eye gaze directions with the Interact Software [2], between looking at the interaction partner, looking at the object and looking anywhere else, the same definitions, as in section 3.2.2.2, were used (Fig 3.8).

3.2.3.4 *Metrics*

For quantifying motionese and contingency, five variables, related to the 2D hand trajectories derived from the videos and the eye gaze bout annotations produced with Interact, were computed.
Motionese was measured in terms of velocity and range as defined in [126].
The contingency of the interactions was quantified in terms of variables related to eye gaze, as defined in [14], for measuring interactiveness, by the *total length of eye gaze bouts to interaction partner*, the *total length of eye-gaze bouts to object* and the *total length of eye gaze bouts elsewhere* as in section 3.2.2.3.

VARIable	ACI		AAI		ARI		ACI vs AAI	ACI vs ARI	AAI vs ARI
	M	*SD*	*M*	*SD*	*M*	*SD*	*Z*	*Z*	*Z*
velocity a1	3.58	0.81	4.72	1.39	2.08	0.86	−2.394**	−3.668***	−3.747***
velocity a2	4.19	1.84	6.39	1.71	2.59	0.87	−2.535**	−2.792**	−3.982***
velocity a3	6.62	2.43	11.78	2.95	3.73	1.51	−3.098***	−2.956**	−3.982***
range a1	4.22	2.49	3.41	0.72	6.29	5.53	−0.211	−1.369+	−1.288+
range a2	2.19	0.48	1.88	0.25	2.72	0.97	−1.549+	−1.314+	−2.635**
range a3	1.57	0.37	1.35	0.09	2	0.56	−1.479+	−2.409**	−3.396***
total length eye-gaze to i.p. in	10.86	14.52	6.65	7.15	6.65	7.15	−0.833	−1.419+	−0.76
total length eye-gaze to i.p. a1	27.81	25.02	9.01	16.92	9.25	11.38	−2.2*	−1.882*	−0.97
total length eye-gaze to i.p. p1	24.19	28.17	3.7	9.71	7.35	8.78	−1.853*	−1.03	−1.634+
total length eye-gaze to i.p. a2	15.39	16.67	2.42	4.44	3.16	4.81	−2.054*	−2.066*	−0.244
total length eye-gaze to i.p. p2	33.73	24.63	2.61	7.09	2.69	5.9	−3.055***	−3.306***	−0.082
total length eye-gaze to i.p. a3	23.05	23.09	4.37	8.71	6.2	10.48	−2.273*	−2.292*	−0.384
total length eye-gaze to i.p. su	43.8	23.81	27.55	7.43	19.66	13.65	−0.493	−2.793**	−1.878+
total length eye-gaze to o. in	69.29	29.43	82.32	22.47	62.65	8.7	−1.353+	−1.15	−2.817**
total length eye-gaze to o. a1	70.94	22.72	89.52	16.69	83.21	13.46	−2.1*	−1.213	−1.155
total length eye-gaze to o. p1	60.95	26.97	88.99	23.87	68.36	25.95	−2.273*	−0.714	−2.097*
total length eye-gaze to o. a2	82.68	18.18	96.2	8.19	92.43	7.85	−2.198*	−1.308+	−1.533+
total length eye-gaze to o. p2	65.02	25.55	97.39	7.09	80.23	22.36	−3.055***	−1.503+	−2.092*
total length eye-gaze to o. a3	76.95	23.25	95.63	8.71	87.23	13.77	−2.273*	−1.252	−1.721*
total length eye-gaze to o. su	55.79	22.63	52.71	31.88	57.92	17.94	−0.352	−0.109	−0.527
total length eye-gaze e. in	20.89	29.12	11.03	18.15	34.93	9	−0.624	−1.984*	−3.127***
total length eye-gaze e. a1	1.91	4.75	1.48	4.67	7.53	10.61	−0.52	−1.625+	−1.919*
total length eye-gaze e. p1	16.09	19.93	7.32	23.14	24.29	26.94	−1.501+	−0.812	−1.952*
total length eye-gaze e. a2	2.51	3.9	1.37	4.34	4.41	7.42	−1.178	−0.371	−1.604+
total length eye-gaze e. p2	2.38	5.35	0	0	17.08	20.59	−1.382+	−1.879*	−2.551**
total length eye-gaze e. a3	0.74	1.67	0	0	6.57	12.94	−1.382+	−0.877	−1.803*
total length eye-gaze e. su	1.09	2.31	7.65	11.74	22.42	15.92	−1.091	−3.507***	−2.267*

Table 3.3: Results of Mean, Standard deviation, Mann-Whitney U test, +p <0.1, ∗p <0.05, ∗ ∗ p <0.01, ∗ ∗ ∗p <0.001, interaction partner (i.p.), object (o.)., else (e.). su = sum = summary, in = intro = introduction

3.2.3.5 *Results*

A non-parametric test (Mann-Whitney U test) was performed on all pairs of samples, ACI vs. AAI, ACI vs. ARI, and AAI vs. ARI. Table 3.3 shows the results of the study.

Motionese

For the motionese metrics, the quantitative results can be found in section 3.3. The results revealed the following:
For the *velocity* metric, which is calculated for each sub-action and takes into account the hand movement during the transportation of the respective block, the results showed significant differences in all three sub-actions, for all pairs of conditions. These results clearly showed that, in AAI, hand movements were faster than in ACI and ARI and additionally that hand movement was slowest in the ARI condition. For all conditions, the mean values increased for the consecutive sub-actions: velocity in sub-action a1 < velocity in a2 < velocity in a3. In ARI, the rate, in which the mean values increased, was lowest and, in AAI, the rate was highest. The latter was specially noticeable for the last sub-action a3.
The *range* metric suggests that ARI exhibited the greatest range for each sub-action and therefore movement was most exaggerated. Also, range was greater in ACI than in AAI. The ACI vs. AAI results revealed no significance, but a trend for sub-actions a2 and a3. The ACI vs. ARI, results for sub-action a3 showed significance and, for a1 and a2, they showed a trend. In AAI vs. ARI, sub-actions a2 and a3 revealed significance, whereas a1 again showed a trend. Again, it was found that, in ARI, the first sub-action a1 had the highest range value of all sub-actions over all conditions. Looking at this metric over time, range decreased rapidly to about one half for sub-action a2 and a bit more for the last sub-action a3. For the other conditions however, the rate of change, i.e. the decrease, was not that drastic. For more details, see [65].

Contingency

Most interestingly, the results for eye gaze showed a completely different picture (Table 3.3). For the *total length of eye-gaze bouts to interaction partner*, they showed that, in ACI, significantly more time was spent gazing at the interaction partner than in AAI and ARI. Differences between AAI and ARI were not significant. Looking at this metric over time, it is interesting to notice that in all three conditions, most time of gazing at the interaction partner was spent in the summary part of the action.
For the metric *total length of eye-gaze bouts to object*, values were significantly lower in ACI than in AAI and ARI, where differences between AAI and ARI exhibited that values were significantly lower in ARI.

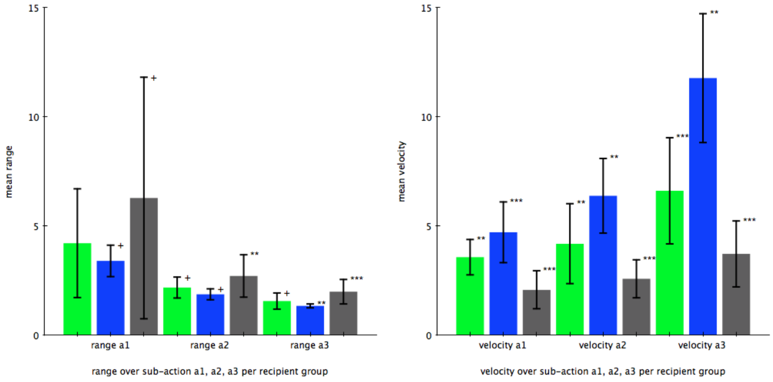


Figure 3.11: This graph shows the range of hand movement in the three different sub-actions on the left. Green represents the hand range of the participant towards the pre lexical children, blue the range of movement towards adults and gray the range of movement towards the robot. On the right, the mean velocity of hand movement in the three different sub-actions can be seen for the "Minihausen"-task. Green represents the hand velocity of the adults towards the pre lexical children, blue the velocity of movement towards adults and gray the velocity of movement towards the robot.

The *total length of eye-gaze bouts elsewhere*, that measures the percentage of time gazed neither to interaction partner nor object, revealed that, most of the time, gazing somewhere else was spent in the ARI condition, followed by ACI. The differences between ACI and AAI could be a result of the design of the study, because the AAI followed the ACI, so that instructions and an experimenter were not anymore needed to help in the demonstration of the task, because it has already been shown once. Additionally, in all conditions, the gaze was elsewhere mostly in p1 and p2 and not during the transportation of the cups in a1, a2 and a3.

3.2.3.6 *Conclusion*

To conclude, ostensive signals were found in tutoring situations in adult-robot interaction. On one hand, the results for range and velocity showed significantly exaggerated hand movements that were clearly distinguishable from those observable in adult-adult interactions and that were even more accentuated than the hand movements in child-directed tutoring. Thus, ostensive stimuli were present in robot tutoring. Those however changed over time as we have seen: range of motion decreased drastically, whereas velocity increased slowly. Therefore, a hypothesis is formulated that the reason for this lies in the behavior of the learner which shapes the behavior of the tutor, as stated for eye gaze behavior and hand movements by Pitsch et al. [92]. This process could be interpreted as an alignment process, where the

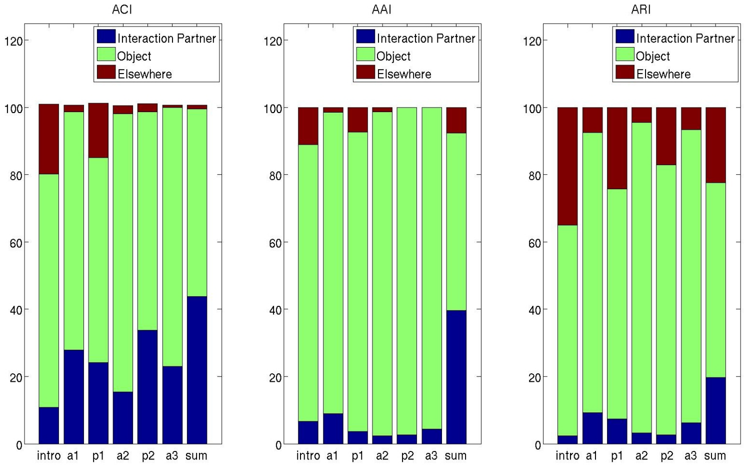


Figure 3.12: This graph shows the total length of eye-gaze bouts to the interaction partner, the object and somewhere else (y-axis) over time: all seven action parts are displayed (x-axis) for ACI (left), AAI (middle) and ARI (right) condition [65].

tutor starts by clearly signaling her/his intention of tutoring the infant. This signal decreases during the ongoing interaction while the tutor captures the infant's attention and while observing an understanding process of the infant. Thus, resulting behavior may be described as consisting of fragmentary cues rather than the complete and exaggerated signal. On the other hand, our results revealed that, in order to create a contingent interaction with the partner, the learner needs to produce suitable feedback. This means that, although the tutor's hand movements in robot-directed tutoring seem to be even slower and less round than in child-directed tutoring, the tutor's eye-gazing behavior in robot-directed tutoring is suggestive of a lack of appropriate social signals on the recipient's side: the percentage of time the interaction partner was observed by the tutor was much lower in ARI than in ACI. The ostensive signals considered here appear practical for the robot in order to detect situations in which it is being tutored, but it is argued that a robot cannot make use of an important ostensive stimulus such as contingency without providing the "right" signals for the interactional construct.

In more detail, it was found that, already from the introduction, the eye-gaze behavior in the ARI situation was rather similar to that of the AAI situation, with less time of the eye-gaze being spent on the interaction partner. This is congruent with previous findings from [126]. If it is hypothesised that eye-gaze is also being used in order to check for understanding in the partner, the eye-gaze behavior, directly after the end of a sub-action, becomes relevant. Indeed, it can be seen that the eye-gaze lengths in both pauses p1 and p2, were significantly longer in ACI as opposed to AAI. Thus, the parents appeared to look for understanding in their infants. Interestingly, the behavior in ARI tended to be similar to the one in AAI – indicating that adults behaved differently towards robots. However, in p1, a trend for the eye-gaze lengths to be significantly longer in ARI, as opposed to AAI, was shown. This might indicate that the subjects were watching out for signs of understanding in the robot as well. Yet, this behavior dramatically changed in p2 where the eye-gaze length was again decreased to the level of AAI, whereas it was even slightly increased in ACI. This may be interpreted as a reaction to missing signals of understanding from the robot. Finally, in the summary part of the action, the overall eye-gaze length towards the robot became significantly shorter than in ACI and AAI.

In order to confirm these results and their interpretation, further analysis of the joint eye-gaze behavior is planned. The hypothesis is that the robot is not able to establish mutual gaze, especially in the pauses, which leads to the increase of eye-gaze towards the robot.

3.2.3.7 *Outlook*

These findings suggest that ostensive signals are present in human-robot tutoring situations and may be used for the robot to learn. However, in order for the robot to elicit a contingent interaction, it needs to provide ostensive signals that indicate its understanding. Based on observations of the infants' behavior, these ostensive signals have to pertain to attention. That is, the robot has to provide eye gaze that signals attention and establishes joint attention as well as shared attention. Another behavior of the infants that was not modeled in the ARI condition was their attempts to reach and grasp the demonstrated objects. Further analysis needs to be carried out in order to reveal the pattern of these reaching gestures. Preliminary investigation of the data suggests that they are far from random but only appear at the end of the demonstrated actions. If this is true, the reaching gestures could be interpreted as a signal that the infant has understood the goal of the action, or at least, the end of the action. Further signals that can be observed from the infants are facial expressions. Again, systematic analysis needs to be carried out, but preliminary results suggest that emotional feedback indicates affective reactions to the objects themselves, but also to the attention grabbing behavior of the tutor and the reaching of the goal.

3.2.4 *Embodiment Corpus*

In this experiment with the iCub, 31 adults (17 females and 14 male) took part (Table 3.4). 14 participants out of that group also belonged to the RDIE group (see above). These 14 participants were divided into two groups, in order to evaluate two different behaviors of the robot.

8 participants interacted with the iCub while it was performing the *NoHead* behavior and 6 participants tutored the iCub while it was performing the *Head* behavior (section 3.1.2 and below).

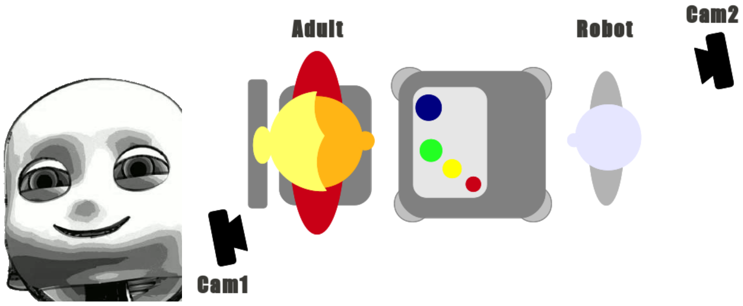


Figure 3.13: The left picture shows the iCub robot. The right one shows the iCub-directed Interaction Setting. There are three cameras recording the scene. The subject is seated across of the robot and the object is put on the table in front of the tutor.

Condition	*Head*	*NoHead*

Participant's age	29 - 63 years;
	median = 31.5 years
		26 - 64 years;
		median = 36 years

Number of participants	6	8
Gender of participants	3 female , 3 male	5 female , 3 male
Number of participants with children	1	4

Estimated age of the robot	0.2 - 5 years;
	median= 2.5 years
		1 -8 years;
		median = 2.3 years

Table 3.4: Participants age, number of participants with children, number of participants, gender of participants and estimated age of the robot by participants which participated in the embodiment study.

The participants were instructed to present the same 6 tasks to the iCub robot, as they did before in the RDIE (section 3.2.1).

In addition, the participants were asked to present several sentences about tasks with the help of toys, to the robot:

- The lion hands the ball to the rabbit. A lion, a rabbit puppet and a ball was handed to the participants, to perform this task.
- The rabbit is rolling the ball to the lion. A lion, a rabbit puppet and a ball was handed to the participants, to perform this task.
- The paper will be folded. A piece of paper is handed to the participants.

- The paper is folded. A piece of paper is handed to the participants.
- A salt stick will be broken. A salt stick.
- A salt stick is broken. A salt stick.
- The car moves to the right box. A toy car and two boxes.
- The hedgehog places the match on top of the box. A puppet hedgehog, a match and a box.
- The hedgehog places the match under the box. A puppet hedgehog, a match and a box.
- The hedgehog oscillates a yo-yo towards the car. A puppet hedgehog, a yo-yo and a toy car.

The robot's behavior was controlled by the same salience system that was used for the Ackachan experiments, but with two different controlled behaviors.

- The first behavior was that only the robot eyes were following the most salient point of the scene (*NoHead*).
- In the second behavior the whole head and the eyes of the robot were following the most salient point of the scene (*Head*).

3.2.5 *Analysis if embodiment's effect on tutoring behavior*

Compared to humans, robots have very different appearances and embodiments. There might be a difference in the acceptance related to the embodiment. To target this question, a study will be presented later on, where a simulated robot (Ackachan) and a physically embodied robot (iCub) will be compared. The difference in embodiment between these two systems, by following the minimal definition of embodiment by Dautenhahn et al. [26], will be quantified. "... a system S is embodied in an environment E if perturbatory channels exist between the two. That is, S is embodied in E if for every time t at which both S and E exist, some subset of E's possible states with respect to S have the capacity to perturb S's state, and some subset of S's possible states with respect to E have the capacity to perturb E's state" [26]. In respect to this definition, there are some not controllable values in the algorithm for measuring the degrees of embodiment (DOM) for the systems, like the environment E. Possibly a value can be given that represents the difference in the degrees of embodiment (DDOM), because of the use of the same environment in all three different conditions of the experiments. However, the three different conditions presented in the two experiments were evaluated with regards to the eye gazing behavior of the human tutor that is teaching the robot some manipulating tasks. In these three different conditions, the gazing behavior of the two robots was designed in a way similar to the one done by Farroni et al. [31] that demonstrated a face, in different conditions, to 4 to 5 months old infants. In their study, they found out that even young infants show a faster saccadic reacting time when there is a shift in the demonstrated face, than if there is a shift in the eye-gazing behavior.

Concerning this difference in perception of humans, three conditions have been taken into account, the Ackachan simulated robot shifting only the eyes (Ackachan condition), the iCub robot also shifting only the eyes (*NoHead* condition) and the iCub robot moving the whole head and the eyes (*Head* condition).

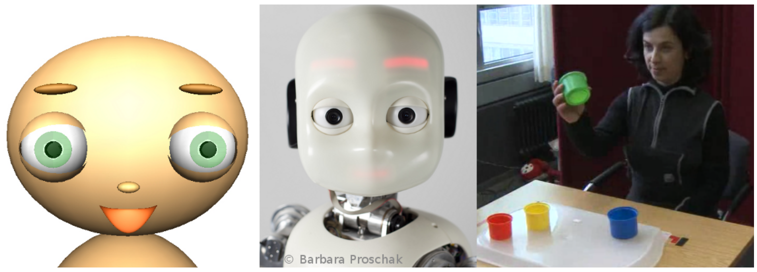


Figure 3.14: The left picture shows the simulated robot. The middle pictures shows the iCub robot. The right picture shows one of the participating tutors.

The dependent variable was the eye gaze behavior of the tutor. If there is a difference in how the tutor is perceiving the robot and how the tutoring behavior shown is different between a physically embodied robot and a simulated robot, a significant differences in the eye gazing behavior of the tutor would be found. If there are differences, as proposed by Farroni et al. and Dautenhahn et al., in the perception of the gazing behavior on the one hand and the DOM on the other hand, significant differences between the *NoHead* and the *Head* condition would be found. "The discovery that another agent's gaze is a cue worthy of monitoring, relies on the infant's ability to detect the contingency structure in interactions with that agent" [32]. The contingency of the interactions was quantified in terms of variables related to eye-gaze, as defined in [14] for measuring interactivity. Brand et al. [14] found that infants received significantly more eye gaze bouts per minute, so that the frequency of eye gaze bouts to the interaction partner was significantly higher in Adult-Child-Interaction (ACI) than in Adult-Adult-Interaction (AAI). The total and average length of eye-gaze bouts to the interaction partner in their study was significantly greater in ACI than in AAI. Equivalent metrics were calculated for the eye gaze on the demonstrated object. The values for frequency of eye gaze bouts to the object, average length of eye-gaze bout to object, and total length of eye gaze bouts to object, as the percentage of time of the action spent gazing at the object, were obtained. Also, three values for the time of the action spent gazing at something else were measured, but none of the participant gazed somewhere else.

3.2.5.1 *Experiments*

Following the idea of the difference in the degree of embodiment (DDOM), the degrees of freedom (DoF) for the two setups are presented here. Also, the major differences for the three conditions are shown (section 3.2.4).

Dautenhahn et al. developed a formula for calculating the DOM :
$\mathrm{DOM}_{S,E} = f(x, y, t)$ [26]
where the DOM of a system *S*, in respect to an environment *E*, is calculated by a function f of the vectors *x* and *y*, and the time *t*. As the environment E is the same, in all of the experiments the differences of the systems are defined by the difference of the robots. These differences are defined by the vectors *x* and *y*, where *x* describes the number of sensors, the detected modalities of the sensors and the channels of information provided by the sensors. This is, as will be shown the same for all three conditions. *Y* describes the DoF of the robot. However, the DoF could give an idea of the DDOM in this special case. To compare the two robots presented above, the number of used DoF's of the iCub platform were reduced to 6 for the *Head* movement condition and 3 for the *NoHead* movement condition. A large number of DoF was not used, allowing for extra capabilities in the future (section 3.2.5).

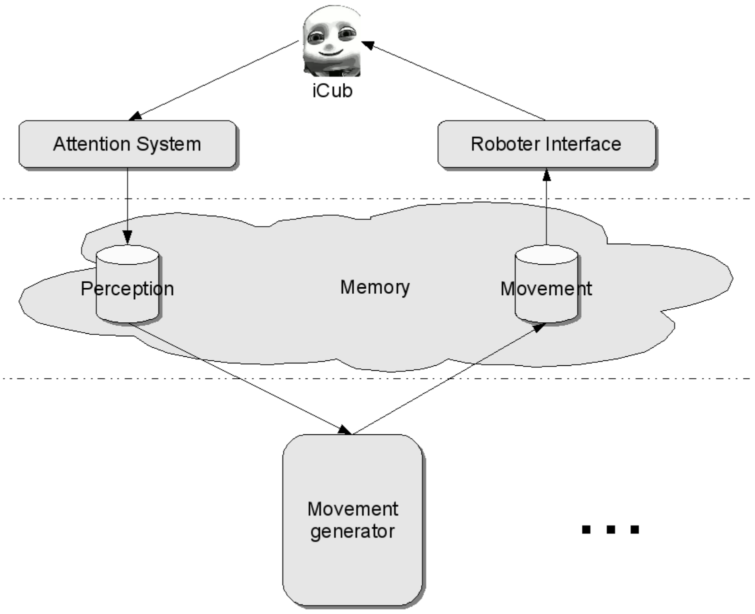


Figure 3.15: The figure shows the software components that were used with the attention system.

The software components, shown in Fig. 3.15, were the *Attention System*, the *Roboter Interface*, the *Memory*, and the *Movement generator*.

- The *Attention System* was the same as the one used for the *Ackachan setting* (see Fig. 3.1).
- The *Roboter Interface* was controlling the connection towards the robot through YARP .
- The *Memory* stored the perception of the visual attention system and the generated movements until they were produced by the robot.
- The *Movement generator* converted the given 2D positions of the salient point to a movement of the robot.

In the two different conditions (*Head* and *NoHead*), only the *Roboter Interface* was connected to different controlling modules.
In the *NoHead* condition, the iCub robot was using only the eyes to follow the salient point in the scene. The eyes were controlled by the same module as in the other condition but the movement of the neck was disabled. In the *Head* attention system, the whole head of the iCub, including eyes and neck, was following the salient point in the scene in this condition. The eyes and neck movement were inspired by the ideas of Lopes et al. [69]. They defined a saccading movement where the head is following the eyes.
Data from a group of 14 participants were collected in both of the two different experiments. The first one was conducted using the simulated robot called Ackachan [126]. After 18 months, the participants were again invited to the experiment with the iCub robot. Both robots were equipped with the same visual attention system. In the Ackachan experiment, there were 31 adults (14 females and 17 male) participants. Out of this group, 14 participants (9 female and 5 male) participated in the iCub experiment.
In the first experiment, the participants were invited to demonstrate several tasks to the simulated robot (Fig. 3.4). For the analysis of this experiment, only the cups stacking task was chosen. The virtual, infant-like robot was equipped with a saliency-based visual attention system. According to the system, the robot's eyes would follow the most salient point in the scene which was computed by color, movement, and other features [82].
In the iCub experiment, 31 adults (17 females and 14 male) took part. 14 participants from this group participated again. These 14 participants were splitted up into two groups to evaluate two different behaviors of the robot. 8 participants interacted with the iCub while it was performing the *NoHead* behavior and 6 participants tutored the iCub while it was performing the *Head* behavior.

3.2.5.2 *Feature extraction and data analysis*

For analysing the data, the actions of the cups stacking task were marked in the video in the same way as in section 3.2.2. Additionally, the sub-actions (a1-a3) of grasping one cup until releasing it into the end position (Fig. 3.6) were marked in the video. The dependent variable "eye gaze" was annotated with Interact in three categories: looking at the interaction partner, looking somewhere else and looking at the object (Fig. 3.8).

The same definition of contingency, as explained before in section 3.2.2.3, was used, along with the same annotation strategies and metrics.

3.2.5.3 *Results*

For all metrics, a *student t-test* was calculated. For equating the Ackachan vs. *Head* and the Ackachan vs. *NoHead*, a paired *student t-test* was used. A paired t-test, could be used because there was an intra subject comparison.

Ackachan vs iCub

Frequency of eye-gaze bouts to interaction partner and *to object* were normally distributed for the paired sample of iCub, with head movement, paired with Ackachan and iCub, without head movement, paired with Ackachan.

The paired sample of iCub, with head movement, paired with Ackachan and iCub, without head movement, paired with Ackachan were normally distributed for the *average length of eye-gaze bout to interaction partner* and the *average length of eye-gaze bout to object*.

Also, the *total length of eye-gaze bout to interaction partner* and the *total length of eye-gaze bout to object* for the paired sample of iCub, with head movement, paired with Ackachan and iCub, without head movement, paired with Ackachan were normally distributed.

The results for the paired sample of iCub, without head movement, paired with Ackachan concerning the *frequency of eye-gaze bouts to interaction partner* (M = 6.35, SD = 1.52) and *to object* (M = 3.18, SD = 1.11), showed no significant differences. This shows that there were no more eye-gaze shifts per minute for the *NoHead* vs. the Ackachan condition.

For the paired sample of iCub, with head movement, paired with Ackachan, the *frequency of eye-gaze bouts to interaction partner* (M = 6.12, SD = 8.12) , no significant differences were found, but, for the *frequency of eye-gaze bouts to object* (M = 5.61, SD = 2.23, t(5) = 2.51, p = 0.054), there were significant differences. The tutors gazed more towards an object in the *Head* condition than in the Ackachan condition.

The difference in the *average length of eye-gaze bout to interaction partner* (M = 3.64, SD = 1.18) and *to object* (M = 8.89, SD = 4.00) was not

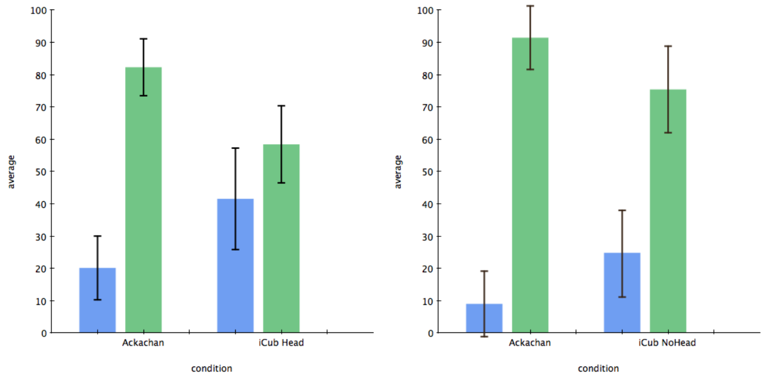


Figure 3.16: The graphs show the total length of eye-gaze bout to interaction partner (blue), to object (green) and to something else (yellow) for the two pairs of paired conditions. On the left, the paired sample of Ackachan with iCub in the, with head movement condition. On the right, the paired sample of Ackachan paired with the iCub without head movement condition.

significant for the paired sample of iCub, without head movement, and Ackachan. The averages of the gazing bout were nearly the same for *NoHead* and the Ackachan condition.

In the paired sample of iCub, with head movement, paired with Ackachan, concerning the *average length of eye-gaze bout to interaction partner* (*M* = 4.84, *SD* = 8.70), there was no significant difference, but for the *average length of eye-gaze bout to object* (*M* = -1.69, *SD* = 1.03, t(5) = -3.998, p = 0.010) there was a highly significant difference. The average length of the gaze bout towards the object was much higher in the Ackachan than in the *Head* condition.

There was no significant difference in scores for the paired sample of iCub without head movement, paired with Ackachan, concerning the *total length of eye-gaze bout to interaction partner* (*M* = 8.4, *SD* = 1.24)(Fig. 3.16).

Testing the *total length of eye-gaze bout to object* for the paired sample of iCub, without head movement, paired with Ackachan, no significant differences were found (*M* = -8.8, *SD* = 1.22)(see 3.16).

However, for the *total length of eye-gaze bout to interaction partner* of iCub with head movement, paired with Ackachan significant differences were found [*M* = 2.27, *SD* = 8.62, t(5) = 2.63, p = 0.046](Fig. 3.16). In total, the tutors looked more to the iCub in the *Head* condition than towards the Ackachan.

Total length of eye-gaze bout to object of iCub, with head movement, paired with Ackachan, significant differences were also found (*M* = -2.41, *SD* = 8.14, t(5) = -2.961, p= 0.031) (Fig. 3.16). In total, the tutors looked less at the object in the *Head* condition than in the Ackachan condition.

iCub *NoHead* vs. iCub *Head*

The *frequency of eye gaze bouts to interaction partner* and *to object*, the *average length of eye gaze bout to interaction partner* and *to object* and the *total length of eye gaze bout to interaction partner* and *to object*, were tested for the unpaired samples of the iCub with head movement and the iCub without head movement condition. They were distributed in a normal manner, as expected.

For the *total length of eye-gaze bout to interaction partner* (*M* = 2.1, *SD* = 1.63, t(12) = -2.212, p = 0.047) and *to object* (*M* = 7.85, *SD* = 1.63, t(12) = 2.214, p = 0.047), no significant differences were found but a trend. In total the tutors looked more to the iCub in the *Head* than in the *NoHead* condition and less towards the object in the *Head* than in the *NoHead* condition.

There were no significant differences shown in any other results.

To summarise, significant differences were found in the frequency, the average length and the total length of the eye-gaze bouts towards the object in the *Head* vs. the Ackachan condition. Also, a significant difference, in the total length of eye-gaze bouts towards the interaction partner, was found in the *Head* vs. Ackachan condition. Concerning the findings in the *Head* condition, the iCub was gazing longer at the object and the shifts of the gazing were higher towards the object in this condition. In the Ackachan condition, it was found that, during the task, the tutor was looking longer towards the object than in the *Head* condition. The data indicate that there was an increase of the acceptance of the robot as an interaction partner in the *Head* condition, because there was more "checking" towards the iCub. For the *NoHead* vs. Ackachan, no significantly differences were found at all. In the comparison between the *NoHead* and the *Head* condition, the same outcome was found for the total length of eye gaze bouts, as in the Ackachan vs. *Head* comparison. The tutors were found to look more towards the iCub in the *Head* condition. It could be argued that the results were influenced by the variability in the behavior of the different participants. However, this was true also for the comparison of the *Head* vs. Ackachan condition, which indicates a increase in the acceptance of the interaction partner in the *Head* condition, compared to the other two conditions.

To conclude, there were 4 DoF for the Ackachan condition, 3 for the *NoHead* and 6 for the Head condition. Also, the results revealed that the Ackachan and *NoHead* conditions were perceived in a very similar way by the tutor.

This fact led to considerations about the differences in the perception of an physical robot like the iCub and about possible cues that guide the tutor's perception of the robot. With respect to the question whether people teach an actual robot in a different manner than a simulated one, the conclusion was made that a simulated robot is taught differently. However, this might be due to the DOM, which

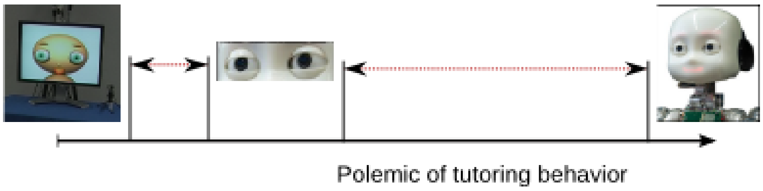


Figure 3.17: The simulated robot is at the one end of the scale and the robot condition, where the whole head of the iCub robot is targeting the salient point, is the other end of the scale.

every robot has. One of the important findings is that the variation in the tutoring behavior is induced by the difference of how the robot is perceived.

3.2.6 *Can state-of-the-art saliency systems model infant gazing behavior in tutoring situations?*

7 state-of-the-art saliency systems were compared and quantified in terms of natural infant tutor interaction in order, to find out why the saliency system that was used before did not work as an adequate gazing strategy, despite being based on an idea that many researchers support. This analysis was also published in [123].
The 7 different systems were compared to the gazing behavior of infants between the age of 8-11 months. The infant's gazing behavior and the interaction structure of the tutoring situation were manually annotated. Based on this annotation, images were selected as input for the saliency systems.

From the Motionese Corpus (section 2.3.1), the gazing behavior of 12 families with 8 to 11 months old children was analysed, as the main feedback and controlling capabilities of those infants were based on the gazing behavior [128]. The cups stacking task was selected, because it had been analysed repeatedly [64],[126] and detailed knowledge of the interaction had been gained.
Several pictures were selected based on an action segmentation (Fig. 2.8). The pictures represented the beginning and ending points of the cups in the cups stacking task. These points were chosen because the scene is changing at them.

3.2.6.1 *State-of-the-art Saliency Systems*

The *Frequency tuned saliency model* [4] is perhaps the simplest of all the existing saliency systems. The absolute difference of each pixel to the image mean is accounted as its saliency value. Achanta et al. [4] recommend to decompose a given input image into L*, a*, b* color space and perform the aforementioned operation on each of the color plane and fuse the results to calculate the final saliency map.

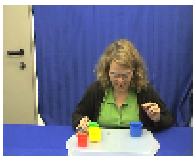
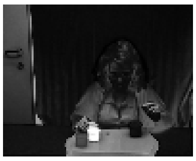
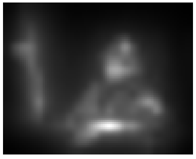
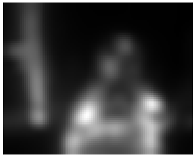
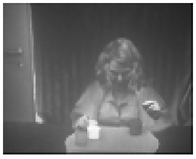
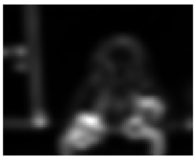
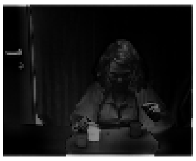
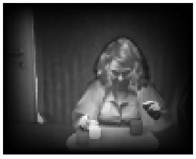
	
Original image	Frequency-tuned saliency 3.2.6.1
	
Graphbased saliency 3.2.6.1	Multiresolution saliency 3.2.6.1
	
Random center-surround 3.2.6.1	Local steering kernel 3.2.6.1
	
Symmetry 3.2.6.1	Random rectangluar regions 3.2.6.1

Table 3.5: An input image and the resultant saliency maps.

Graph-based visual saliency models [45] envisage the input image as a complete graph with each pixel as its nodes. A Weber's [20] law-based dissimilarity metric is employed to calculate the dissimilarity between any two given pixels. A normalising function is further employed to calculate the final saliency map. Experiments carried out in [45] and [124] have shown that the method has high performance for human eye gaze and also for object segmentation.
The *multi resolution saliency map* [50] is perhaps the most cited saliency system to date. The authors propose a computational model for the theoretical framework presented by Koch and Ulman [58] for visual attention. An input image is analysed with opponent color maps, pixels gradient and orientation at different scale spaces. A winner-takes-all (WTA) network fuses these multiple maps into a single saliency map.
Local steering kernel-based saliency [105] is based on the center-surround paradigm, which is employed to calculate the saliency of a selected region. Seo et al. propose a novel local steering kernel which is employed to calculate the similarity between a center region and the surrounding patches. The methodology is inherently robust to image brightness and contrast changes and has very few parameters that are required to be fine-tuned.
Symmetry based saliency [3] is taking rectangular regions of interest

centered on a pixel and computes first order moments of features in order to calculate the saliency of the given pixel. The methodology is among the best in terms of programming simplicity. The authors also make claims regarding the biological plausibility of the model.
Random center-surround pattern based saliency [125] is a methodology that employs a biologically plausible dissimilarity metric which is employed to calculate the contrast between any two random pixels on the input image. The contrasts are updated and normalised to generate the final saliency map. This methodology is shown to have state-of the-art performance in the task of salient region detection.
For the *Random rectangular regions of interest based saliency* [124], the same authors as in [125] propose to compute frequency-tuned salient region detection [4] on random rectangular regions of an image for a large number of times and then sum them to calculate the final saliency map. Such a formulation is shown to have excellent correlation with human eye gaze and has good performance for the task of salient regions, as shown in their experiments. The methodology also has only two parameters that require fine-tuning, reducing the implementation complexity.

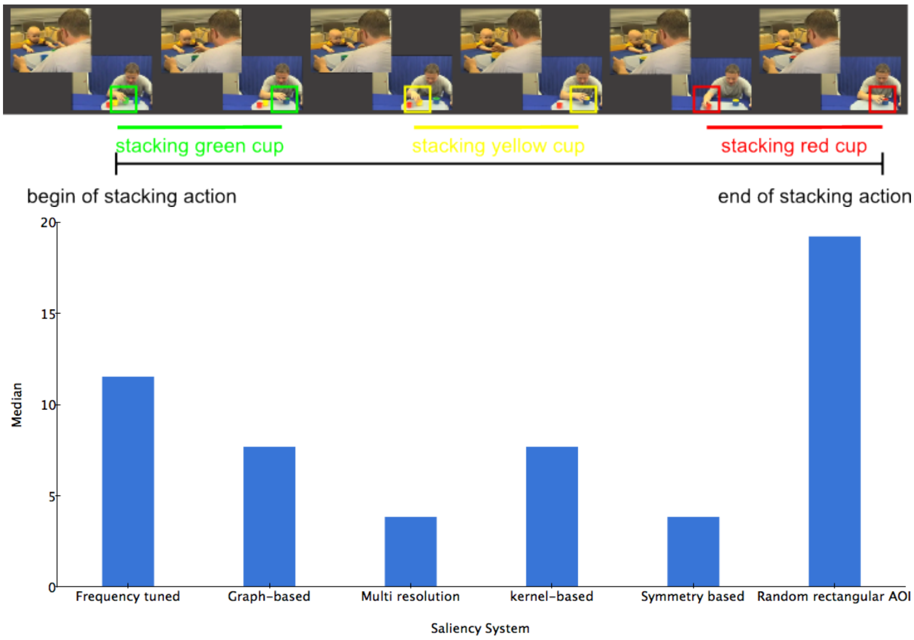


Figure 3.18: The images shows the action segmentation of the cups stacking task. The results of the performance of the different saliency systems, *Frequency tuned saliency model*, *Graph-based visual saliency models*, *The Multi resolution saliency map*, *Local steering kernel-based saliency*, *Symmetry based Saliency*, and *Random rectangular regions of interest based saliency* are presented in the graph.

3.2.6.2 *Data*

The beginning and concluding snapshots of 12 motionese videos were cropped and parsed into the seven aforementioned saliency systems. A 15 x 15 region centered on the maximally salient point was chosen as the focus of attention (Table 3.5).

Results: Results are summarised in Fig. 3.18. The results were all under 20% accuracy. The percentages describe the matching accuracy of a given saliency system to the child's eye gazing behavior. It seems that the methodology of Seo and Milanfar [105] had the best performance in matching the gazing behavior of a 8-11 month old child.
For modeling an adequate system, the initial thought was about an adequate feature set. Then, it was considered to model the contingency, in order to relate these features to each other, as contingency relates to both interaction partners.

3.2.7 *Summary*

In this section, the attention mechanism of a child like robot was studied. As described in section 3.1, a saliency system was used to drive the attention mechanism of two robots.
In the first study the simulated robot Ackachan was interacting with an adult tutor. These interactions were compared with adult-child and adult-adult interactions. For the analysis of these interactions, the motionese features were used in order to compare the hand trajectories of the adult tutor, as well as contingent gazing behavior, in order to measure the differences in the interaction (Fig. 3.19). The results of this study showed that, for the simulated robot, the motionese features are even more exaggerated than compared to a child. But for the contingent gazing behavior, the simulated robot could not even induce the same behavior as towards an adult. These differences in the gazing behavior towards the simulation were studied further in a second analysis of another task from the same corpus. For the motionese features, as well as for the gazing behavior, the findings were the same as before.
In both studies, the results of the gazing behavior were not showing the expected results.
This is why in the next part of the section, the focus was on studying the gazing behavior and, specifically, how a more contingent gazing behavior towards a robot could be induced. Another study was carried out to analyze if the embodiment of a robot could change the gazing behavior of an adult tutor (Fig. 3.20). The results suggested that it is not the embodiment, but the degrees of freedom used to express the focus of attention, that can manipulate the gazing behavior of the adult, towards a more contingent one. Finally, the question was posed, if using a saliency system as an attention mechanism is sufficient. To answer this, a final analysis with 6 different saliency systems was

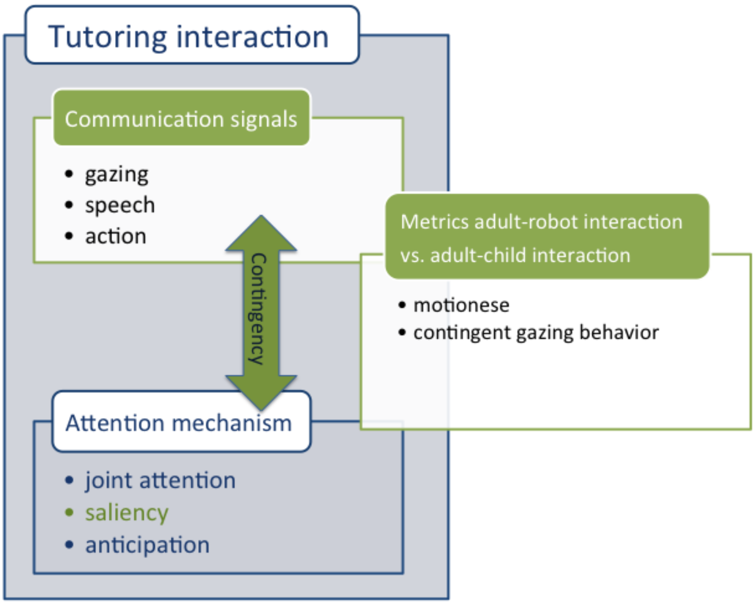


Figure 3.19: The saliency mechanism used as an attention mechanism for a simulated robot.

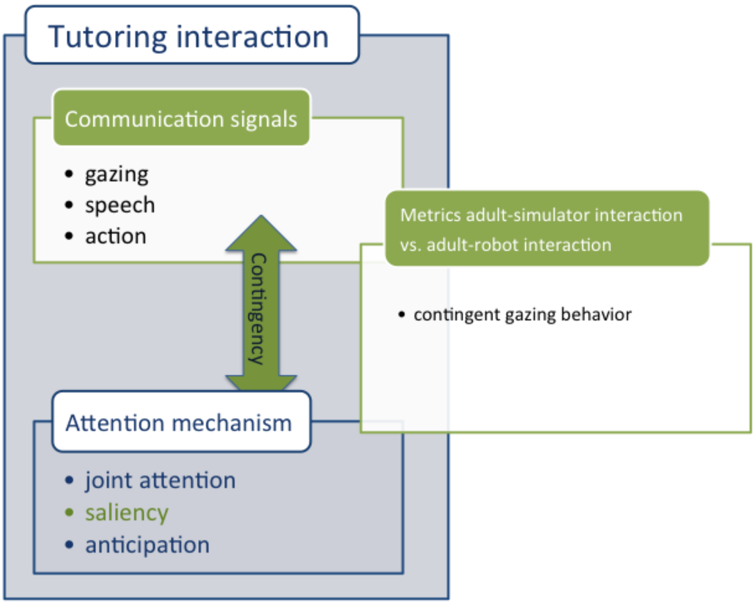


Figure 3.20: The saliency mechanism used as an attention mechanism for a simulated and an embodied robot.

conducted to verify their capability for reproducing child-like gazing behavior (Fig. 3.21).

The results of this study suggested that using only a saliency system as a attention mechanism is not enough to produce child like gazing

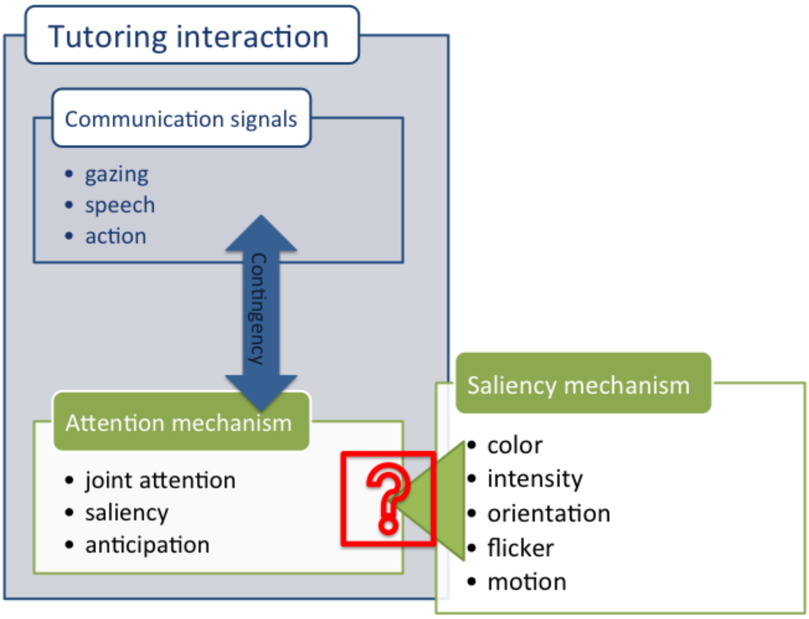


Figure 3.21: Is a saliency mechanism an sufficient mechanism as an attention mechanism for a robot?

behavior.

Overall, the results in this section suggested that, to create a contingent interaction in a tutoring situation between a robot and a human, several aspects have to be taken into account. An effective robot behavior that induces contingent tutoring behavior, but also induces motionese hand trajectories and motherese-like speech in the tutors, needs to take care of a highly responsive attention mechanism and a sophisticated feedback strategy.

4 A NEW MODEL FOR DETECTING A TUTORING SITUATION

Based on the results presented in the section 3.2, the eye gazing strategy between a human and a robot is not similar as the one towards a child. It is more like the behavior towards an adult. But what causes this effect? Is there an interaction loop needed between the interaction partners?
Argyle and Cook [5], in their chapter about measurement of gaze, argued that "the most important aspects of gaze are total amount of gaze, amount of mutual gaze, the timing of glances, pupil dilation, and amount of eye-opening." In their chapter, "Gaze as part of the sequence of interaction", they argue that "one person's gaze would affect another's – either by reinforcement or though imitation: in each case an increase in A's looking should produce an increase of B's." In the same chapter, they refer to a study where one person systematically varied his gaze, and the other person's gaze was measured. The results of this study showed that "people are more likely to match another's gaze pattern than to compensate for too much intimacy. Response matching has been found for a wide range of verbal and non-verbal aspects of interaction..." Following their argumentation, they go on to speak about longer sequences of interaction – social episodes. "Much of social behavior consist of longer interaction sequences in which each interactor knows his part, and there is close coordination of the moves by different interactors. Such sequences appear to be rule-governed, and are similar to a game, in that all must keep to the rules." So what are these rules? And why are they broken in the communication between the system presented before and the human tutor? It seems to be a promising way to look onto the recipients gazing behavior. Considering this behavior, the system needs to be sensitive to the tutor's gaze and has to react accordingly. One way of reacting accordingly could be achieved by imitating the gazing behavior of the tutor [5].

4.1 RELATED WORK: CONTINGENCY IN OTHER WORK

There seems to be a disharmony between the expectations of the current research (considering saliency-based gazing behavior as a sufficient capability) and the behavior of a naive user. Looking back at sections 2.1.2 and 2.2.2, it can be seen that saliency seems to be only a part of the attentive behavior of an infant. The concept of a contingency mechanism, that provides the opportunity to detect and guide the structure behind a tutoring situation, has been found as an additional mechanism. In this section, a few approaches targeting the concept of a contingency detection mechanism, are presented.

4.1.1 *An Infomax controller for real time detection of social contingency*

In an approach proposed by Movellan [80], a real time Infomax controller was implemented in a humanoid robot, in order to detect people using contingency information. They defined the contingency approach as a reactive, not cognitive, continuous "dance" of actions and reactions with the world, rather than a turn-taking inferential process like chess-playing. The concept they followed is related to the concept proposed by Watson [130], that infants use contingency information to define and recognize human beings. Movellan [80] is referring to an experiment done in 1986, where 10-months old infants were tested for using contingency information to detect novel social agents. This experiment was conducted in a way that there were two groups interacting with a robot that did not look quite human. In the first group, the robot was responding to the environment with a behavior that simulated the contingency properties of human beings. In the second group, the robot was using the same temporal distribution of lights, sounds and turns as in the other condition, but the robot was not responsive. From the first group of this experiment, Movellan selected one infant, referred to as Baby-9, as a basis for an analysis of the vocalisation of this participant. An Infomax controller was implemented in order to learn a causal model of detecting a social contingency. The focus was on how to schedule the behavior of the robot's sensor in real time, in order to maximize the information received about the presence or absence of social agents. Based on this implementation, Movellan was trying to simulate the first 43 seconds, in terms of vocalisations, of Baby-9, to show that the system which was developed could be used to simulate and detect contingent behavior. For achieving this, five parameters had to be set: The sampling period for the time discretisation, the self-delay parameters and the agent delay parameters. To get the two latency parameters for the agent, they asked 4 people to interact with an animated character on a computer. These 4 people had an age ranging from 4 to 35 years. An optimal encoder, to binaries the activity of an auditory sensor, was

used. 150 trials, where each trail started with a vocalisation of the animated character and ended after 4 seconds, were recorded and analysed for each participant. A response to the vocalisation from the animated character was found, about 1200 to 1440 ms after the end of the vocalization from the animated character. Following these results, the five parameters, needed for a simulation on the optimal controller, were set to $\Delta t = 800ms$, $\tau_1^s = \tau_2^s = 0$, $\tau_1^a = 1$; $\tau_2^a = 3$. So, the human audio response was expected to occur within 800 to 2400 msec. Finally, the results of the simulated episode of 43 seconds of the optimal controller, were compared with the 43 seconds of Baby-9. The optimal controller produced 6 vocalisations and Baby-9 produced 7 and the average interval between vocalisations was 5.92 seconds for the simulation and 5.833 seconds for Baby-9.
These results refer only to audio signals and they show that, with such an implementation, the question whether there is another responsive agent around can be answered.

4.1.2 *Vision-based contingency detection*

Lee et al. [60] presented a vision based detection of a contingent response by a human. They defined contingency as "a change in an agent's behavior within a specific time window in direct response to a signal from another agent". In terms of Gergely and Watson [42], they were referring to "temporal contingencies". They used a similar approach as shown in section 4.1.1 by Movellan [80]. The only difference was the input signal they were using. Instead of audio signals, a visual input from a stereo camera setup was used. Based on the stereo images, they calculate region-based dense optical flow as developed by Werlberger et al. [131], to create a motion vector map. Then they eliminated the background motion by using the depth information. On the basis of these two steps, they calculated motion segments using an adapted graph-based color image segmentation method, introduced by Felzenszwalb and Huttenlocher [34]. After obtaining these motion segments, they removed segments that had small motion magnitude as well as those that had a large depth. The dimensions of the data were then further reduced with a Principal Component Analysis (PCA) and a Non-negative Matrix Factorization (NMF). In order to model temporal events, distances between groups of consecutive reduced frames were computated called "clips". The distance matrix between these clips was calculated in order to estimate the dissimilarity between behaviors and construct a dissimilarity graph. That way, the system would detect a contingent response to the robot's while the robot was signaling and after the signaling. A study with 43 cases was conducted, 20 contingent and 23 non-contingent ones. 2 gestures for the robot were used, waving or beckoning. Based on these data, a longer amount of delay time was found, about 5000 milliseconds on

average. It was found that contingent and non-contingent behavior could be distinguished on this vision based detection.
By transferring this concept to eye-gazing behavior, the hope is to incorporate the contingency structure of a tutoring situation and detect the user as a tutor.

4.1.3 *Summary*

In this section, the related work concerning the measurement of contingent behavior was presented. An implementation of a vision based contingency detection was discussed.

4.2 MODELING: DERIVING A FEATURE SET FOR A TUTORING SPOTTER

The function of the interactional regularities has been investigated in approaches towards natural pedagogy [24], [25]. Senju and Csibra [104] have shown that children follow social information conveyed by the direction of the eye gaze (i.e., they look where somebody else is looking). This is even more reliable when both, eye-contact and motherese (child-directed speech), is used to address the child to transfer social information. That way, the social information seems to be framed in ostensive cues that also provide a sequential organisation of the information conveyed: the tutor addresses the child and the child sends feedback to her or his attention focus [128], [28], [39]. For robotic research, that is taking its inspiration from developmental approaches, it is essential to penetrate the concrete mechanisms of such reciprocal contribution. So far, the systems are rather reactive, which means that even though they are able to process the input and take advantage of it [111], they provide less feedback and are not able to support the interactional loop [68]. Thus, the motivation for the following contingency system is to take advantage of this social tutoring interaction, so the system can learn within this interaction [98]. To accomplish this, the system shall be equipped with mechanisms that make it sensitive to the signals of the tutor and with a feedback mechanisms that is presenting the attention focus of the system to the tutor. Contingency, as a mean to learn in human-robot interaction, has increasingly received attention in recent robotics research. Movellan [80] has been one of the first who discovered the potential of contingent interaction to achieve and maintain infant's motivation and attention. Although the interaction was reduced to sounds, he showed that the system evoked not only continuing attention from infants, but also that the temporal pattern of signals produced by the system resembled that of infants probing their environment for contingent reactions. The production of contingency has also been used in different applications where contingent robot behavior initiated and maintained interaction [60], [135]. Sumioka and colleagues [110] were able to show that, by making use of cause-and-effect relationships between sensory and motor data, gaze following and alternation could be learned in interaction with a teacher. Importantly, these approaches use quite different operationalisations of contingency. On one hand, single instances of contingent events are used. For example, contingent robot behavior can be designed by simply reacting within a certain time-window to an observed behavior (e.g., [135]) or contingency can be observed by detecting, for example, changes in behavior within a certain time-window [60], [80]. However, in order to make use of contingent cues for learning in interaction, statistical metrics - generally based on entropy [80] - are needed in order to allow for robust learning. Yet, none of these approaches made use of contingency to discriminate between teaching and non-teaching situations and, thus, facilitate interaction and learning at the same time. On the lookout for the concrete mechanisms of a reciprocal interactive contribution, it seems that there is more than just additional social information that the child can take advantage of. It seems that, from very early on, children are biased towards such interactional exchanges. Here, the focus is on two aspects that contribute to such biases: children's preference to look at faces and their preference for contingent actions. These two aspects are considered as crucial and linked to each other and were chosen for subsequent models of tutoring spotter. In the following, these aspects are elaborated and evidence for their significance is provided from the developmental psychology.

4.2.1 *Model of contingency detection*

Watson [130] describes contingency as a relation between a behavior and a subsequent stimulus occurring between two interaction partners serving as a powerful social signal. Thus, the detection of contingency can be viewed as a quantitative metric providing hints about the involvement of the interaction partners and the acceptance of a robot as a social learner [66]. The system is motivated by the scenario, in which the robot is learning about objects, their properties and the actions possible with them. In previous developmental studies [98], it has been shown that such a scenario is linked to particular strategies of the parents, i.e., how they talk about actions and present the objects to their children. In general, it has been shown that, in comparison to an action performed towards another adult, the action performed towards a child is modified [126], [66]: The movements are performed in a tight temporal synchrony with the speech [43] and are shorter, which results in less roundness and more pauses between the individual segments [13], [98]. It seems that young infants learn word-object relations within a tightly coupled interaction between infants' perception, joint attention and specific properties of caregivers' naming [72]. Therefore, for the purpose of the interaction within this scenario, the operationalisation of detection of contingency links children's preference for faces (eye-gaze module) with their preference for a particular temporal pattern of actions (temporal contingency module) and some particular action modifications (looming module). Therefore, the iCub was equipped with additional sensors that allow for analysis of the current interaction with regards to the gazing and looming behavior of the tutor and the robot. Thus, the contingency detection is calculated based on temporal co-occurrence of visually detected ostensive signals of human and robot behavior.

4.2.2 *Modeling Behavior*

In this section, the modeled feedback strategies of the robot are described. The system was overall responsive to the tutors' behavior with a reaction time of 300 ms.
Keller et al. [54] showed that, using a sampling interval of 2 ms and Watson's method of contingency analysis across multiple communication modalities, in a face-to-face interaction, mothers respond to infants with contingencies within short intervals of less than 1 s. These findings correlated with the findings presented by Stern [108] concerning gaze and head orientation, recorded with 16 mm film, 24 frames per second. Van Egeren et al. [122] found that mother and infant contingencies were organised within a 3 s window, using a sampling interval of 1 s and an odds ratio method of contingency analysis. In contrast, Cohn and Beebe [22] found that most mothers and infants responded to each other with contingencies of less than 0.5 s, using a sampling rate of 1/12 s. Thus, a sampling interval for contingencies have been documented within a 0.5-3 s window. Within this time frame (200-334 ms), a human would expect a respond after an ostensive signal . For more details about the implementation see section 4.2.3. The system reacts on the following 4 behaviors of the tutor:

- Reaction Pattern 1 (RP-1): system detects participant-gazes-at-elsewhere and reacts by gazing at random locations
- Reaction Pattern 2 (RP-2): system detects participant-gazes-at-object and reacts by directing its gaze at the object
- Reaction Pattern 3 (RP-3): system detects participant-gazes-at-robot's-face and reacts by directing its gaze to the co-participant
- Reaction Pattern 4 (RP-4): system detects participant-looms-in-the-object and reacts by performing a pointing/looming gesture towards the detected location of the pointing.

4.2.2.1 *Hand trajectories*

The importance of the hand trajectories of the tutor (section 2.2.3) has been showed. So far, the model invented was just reactive toward a specific trajectory class (looming behavior). However, by using the Kinect sensor as a tracking device, it was easy to extend the system by taking more trajectory classes into account.

4.2.2.2 *Object trajectories*

The object trajectories were tracked in order to know if the tutor was looking at the direction of the object. The object was tracked by an ARToolkit marker. The ARToolkit system was returning the 3D coordinates of the marker.

4.2.2.3 *Looming behavior*

According to the findings presented in section 2.3.2, looming behavior of the tutor has been formulated - while holding an object - as a single-handed movement towards the robot and therefore an approach at a certain distance to the robot. In addition, if the human tutor was moving an object by hand towards the robot and reaching the Dmin (Fig. 4.1), the robot was responding by trying to point at the object.

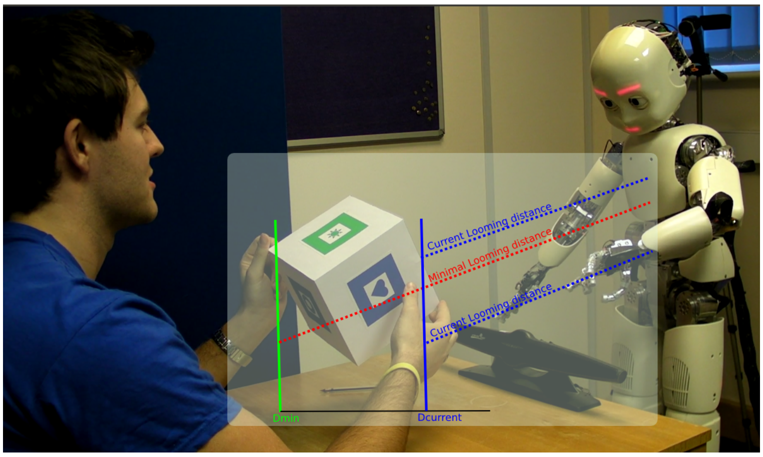


Figure 4.1: Looming behavior: **Dmin** is the minimal distance that must be reached between an object and as hand of the tutor to activate the pointing behavior of the robot. **Dcurrent** represents the current distance between the hand and the object detected by the robot.

4.2.2.4 *Eye gaze*

For the purpose of the interaction within this scenario, the creation of a system for detecting contingency linked children's preference for faces (eye-gaze module) with their preference for a particular temporal pattern of actions (temporal Contingency module) and some particular action modifications (looming module). Therefore, the iCub was equipped with additional sensors that supported the analysis of the interaction with regards to the gazing and looming behavior of the tutor and the robot. Thus, the contingency detection was calculated based on temporal co-occurrence of visually detected ostensive signals of human and robot behavior. The classification of the eye gaze was obtained by geometrical calculations, resulting from locating the intersection point between gazing orientation and the object plane or the face plane of the robot. In other words, the eye gaze module was detecting whether the tutor was looking towards the object (Fig, 4.2b), towards the face of the learner (Fig. 4.2a) or somewhere else (Fig. 4.2c).

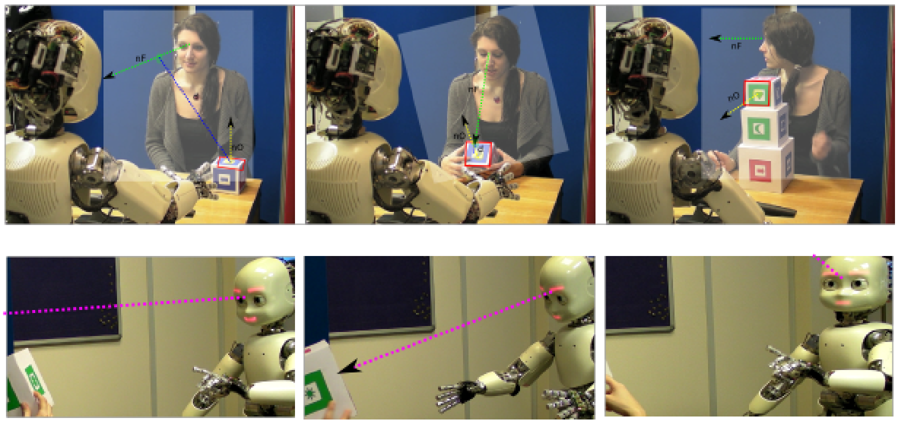


Figure 4.2: a) looking at the interaction partner, b) looking at the object and c) looking somewhere else

4.2.3 *Implementation of the model*

The structure of the robot system is summarised in Fig. 4.3. The iCub robot was connected via YARP [74] with the system, storing and exchanging data.

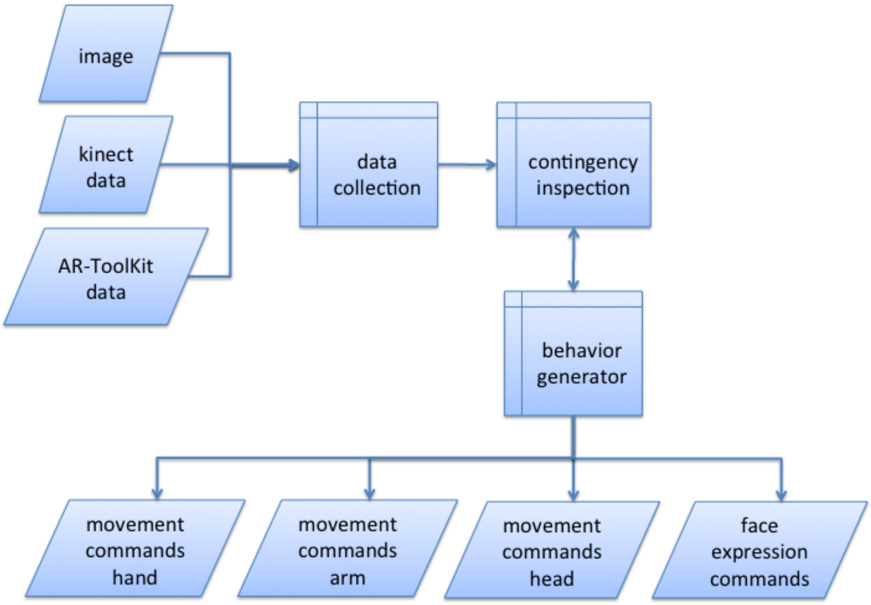


Figure 4.3: The system is structured into three components, the robots behavior, the collected data of the tutors behavior, and the contingency calculation.

The model was implemented in Java.

4.2.3.1 *Modules*

The tutoring spotter system is structured into 3 main modules, the behavior generator module, the data collection module and the contingency observer module. Each of these modules is a container for several object classes, each of them focusing on subtasks of the system.

The behavior generating module

This part of the system is generating the feedback behavior of the robot. The module consists of six classes. These classes are responsible for each part of the robot that is used (face expressions, left arm mover, left hand mover and head mover) and the transformation of the coordinates from the internal representation of the system to the coordinations (and values) needed to control the robot. Finally, a thread class organises the timing of each feedback produced (see Fig. 4.4).

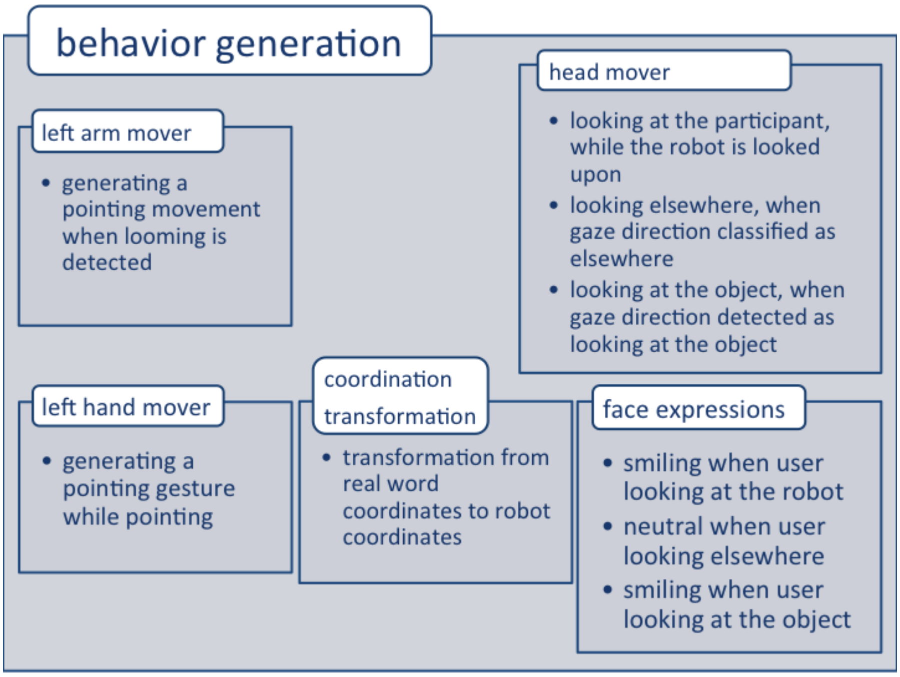

behavior generation
left arm mover
• generating a pointing movement when looming is detected
head mover
• looking at the participant, while the robot is looked upon
• looking elsewhere, when gaze direction classified as elsewhere
• looking at the object, when gaze direction detected as looking at the object
left hand mover
• generating a pointing gesture while pointing
coordination transformation
• transformation from real word coordinates to robot coordinates
face expressions
• smiling when user looking at the robot
• neutral when user looking elsewhere
• smiling when user looking at the object

Figure 4.4: The structure of the behavior generation.

The data collection module

The data collection module consists of six classes. These classes represent the five sensor inputs. The hand data collector is collecting data from the hands of the human tutor by using the skeleton provided by the Kinect sensor. The arm decider is getting the positions of both arms of the human tutor provided by the Kinect sensor and decides if the tutor is producing a looming action with one of the arms. The object data collector is collecting the input data from the ARtoolkit tracking of the markers on the objects and the data collected by the arm calculator to verify if the one, that is presenting the looming action, is the arm closest to one of the objects detected. The face data collector is collecting the output of the faceAPI, a commercial face tracking system. The gaze decider is calculating, based on the data collected by the face data collector and the object data collector, if the human tutor is looking at the robot, at the object or at somewhere else. The result of all these classes is sent to the behavior class. A thread class organises the polling from the sensors and sends the calculated behavior to the contingency module and to the behavior generator module (Fig. 4.5).

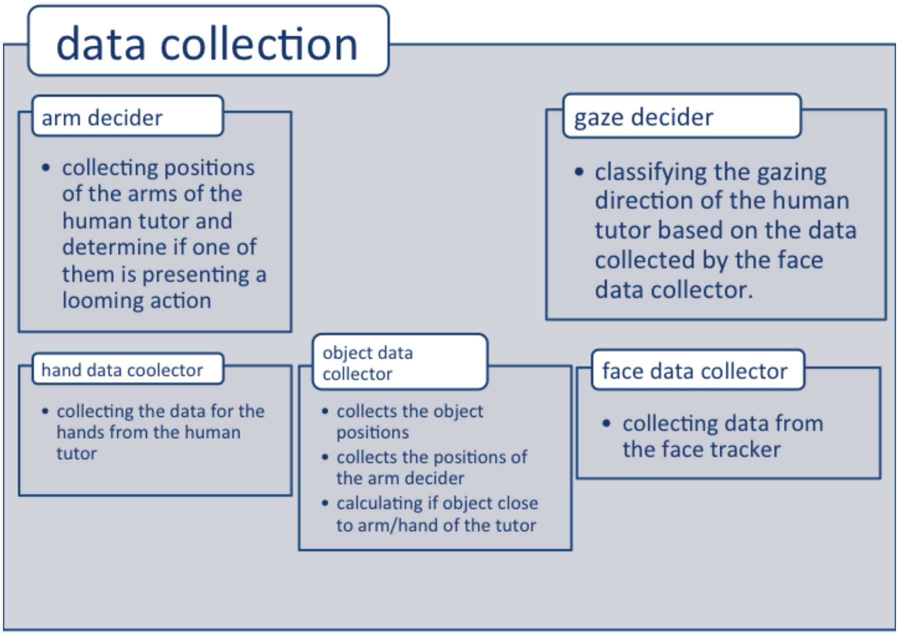

data collection
arm decider
• collecting positions of the arms of the human tutor and determine if one of them is presenting a looming action
gaze decider
• classifying the gazing direction of the human tutor based on the data collected by the face data collector.
hand data coolector
• collecting the data for the hands from the human tutor
object data collector
• collects the object positions
• collects the positions of the arm decider
• calculating if object close to arm/hand of the tutor
face data collector
• collecting data from the face tracker

Figure 4.5: The structure of the data collection module.

The contingency module

The contingency module is notified whether the tutor is gazing at the robot, at the object or somewhere else and whether the tutor is presenting the object to the robot by holding it in a certain distance towards the robot or not. In addition, the ongoing behavior of the robot is captured by the contingency module. For the measurement of contingency, both interaction partners are taken into account (Fig. 4.6). Contingency is measured by the necessity and the sufficiency index (section 5.2.2). According to Watson, the necessity index describes the forward probability of a consequence given a (hypothesised) cause. From the robot's perspective, this refers to the probability that the subject's gaze is focused on a certain object X, given that the robot

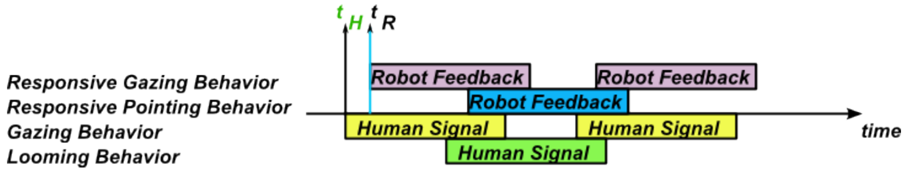


Figure 4.6: The graph presents a sample of the event sequences of robot's and human's behavior, in the responsive behavioral system of the robot. The difference between t_H and t_R is, at most, 300 ms.

had previously been looking at X. The sufficiency index measures if there are also other sources influencing the subject's gazing behavior: given that the subject's gaze is towards X, the probability that the robot had previously been looking at X is calculated. In the interaction, that would imply that the necessity and the sufficiency indices for the subject's behavior are calculated as follows:
Necessity and sufficiency index are non-symmetric. The above description is a computation from the robot's perspective and measures the contingency in the behavior of the subject towards the robot. The overall contingency is then computed as a product of these two variables: Note that the value for the contingency lies between 0% and 100%, where 100% means perfect contingency (like for example in a mirror reflection) and 0% means no contingency at all. The sufficiency in the systems set up is rising if the tutor is looking at the robot or the object and if the tutor is showing looming behavior. The sufficiency is falling if the tutor is looking somewhere else or not showing looming behavior. In the scenario used, the necessity is computed on the robots behavior and represents the responding behavior of the robot: is rising in the case that the robot is looking at the tutor or the object and when it is pointing at an object. It is falling in the case of the robot looking somewhere else or not showing pointing behavior. The whole calculation is event driven [27]. The classification is done based on the detected behavior of the tutor, with classified behavior being measured as one event (Fig. 4.7).

Each classification is, within the system, rated as a time dependent event and, to calculate the contingency, the following calculation takes place:

$$Contingency = S(t) * N(t)$$

where S is the sufficiency, which gives a information with regards to the question, if the human tutor is looking or looming towards the robot, because the robot did that before.

$$S = \frac{\sum_t (PER)}{\sum_t (ER)}$$

where the nominator $\sum_t(PER)$ is the sum of positive behaviors towards the robot, the number of gazes towards the robot's face or the

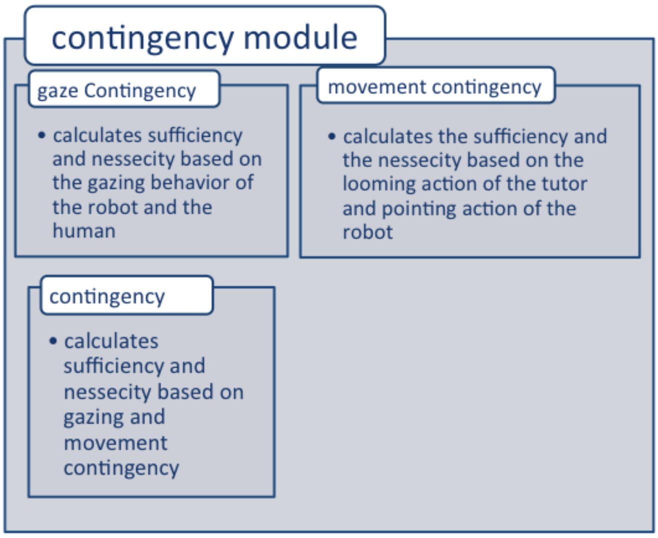

contingency module
gaze Contingency
• calculates sufficiency and nessecity based on the gazing behavior of the robot and the human
movement contingency
• calculates the sufficiency and the nessecity based on the looming action of the tutor and pointing action of the robot
contingency
• calculates sufficiency and nessecity based on gazing and movement contingency

Figure 4.7: The structure of the contingency module.

objects and the number of looming behaviors towards the robot, over time and the denominator $\sum_t(ER)$ is the number of all events towards the robot, over time.

The necessity is giving information with regards to the question if the robot is looking towards the human and if it is pointing towards the object, because the tutor did that before.

$$N = \frac{\sum_t(PEH)}{\sum_t(EH)}$$

Where the nominator $\sum_t(PER)$ is the sum of positive behaviors towards the tutor, the number of gazes towards the tutor's face and the objects and the number of pointing behaviors, over time and the denominator $\sum_t(EH)$ is the number all events towards the human tutor, over time.

An event is created by each processed image (if it is a positive event PER+1 or PEH+1).

In the contingency corpus, there were two different behavior strategies of the robot, presented to two groups of participants. One group of participants was interacting with a robot that was randomly presenting one of the three gaze classes and randomly presenting a pointing or non-pointing behavior. Both behaviors were decoupled. In the other group that was interacting with the robot, the robot behaved after a predefined imitation behavior, the tutoring spotter behavior.

For the implementation of this definition of contingency, it was chosen to take both interaction partners into account. As the system itself was event driven, the timeline had a frame rate of 25 fps, thus the contingency metrics had also to be adopted, in order to be calculated on a discreet scale. So, the definition was following this approach. Overall, in every frame there was the possibility for two events for the robot and two events for the human. One event for the gazing class of the robot, one for the gazing class of the human, one for the status of the arm/hand robot (pointing or no pointing) and one for the human having a status of looming or not looming. All events were collected over time, from the begin of the experiment till the end.

4.2.4 *Summary*

In this section, the implementation and design of a robotic system that is based on the knowledge gained from the studies, was presented. This system was more likely to sustain a contingent interaction with a human tutor. The model presented was focusing on a highly responsive feedback mechanism, that controlled the gazing behavior, as well as a pointing behavior of the iCub robot (Fig. 4.8).

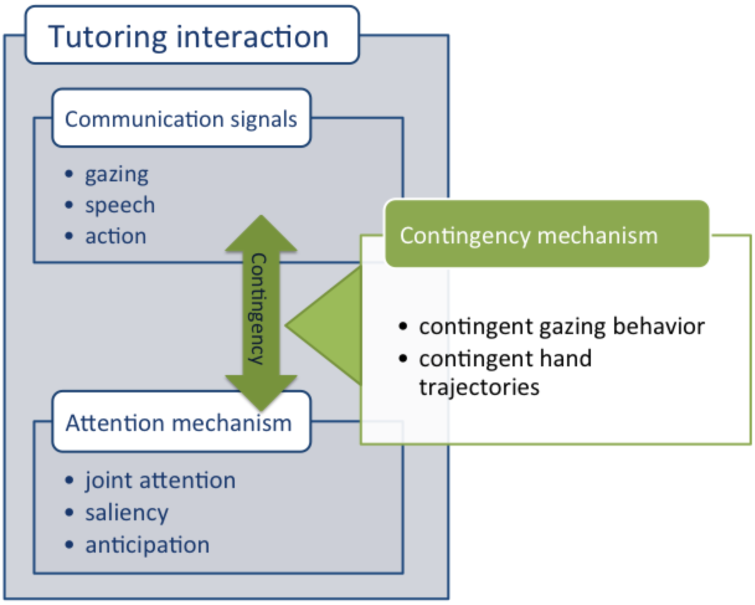


Figure 4.8: Implementing a tutoring spotter module that induce contingent tutoring behavior and is highly responsive.

In the previous chapters, inducing a contingent gazing behavior was presented as being more difficult, compared with the motions features. But as the tutoring scenario, which is targeted here, is highly restrictive, regarding the robot's gazing behavior, a joint attention mechanism could be implemented, that is supported by a contingent response on the tutors gazing behavior. In the next section, the results of a study with naive users tutoring this new model will be presented
.

4.3 ANALYSIS: THE TUTORING SPOTTER IN A INTERACTION WITH A HUMAN TUTOR

4.3.1 *Studying the tutoring spotter*

These experiments were done in cooperation with the University of Hertfordshire [63], [67]. In them, the topic of the interaction between a human tutor and the iCub robot was changed to 3 different sized cubes with colored markers, on every side. There were two tasks for the tutor to present to the robot. In the firt one, the tutors were asked to present the different cubes and to explain the colors and shapes of the markers.

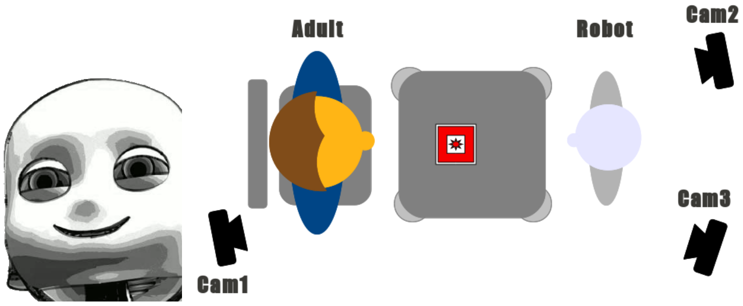


Figure 4.9: In the left figure is the iCub's face, on the right one is an overview of the setting.

Condition	Contingency behavior	Random behavior

Participants age	21-34 years;
	median= 24 years
		21-69 years;
		median = 25 years

Number of participants	12	13
Gender of participants	9m; 3f	8m; 5f
Number of participants with children	1	4

Estimated age of the robot	1 - 8 years;
	median= 5.5 years
		1 -12 years;
		median = 4 years

Table 4.1: Participants age, number of participants with children, number of participants, gender of participants and estimated age of the robot assumed by the participants that took part in the University of Hertfordshire and Bielefeld University study.

For the second task, the participants were asked to show how to stack these different sized cubes onto each other. 25 participants took part in the study. They were divided into two groups. Each group saw a different behavior of the robot. The participants interacted with the robot twice, with a break of 7 days in between. The participants got to know the iCub by the name DeeChee. All participants were native English speakers with the age ranging from 21 to 69 years (Table 4.1). Most of the participants were students or administrative staff at the University of Hertfordshire. The participants were instructed as follows:

> *Your task today is to teach something new to DeeChee. Today and on subsequent days, you will be asked to play with DeeChee. In subsequent sessions with DeeChee, DeeChee may or may not make verbal responses.*
>
> - *The DeeChee is equipped with a set of sensors, so that it is connected to our world.*
> - *You have a number of coloured boxes with patterns on them in a basket next to you.*
>
>
>
> *Your job is to play with DeeChee. You are welcome to talk to DeeChee, to use gestures, and you should show the patterns and the colours of the boxes to the robot.*
> *There will be two short tasks for you: I will give you the instruction for the first task now, the task will take 2 minutes. Then I will come back and give you the second task.*
>
> - *Your first task : Please present the pattern and the colours of the boxes to DeeChee. In doing this, please make sure to indeed use all the boxes (Fig. 4.10).*
> - *The second task : Please teach DeeChee how to stack these different boxes. Please use a different colour and a different pattern for each box (Fig. 4.10).*

The participants were seated in front of a table looking towards the robot (Fig. 4.9). The experimenter was sitting in the room in order to take care of the robot. Three cameras recorded the scene. Participants had the possibility to use three differently sized boxes covered with ARToolkit markers.

4.3.2 *Random behavior*

For the random behavior setup an implementation, based on the tracking of the objects and the face of the participant was used to control the iCub's behavior. The robot was capable of looking at the participant's face, the object or somewhere else and it was using pointing or non-pointing gestures.

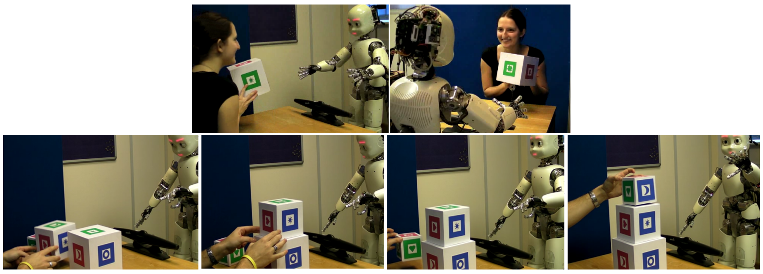


Figure 4.10: In the first 2 pictures of the participants of the University of Hertfordshire and Bielefeld University study presenting an object to the iCub robot (Task 1).In the other pictures, the stacking task, presented to the iCub robot, by the participant, is shown (Task 2).

4.3.2.1 *Gazing feedback:*

The random behavior was saturating a 'boredom' filter if the same face or object was seen for too long and the robot was switching to random gazing. This meant that the robot was changing the gazing behavior based on a timing.

4.3.2.2 *Pointing feedback*

In the random behavior condition, the robot was tracking the object and was occasionally (on a random basis) pointing at the object. The robot was making no use of the contingent feedback of the tutor in this condition.

4.3.3 *Analysing the implemented contingency behavior*

To evaluate the implemented contingency behavior (proposed in section 4.2.1), the method of Ethnomethodological Conversational Analysis (EM/CA) [67], [103], [116], [91] was used. This is a qualitative method that was used as a first step instead of a quantitative analysis, because the approach used here was based on the results of the previously gained knowledge. The qualitative approach promises a highly detailed and inductive result. It helps to, not only answer the question whether the system is scaling the behavior adaptation of the tutor, but also how the system is influencing the behavior of the tutor. The CA is targeting questions like, how is the participant organizing its interaction locally or how does the tutor engage in the tutoring action (a) when the robot reacts appropriately or (b) when it does not react appropriately. The CA is a micro-analysis that investigates how robots and tutors are responding to each other.
The results were gained based on the behavior of Reaction Patterns (RP) of the implemented contingency behavior. Robots detect (Fig. 4.2 and 4.1):

- participant's gazes at elsewhere (RP1)
- participant gazes at object (RP2)
- participant gazes at robot's face (RP3)
- participant showing looming behavior (RP4).

4.3.3.1 *Data and Analysis*

Timeline-based data (video, audio, logging of robot's perception and robot's internal states) were obtained, as well as questionnaires filled out by the participants after their interaction with the iCub in the University of Hertfordshire (section 4.3.1). The timeline based data were combined using the annotation tool ELAN [16]. A manual annotation was performed. Based on these combined primary and secondary data, an Ethnomethodological Conversation Analysis (CA) [92], [103] was performed. The CA is linking the system level and the user's perspective of an interactional frame and enables to close the loop between technical implementation and user studies [61]. It was hypothesised that the contingent interaction will elicit more tutoring behavior from the tutor, resulting in the iCub being perceived as more human-like.
In order to access the system's performance, a two-step approach was taken. In the first step, qualitative analysis was performed for two cases of interaction: In the first case, the tutor spotter system was working well in an interaction with a tutor, while in the other case, the tutor spotter system was not working appropriately and the robot was not able to spot the tutor. For these two cases, *sequential analysis* was used, allowing for micro-analytical insights into the sequential structure of the interaction.
The *sequential analysis* method is used for investigating the close interrelationship between robot's and tutor's actions. It helps answering the question of how they respond to each other, in terms of structural features of the interaction. With this approach, the participant's view could be reconstructed by looking into the user's perception and understanding of the robot's actions.
In the second step of the approach, analysis of the questionnaires was performed, using data from all 12 participants.

4.3.3.2 *Results: Participant's Engagement*

The participant's engagement was evaluated based on the system's behavior functioning in the concrete interaction with two different participants: In one case the system was able to engage in a responsive, contingent interaction with the tutor (VP004) while, in the other one, a contingent interaction does not appear (VP007). Therefore the performance of the system and the effects of a contingent vs. a non-contingent robot behavior on the tutor's engagement and presentation of a task, was studied. For VP004 and VP007, the first 20 seconds of the interaction, between participant and the robot, were analysed and their implications, for the tutor's engagement in the trend of the interaction [92], [87], compared. The analysis was based on the four different RPs. The results of the CA were published in [67] and [63]. Only a short example is presented here, based on the results of the RP3 in the following.

System performs contingent behavior (VP004)

The results of the CA for this participant were dealing with the explanatory part of the interaction (16.2-19.9 seconds, Fig. 4.11) with the RP3.
In this sequence, the participant was describing the object. She redirected her gaze to the object and rotated it as to bring it in a position that allowed both participants - robot and herself - to look at a particular side (the green cross) and then explained: "so THIS is (-) GREE:N," and pointed to the cube's green area. At the end of this utterance, she gazed at the robot (#18.3) and thereby transferred this information to the robot. In structural terms, she created a slot where, in human-human interaction, a recipient's acknowledgement was expected [9]. Indeed, the robot reacted by lifting its head, gazing and smiling at the tutor (#19.3). This conduct was triggered by the contingency module using RP-3: While rotating the cube, the tutor briefly gazed at the robot's face, which the system detected correctly (#16.387, #16.515, #17.215) and launched the gaze-reciprocating behavior. Shortly after this (#18.576, #18.860), the system also detected "participant-gazing-at-object", so its eyeballs started to move quickly between the tutor's face and the object. The tutor reacted to this conduct by waiting for about 0.9 seconds, then adding the deictic "HERE," accompanied by a new pointing gesture to the cube. Thus, she interpreted the robot's reaction as appropriate in terms of its timing and the type of action produced. At the same time, she interpreted its eye movements as a searching activity to which she was providing help for the system to focus better on the relevant location. In this interactional micro-coordination, the tutor treated the system as being responsive on a very fine-grained level, orienting to its conduct as sequentially appropriate. She also assumed that the system was able to react on her additional support.

The transcript of the first 14 to 40 seconds interaction of the participant VP004 with the iCub robot can be found in Transcript 1.

The CA results showed some implications for participant's further engagement.
The analysis of the first 25 seconds revealed that the contingency module enabled the robot to engage in an interaction with the human tutor, in which, not only all four implemented contingency patterns work as assumed, but - more importantly - the participant accepted the robot's conduct as appropriate and responsive:

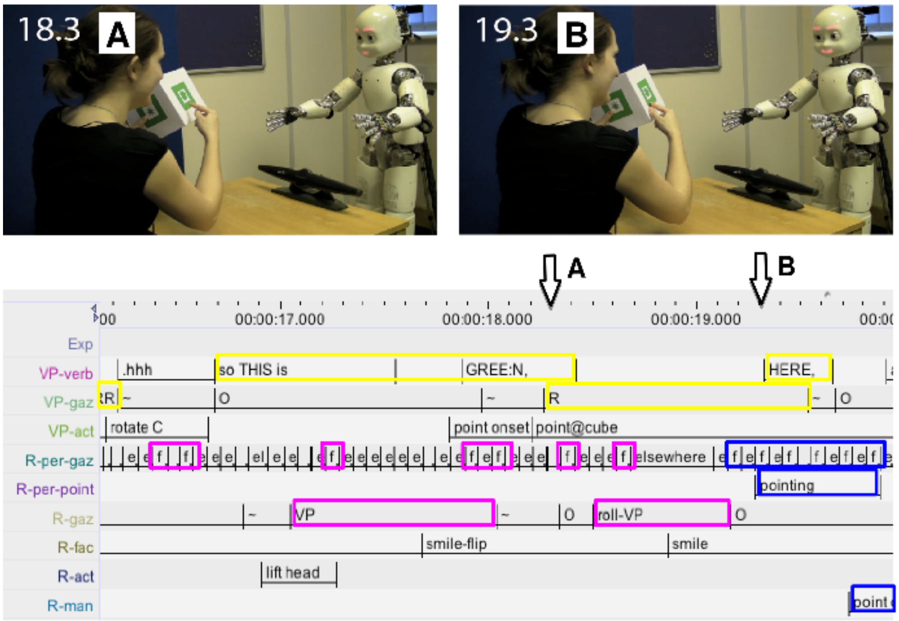


Figure 4.11: VP004 - Transcript of interaction (seconds 16.0-20.0)

- She explicitly acknowledged the robot's responsive behavior (laughter).
- She attributed the capability of "seeing" to the system.
- She realized a form of presentation that was closely oriented towards the robot: she planned her utterances in a way that they projected occasions for the robot to produce recipient feedback. That way, she attributed to the system the ability of being responsive.

This had implications for the pursuit of the interaction (as shown in the transcript regarding her verbal actions, Transcript 1): Having experienced the robot as a reactive system, the participant continued to present the task in a way that was highly oriented towards the system's actual conduct and displayed states and capabilities: she used short sentences, with a simple repetitive syntactic structure ("and" + subject-verb-object (S-V-O)), final rising pitch contour and pauses (ranging between 0.3 and 1.7 seconds) that allowed for the robot's reactions. This way, her presentation was oriented towards the robot and enabled the system to contribute with a responsive conduct at the same time[1].

System performs non-contingent behavior (VP007)
In this case, as can be seen in Fig. 4.12, there was nearly no detection of the RP3. The transcription of the first 5.5 to 34 seconds of the

[1] The system evaluation using the ideas and methods of Ethnomethodological Conversational Analysis, carried out by Karola Pitsch and was published in ([67], [63]).

Transcript 1 VP004 (14.0-40.0)

```
Hallo,
(0.4)
(laughs)
(0.2)
.hhh so THIS is (0.3) GREEN,
(0.9)
HERE,
(0.3)
and you ca:n (.) SEE the (0.2) CROSS in the MIDDLE,
(1.7)
YES,
(0.6)
on this side it it’s green, (.) ALSO,
(1.7)
and (.) you can see it’s (.) a: SUN,
(1.0)
and eh that’s within a WHITE BO:X,
(0.9)
and (.) then a GREEN BO:X,
(0.8)
a SQUARE,
```

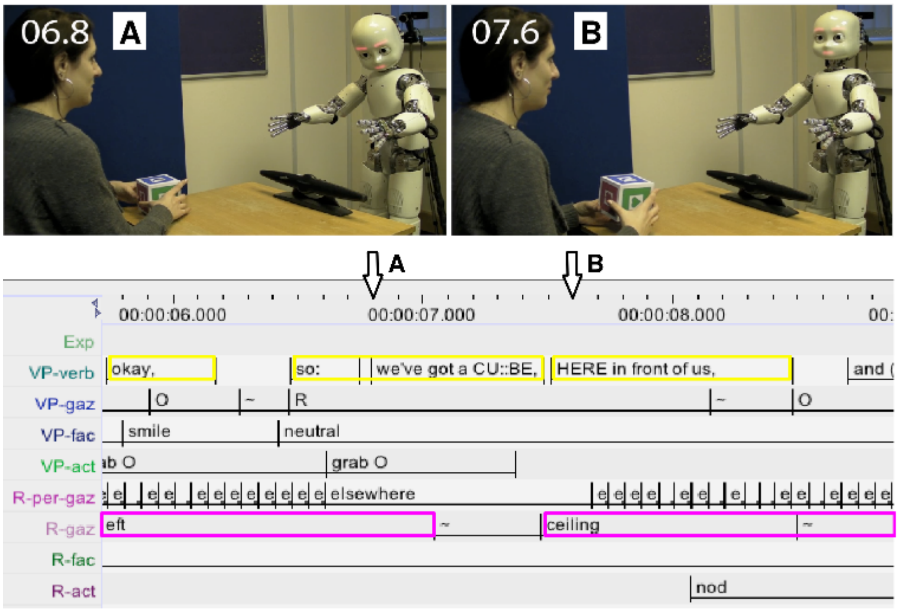


Figure 4.12: VP007 - Transcript of interaction (seconds 05.5-09.1)

interaction, between the participant VP004 and the iCub robot, can be found in Transcript 2.

The CA results showed some implications for the participant’s further engagement.
The suite of the interaction was characterised by a repetitive re-

Transcript 2 VP007 (05.5-34.0)

```
okay,
(0.2)
so (.) we’ve got a CUBE here in front of us,
(0.2)
and (.) it’s got six SIDES,
(0.5)
and always the same- same dimensions,
(0.3)
and (.) so first we look at the top of the CUBE,
(0.6)
so:: (.) this is a (.) a SQUARE, pasted onto the cube,
which is BLUE,
(0.2)
in the outer SQUARE,
(0.5)
we then have a SMALLER square in the MIDDLE, (.) which
is WHITE
(0.5)
and we have a CONTOUR shape of a MOON,
(0.6)
which shows the blue BACKground (.) through (.) so
that’s the TOPside of the CUBE,
```

occurrence of this pattern of the robot’s withdrawal of gaze at those moments, where the tutor addressed her presentation to the robot and an acknowledging recipient feedback would relevantly be in place. Thus, the robot did not engage in an appropriate responsive interaction with the tutor. In comparison to VP004, the tutor adopted a different attitude towards the system during her tutoring behavior:
(i) She visually oriented more and more to the object, while gazing less at the robot. This means that the robot had lost the opportunity to participate as ’co-participant’ in the action presentation [92].
(ii) In her presentation, she used complex syntactical constructions (S-V-O + relative clause ("which")), with fewer and shorter pauses than the tutor in VP004 (ranging from 0.2 to 0.6 seconds). This not only made it more difficult for the system to understand the tutor and to discriminate actions, but it also limited the opportunities for the robot to give feedback.

4.3.4 *Questionnaire*

In order to verify the qualitative findings for the group of participants, the questionnaires given to the participants, after interacting with the iCub, were analysed. In these questionnaires, the participants were asked about their impression of the robot and the interaction with

it. On a scale from 1 to 5, participants ranked, for example, how independent they could see the robot. In analysing the questionnaire data, a Spearman correlation of their answers, with the success of the designed contingent feedback that the robot gave during the first 30 seconds, were calculated. Note that on by the first 30 seconds targeted because, our qualitative analysis's results yielded that they were crucial for the user's impression.
Significant results were obtained for the correlation between the contingent feedback within the first 30 seconds and the answers of the users about their perception of the robot, suggesting that when the robot's behavior was contingent, it appeared more human-like to the tutors than when its behavior was less contingent. The questionnaire and the results can be found in the table 4.2.

These findings suggest that there is a relation between perception of the robot on one hand, and the contingency metric and resulting contingent behavior of the robot, on the other hand.

4.3.5 *Summary*

In this section, a CA of the tutoring spotter system was presented (Fig. 4.13). The results of the analysis suggest that the tutoring interaction can benefit from the contingent and highly responsive behavior of the system implemented.

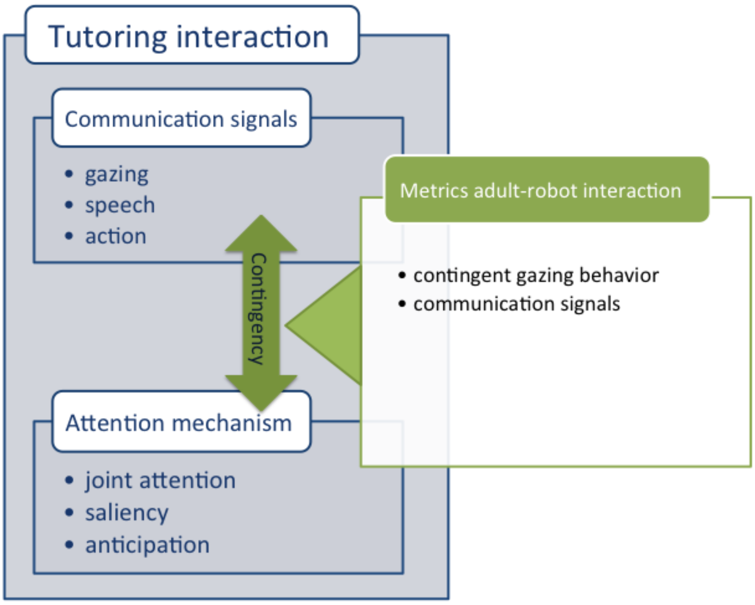


Figure 4.13: Conversation analysis of the tutoring spotter system.

Question	Mean value	Correlation with 30 seconds of iCub interaction in the contingent behavior	
Age of the participant	25.14	Correlation coefficient Sig. (2-tailed) N	-0.49 0.26 7
Did you like to interact with DeeChee?	3.29	Correlation coefficient Sig. (2-tailed) N	-0.46 0.30 7
Did you like DeeChee?	3.57	Correlation coefficient Sig. (2-tailed) N	-0.61 0.14 7
Was DeeChee lovely?	3.43	Correlation coefficient Sig. (2-tailed) N	-0.67 0.10 7
How old is DeeChee?	4.57	Correlation coefficient Sig. (2-tailed) N	0.73 0.06 7
Do you think that DeeChee was acting independently?	3.43	Correlation coefficient Sig. (2-tailed) N	0.00* 0.02 7
Do you think that DeeChee was behaving humanly?	3.00	Correlation coefficient Sig. (2-tailed) N	0.00* 0.05 7
Are you familiar with virtual communication partners (computers, roboters, ECAs)?	1.43	Correlation coefficient Sig. (2-tailed) N	0.61 0.14 7

Table 4.2: Participants age, number of participants with children, number of participants, gender of participants and estimated age of the robot assumed by the participants which participated in the Robot-Directed Interaction Experiment.

4.4 DISCUSSION AND FURTHER DIRECTIONS

As it was shown, the system was flexible and it was easy to add more features, to either enhance the contingency or to design feedback strategies, based on the detection of contingent behavior the robot. But at this, stage there was also the need to make the robot's detection behavior more stable. Thus, there was a refinement for the next study. There were also some possibilities for adding, in the future, processing extensions to the feature set. One of the next steps could be to add a reaction on the user's speech to the system, as speech can be a very powerful ostensive cue [25], [104] that even young children respond to. Thus, a keyword spotting system (section 5.3.2), considering what children are reacting to [128], would be an appropriate supplement to this model, making the robot even more sensitive to the ostensive state of the tutor. As to the feedback design, in research on feedback behavior [128], there is evidence for the need of anticipation in the gaze (section 5.3.1) of the robot (e.g. gazing at the target object). The robot could signal what it knows about the action by looking, for example, at the temporal position of an action. A future step could be to modify the scenario presented before, in order for the tutor to get feedback regarding the learning of the actions by the robot.

The initial results allowed an elaboration on the development of the Tutoring Spotter system, which seemed to be a promising module with an extendable feature set, for facilitating interaction within a tutoring scenario.

5 REVISING THE TUTOR SPOTTER

In the previous chapter, a first implementation of the tutoring spotter was described. As was shown in the analysis, the system had some remaining issues that needed to be addressed. In this chapter, the results of the contingency corpus (section 5.1) study are presented, where the refined tutoring spotter system was tested in an interaction with a number of human tutors. The evaluation will focus on the eye gaze analysis, but also the results for the usability of the system and the speech level of the contingency corpus.

5.1 CONTINGENCY CORPUS

A study to compare the differences in tutoring behavior of human tutors towards the same robotic platform, the iCub, within two different behavior constraints, was conducted, in order to evaluate the previous results. In one condition, the iCub robot was controlled by a contingent reaction pattern and, in the other condition, the robot's behavior was random.

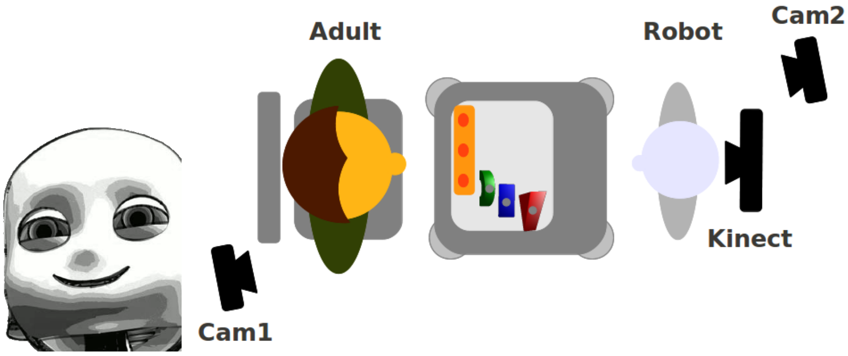


Figure 5.1: In the left figure, the iCub's face. In the right figure, an overview of the setting.

In the study, the participants (Table 5.1) were asked to perform a task divided into 3 different parts. In the first part, they were asked to perform the presented motion tasks (section 3.2.1). In the second part, they were asked to present to the robot several sentences about a task with the help of toys. In the third part, 3 videos were shown to the participants and they were asked to tell the story of the clip to the robot. For the second part, the sentences used were:

- The lion hands the ball to the rabbit. A lion, a rabbit puppet and a ball was handed to the participants, to perform this task.

Condition	Contingency behavior	Random behavior

Participants age	19-68 years;
	median= 24 years
		20-55 years;
		median = 25.5 years

Number of participants	19	19
Gender of participants	7 male; 12 female	9 male; 10 female
Number of participants with children	3	3

Estimated age of the robot	1 - 10 years;
	median= 5 years
		0.3 -12 years;
		median = 4 years

Table 5.1: Participants' age, number of participants with children, number of participants, gender of participants, and estimated age of the robot, for participants the participated in the contingency study.

- The rabbit is rolling the ball to the lion. A lion, a rabbit puppet and a ball was handed to the participants, to perform this task.
- The paper will be cracked. A piece of paper is handed to the participant.
- The paper is cracked. A piece of paper is handed to the participant.
- A salt stick will be broken. A salt stick.
- A salt stick is broken. A salt stick.
- The car moves to the right box. A toy car and two boxes.
- The hedgehog places the match on top of the box. A puppet hedgehog, a match and a box.
- The hedgehog places the match under the box. A puppet hedgehog, a match and a box.
- The hedgehog oscillates a yo-yo towards the car. A puppet hedgehog, a yo-yo and a toy car.

For the third part, the following instruction was given:

- Please retell the following three videos to the iCub robot.

Between each part, the participants were asked to stand up, in order to recalibrate the kinect body tracking system. In the first and second parts, the robotic system with the contingency condition was reacting on the objects and the users gazing behavior. A new gaze detection system was used for the contingency condition, as the system used before (that was based on FaceAPI) had very poor performance (3.4%). For more details about the system, see sections 4.2 and 5.1.1.

The random behavior of the robot was similar to the one presented in section 4.3.2. The difference was that there was no face detection, but the position of the participants' heads was found by using the kinect sensor. Thus, this random behavior created based on the timing, either of gazing of the robot towards the participant, the object or randomly arround and randomly a pointing gesture towards the object.

5.1.1 *Gaze direction detection*

The gaze direction detection system was based on a face detection model called ENCARA [19], that uses an active appearance model (AA model). This model is searching for faces in the image and, if a face is found, it tries to fit a pre-trained statistical model of a face to it. That model gives the positions of the eyes, the mouth, the center of the face and the edges of a bounding box around the face. Out of these feature points, the system calculates three different classes of gazing behavior, which are defined as "object", "interaction partner" and "elsewhere" (Fig. 5.2).

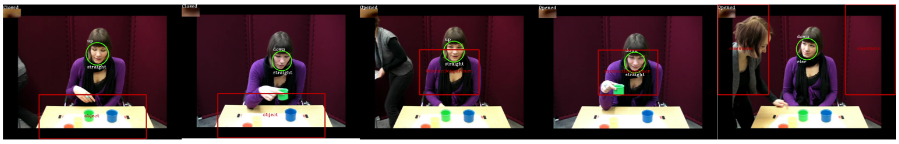


Figure 5.2: One of the participants explaining the stacking cups task to the iCub robot. The classification of the gaze direction detection can be seen in the pictures.

In more detail, the gaze direction detection system was consisting of 6 modules, the eye crop module, the eyelid tracker module, the head movement tracker, the gaze estimator, YARP data import and YARP data export module.

The eye crop module crops the input image to a smaller image that contains only the right eye, by using the data provided by the ENCARA system. In this cropped image, the colors are compared to skin color.

The eyelid tracker is designed for deciding whether an eye is opened or closed. It uses the output of the eye crop module. This module scans the area in this cropped image, that is not consisted of skin color (the eye ball) and creates a line from the most right to the most left point. By doing so, it creates two subregions in the image. Finally, the skin color in these subregions is compared, in order to estimate if the eye is closed or open.

The head movement tracker module is designed to decide whether an eye is opened or closed. This module is taking the results of the eyelid tracker and from ENCARA, in order to calculate where the face is looking: down, up, straight down or straight to the camera. This is

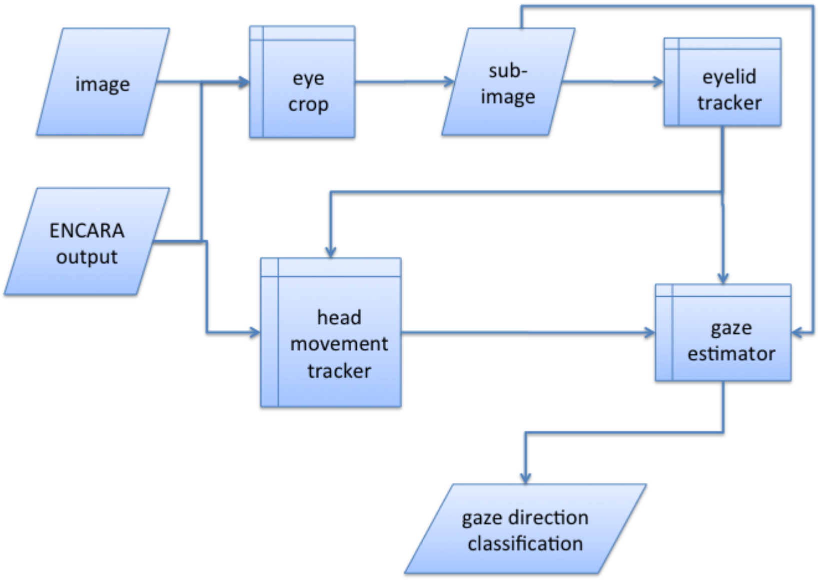


Figure 5.3: Diagram of the software components that were used as plugins for the iceWing [62] framework.

done by taking the orientation of the head into account.
The gaze estimator takes all the information from the head movement tracker, eyelid tracker and eye crop module and estimates a gaze direction. This module classifies the output of the other modules into looking at the object, at the interaction partner or looking somewhere else.
The last two modules are dealing with the data input/export aspects of the tutoring spotter system, by feeding the images to the gaze detector (YARP data import module) and exporting the result of the gaze detector (YARP data export module) to the tutoring spotter system.
The design of the gaze direction detection system supports the use of different modules separately. That gives the ability to also have an insight in the influence of the eye movement, as a feature for the classification of the gazing direction. The system can be seen in Fig. 5.3.

5.1.2 *Summary*

In this section, the contingency corpus was introduced. This corpus was using a new and more precise gaze direction detection system. This system had support for the tutoring spotter system and enhanced the contingent gazing behavior of the iCub (Fig. 5.4).

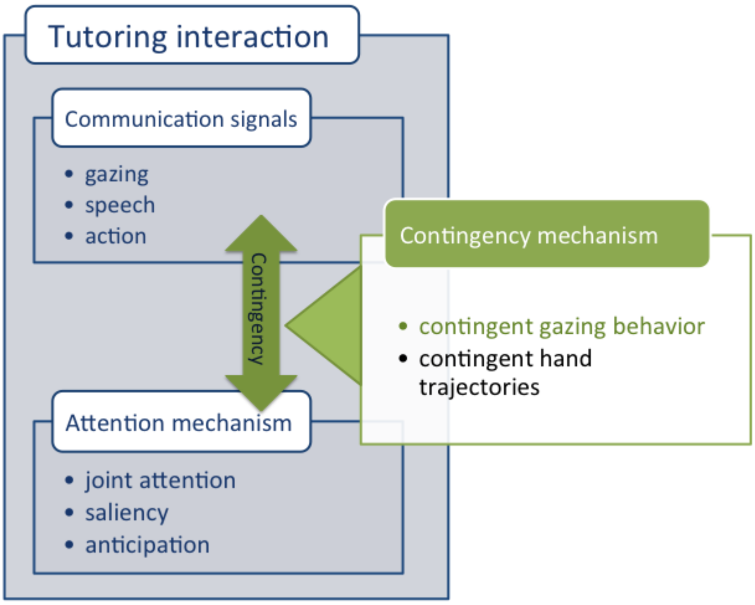


Figure 5.4: Improvements in the tutoring spotter system.

The refinement of the tutoring spotter system was concentrated on the stabilisation of the gaze classification and also, the system in general.

5.2 EVALUATION: THE TUTORING SPOTTER 2.0

In this section, the results of the evaluation of the contingency corpus (section 5.1), using the tutoring spotter system version 2.0, are presented. For this study, the tutoring spotter system was adapted in a way that there were not any markers on the objects, but on the tray where the objects, that were presented to the participants, were. Also, for the face tracking module, the model presented in section 5.1.1 was used, instead of the commercial FaceAPI tracker.

5.2.1 *Eye gaze Results*

The analysis of the eye gaze data of the tutors' gazing behaviour was done offline. The classification of the data was done online (by capturing using the Kinect). There was also a manual annotation created, as in Chapters 2 and 3. The quality of the automatic detection system was calculated by comparing the automatic detection with the manual annotation. Afterwards the following eye gaze metrics were calculated:

- *Frequency of eye-gaze bouts to the interaction partner/object/elsewhere*, i.e., eye gaze bouts per second.
- *Average length of eye-gaze bout to the interaction partner/object/elsewhere*, average length of gaze bouts towards the interaction partner/object/elsewhere.
- *Total length of eye-gaze bouts to interaction partner/object/elsewhere*, as percentage of time of the action spent gazing at the interaction partner.

For the manual annotation only the cups stacking task was taken into account. The tasks were automatically separated by the automatic detection of the ARtoolKit markers, which were captured during the experiment. Each task was marked with a different symbol.

5.2.1.1 *Manual annotation vs. online classification*

To get an idea of how accurate the automatic eye gaze detection system and classification of gazing behavior, as well as the looming detection, were in this study, a manual annotation of the gaze was done. Then, the manual classification results were compared with the results of the automatic classification. It was found that the accuracy of the detection of the gazing class *looking at the interaction partner* had a mean value across all participants of 24.81%. The accuracy of the detection of the gazing class *looking at the object* was 65.28% accurate and the classification of when the participants were *looking towards something else* was 6,78% accurate (Fig. 5.5). For the looming detection, a mean value, across all participants of 9.16%, was found. During the annotation, 12 participants were excluded , because of the way the system was performing and because of participants performing the task incorrect. The new gaze detection system had a performance of 32,2% of successful detections, an almost ten-fold increase over the previous system that was based on FaceAPI (3.4%). Despite this, it was decided that this accuracy was not enough, thus the manual annotation was used for the following evaluation.

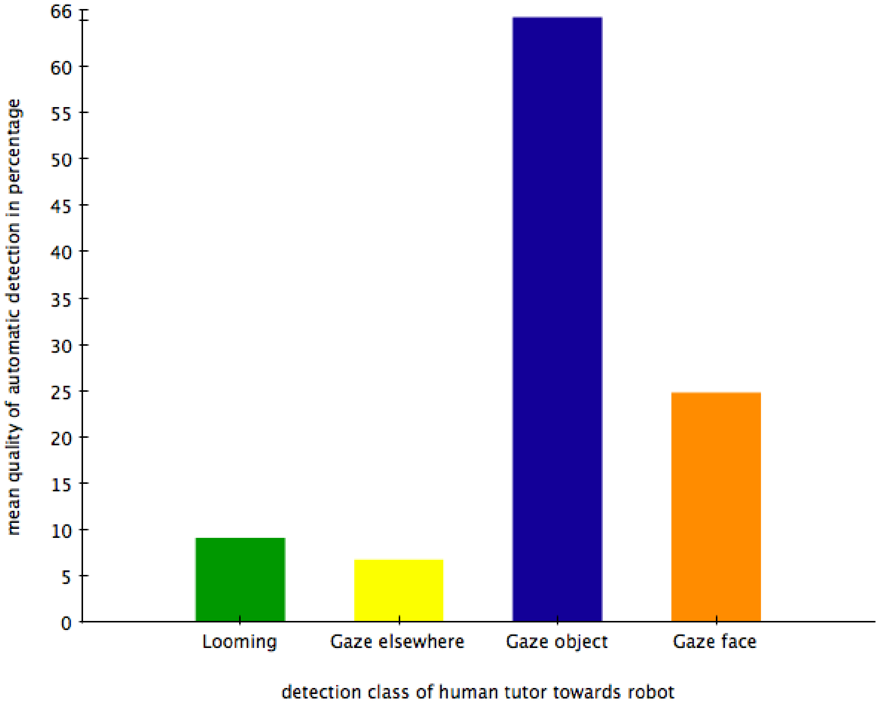


Figure 5.5: The graph shows the percentage for correct detections of the gazing behavior for the combined results of the tutors presenting the cups stacking task to the iCub Robot. The automatic detection was compared to a manual annotation.

5.2.1.2 *Results of the cup stacking task*

An one-way ANOVA was performed on the manually annotated eye gaze data.

The *average length of eye-gaze bout to the interaction partner* showed no significant difference between the two conditions (F(21,4)=0.99, p=0.32).

The *average length of eye-gaze bout to the object* showed no significant difference between the two conditions (F(1,24)=1.66, p=0.21).

The *average length of eye-gaze bout elsewhere* showed no significant difference between the two conditions (F(1,24)=0.83, p=0.83).

The *frequenzy of eye-gaze bout to the interaction partner* showed no sig-

nificant difference between the two conditions (F(1,24)=0.34, p=0.57). The *frequenzy of eye-gaze bout to the object* showed no significant difference between the two conditions (F(1,24)=1.8, p=0.19).
The *frequenzy of eye-gaze bout elsewhere* showed no significant difference between the two conditions (F(1,24)=0.15, p=0.71).
The *total length of eye-gaze bout to the interaction partner* showed a significant difference between the two conditions (F(1,24)=0.25, p=0.62). The *total length of eye-gaze bout to the object* showed no significant difference, but a trend, between the two conditions (F(1,24)=0.25, p=0.62). The *total length of eye-gaze bout elsewhere* showed no significant difference between the two conditions (F(1,24)=0.05, p=0.83).
There were no significant differences found between the conditions in the gazing behavior. The results of the mean values of the average length of eye gaze bout can be seen in Fig. 5.6.

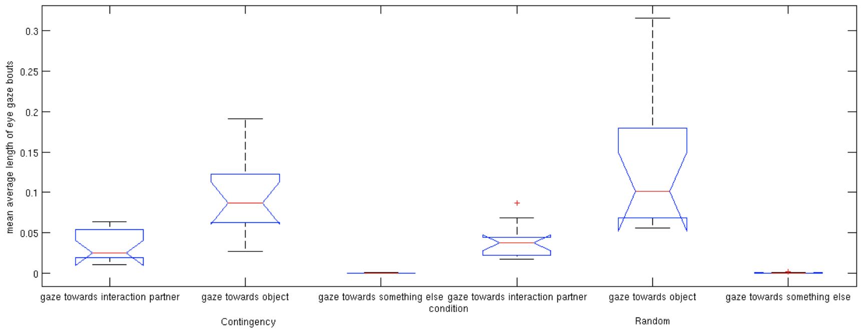


Figure 5.6: The graph shows the mean values of the average length of eye gaze bouts of the tutor.

5.2.2 *Contingency Results*

As described in Chapter 4, the system detects three different eye-gazing classes and whether the human is presenting a looming behavior or not and calculates the contingency based on these classifications. Based on the manual annotations performed on the cups stacking task, the contingency values over this task were re-calculated in the same way as in section 4.2.3.1.

5.2.2.1 *Results of the stacking cups task*

There was an one-way ANOVA performed on the contingency values for the cups stacking task (Fig. 5.7). There was a significant difference found between contingency and the random condition (F(1,18)= 10.29, p=0.005).

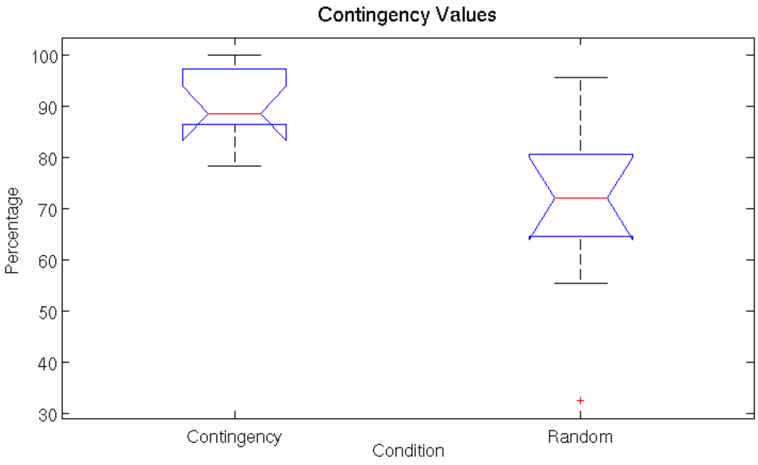


Figure 5.7: The results of the contingency mean values for the cups stacking task can be seen here.

These results show that the contingency was much higher in the contingent interaction than in the random one.

5.2.3 *Speech results for the cups stacking task*

The same annotations as before (sections 2.10 and 2.3.2), were done, in order to analyse the speech, based on the transcription of the cups stacking task. The goal was to investigate whether there is a difference in the use of attention getters, naming of the robot, and the use of *MANNER* and *PATH* descriptions of the task. For the attention getters and naming behavior in the two different conditions, the following results were found.

Condition	Number of utterances	Attention Getter	Naming	Other
Contingency	54	~1.82%	~12.96%	~85.19%
Random	77	~10.39%	~2.60%	~87.01%

In the contingency condition, there was significantly more naming of the robot than attention getting. This result was inverted for the random condition (M=0,10, sd=0,298: contingency condition), (M=0,02, sd=0,148: random condition; t (contingency naming vs. random naming)= 2,086; p=0,039).

The results for the attention getters were significantly different (M=0,01, sd=0,118: contingency condition), (M=0.09, sd=0.286: random condition); t (contingency attention getter vs. random attention getter)=-2,086; p=0,039).

The results for the other words were not significantly different (M=0.64, sd=0.484: contingency condition), (M=0.74, sd=0.439: random condition).

For more details, see Fig. 5.8.

An interpretation of this result could be that the participants recognised less attention shown by the robot in the random condition, at it was intended.

Condition	Number of utterances	*MANNER*	*PATH*	Other
Contingency	54	~18.52%	~40.74%	~40.74%
Random	77	~19.48%	~45.45%	~35.06%

In the description of the task, in terms of the *MANNER* and *PATH* constructions used by the participants, there was no big difference, which is also following the idea that only over different tasks (more *MANNER*-oriented or more *PATH*-oriented tasks) there is a big difference.

The results for the more *MANNER*-oriented utterances were not significantly different (M= 0.14, sd=0.348: contingency condition), (M=0.17, sd=0.375: random condition).

The results for the more *PATH*-oriented utterances were not signifi-

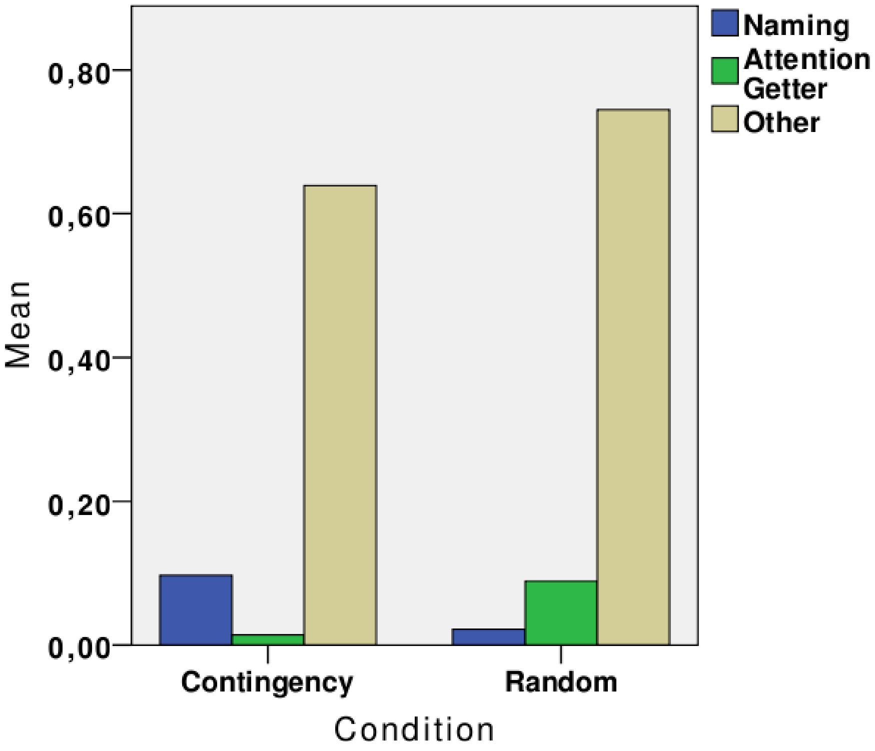


Figure 5.8: Mean values of occurrence of utterances including naming, attention getting or other meaning,s in the two conditions.

cantly different (M=0.31, sd=0.464: contingency condition), (M=0.39, sd=0.490: random condition).

5.2.4 *Questionnaires Results*

The results of the questionnaires that were given to the participants after the interaction with the iCub robot, are presented in this section. The questionnaires comprised of three parts (appendix A.2.1). The first 11 questions were open questions, followed by 9 questions about the personal data of the participants and ending with 10 questions regarding the system usability scale (SUS, [15]).

5.2.4.1 *Results of SUS*

The SUS questionnaire, a questionnaire proposed by John Brooke [15], is used to rate the usability of a system and is using a simple, ten-item attitude Likert scale, giving a global view of subjective assessments of usability. It is mostly used in usability engineering of electronic office systems. The results, as shown in Fig. 5.9, are not significantly different between the two conditions (contingency condition: M=60.47, sd=17.824; random condition: M=62.74, sd=14.383).

The results were marginal in both cases, in terms of the scale proposed by Bangor et al.[7] (Fig. 5.10). The contingency condition was bordering on high-marginal.

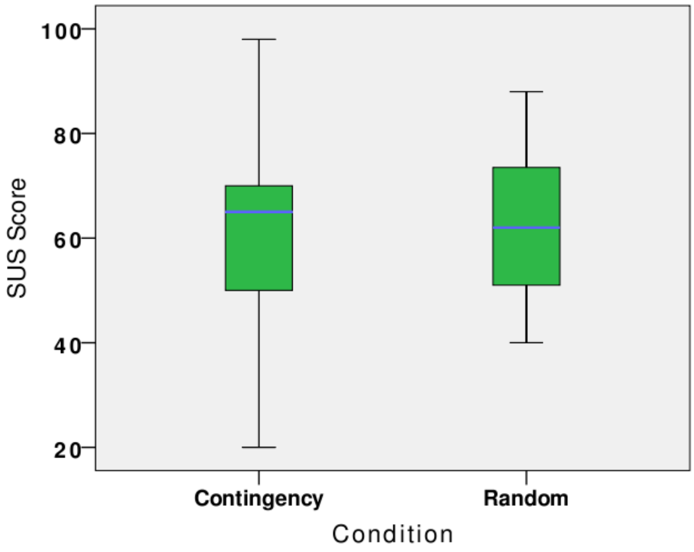


Figure 5.9: Blue bars are the median values, green bars are the ranges and black bars are the percentiles.

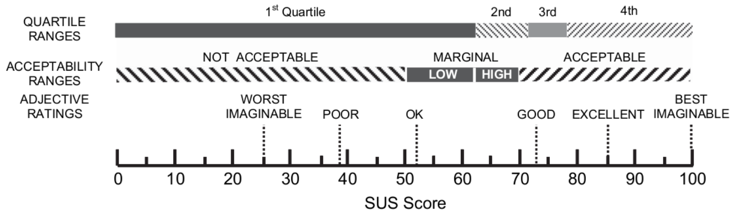


Figure 5.10: A comparison of mean SUS scores by quartile, adjective ratings, and the acceptability of the overall SUS score [7].

5.2.4.2 *Results-open questions*

The mean results of the open question part of our questionnaire can be seen in Fig. 5.11. For the question *Did you like the interaction with iCub?*, no significant difference was found between the two conditions (contingency condition: M=3.63, sd=1.116; random condition: M=4.16, sd=1.015).

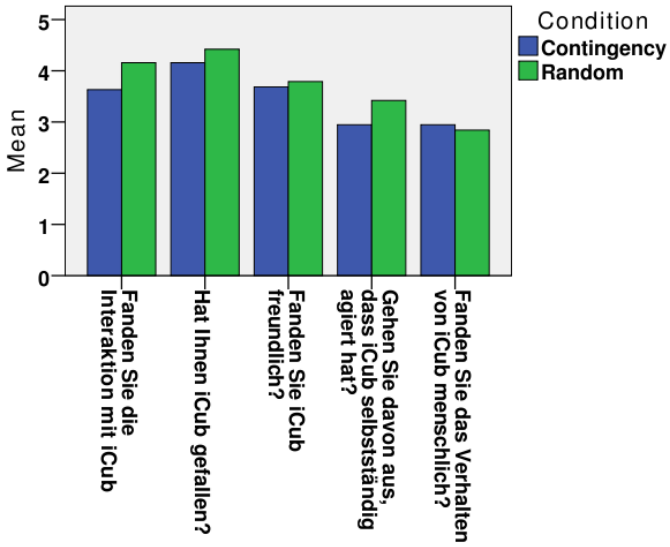


Figure 5.11: The questions are from left to right: Did you like the interaction with iCub?, Did you like iCub?, Was iCub friendly?, Was iCub behaving autonomous?, Was iCub behaving human like?. The rating was always from 1 to 5 with 5 being the most positive.

For the question *Did you like iCub?*, no significant difference was found between the two conditions (contingency condition: M=4.16, sd=1.119; random condition: M=4.42, sd=0.769).
For the question *Was iCub friendly?*, no significant difference was found between the two conditions (contingency condition: M=3.68, sd=1.003; random condition: M=3.79, sd=0.976).
For the question *Was iCub behaving autonomous?*, no significant differ-

ence was found between the two conditions (contingency condition: M=2.95, sd=0.970; random condition: M=3.42, sd=1.071).
For the question *Was iCub behaving human like*, no significant difference was found between the two conditions (contingency condition: M=2.95, sd=1.079; random condition: M=2.84, sd=1.015).

5.2.5 *Discussion*

In this section, the results of the evaluation of the improved tutoring spotting system were presented. The metrics used for this evaluation were structured into measuring the gazing behavior, the created contingency and analysing the speech (Fig. 5.12). The results of the evaluation of the gazing behavior, of a human tutor facing the tutoring spotter system 2.0, were compared to the random behavior of the iCub. The comparison suggested that, in the tutoring spotter system, the naive human tutors were tending to spend more time looking at the robot than in the other condition. The results of the contingency metrics suggested that, in the tutoring spotter system 2.0 compared to the random behavior of the iCub, the interaction was more contingent between both interaction partners. Thus, the human and the robot were more responsive to each other than in the random condition. Finally, the results of the speech analysis showed that the tutors were using less attention-getting words and more naming (of the robot) in the tutoring spotter system 2.0, compared to the random behavior of the iCub. That could suggest that the robot, equipped with the

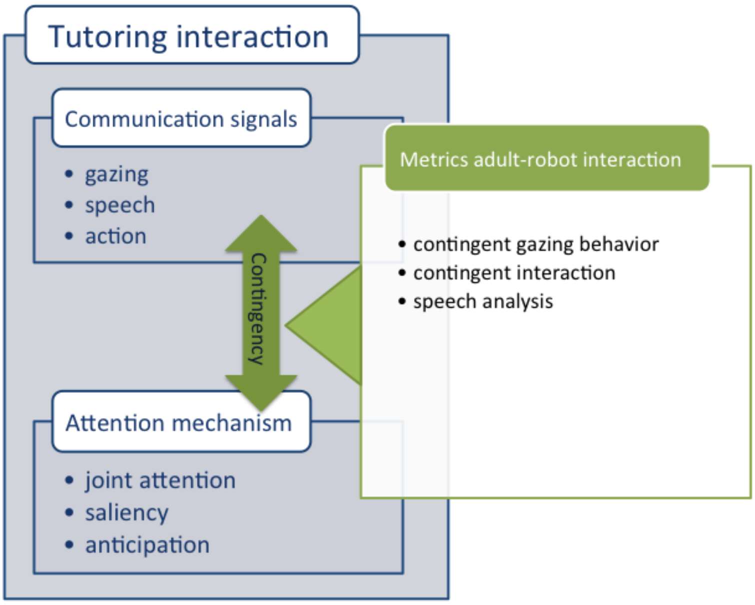


Figure 5.12: Evaluation of the improved tutoring spotter system.

tutoring spotting system, was more attentive to the interaction than the random behavior. Overall, the results revealed that the tutoring spotter system 2.0, compared with the random behavior of the iCub, was more focused on the interaction.

5.3 ANALYSIS: FURTHER PROMISING FEEDBACK STRATEGIES

In this section, some more promising feedback mechanisms that could be included in the system presented before, are discussed.

5.3.1 *Anticipation as a feedback mechanism*

As already mentioned in sections 2.1.2 and 2.3.4, an attention system requires an anticipatory behavior in order to give adequate feedback. Targeting the problem of anticipation, there are differences in the objects concerning their use and the demonstration of that use (*MANNER*- and *PATH*-oriented tasks). Thus, a starting point in the implementation of such an anticipation mechanism could be to verify and learn to classify these differences. For this purpose, the object trajectories of the cup stacking and saltshaker tasks of the motionese corpus are taken into account. As the average decision for anticipatory gazing behavior is made in the first 0,6 seconds of the cups movement 2.3.4, only the first 15 frames of every trajectory taken into account. The cup stacking task produced trajectories that had a more horizontal than vertical movement towards the side of the blue cup. The trajectories from the saltshaker task seemed to be oriented almost exclusively in an upwards direction. Based on these findings, a single-layer feedforward neural network with online back-propagation, was trained with the data of 20 trajectories (length of 15 frames) in the *PATH*-oriented task and a set of 42 trajectories (length of 15 frames) in the *MANNER*-oriented task (Fig. 5.15).

With this data set, the results of a cross-validation, with 25 runs and 10 sub-datasets, showed a mean prediction accuracy of 0.75 and a deviation of 0.1757. Even though the results were not perfect, the classifier should be able to classify a number of trajectories correctly and if a larger number could be used as input for the training, the performance would may improve. With the help of this classifier, an anticipation, for at least the presentations of the two tasks, could be simulated. Another idea for solving this problem, based on vision, would be to use reinforcement, but a project involving this has just started and its results will come in the future. These issues will be subject to further research.

5.3.2 *Keyword spotting*

As was shown in sections 2.2.4, 5.3.3 and 2.10, the language used by the parents transmits important information. But how could speech support a robotic system in order to get accepted by a human tutor in a more child-like way? On one hand, the obvious answer could be that speech is a prioritised communication signal between humans. But on the other hand, the children targeted in the first part of this

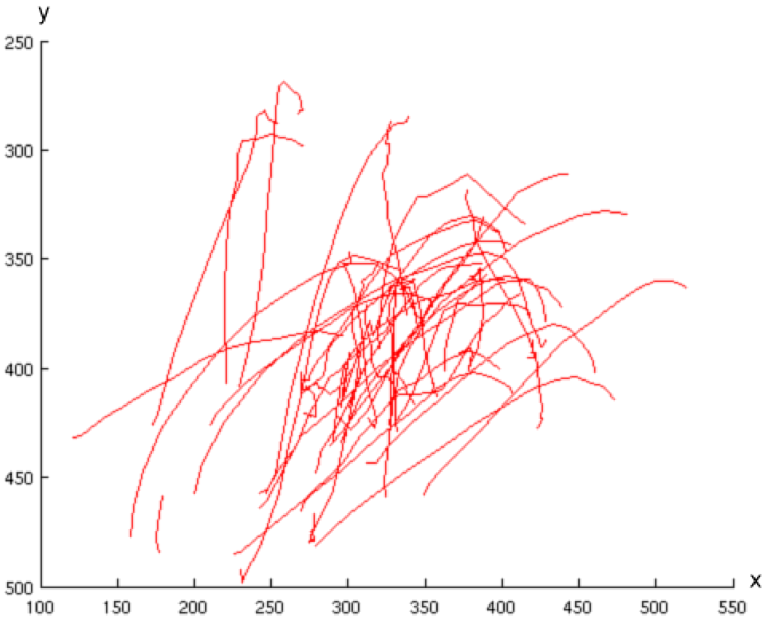


Figure 5.13: Visualised trajectories recorded from the cup stacking task (in red). Only the first 15 coordinates of every recorded trajectory are visualized here (the first 0,6 seconds).

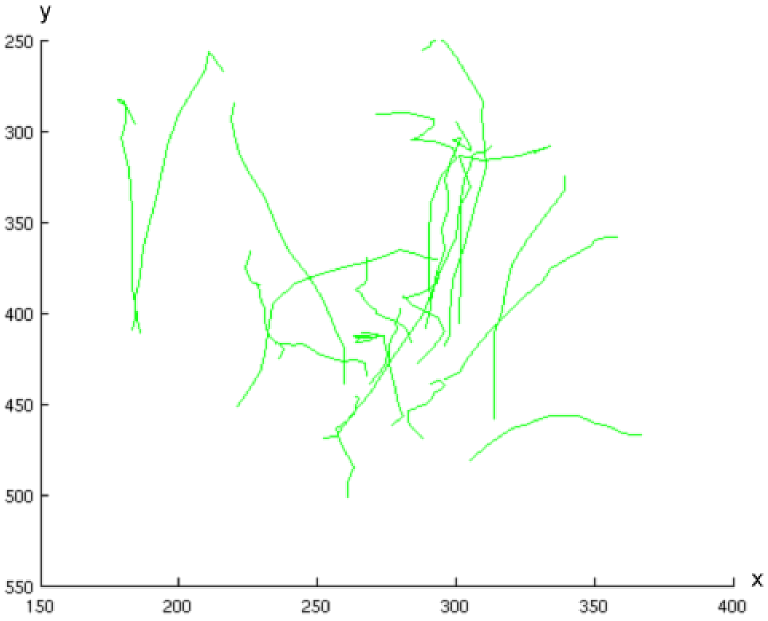


Figure 5.14: Visualised trajectories recorded from the saltshaker task (in green). Only the first 15 coordinates of every recorded trajectory are visualized here (the first 0,6 seconds).

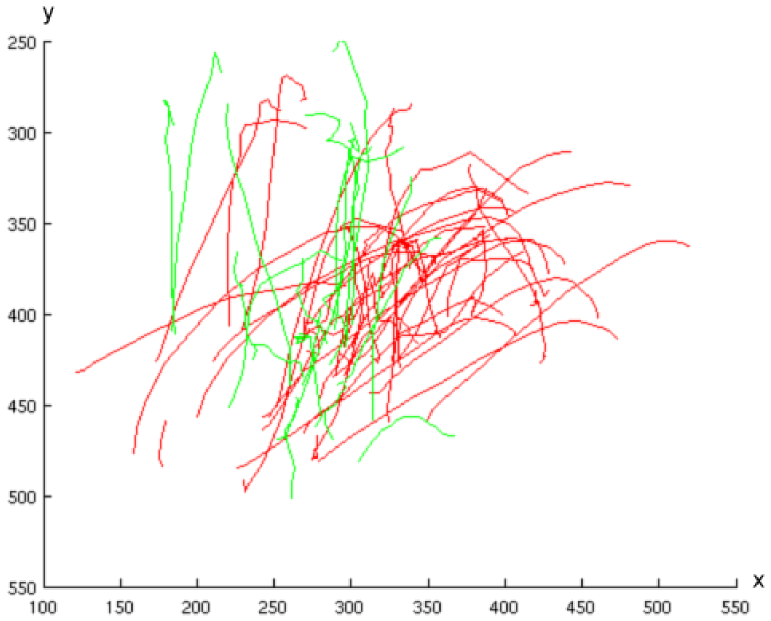


Figure 5.15: In the graph, both tasks (cup stacking and salt shaker) were visualized in a single coordinate system. Horizontal and vertical axes in the visualizations correspond to the x- and y-axis in a coordinate system. Only the first 15 coordinates of every recorded trajectory are visualised here (the first 0,6 seconds).

thesis were at the beginning of their speech acquisition stage. Thus, one could argue that handling speech is not needed for finding the beginning of a tutoring situation. But, as presented in section 2.10, all parents showed a difference in their utterances, depending on the age of the child they were addressing and they used speech, even in the pre-lexical group of the motionese corpus. So, it could be helpful to detect at least some speech or even an audio signal in the system. Searching for an adequate level of speech detection using the idea of keyword spotting, will be the focus next . Keywords (section 2.3.2), in the context of the motionese corpus, could be attention getters (like the children's name or 'look here') or meaningful words that are transferring important information about the task to be learned. Overall, the signals being sought in the behavior of the tutor, were guiding the attention by being exaggerated. For the cup stacking task:

Pre-lexical children	number of utterances	Attention getter
Green cup	50	19 (~38.00%)
Yellow cup	43	9 (~20.93%)
Red cup	34	2 (~5.88%)
Early lexical children	number of utterances	Attention getter
Green cup	36	3 (~5.78%)
Yellow cup	35	4 (~11.43%)
Red cup	33	5 (~15.16%)
Lexical children	number of utterances	Attention getter
Green cup	35	8 (~22.86%)
Yellow cup	16	3 (~18.75%)
Red cup	17	2 (~11.76%)

and for the salt shaker task:

Pre-lexical children	number of utterances	Attention getter
Salt shaker	97	38 (~39.18%)
Early lexical children	number of utterances	Attention getter
Salt shaker	71	24 (~33.80%)
Lexical children	number of utterances	Attention getter
Salt shaker	45	10 (~22.23%)

Looking back to the results of the looming actions (section 2.3.2), it can be seen that most of these attention getters were co-occurring with a looming action. This could provide another way of detecting the attention getter. Instead of detecting speech in a audio signal and limiting the vocabulary, speech could be detected only while a looming action was presented and this audio signal was being learned as a representation of an attention getter.

5.3.3 *Levels of embodiment: linguistic analysis of factors influencing HRI*

Revisiting the debate regarding embodiment and if it affects the tutoring behavior, small differences were found in the Akachan vs. the iCub, using only its eyes for looking at the most salient point (section 3.2.4 and the following). In the seventeenth century, the principle of mind/body dualism was developed –"there has been a common assumption within philosophy and the other more recent cognitive sciences that the mind can be studied without recourse to the body, and hence without recourse to embodiment" [29]. This principle has been contrasted with the empiricist view, that "the human mind – and therefore language – cannot be investigated in isolation from human embodiment" [29]. It is assumed that there are differences in the tutor's speech towards the three different embodied robotic systems (Ackachan, iCub *NoHead* and iCub *Head* condition). To test this, an analysis was performed on the tutor's speech data[1].

5.3.3.1 *Data acquisition and analysis*

The produced utterances of the participants were manually transcribed and an analysis of the syntax was performed. Based on these data, a linguistic analysis was carried out, using the constraint-based parser described in [40]. The constraint-based parser performs a morphological classification and syntactic and referential dependency analysis on the word level. The system assigns each dependency to one of 35 syntactic classes. Using the results of this system, a quick computation of basic frequency counts, can be conducted. This quick computation can result in Mean Length of Utterance (MLU) [101] or category distribution, in support searches among inflected words for their stems, or for the syntactic roles of words. With this classification, a distinction between subjects, direct objects, and indirect objects, or between active and passive voice, can be easily retrieved [46]. Three factors, concerning the linguistic analysis, were evaluated: verbosity, complexity and interactivity. The term linguistic verbosity concerns the amount of speech presented to a communication partner, in this case one of the three robots. The metrics represent the effort that the speakers puts in each task and the amount of information that the tutor considers necessary for the robot to understand the presented task. The number of different words indicates the suspected competence level of the robot. The total number of words for further analysis and the number of different words per tutor in each of the six tasks (Fig. 2.7) were counted, as well as the number of utterances per task. For more details see [35].

5.3.3.2 *Results*

Overall, the linguistic features remained the same in all three conditions. There were no differences in the amount of speech (i.e., the verbosity metrics). Also, the complexity metrics were very similar for the different conditions. But there were significant differences among all three conditions, a fact that indicates that there were different effects concerning the degrees of freedom. Some of these differences were only significant between two out of the three conditions. It turned out that, the use of the robot's name, showed a significant difference in all conditions. The mean in the Ackachan condition is M= 0.0,1 (sd= 0.03), in the iCub *NoHead* condition M= 0.04 (sd= 0.05) and in

[1] In cooperation with Kerstin Fischer and Kilian Foth.

the iCub *Head* condition M=0.16 (sd= 0.13); t (Ackachan condition vs. iCub *NoHead* condition) = -2.29, p < 0.03; t (Ackachan condition vs. iCub *Head* condition) = -5.69, p < 0.001; t (iCub *NoHead* condition vs. iCub *Head* condition) = -2.26, p < 0.05. The results indicated that the robot's name was uttered more in the iCub *Head* than in the other two conditions.

Simulated versus physical robot (eyes only)

Comparing the Akachan condition with the iCub *NoHead* condition, some features like the number of instances of the personal pronoun 'we', (Akachan condition M=0.04, sd=0.064; iCub *NoHead* condition: M=0.12, sd=0.15, t=-2.29, p< 0.03), the number of modal particles (Akachan condition M=0.10, sd=0.07; iCub *NoHead* condition M=0.21, sd=0.15, t=- 2.73, p< .01), the number of direct objects (Akachan condition M=0.03, sd=0.03; iCub *NoHead* condition M=0.46, sd=0.24, t=-9.72, p< 0.001), and the number of utterances containing a copula, i.e. a form of 'to be' as the main verb of the sentence (Akachan condition M=0.04, sd=0.04; iCub *NoHead* condition M=0.21, sd=0.16, t=-4.99, p< 0.001) reveal that there was a significant difference in the used utterances towards the two robots. A tendency (t=1.96, p= .058) for more imperatives was found in the iCub *NoHead* condition (M=0.05, sd=0.07), compared to the Akachan condition (M=0.02, sd=0.02).

Physical robot (eyes only) versus physical robot (eyes and head)

In this comparison, significant differences, that point to an influence of the amount of degrees of freedom of the robot, were found. The vocative and also the amount of passive constructions employed (iCub *NoHead* condition M=0.04, sd=0.03; iCub *Head* condition: M=0.08, sd=0.01, t=2.19, p< .05), showed significant difrences. There was a statistical tendency towards more direct objects (iCub *NoHead* condition M=0.46, sd=0.24; iCub *Head* condition M=0.26, sd=0.12, t=1.87, p< 0.09).

Simulated versus physical robot (eyes and head)

While comparing the interactions with the Akachan with the interaction with the iCub *Head*, significant differences were found in the vocative and the amount of understanding checks employed (Akachan condition M=0.001, sd=0.004; iCub *Head* condition: M=0.017, sd=0.04, t=-2.35, p<0.03). Furthermore, there were significantly more direct objects in the iCub *Head* condition than in the Akachan condition (Akachan condition M=0.03, sd=0.03; iCub *Head* condition: M=0.25, sd=0.12, t=-9.35, p<0 .001). There was a tendency for increased use of the imperative in the iCub *Head* condition (Akachan condition M=0.016, sd=0.03; iCub *Head* condition: M=0.045, sd=0.07, t=-1.80, p= 0.08). There were also significantly more uses of the copula in the iCub *Head* condition than in the Akachan condition (Akachan condition M=0.037, sd=0.04; iCub *Head* condition: M=0.19, sd=0.05, t=-6.43, p< 0.001). In addition, there were significantly more modal particles in the iCub *Head* condition than in the Akachan condition (Akachan condition M=0.10, sd=0.07; iCub *Head* condition: M=0.19, sd=0.12, t=-2.44, p< 0.02). Finally, the evaluation also showed that the users' mean length of utterance was significantly shorter in the iCub *Head* condition than in the other two conditions (Akachan condition: M= 8.4, sd 2.7; iCub *NoHead* condition: M= 8.2, sd= 4.0; iCub *Head* condition: M= 6.0, sd= 2.3; t= 0.167, p< 0.05).

5.3.3.3 *Implications*

Overall, the different linguistic behaviors supported the results presented in section 3.2.5. The two factors that were influencing the interpersonal relationship, between human user and the robot and the one which is referring to the amount of what tutoring the robot received, indicate that there could be different ways of understanding the robotic agents [35]. The results show that, not only the physical embodiment but also the degrees of freedom that are used for the interaction, influence the interaction with a human tutor. These results implicate that the robot should use its degrees of freedom in a way that is in accordance with its capability.

5.3.4 *Summary*

In this section, further promising feedback strategies, that could be added to the tutoring spotter system, were analysed (Fig 5.16). First, an anticipatory gazing mechanism was discussed and an early solution for creating anticipatory gaze was presented. Following that, a keyword spotting system was discussed and initial results were presented. It was argued that a combination with an action detection mechanism could improve the development of a keyword spotter system.

Finally, a deeper analysis of how the embodiment affects the language, used in human-robot interaction studies, was presented and it was shown that the difference found in section 3.2 could also be found in the influence of the language. The results for the language corresponded with the results of the analysis of the gazing behavior for this data set and showed that there were more understanding checks in the condition which used the iCub robot (*Head* condition), compared with the simulated robot.

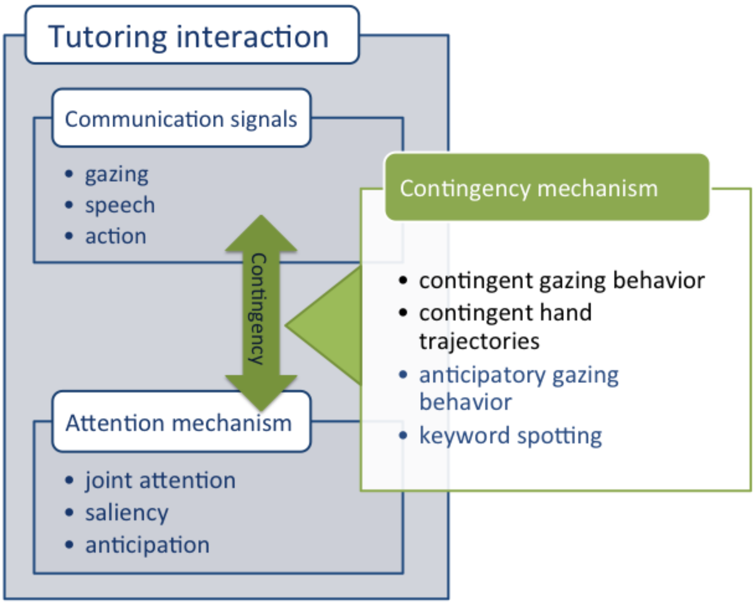


Figure 5.16: The system could be improved with further feedback mechanisms.

6 CONCLUSION

6.1 SUMMARY

In this thesis, the steps towards a tutoring spotter system implementation were presented. Knowledge from developmental psychology was transferred to a tutoring scenario. The combination of a controllable learner (the robot) with a controlled and comparable setup to verify the observed features in ACI, competing with the results of HRI, was used.

Based on the results of several studies, the feedback strategies of the human tutors were filtered out and a feature set of behavior classifications, that could help to implement a robotic system that is able to detect a tutoring situation, was constructed. The tutors' behavior, which was highlighted by OS, was quantified in metrics presented in Chapter 2. But as a tutoring interaction is a dyadic interaction, the focus was not only on the behavior of the tutor, but also on the learners' behavior and, by extension, it's attention capabilities.

Therefore, a simple attention system was initially implemented, in order to get a placement on the continuum of the behavioral adaptations a human tutor was displaying towards this system, in contrast to what an adult tutor is displaying towards a child. The already facilitated metrics were transferred towards the interaction between a human and a robot and other sources, like the dimension of embodiment of a robot, were tested. The capacity of the metrics for generalisation was analysed and the implemented saliency system was compared to other saliency systems and the gazing behavior of children.

Based on the results, the system was redesigned in a way that the behavior adaptation of the tutor was modified towards the behavior adaptation shown towards a child. Using this knowledge, a new system was implemented (tutoring spotter 1.0) by taking the structuring history of the tutor's behavior into account, in order to create a history-based attention system. That system considered the behavior of both interaction partners and used the contingency metric to calculate the structure of the interaction. That first model was revised after the results of the study presented in section 4.3.1.

That revision (tutoring spotter v.2.0) was evaluated and allowed for further analysis of other interesting feedback strategies.

The tutoring spotter is a stable system and could help the robot to learn from interaction with a human tutor. Therefore, it enables robots to adapt their behavior towards an interaction partner more easily and allows for improving the human robot interactions over time.

6.2 CAPABILITY AND PROBLEMS

The implemented system could be used in a robotic system in order to provide an online analysis of the system performance and to create a more contingent interaction with a naive user. The current gaze tracking system, even though it is modular and combining different approaches of gaze classification, has some disadvantages. This is attributed to the current state of gaze tracking systems, but could partly be improved by implementing another object tracker. The implemented system could also be used to learn a borderline condition, when a human is not only interacting with the robot but also is starting to teach something to the robot.

6.3 FUTURE WORK

The feedback strategies are important in order to create an interaction loop, between the two interaction partners and to create a contingent interaction. Further extensions of the implemented model could be made to include linguistic cues and anticipative gaze as feedback strategy. The system could also benefit from learning verbal attention getters, based on the looming behavior detection. Based on this system, the human attention system could be studied more deeply. This could help to create user interfaces that are capable of learning from naive users, without further instructions of the user.

Also, it would be interesting to change the scenario, to see if the system can be improved in order to be capable to adapt to new situations. In general, adaptations for other scenarios should be available to help testing the robotic system, with regards to its interaction capabilities. Finally, the potential of the interdisciplinary research approach, which was followed to create this model, enriched the overall results and will be needed for continuing this work.

Part I

APPENDIX

A

APPENDIX

A.1 MOTIONESE CORPUS

A.1.1 *Speech analysis of Manner and Path*

All utterances highlighted in pink are related as path-oriented, all highlighted in green are treated as manner-oriented:

VP005	AC	3	Action1	1,4	10,17	adult	0,99842	2,0317	alicia guck mal-						
						adult	2,6607	4,0983	das ist der salzstreuer-						
						adult	4,8396	6,2997	aber nicht zu hause nach machen""						
						adult	8,9952	9,8937	einfach umkippen-						
						adult	9,8937	10,4778	guck mal-						
VP056	AC	3	Action1	0,69	4,2	parent	0,69144	1,5531	guck mal	maeuschen.					
VP056	AC	1	Action1	2,73	7,93										
VP040	AC	3	Action1	1,12	8,29	parent	1,3438	1,8406	schau mal.						
						parent	4,4071	5,72	guck mal	was da rauskommt.					
VP040	AC	1	Action1	7,17	17,58	parent	7,2996	8,2776	hier guck mal	salz.					
						parent	10,4738	10,9713	guck mal.						
VP014	AC	1	Action1	1,85	17,68	parent	3,8133	5,1914	schau her	hier haben wir einen salzstreuer-					
						parent	6,8952	7,5359	das sind-						
						parent	9,127	9,8307	jerome.						
						parent	11,3764	11,9593	hallo.						
						parent	12,6776	14,439	hey da sind lcher und das muss man umdrehen-						
						parent	15,8361	16,6658	da kommt was raus.						
VP010	AC	1	Action1	2,6	14,11	parent	5,1716	6,1648	guck mal was da drin ist.						
						parent	9,0149	9,3969	guck mal.						
						parent	11,8139	12,8008	salz liebt sie-						
						parent	13,0412	13,9521	findet sie ganz toll.						
VP010	AC	3	Action1	2,36	11,9	parent	3,4076	5,0066	nein //*//.						
						parent	11,6148	13,32	oi	du auch einmal?.					
VP028	AC	1	Action1	13,903	19,973	parent	13,6757	14,7223	hoerst du vielleicht doch noch nicht	ne?.					
						parent	15,1599	16,1379	hm? einmal streuen.						
						parent	16,6267	16,8841	&.						
						parent	17,0043	19,6381	klopfen	klopfen	klopfen	klopfen	klopfen	klopfen	klopfen.
						parent	19,7676	20,4968	und dann kommt was raus.						
VP028	AC	3	Action1	3,05	13,39	parent	2,8595	3,8354	was ist das denn hier?.						
						parent	6,7554	7,1799	guck mal hier.						
						parent	9,2682	9,6611	guck mal.						
						parent	11,6607	13,3527	eins	zwei	drei	vier	fuenf.		
VP052	AC	1	Action1	0,75	3,62										
VP052	AC	3	Action1	2,24	7,91	parent	1,3841	4,1758	sieh mal das ist ein salzstreuer	da ist salz drin.					
						parent	4,8788	6,1289	kann man den umkippen-						
						parent	6,3265	7,3928	dann kommt das da raus.						
VP001	AC	3	Action1	2,23	7,53	parent	4,8978	5,3208	und.						
VP001	AC	1	Action1	2,56	10,15	parent	3,2875	4,0209	guck mal.						
						parent	7,7168	8,5509	was machen wir damit?.						
						parent	9,3993	10,4635	so machen wir.						
VP006	AC	1	Action1	1,51	15,73	adult	3,6031	4,6297	gerry guck mal -						
						adult	5,2829	6,8694	das ist ein salzstreuer-						
						adult	6,8694	7,9192	und dadrin-						
						adult	7,9192	9,1467	da befindet sich das salz-						
						adult	9,1467	9,7857	guck						
						adult	10,2289	11,0688	und das kommt jetzt-						
						adult	11,2322	11,5768	& -						
						adult	11,5768	12,3287	oh-						
						adult	12,3287	12,8653	guck mal-						
						adult	12,8653	13,7966	das kann da raus-						
VP007	AC	1	Action1	1,04	9,98	adult	1,6383	2,5365	jasmin guck mal hier -						
						adult	2,6648	3,4114	da ist salz drin-						
						adult	3,4114	3,493	siehste						
						adult	3,493	4,6595	siehste-						

						adult	4,6595	5,6511	das weisse zeug dadrin-	
						adult	5,6511	6,4093	guck mal hier-	
						adult	6,5726	7,4008	wenn ich das umdrehe-	
						adult	9,3489	10,7837	siehste dann kommt das raus.	
VP002	AC	1	Action1	1,58	6,25	parent	1,4546	1,833	zt.	
						parent	2,0222	2,6744	&.	
VP002	AC	3	Action1	1,72	9	parent	3,8338	4,6996	guck mal hanna-	
						parent	5,1468	5,5498	guck mal hier.	
						parent	6,1129	6,5104	&.	
VP008	AC	1	Action1	2,7	24,25	parent	2,4788	3,3169	salzstreuer.	
						parent	4,6798	5,7473	ein salzstreuer.	
						parent	6,5874	6,9728	oh.	
						parent	7,6396	8,8872	was kann man damit machen?.	
						parent	9,3185	9,7406	hm?.	
						parent	10,5781	11,5844	was kann man damit machen?.	
						parent	11,6907	13,0304	den brauch mama //*//-	
						parent	13,1011	14,1716	fuer ihr ei	guck mal.
						parent	14,7526	15,6672	so.	
						parent	17,9644	18,411	guck-	
						parent	18,5089	19,9006	wie da salz raus kommt	siehst du?.
						parent	22,1037	22,5075	guck.	
						parent	22,7733	24,6636	auf das ei drauf machen zum beispiel-	
VP008	AC	3	Action1	3,7	9,15	parent	3,794	5,0833	gucke mal ein salzstreuer-	
						parent	6,0278	7,6291	da kommt das salz raus	guck mal-
						parent	7,6839	8,4734	&.	
VP023	AC	3	Action1	1,45	7,59	parent	1,4258	2,4922	so gibt es salz.	
						parent	2,964	4,4892	guck mal	das wuerde manu gefallen.
						parent	7,3363	8,9261	& hast du das gesehen?.	
VP023	AC	1	Action1	3,11	13,52	parent	1,8069	3,1194	wie das salz gestreut wird.	
						parent	6,1366	6,4111	ida.	
						parent	7,535	8,3586	das ist salz.	
VP025	AC	1	Action1	2,01	14,15	parent	2,4846	3,5227	luca. guck mal hier.	
						parent	3,617	4,7666	da guck mal	da ist was drin.
						parent	5,2299	5,9162	guck mal.	
						parent	6,195	6,6417	&.	
						parent	7,856	9,4517	guck mal	da ist wa drin.
						parent	9,8853	12,1844	& und die kommen da auch raus	so.
VP025	AC	3	Action1	1,53	8,4	parent	1,9427	3,3108	maeuschen	guck mal. &.
VP018	AC	1	Action1	2,13	13,26	parent	1,656	2,6304	jonas guck mal-	
						parent	7,3475	7,7496	hm-	
						parent	8,4821	9,4565	das ist salz.	
						parent	12,5729	12,8745	hier.	
VP018	AC	3	Action1	2,79	6,61	parent	3,515	4,0758	guck mal-	
						parent	4,736	6,8231	jonas und da drauf hauen kann.	

VP054	AC	3	Action1	2,43	5,28							
VP011	AC	1	Action1	8,85	14,24	parent	8,3209	8,9063	Julia?.			
						parent	9,0875	9,9067	Guck mal hier.			
						parent	11,1141	13,1893	Das kennst du doch von zu Hause //*//.			
						parent	14,0506	14,3575	/*/.			
VP011	AC	3	Action1	1,41	4,87	parent	2,2144	2,6768	Guck.			
VP017	AC	1	Action1	5,08	16,15	parent	6,7736	8,3782	Das ist Salzstreuer.			
						parent	9,4112	12,8181	Da	da kommt ueberall das Salz raus	aber man sollte versuchen%	
						parent	13,2625	15,3679	Ich zeig dir das mal	was du gleich machen sollst.		
						parent	16,0052	17,6757	So	das Salz da raufstreuen.		
VP021	AC	1	Action1	3,39	13,58	parent	4,3671	4,7604	Guck.			
						parent	5,2005	5,8038	Salz.			
						parent	6,0127	6,8127	Fuers Ei.			
						parent	7,4877	7,8837	Guck.			
						parent	10,4567	10,9563	Salz.			
VP021	AC	3	Action1	1,53	8,03	parent	1,3488	2,9643	Ah ja	ein Salzstreuer.		
						parent	3,6328	4,0522	Ja.			
						parent	4,8735	5,2705	So.			
						parent	5,7342	6,6056	Tuk tuk tuk.			
						parent	6,8886	7,8308	Da kommt Salz raus.			
VP060	AC	1	Action1	1,99	5,17	parent	1,1366	2,2509	/*/ zeigen wir mal.			
						parent	3,5556	4,6064	Ma Salz da drauf streuen.			
VP060	AC	3	Action1	3,27	6,07	parent	2,4507	4,396	Das ist ein Tablett und das ist ein Salzstreuer.			
						parent	4,396	4,6952	Guck.			
						parent	4,6952	6,0499	Jetzt kommt hier Salz drauf.			
VP053	AC	3	Action1	1,25	9,43	parent	1,2816	2,9358	Guck mal	hier ist ein Salzstreuer.		
						parent	3,3364	5,6368	Und wenn ich den nehme und umdreh-			
						parent	6,6189	10,1987	und dann f hrt ein bi§chen Salz raus	und beim Sch tteln fllt noch mehr Salz raus.		
VP053	AC	1	Action1	2,09	13,24	parent	1,6211	2,5462	Guck mal	Simon	hier.	
						parent	2,8182	3,7977	Da is was drin.			
						parent	4,4234	5,5661	Salz is da drin.			
						parent	5,6342	6,1647	Siehste das?.			
						parent	6,8585	7,8379	Guck mal	und das kann man%		
						parent	8,9534	10,7354	&Guck mal	das kommt da raus.		
						parent	10,8035	11,9733	Wenn ich da drauf haue-			
						parent	12,9256	14,2043	kommt vorne Salz raus.			
VP027	AC	1	Action1	2,91	10,35	parent	4,9829	5,5273	Tipp ich mal.""			
						parent	5,5273	9,8933	Also	machen wir zu Hause zwar nicht so mit dir ne. Also	Tom-Niklas	wenn man das hier umdreht-
VP027	AC	3	Action1	3,36	6,51	parent	1,4577	6,4077	Hey	das kennst du schon	ne? Hier kannst du endlich mal mit dem Salz rumoelen.	
VP035	AC	1	Action1	3,02	8,14	parent	2,3252	4,3646	guck mal	torben	das ist ein salzstreuer.	
						parent	5,5699	8,2417	das kennst du auch schon. guck mal	dreht man so rum-		
VP024	AC	3	Action1	2,37	6,27	parent	2,4882	6,1609	marvin wir kochen wir beide kochen	was machen wir denn da immer?. da nehmen wir salz und-		
VP024	AC	1	Action1	4,62	11,91	parent	5,4554	7,2828	da kennst du aus deiner kueche	ne.		
						parent	8,3635	9,9411	das ist der salzstreuer.			
						parent	10,9434	12,7056	und dann macht man so	ups-		
VP050	AC	1	Action1	4,927	9,827	adult	5,6	7,3	kuck mal kimmi	hier ist salz drin.		
						adult	7,9	8,3	salz.			
						adult	8,8	9,7	wenn du das umdrehst-			
VP055	AC	3	Action1	1,6	3,71	adult	2,4	3,5	den salzstreuer			
						adult	3,6	5,3	den brauche ich dir gar nicht zeigen eigentlich.			
VP055	AC	1	Action1	6,27	16,7	adult	5,707	6,307	wie ist denn das?			
						adult	9,307	9,907	salzstreuer.			
						adult	11,9	12,7	//'*//			
						adult	12,7	15,2	ich zeige dir mal	papa zeigt dir noch /*/ was	wie	was man mit machen kann.
						adult	16	17,1	hier	wenn du auf den kopf stellst		
VP033	AC	3	Action1	1,28	5,71	parent	1,5223	2,3202	das ist ein streuer-			
						parent	2,5613	3,5919	und da ist salz drin.			
						parent	4,2735	5,2044	und hier das salz-			
VP033	AC	1	Action1	3,38	12,07	parent	4,6754	5,3709	horch mal.			
						parent	5,7661	6,5249	was ist das?-			
						parent	7,7421	9,7814	salz	sag mal salz.		

						parent	11,1242	13,4864	soll ich das mal hier drauf streuen? pass mal auf.		
VP045	AC	1	Action1	4,05	8,04	parent	5,8529	6,9662	guck mal	was da raus kommt.	
VP045	AC	3	Action1	2,13	3,91	parent	1,9986	2,6416	guck mal hier.		
VP041	AC	1	Action1	5,6	18,465	parent	6,1525	7,3684	patricia	das ist salz-	
						parent	7,8601	8,6361	gucke	das ist salz.	
						parent	10,903	11,5019	hoerst du?.		
						parent	13,9848	15,4488	weisst du	wie man das macht?. guck mal.	
						parent	15,9927	18,6995	wenn man da jetzt zum beispiel ein brot liegen hat	ne	dann kann man da-
VP041	AC	3	Action1	4,46	6,2	parent	5,0619	6,9824	einfach salz hier drauf machen.		
VP042	AC	3	Action1	0,5	4,19	parent	1,6682	2,8926	hmm.		
VP042	AC	1	Action1	1,93	2,87	parent	2,047	2,4175	(lachen)		

VP003	AC	1	Action1	2,07	5,27	parent	0,91841	2,3739	guck mal das ist ein salzstreuer schatz.			
						parent	3,2451	5,338	und aus einem dieser lher kommt das salz raus.			
VP031	AC	3	Action1	10,96	13,408	parent	11,5896	12,703	guck mal	da ist das salz	ne.	
VP031	AC	1	Action1	2,16	8,6	parent	2,3762	2,7426	mhm.""			
						parent	3,7004	4,9218	guck mal hier sind lher drin.			
						parent	5,1957	6,2564	kennst du von zuhause	ne?.		
						parent	6,7265	7,8584	unsern salzstreuer?.			
VP038	AC	1	Action1	1,91	7,63	parent	4,4814	5,555	das kennst du	ne?.		
						parent	7,4133	7,8812	was ist das?-			
VP038	AC	3	Action1	2,43	6,706	parent	2,4127	3,1081	was ist das?.			
						parent	3,6597	5,0231	ist da salz drinne?.			
						parent	5,8932	7,0717	was kann man damit machen?.			
VP051	AC	3	Action1	7,59	15,76	parent	7,317	8,2692	das ist der salzstreuer-			
						parent	9,2153	12,2306	und jetzt schtteln wir den salzstreuer	wir drehen den salzstreuer um-		
						parent	13,1367	15,7972	und jetzt kommen ganz viele kleine kugeln raus.			
VP047	AC	3	Action1	4,01	5,42	parent	3,7864	5,3369	hier kann man mit dem salzstreuer-			
VP047	AC	1	Action1	2,41	3,89	parent	2,5358	4,5008	guck mal	das ist das	was du zuhause nicht darfst	
VP012	AC	1	Action1	1,93	10,37	parent	1,5845	1,9528	schau mal.			
						parent	2,4689	3,4401	hier ist salz drin-			
						parent	3,8256	5,1489	und wenn du das salz da aus diesem-			
						parent	6,2158	9,6826	behaelter raus haben moechtest	dann musst du das einmal auf den kopf drehen-		
						parent	10,2395	11,8534	und dann kommt das da raus. hast du das gesehen?.			
VP044	AC	1	Action1	3,16	4,71	parent	3,2024	3,8519	ja.			
VP044	AC	3	Action1	2,81	6,17	parent	2,3083	4,0587	guck mal	wenn du jetzt hier den salzstreuer nimmst-		
						parent	4,6557	5,4301	ne kennst du-			
VP009	AC	1	Action1	6,52	8,97	parent	6,149	8,1898	und da musst du ganz vorsichtig-			
						parent	8,6157	10,5154	salz rausmachen	sonst kommt da ganz viel raus.		
VP009	AC	3	Action1	2,37	4,15	parent	2,1363	2,4029	so.			
						parent	2,5823	3,6786	da kann man das so raus-			
						parent	3,8837	4,4921	streuen.			
VP030	AC	1	Action1	1,92	16,38	parent	3,3321	4,1638	weisst du	was das ist?.		
						parent	4,9599	5,5959	was ist das?.			
						parent	6,4013	7,0147	was ist das?.			
						parent	7,5265	8,1083	kennst du das?.			
						parent	9,0416	10,9719	salz? und wofuer brauchen wir das salz immer?.			
						parent	11,0773	11,7773	weisst du das noch?.			
						parent	12,3004	12,8785	aha.			
						parent	13,315	15,0697	auf unser ei machen wir das immer drauf	ne?.		
						parent	15,8901	16,3301	guck mal-			
VP032	AC	1	Action1	5,43	7,07	parent	5,3296	6,5034	ich zeig dir wie das geht.			
VP063	AC	1	Action1	2,56	4,18	parent	1,4889	4,0147	weisst du	wie salz gestreut wird? ja	ne. guck mal	einfach soo.
VP036	AC	1	Action1	1,88	7,63	parent	1,8866	2,9109	weißt du was das ist nara?.			
						parent	5,5135	8,9431	ein salzstreuer	ne	guck mal und da kann man so mit streuen.	
VP062	AC	1	Action1	1,94	6,54	parent	1,8403	2,5127	kennst du das?.			
						parent	4,2936	7,6083	das muss man nicht aufmachen. das ist nicht wie bei uns zu hause. guck mal	das kommt gleich raus.		

VP005	AC	3	Action1	1,47	4,79	adult	0,99185	1,8564	alica guck mal hier-	
						adult	1,8564	2,8144	der kleinste ne -	
						adult	2,8144	3,6323	der rote-	
VP005	AC	1	Action1	4,712	6,352					
VP056	AC	3	Action1	2,23	5,23					
VP056	AC	1	Action1	3,065	4,073	parent	2,8718	3,3002	mhm.	
VP040	AC	1	Action1	10,557	13,683					
VP040	AC	3	Action1	4,208	8,436	parent	3,735	4,4467	spatz	guck mal.
						parent	5,4664	7,9096	den grünen stecken wir hier rein.	
VP014	AC	1	Action1	13,19	34,55	parent	14,637	15,8042	so vor kann ich ja gehen	ne?.
						parent	17,1196	17,5457	ach so.	
						parent	18,3795	18,787	mhm.	
						parent	19,3429	20,3619	genau. genau. genau.	
						parent	20,8721	21,7985	da schu mal her	jerome.
						parent	24,9481	26,3932	gut	die sind verschieden groß.
						parent	26,6711	28,5053	obwohl	das wird er wahrscheinlich noch nicht realisieren.
						parent	30,3149	30,7596	jerome.	
						parent	34,2056	35,5766	da kann man die ineinander stecken.	
VP010	AC	1	Action1	5,607	6,746					
VP010	AC	3	Action1	1,67	3,71	parent	2,0065	2,8614	jetzt aber aufpassen	
VP028	AC	3	Action1	5,3	8,35	parent	6,2848	7,1718	guck mal	torin.
						parent	7,451	8,8308	die becher kann man ineinander stecken.	
VP028	AC	1	Action1	2,67	5,43	parent	2,1682	2,6818	hier guck mal.	
						parent	3,3025	3,8589	hallo	hallo.
						parent	4,8541	5,0039	&.	
VP052	AC	3	Action1	9,136	10,252	parent	8,6001	9,3052	guck mal hier!.	
						parent	9,5153	11,0906	den kann man da rein stellen.	
VP052	AC	1	Action1	0,914	2,321	parent	0,56718	1,003	ja.	
VP001	AC	3	Action1	4,127	8,903	parent	5,4388	7,9803	der grüne becher	den packen wir in den-
						parent	8,7931	9,5543	blauen.	
VP001	AC	1	Action1	10,07	13,951	parent	10,9775	13,3436	guck mal	erst nehmen wir den gnen-
VP006	AC	3	Action1	2,11	3,83	adult	2,1398	3,4664	die passen alle ineinader-	
VP006	AC	1	Action1	11,15	16,91	adult	10,9459	12,4647	also jetzt haben wir hier becher-	
						adult	13,1891	14,0069	und die kann man-	
						adult	14,0069	15,8295	ineinander stapeln pass mal auf	
VP007	AC	1	Action1	9,699	13,708	adult	9,6374	10,1047	guck mal der -	
						adult	10,1047	10,4786	grüne-	
						adult	10,4786	11,8806	der ist nen bisschen kleiner als der-	
						adult	12,1376	13,3527	den kann man da rein machen-	
						adult	13,3527	14,1705	dann ist das so-	
VP007	AC	3	Action1	5,83	7,23					
VP008	AC	3	Action1	1,91	3,51	parent	1,435	1,9848	guck mal-	
						parent	2,2562	3,4474	der rote in den gelben-	

VP008	AC	1	Action1	3,19	6,434					
VP023	AC	3	Action1	5,15	8,51	parent	4,9358	5,6051	speedy.	
						parent	7,322	8,7188	guck mal	das ist der kleinste.
VP023	AC	1	Action1	13,27	24,23	parent	13,7073	14,8969	das ist der größte.	
						parent	16,2371	16,5382	/*/	
						parent	17,2761	17,7429	speedy.	
						parent	19,732	20,2741	ida.	
						parent	22,0854	24,6452	dann kommte der grüne	der ist ein bischen kleiner.
VP025	AC	3	Action1	2,651	3,945					
VP025	AC	1	Action1	16,35	18,51	parent	16,4486	16,836	/*/	
						parent	16,985	17,3576	und dann -	
						parent	18,3261	18,9221	dann da rein.	
VP018	AC	1	Action1	8,742	10,382	adult	9,5973	9,9552	kann man-	
VP018	AC	3	Action1	3,23	4,35	parent	2,9646	3,4428	so.	

VP005	AC	3	Action2	9,19	11,07	adult	10,1281	10,5253	und hier-	
VP056	AC	3	Action2	6,23	8,11	parent	5,9571	7,4447	und jetzt der gelbe -	
VP056	AC	1	Action2	4,598	5,333					
VP040	AC	1	Action2	15,954	18,148	parent	17,0187	18,497	jetzt den gelben.	
VP040	AC	3	Action2	10,238	12,656	parent	9,3473	10,8031	& und den gelben-	
						parent	11,2103	12,1034	in den grünen.	
VP014	AC	1	Action2	14,39	37,23	parent	14,637	15,8042	so vor kann ich ja gehen	ne?.
						parent	17,1196	17,5457	ach so.	
						parent	18,3795	18,787	mhm.	
						parent	19,3429	20,3619	genau. genau. genau.	
						parent	20,8721	21,7985	da schu mal her	jerome.
						parent	24,9481	26,3932	gut	die sind verschieden groß.
						parent	26,6711	28,5053	obwohl	das wird er wahrscheinlich noch nicht realisieren.
						parent	30,3149	30,7596	jerome.	
						parent	34,2056	35,5766	da kann man die ineinander stecken.	
						parent	36,3177	37,7628	und so hinein.	
VP010	AC	1	Action2	10,493	12,874	parent	12,1854	13,1652	den da rein.	
VP010	AC	3	Action2	4,63	9,55	parent	4,163	5,4107	erst den gruenen.	
						parent	5,5802	6,2295	guck mal hier.	
						parent	7,3861	8,5359	amelie	guckst du hier?.
						parent	9,1687	10,6587	&der gelbe.	
VP028	AC	3	Action2	8,87	10,23	parent	9,0279	9,5864	eins.	
VP028	AC	1	Action2	5,55	5,9					
VP052	AC	3	Action2	11,054	12,088	parent	9,5153	11,0906	den kann man da rein stellen.	
						parent	11,4207	13,116	und diesen kann man da rein stellen.	
VP052	AC	1	Action2	2,765	3,683					
VP001	AC	3	Action2	11,306	13,747					
VP001	AC	1	Action2	16,028	22,226	parent	16,456	17,123	dann-	
						parent	17,82	20,0459	hallo rasmus	hierher gucken""
						parent	20,575	21,542	dann den gelben-	
VP006	AC	3	Action2	4,79	6,15	adult	3,9725	4,9299	der kommt da rein-	
						adult	5,6001	6,5438	dann der gelbe-	
VP006	AC	1	Action2	17,95	21,55	adult	18,283	20,1523	und jetzt haben wir einen gelben becher-	
						adult	21,1337	21,6477	und der -	
VP007	AC	1	Action2	15,77	19,074	adult	16,6473	17,3717	der gelbe-	
VP007	AC	3	Action2	7,75	8,73					
VP008	AC	3	Action2	4,39	6,14	parent	4,9313	6,1437	der gelbe in den gruenen-	
VP008	AC	1	Action2	8,798	11,94	parent	9,2766	10,6671	und der gelbe-	
						parent	10,7841	11,179	guck-	
VP023	AC	3	Action2	10,15	11,55	parent	10,9013	12,1526	und der kleine kommt hier rein.	
VP023	AC	1	Action2	25,79	28,11	parent	25,2944	26,5141	dann kommt der gelbe.	
VP025	AC	3	Action2	5,219	6,321	parent	2,3982	5,5215	so mäuschen. der grüne becher in den blauen.	
						parent	5,7737	7,0347	dann kommt der gelbe-	
VP025	AC	1	Action2	22,27	24,99	parent	21,7677	22,8107	gelber becher.	
						parent	22,9448	24,4051	guck mal hier. gelber becher -	

						parent	24,5903	24,8288	und -
VP018	AC	1	Action2	12,164	13,228	adult	12,7989	14,1909	das gelbe in das grüne-
VP018	AC	3	Action2	5,31	6,79	parent	5,5436	6,4938	und nochmal.

VP005	AC	3	Action3	13,99	15,71	adult	14,4743	15,7828	und der grüne geht jetzt-		
VP056	AC	3	Action3	8,99	11,03	parent	8,7971	9,9353	und den kleinen noch.		
						parent	10,0706	10,296	guck!.		
VP056	AC	1	Action3	5,931	7,022						
VP040	AC	1	Action3	20,035	22,002	parent	19,8327	21,2788	und dann den roten.		
VP040	AC	3	Action3	15,59	17,69	parent	15,8101	17,009	der rote-		
VP014	AC	1	Action3	37,75	43,35	parent	36,3177	37,7628	und so hinein.		
						parent	38,4668	39,0782	und schau her.		
						parent	41,4497	42,4872	eins	zwei	drei.
						parent	43,0769	43,9106	und da herein.		
VP010	AC	1	Action3	14,348	15,737	parent	14,4552	15,2716	und den-		
						parent	15,6961	16,382	da rein.		
VP010	AC	3	Action3	10,63	11,55	parent	9,1687	10,6587	&der gelbe.		
VP028	AC	3	Action3	10,39	11,59	parent	10,2927	10,8019	zwei.		
						parent	11,0319	11,4425	und -		
VP028	AC	1	Action3	5,94	6,47						
VP052	AC	3	Action3	12,851	13,82	parent	11,4207	13,116	und diesen kann man da rein stellen.		
						parent	13,3834	14,4036	und den roten.		
VP052	AC	1	Action3	4,076	4,968						
VP001	AC	3	Action3	15,711	17,607						
VP001	AC	1	Action3	23,637	27,316	parent	24,1145	25,0268	und dann-		
						parent	25,7566	26,4682	den roten.		
VP006	AC	3	Action3	6,75	7,71	adult	7,2276	8,3765	und der rote.		
VP006	AC	1	Action3	22,79	29,31	adult	21,6477	23,3769	passt da rein weil er kleiner ist-		
						adult	23,3769	26,0173	und noch einen ganz kleinen becher und der ist rot.		
						adult	26,2976	27,3258	soll ich den da auch rein tun.		
						adult	27,728	28,1946	eins-		
						adult	28,3346	28,8012	zwei-		
VP007	AC	1	Action3	19,96	22,195	adult	20,1523	21,3673	und der rote-		
						adult	21,3673	22,7226	passt in den gelben.		
VP008	AC	3	Action3	6,83	8,35	parent	6,8493	8,4599	der gruene in den blauen.		
VP008	AC	1	Action3	13,589	15,681	parent	13,8337	15,6756	und dann hier der rote-		
VP023	AC	3	Action3	12,83	14,23	parent	13,7239	14,8297	und der kommt hier rein.		
VP023	AC	1	Action3	29,51	32,31	parent	29,7666	30,6701	der ist am kleinsten-		
VP025	AC	3	Action3	7,278	8,066	parent	7,5391	8,9553	und der rote.		
VP025	AC	1	Action3	27,35	31,47	parent	27,5939	29,2181	& roter becher -		
VP018	AC	1	Action3	14,527	15,337	adult	14,9864	15,7221	und das rote.		
VP018	AC	3	Action3	7,35	8,67						

VP054	AC	3	Action1	2,43	3,43	parent	2,8469	6,2705	Das haben wir zu Hause auch und du findest es ganz toll.			
VP054	AC	1	Action1	10,32	15,35	parent	8,515	10,4681	Den blauen Becher haben wir zu Hause auch	he?.		
						parent	12,0754	12,5248	Ja.			
						parent	13,0087	13,7692	Und den grnen-			
						parent	15,1841	16,2753	Den schmei§en wir da rein.			
VP011	AC	3	Action1	3,89	4,89							
VP011	AC	1	Action1	1,18	2,45							
VP017	AC	1	Action1	5,97	7,63	parent	6,7962	9,8693	Die koennen wir alle ineinander stellen.			
VP021	AC	1	Action1	1,96	4,57	parent	2,1928	3,3153	So	aufraeumen.		
VP021	AC	3	Action1	3,32	5,75	parent	3,9573	4,3193	Na.			
						parent	4,6953	5,9486	//*//.			
VP060	AC	1	Action1	5,6	6,59							
VP060	AC	3	Action1	1,37	2,87	parent	1,3723	3,4279	Jetzt kommt der grne da rein.			
VP053	AC	3	Action1	3,83	5,7	parent	3,3292	4,0022	machs dir erst vor.			
						parent	4,3867	5,4442	roter Becher-			
						parent	5,5596	6,5786	und den gelben.			
VP053	AC	1	Action1	2,88	4,57	parent	1,6377	3,0879	Das hast du auch schon mal gesehen	ne guck mal.		
						parent	4,2198	5,0001	Der kommt hier rein.			
VP027	AC	3	Action1	3,31	4,88	parent	3,1793	4,0769	das kennst du	ne?.		
						parent	4,6358	6,0386	Den grnen da rein.			
VP027	AC	1	Action1	6,13	8,29	parent	6,2215	7,361	Dann kommt der gme-			
						parent	7,8024	8,478	-da rein.			
VP035	AC	1	Action1	8,94	9,73	parent	9,3553	10,2056	so.			
VP024	AC	1	Action1	8,99	9,75	parent	9,5085	9,9157	ja.			
VP024	AC	3	Action1	16,6	17,63	parent	16,8116	19,7204	und jetzt in den blauen	ist wie zu hause	aufrumen	ne.
VP050	AC	1	Action1	14,82	16,99	parent	15,6203	16,992	und der passt in den grossen blauen.			
VP055	AC	1	Action1	14,94	17,96	adult	14,9	15,4	der gruene-			
						adult	17,3	18	kommt in den blauen.			
VP055	AC	3	Action1	5,43	6,59	adult	5,7	7,1	hier kommt der gruene rein.			
VP033	AC	3	Action1	5,61	8,02	parent	6,0025	7,6353	den becher stecken wir jetzt da rein.			
VP033	AC	1	Action1	20,3	23,73	parent	20,5505	21,7829	den gruenen becher-			
						parent	22,9391	24,2649	tu ich jetzt in den blauen.			
VP045	AC	1	Action1	10,96	12,69	parent	10,2834	11,2187	guck mal maeuschen-			
						parent	12,4472	12,9336	den-			
VP045	AC	3	Action1	6,36	7,37							
VP041	AC	3	Action1	5,47	6,75	parent	5,6527	7,3066	der passt in den gelben.			
VP041	AC	1	Action1	7,45	10,79	parent	7,4418	8,6844	wie man die zusammen stem-			
						parent	8,807	9,264	stellt.			
						parent	10,2217	11,1968	der gruene-			
VP042	AC	1	Action1	12,98	14,12							
VP042	AC	3	Action1	3,08	4,08	parent	1,6566	3,8787	dann guck mal	werden wir den	den einen nehmen wir-	

VP054	AC	3	Action2	3,83	4,48					
VP054	AC	1	Action2	20,44	26,87	parent	21,5765	22,1881	Den gelben-	
						parent	24,8316	25,4192	Guck mal	Lukas.
						parent	25,6223	26,2877	Der gelbe Becher.	
VP011	AC	3	Action2	5,37	6,24	parent	5,7734	6,7895	Kannst du das auch?.	
VP011	AC	1	Action2	3,88	4,82					
VP017	AC	1	Action2	8,29	8,93					
VP021	AC	1	Action2	5,3	7,17	parent	5,4498	7,4003	Wie beim Aufraeumen	alles ineinander packen.
VP021	AC	3	Action2	7,49	10,5	parent	6,0532	7,5083	//*//.	
						parent	7,8982	8,7477	Du willst selber	ne?.
						parent	9,43	10,9061	Erst zeigen	der gelbe-
VP060	AC	1	Action2	7,24	8,15					
VP060	AC	3	Action2	3,24	3,9	parent	1,3723	3,4279	Jetzt kommt der grne da rein.	
						parent	3,4279	4,4556	Der gelbe.	
VP053	AC	3	Action2	7,23	9,39	parent	7,3862	9,9434	der gelbe ist so klein	der passt in den grnen.
VP053	AC	1	Action2	6,51	7,46					
VP027	AC	3	Action2	6,11	7,39	parent	6,936	8,3069	Den gelben da rein.	
VP027	AC	1	Action2	9,46	11,19	parent	9,1101	10,1583	Der gelbe-	
						parent	10,65	11,3924	da rein.	
VP035	AC	1	Action2	6,7	8,85	parent	6,1786	7,3576	kommen in die grossen.	
						parent	7,6079	7,9223	ne.	
						parent	8,1581	8,8797	so-	
VP024	AC	1	Action2	7,71	8,74	parent	7,1478	8,717	guck mal	die knnen wir ineinander stecken.
VP024	AC	3	Action2	14,32	15,43	parent	14,696	15,947	in den grnen-	
VP050	AC	1	Action2	12,34	13,7	parent	12,4209	14,4322	hm der passt wieder in diesen hier.	
VP055	AC	1	Action2	33,09	34,45	adult	33,2	35,2	jetzt koennen wir den grossen gruenen in den blauen	
VP055	AC	3	Action2	7,37	8,39	adult	7,8	8,7	der gelbe	
VP033	AC	3	Action2	13,11	15,04	parent	13,1963	14,0702	den becher-	
						parent	14,3663	14,9489	hier rein.	
VP033	AC	1	Action2	26,65	29,96	parent	26,4407	28,3456	und den gelben becher-	
						parent	29,2805	29,9859	tu ich in den gruenen.	
VP045	AC	1	Action2	14,1	15,84	parent	14,9581	16,3548	&und den gelben-	
VP045	AC	3	Action2	8,14	9,07	parent	7,9912	8,7657	einmal-	
VP041	AC	3	Action2	9,12	10,71	parent	7,9804	9,1226	und der gelbe-	
						parent	9,6343	11,1696	passt in den gruenen.	
VP041	AC	1	Action2	13,31	14,89	parent	13,5202	14,2056	der gelbe-	
						parent	14,6569	15,7156	kommt auch da rein-	
VP042	AC	1	Action2	10,03	12,22	parent	10,5752	12,5478	der becher da rein-	
VP042	AC	3	Action2	6,78	8,18	parent	6,3992	7,4841	ne guck mal hier hin..	
						parent	7,6556	8,8562	den machen wir da rein.	

VP054	AC	3	Action3	4,87	5,41					
VP054	AC	1	Action3	31,52	34,39	parent	32,1977	33,1793	Lukas	der rote Becher-
VP011	AC	3	Action3	6,78	7,57	parent	5,7734	6,7895	Kannst du das auch?.	
VP011	AC	1	Action3	5,86	6,68					
VP017	AC	1	Action3	9,38	10,37	parent	6,7962	9,8693	Die koennen wir alle ineinander stellen.	
VP021	AC	1	Action3	7,83	8,86	parent	8,5656	9,1723	Jaha.	
VP021	AC	3	Action3	11,33	13,25	parent	12,3823	12,8697	der-	
VP060	AC	1	Action3	9,04	10					
VP060	AC	3	Action3	4,44	5,02	parent	3,4279	4,4556	Der gelbe.	
						parent	5,0038	5,8945	Und der rote.	
VP053	AC	3	Action3	10,49	12,38	parent	10,4241	12,1545	der gme ist kleiner als der blaue-	
VP053	AC	1	Action3	9,73	11,34	parent	9,2874	10,0387	Und dann kommt das-	
						parent	11,264	11,6331	hier rein.	
VP027	AC	3	Action3	9,88	11,02	parent	9,7626	10,1936	Und-	
						parent	10,3359	11,4052	den roten-	
VP027	AC	1	Action3	12,52	13,5	parent	12,2866	13,1805	Und der rote-	
VP035	AC	1	Action3	4,93	6,57	parent	5,1783	6,0572	und die kleinen-	
						parent	6,1786	7,3576	kommen in die grossen.	
VP024	AC	1	Action3	6,01	7,46	parent	7,1478	8,717	guck mal	die knnen wir ineinander stecken.
VP024	AC	3	Action3	11,36	13,46	parent	11,2318	12,2497	den roten-	
						parent	12,8323	13,5493	in den gelben-	
VP050	AC	1	Action3	9,15	10,98	parent	8,8671	10,1463	der becher ist ganz klein-	
						parent	10,3852	11,464	der passt in diesen rein.	
VP055	AC	1	Action3	35,8	36,71	adult	36	37,1	den gelben in den gruenen	
VP055	AC	3	Action3	9,01	9,95	adult	9,4	10,4	und der rote	
VP033	AC	3	Action3	16,49	19,15	parent	16,4163	17,363	und den becher-	
						parent	17,9033	18,5283	stecken wir-	
						parent	18,674	19,1109	da rein.	
VP033	AC	1	Action3	32,45	42,65	parent	32,7859	36,8743	und was ist das fuer ein becher	ein roter becher. du mchtest ihn gerne haben. soll ich ihn auch da rein tun?.
						parent	37,8007	40,6452	ich tu den einmal jetzt da rein. und dann kriegst du es gleich. ja pass mal auf.	
						parent	41,1807	43,7938	den roten tu ich jetzt in den gelben und &-	
VP045	AC	1	Action3	17,53	19,06	parent	18,1029	19,8925	&der rote passt da auch rein.	
VP045	AC	3	Action3	9,93	10,83	parent	9,5263	9,9352	so-	
VP041	AC	3	Action3	12,5	14,23					
VP041	AC	1	Action3	17,42	18,68					
VP042	AC	1	Action3	4,05	8,21	parent	2,9227	4,1046	guck mal	finnja-
						parent	4,629	6,4351	ein kleiner becher-	
						parent	7,2674	8,6158	da rein.	
VP042	AC	3	Action3	10,65	11,48					

VP003	AC	3	Action1	10,19	12,19	parent	8,5213	10,9896	den?. //*//	exp		8,437	11,1307	text	
VP003	AC	1	Action1	2,85	5,43										
VP031	AC	1	Action1	9,02	11,12	parent	8,7468	10,1646	guck mal	hier ist der grosse-					
						parent	10,3882	12,033	da kommt der gruene rein	ne?.					
VP031	AC	3	Action1	4,36	5,55	parent	3,0992	5,2287	die kann man ineinander stapeln. gross-						
VP038	AC	1	Action1	5,4	7,81	parent	7,489	7,9599	wupp.						
VP038	AC	3	Action1	12,76	14,11	parent	10,7896	13,0587	ich zeig dir das mal. guck mal	das ist ein blauer becher-	x		13,8164	15,2798	da muss der gruene rein.
						parent	10,3882	12,033	da kommt der gruene rein	ne?.					
VP051	AC	3	Action1	8,35	9,77	parent	8,5142	9,4259	nehmen wir zuerst den-						
VP051	AC	1	Action1	17,01	19,64	parent	16,0937	17,1823	und dann kann man den-						
						parent	17,6088	20,0071	becher	den gruenen	in den blauen stellen.				
VP047	AC	3	Action1	4,9	5,43										
VP047	AC	1	Action1	2,95	3,76	parent	2,5529	3,2204	guck mal-						
VP026	AC	1	Action1	4,11	5,69	parent	4,5236	4,7741	//*//						
						parent	5,1467	6,2876	also	erst der grne-					
VP012	AC	3	Action1	5,23	7,69	parent	3,3805	5,2557	//*//						
						parent	6,0558	9,2001	der gruene becher kommt in den blauen becher.der ist kleiner als der blaue.						
VP012	AC	1	Action1	8,52	14,95	parent	10,7611	14,244	der gruene kommt in den blauen. wir stapeln die becher jetzt ineinander.						
VP044	AC	3	Action1	9,41	10,49	parent	9,5344	10,5536	guck mal	dann nehmen wir den-					
VP009	AC	1	Action1	13,88	15,06	parent	13,3192	14,2004	dann geb ich sie dir-						
						parent	14,69	15,4145	nochmal.						
VP009	AC	3	Action1	8,42	9,75	parent	8,6931	9,1357	das geht so.						
VP030	AC	1	Action1	21,86	24,3	parent	22,4927	23,3144	und den gruenen-						
VP032	AC	1	Action1	5,01	6,49	parent	4,8167	5,1393	ah.						
						parent	5,5883	5,9139	guck mal.						
VP036	AC	1	Action1	7,26	12,67	parent	6,3785	8,9274	und der ist so gross	da muessen wir einen kleinen rein tun.					
						parent	8,9274	10,9309	einen% weisst du	was das fuer eine farbe ist?.					
						parent	11,6666	12,4788	gruen	ne?.					
VP036	AC	3	Action1	4,83	5,91	parent	3,7708	5,8668	gucke	dass der gruene in den blauen-					
VP062	AC	1	Action1	2,96	8,5	parent	2,3967	3,0197	pass mal auf.						
						parent	3,7269	4,5625	so	der ist unten	ne?.				
						parent	4,6498	4,9439	ja.						
						parent	5,0487	5,9575	der soll stehen bleiben	ne?"".					
						parent	6,5951	6,9795	guck mal-						
						parent	7,1512	8,0363	den gruenen-						
VP062	AC	3	Action1	7,28	12,69	parent	7,0568	8,7543	jetzt kannst du diese becher-						
						parent	9,3876	13,4021	die hier stehen in den grossen blauen packen. guck mal	den gruenen in den blauen-					

VP003	AC	3	Action2	13,367	14,497	parent	12,7194	13,5943	grn in blau-			
						parent	14,308	14,9987	gelb in grn-			
VP003	AC	1	Action2	6,1	8,42	parent	0,97194	8,0957	guck mal schatz	die sind unterschiedlich gro§ die becher	siehst du das? das ist ein kleiner becher	der kann in den gr§eren becher rein und dieser gr§ere becher kann dann in den becher rein-
VP031	AC	1	Action2	12,32	13,75	parent	12,656	13,9303	und dann wird das immer kleiner.			
VP031	AC	3	Action2	6,36	7,35	parent	6,0325	7,2051	mit immer kleiner-			
VP038	AC	1	Action2	9,64	11,33	parent	10,0347	10,2881	warte.			
						parent	10,6398	11,0671	guck mal hier.			
VP038	AC	3	Action2	14,99	16,38	parent	13,8164	15,2798	da muss der gruene rein.			
						parent	15,633	17,7702	da muss der gelbe rein-			
VP051	AC	3	Action2	11,26	12,49	parent	10,9435	11,6889	dann nehmen wir den-			
VP051	AC	1	Action2	21,36	23,23	parent	21,0957	22,1395	und den gelben-			
						parent	22,4425	23,4526	wieder da rein.			
VP047	AC	3	Action2	5,85	6,31							
VP047	AC	1	Action2	4,87	5,68							
VP026	AC	1	Action2	6,55	7,1							
VP012	AC	3	Action2	9,2	10,63	parent	6,0558	9,2001	der gruene becher kommt in den blauen	becher.der ist kleiner als der blaue.		
						parent	9,4197	12,3366	der gelbe kommt in den gruenen becher.	der ist noch kleiner.		
VP012	AC	1	Action2	15,97	18,07	parent	15,989	18,1754	und dann kommt der gelbe in den gruenen-			
VP044	AC	3	Action2	11,98	13,31	parent	12,3485	14,146	und packen da alle kleineren rein.			
VP009	AC	1	Action2	16,43	17,87							
VP009	AC	3	Action2	10,99	12	parent	10,4153	11,2021	der grne-			
VP030	AC	1	Action2	16,856	20,556	parent	17,5721	18,5132	den gelben-			
						parent	19,7634	20,8712	mit dem roten in den-			
VP032	AC	1	Action2	8,53	9,47							
VP036	AC	1	Action2	14,66	19,96	parent	13,5416	16,9056	ja	du darfst das gleich auch ausprobieren. und was ist das fuer ne farbe? weisst du das?.		
						parent	17,8838	18,6319	gelb	ne?.		
VP036	AC	3	Action2	6,63	8,1	parent	7,1069	8,0653	der gelbe in den-			
VP062	AC	1	Action2	9,7	11,58	parent	9,5339	10,1399	da rein.			
						parent	10,5767	11,5873	den gelben-			
VP062	AC	3	Action2	13,33	15,1	parent	9,3876	13,4021	die hier stehen in den grossen blauen pa	den gruenen in den blauen-		
						parent	14,1583	15,8879	den gelben in den gruenen-			

VP003	AC	3	Action3	16,267	17,437	parent	16,5183	17,7155	und rot in gelb.		
VP003	AC	1	Action3	9,12	11,09	parent	9,1664	10,5478	dann kann man den noch da rein stellen.		
VP031	AC	1	Action3	14,45	15,68	parent	15,0036	16,5911	kennen wir ja von zu hause	ne?.	
VP031	AC	3	Action3	8,12	9,2	parent	8,0671	9,1972	und zum schluss der kleinste.		
VP038	AC	1	Action3	12,69	14,76	parent	12,8182	13,4264	und? weiter?.		
						parent	13,656	14,4248	gucke mal. und-		
VP038	AC	3	Action3	18,34	19,61	parent	17,9386	19,2727	und da muss der-		
VP051	AC	3	Action3	13,37	15,07	parent	13,5744	15,6967	und dann zum schluss den kleinen.		
VP051	AC	1	Action3	24,97	28,08	parent	24,5917	25,5793	und dann den roten.		
						parent	26,017	29,0697	ganz zum schluss. den ganz kleinen kann man dann auch noch reinstellen	ne?.	
VP047	AC	3	Action3	6,85	7,47	parent	4,0845	6,9575	die du zu hause auch hast	ne? die kann man ineinander stecken.	
VP047	AC	1	Action3	6,84	7,57						
VP026	AC	1	Action3	7,89	8,51	parent	7,5904	8,1894	und der gelbe-		
VP012	AC	3	Action3	12,15	14,68	parent	9,4197	12,3366	der gelbe kommt in den gruenen becher. der ist noch kleiner.		
						parent	12,8294	15,8455	und der rote ist der kleinste von allen. der kommt da noch oben drauf.		
VP012	AC	1	Action3	18,74	20,85	parent	19,6769	21,682	und der rote in den gelben.		
VP044	AC	3	Action3	14,03	15,56	parent	12,3485	14,146	und packen da alle kleineren rein.		
						parent	14,4695	14,8826	niedlich.		
VP009	AC	1	Action3	18,99	19,71	parent	18,2148	19,3775	und der kann da rein-		
VP009	AC	3	Action3	12,8	14,12	parent	13,0182	13,9268	und der kleinste-		
VP030	AC	1	Action3	12,486	15,226	parent	13,0732	13,7886	den roten-		
						parent	14,6118	15,1562	in den-		
VP032	AC	1	Action3	11,33	12,71	parent	11,4992	12,4993	und dann der rote.		
VP036	AC	1	Action3	19,96	22,33	parent	20,3637	23,5927	und einen roten becher	guck mal	die packen wir jetzt alle so zusammen.
VP036	AC	3	Action3	8,59	9,94	parent	8,9022	9,9855	gruenen und der rote in den-		
VP062	AC	1	Action3	12,48	14,36	parent	12,7625	14,4687	und als letztes den kleinen roten.		
VP062	AC	3	Action3	15,77	17,95	parent	14,1583	15,8879	den gelben in den gruenen-		
						parent	16,8116	18,2275	und den roten in den-		

A.1.2 *Looming*

VP					Looming	start	end	parent					Naming	Attention ge	Others	infant				exp				acoustic			
VP007	1	AC	1	Looming	1	3340	5570	1	26.647.902	34.113.606	da ist salz drin-		1			1	52.194.607	57.560.582	babygerÃ¤usc	0				1	26.647.902	37.146.548	salzsteuer!
								0	34.930.167	46.595.329	siehste-					0				0				0	37.146.548	46.478.678	salzstreuer!
								0	46.595.329	56.510.718	das weisse zeug dadrin-					0				0				0			
VP052	2	AC	3	Looming	1	11060	11680	0								1	113.208.134	11680	(babylaute)	0				0			
VP002	3	AC	3	Looming	1	3980	5530	1	38.338.322	46.996.089	guck mal hanna-			1		0				0				1	52.903.152	70.293.104	poltern
								0	51.467.791	55.497.843	guck mal hier.					0				0				0			
VP025	4	AC	3	Looming	0			0								0				0				0			
VP023	5	AC	3	Looming	1	9810	11240	0								0				0				1	105.167.493	11.421.541	poltern
VP025	6	AC	1	Looming	1	3480	8960	1	2.484.607	35.226.582	luca. guck mal hier.			1		0				0				1	32.095.113	4.283.308	salz wird gestreut!
								0	36.170.265	47.666.038	da guck mal	da ist was drin.				0				0				0	48.159.112	95.632.341	salz wird gestreut!
								0	52.298.663	59.161.812	guck mal.					0				0				0			
VP025	7	AC	1	Looming	1	26520	28990	1	253.714.406	267.802.619	guck mal	wie geht das raus?.		1		0				0				0			
								0	27.080.925	275.619.859	hm?.					0				0				0			
VP008	8	AC	3	Looming	0			0								0				0				0			
VP040	9	AC	1	Looming	0			0								0				0				0			
VP010	10	AC	1	Looming	1	5240	9010	1	51.716.088	6.164.803	guck mal was da drin ist.			1		1	74.295.829	78.821.839	babylaute	0				1	61.672.831	83.181.756	salzstreuer!
VP014	11	AC	1	Looming	1	7000	13640	1	68.952.094	75.358.697	das sind-			1		0				0				1	75.358.697	9.127.018	gerÂŠusche
								0	9.127.018	98.306.941	jerome.					0				0				0	107.147.782	126.566.047	gerÂŠusche
								0	113.764.438	119.593.397	hallo.					0				0				0			
								0	12677.61	144.390.168	hey da sind lÂšcher und das muss man umdrehen-					0				0				0			
VP023	12	AC	1	Looming	1	4510	5470	0								0				0				0			
VP023	13	AC	1	Looming	1	27280	28750	1	284.266.038	293.188.131	jetzt probieren wir lieber nicht.				1	1	269.222.872	279.719.203	babylaute	0				0			
VP018	14	AC	3	Looming	0			0								0				0				0			
VP052	15	AC	1	Looming	1	4860	5680	0								0				0				1	51.158.732	54.154.158	geraeusch
VP014	16	AC	3	Looming	1	10880	11860	0								0				0				0			
VP028	17	AC	1	Looming	1	10992	12142	1	107.100.299	115.421.866	hoe	hoerst du das?.		1		0				0				1	111.256.056	131.330.764	salzstreuer wird geschuettelt!
VP028	18	AC	1	Looming	1	27302	35712	1	269.724.473	278.646.565	willst du mal versuchen?.			1		0				1	251.022.394	288.254.973	text	0			
								0	298.792.385	30.222.396	so.					0				0	319.724.988	325.387.085	text	0			
								0	312.089.735	336.024.965	hast du zu der frau geguckt	ne? die steht gar nicht hinter dem vorhang verflucht.				0				0	333.829.226	356.220.247	text	0			
								0	349.786.046	356.134.458	ja	ich glaube auch.				0				0				0			
VP006	19	AC	1	Looming	1	6330	9480	1	52.829.197	68.693.818	das ist ein salzstreuer-		1			0				0				1	46.296.706	67.527.301	salzstreuer !
								0	68.693.818	79.192.464	und dadrin-					0				0				0	72.659.973	78.849.345	salzstreuer !#
								0	79.192.464	91.467.195	da befindet sich das salz-					0				0				0	87.728.573	97.308.792	salzstreuer !
								0	91.467.195	97.856.724	guck					0				0				0			
VP001	20	AC	3	Looming	1	2540	4880	0								0				0				1	22.692.228	42.180.244	salz wird gestreut!
VP001	21	AC	1	Looming	1	2930	8630	1	32.875.114	40.209.289	guck mal.			1		0				0				1	41.647.362	65.375.575	salz wird gestreut; rascheln!
								0	77.167.778	85.508.605	was machen wir damit?.					0				0				0	65.375.575	73.141.172	klopfen auf tisch (kind)#
								0								0				0				0	73.141.172	107.654.937	salz wird gestreut!
VP001	22	AC	1	Looming	1	18830	23790	1	187.494.739	191.886.405	lachen				1	0				0				1	194.681.102	200.070.874	klappern#
								0								0				0				0	214.842.841	218.236.401	auf tisch klopfen#
								0								0				0				0	228.816.323	235.403.822	salz wird gestreut!
VP001	23	AC	1	Looming	1	52440	53430	0								0				0				0			
VP005	24	AC	3	Looming	0			0								0				0				0			
VP056	25	AC	3	Looming	0			0								0				0				0			
					19			12					2	8	2	4				1				12			
					76,00%			48,00%					16,67%	66,67%	16,67%	16,00%								48,00%			

					Looming			Parent						Naming	Attention get	Others	Infant				exp				acoustic							
VP052	1	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP005	2	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP028	3	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP014	4	AC	1	Looming	1	3640	11390	1	4034,4678	5146,0939	schau mal	da sind becher.		0	1	0	0				0				0							
								0	7184,0752	7665,7799	ja.			0	0	0									0							
								0	10820,3955	12747,2142	schau mal	die haben verschiedene grÃ¶ÃŸen.		0	0	0									0							
VP014	5	AC	1	Looming	1	14740	19200	1	14636,9786	15804,1861	so vor kann i	ne?.		0	0	1	1	14488,7618	14803,7226	babylaut	1	16285,8908	20324,7992	text	0							
								0	17119,6104	17545,7338	ach so.			0	0	0									0							
								0	18379,4534	18787,0496	mhm.			0	0	0									0							
VP006	6	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP018	7	AC	3	Looming	1	10520	11360	1	10072,676	10975,8366	ja.			0	1	0	0				0				0							
VP025	8	AC	1	Looming	1	5620	13090	1	5000,985	6997,7004	& luca	guck mal	schau mal.	0	1	0	0				1	10290,7908	13468,4343	text	1	7116,9073	8010,959	becher #				
								0	7504,3297	7951,3556	schau mal.			0	0	0									0							
								0	8070,5624	9769,2606	ach so	die muss ich dran bleiben. ach so.		0	0	0									0							
								0	10652,1715	11382,3137	luca	guck mal.		0	0	0									0							
								0	13051,2102	13557,8395	okay.			0	0	0									0							
VP002	9	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP002	10	AC	1	Looming	1	19110	20620	1	18311,9524	19321,6634	oh jetzt haben wir es kaputt gemacht			0	1	0	1	19316,5895	19555,0639	(babylaut)	1	19478,955	19905,1647	text	0							
								0	19910,2386	20554,6271	guck mal	da		0	0	0									0							
VP002	11	AC	1	Looming	1	21730	22700	0						0	0	0	0				0				1	22121,9164	23014,9271	becher klappern!				
VP002	12	AC	1	Looming	1	24420	27430	1	25218,5815	26010,1137	guck mal hie	maeuschen.		0	1	0	1	27327,9385	27596,8565	(babylaut)	0				1	27084,3901	27327,9385	becher klappern!				
								0						0	0	0									0	26497,2105	26908,1984	geraeusch				
VP056	13	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP008	14	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP006	15	AC	1	Looming	1	5200	9290	1	4730,4719	5688,4939	also pass auf-			0	1	0	0				1	6576,4167	9053,2539	text	0							
								0	5688,4939	6693,2486	guck mal-			0	0	0																
								0	9076,6203	9543,9481	ok.			0	0	0																
VP005	16	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP040	17	AC	3	Looming	1	390	3130	1	144,6465	527,0514	&hm			0	0	1	0				1	1631,7765	4128,0303	text	1	155,2689	643,8973	klappern	acoustic	888,2115	1557,42	klappern
VP028	18	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP007	19	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP052	20	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP001	21	AC	1	Looming	1	24070	25630	1	24114,5497	25026,8092	und dann-			0	0	1	0				0				0							
VP018	22	AC	1	Looming	1	1750	4790	1	827,5954	2836,0721	jonas	guck mal. da	wie dein Turm zuhause.	0	1	0	0				1	2836,0721	6853,0255	text	1	1861,6626	2617,3271	gerÃ¤usch				
								0	3770,7098	5162,7233	ach so	das passt	ja.	0	0	0																
VP056	23	AC	3	Looming	0	No Looming		1						0	0	0	0				0				0							
VP040	24	AC	1	Looming	1	45698	47360	0						0	0	0	1	46967,0603	47360	babylaut	0				0							
VP001	25	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP025	26	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP010	27	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP008	28	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
VP023	29	AC	3	Looming	1	6440	7480	1	7321,9967	8718,7846	guck mal	das ist der kleinste.		0	1	0	1	1094,6505	6710,9019	babylaute	0				1	5605,1115	6740,0017	mit den bechern klopfen!				
VP007	30	AC	3	Looming	0	No Looming		0						0	0	0	0				0				0							
VP010	31	AC	1	Looming	0	No Looming		0						0	0	0	0				0				0							
					13			12						0	8	3	5				6				6							
					41,94%			92,31%						0,00%	66,67%	25,00%	38,46%				46,15%				46,15%							

A.2 contingency corpus

A.2.1 *Questionnaire*

iCub Studie August 2011

Fragebogen zum Experiment

* Erforderlich

Fanden Sie die Interaktion mit iCub *

	1	2	3	4	5	
langweilig?	○	○	○	○	○	interessant?

Hat Ihnen iCub gefallen? *

	1	2	3	4	5	
nicht gefallen	○	○	○	○	○	sehr gefallen

Fanden Sie iCub freundlich? *

	1	2	3	4	5	
nicht freundlich	○	○	○	○	○	sehr freundlich

Wenn iCub ein Kind wäre, was schätzen Sie, wie alt er dann wäre? *

Hatten Sie den Eindruck, dass iCub Ihre Demonstrationen verstanden hat? Wenn nicht, woran denken Sie, dass es gelegen haben könnte? *

Erinnern Sie sich bitte an die Situationen, in denen Sie iCub die Lampe und den Klingelknopf erklärt haben. Haben Sie einen Unterschied in den beiden Situationen festgestellt? *

Gehen Sie davon aus, dass iCub selbstständig agiert hat? *

	1	2	3	4	5	
nicht selbstständig	○	○	○	○	○	komplett selbstständig

Fanden Sie das Verhalten von iCub menschlich? *

	1	2	3	4	5	
nicht menschlich	○	○	○	○	○	sehr menschlich

Welche Aufgaben halten Sie für relativ einfach und unproblematisch für iCub? *

Welche Aufgaben halten Sie für besonders schwierig für iCub? *

Sind Sie mit der Kommunikation mit Kindern vertraut? *

	1	2	3	4	5	
nicht vertraut	○	○	○	○	○	sehr vertraut

Sind Sie mit künstlichen Kommunikationspartnern (Computern, Robotern, ECAs) vertraut? *

	1	2	3	4	5	
nicht vertraut	○	○	○	○	○	sehr vertraut

Haben Sie Kinder? *

□ ja

□ nein

Wie alt sind Sie? *

□

Was ist Ihr höchster Abschluss? *

□

In Informatik? *

□ ja

□ nein

Sind Sie *

□ männlich?

□ weiblich?

Inwieweit treffen die folgenden Aussagen auf Sie zu? *

Bitte kreuzen Sie auf der Skala die Antwort an, die am ehesten Ihrer Einschätzung entspricht! Bitte in jeder Zeile ein Kästchen ankreuzen! Ich...

	trifft überhaupt nicht zu	trifft eher nicht zu	weder noch	eher zutreffend	trifft voll und ganz zu
... bin eher zurückhaltend, reserviert.	○	○	○	○	○
... schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	○	○	○	○	○

	trifft überhaupt nicht zu	trifft eher nicht zu	weder noch	eher zutreffend	trifft voll und ganz zu
… bin bequem, neige zur Faulheit.	○	○	○	○	○
… bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	○	○	○	○	○
… habe nur wenig künstlerisches Interesse.	○	○	○	○	○
… gehe aus mir heraus, bin gesellig.	○	○	○	○	○
… neige dazu, andere zu kritisieren.	○	○	○	○	○
… erledige Aufgaben gründlich.	○	○	○	○	○
… werde leicht nervös und unsicher.	○	○	○	○	○
… habe eine aktive Vorstellungskraft, bin phantasievoll.	○	○	○	○	○

In den nun folgenden Fragen bitten wir Sie, die Aussagen mit 1 „trifft gar nicht zu“ bis 5 „trifft voll zu“ zu bewerten.

Ich denke, ich möchte das System regelmäßig benutzen. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich fand das System unnötig komplex. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich denke, dass das System einfach zu benutzen ist. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich denke, dass ich die Unterstützung einer Person mit technischem Verständnis brauchen würde um das System zu benutzen. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich fand, dass die verschiedenen Funktionen des Systems gut integriert waren. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich fand, dass das System inkonsistent erscheint. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich kann mir vorstellen, dass die meisten Menschen sehr schnell lernen würden mit dem System umzugehen. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich fand das System schwerfällig in der Benutzung. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich fühlte mich sicher im Umgang mit dem System. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

Ich musste vieles lernen, bevor ich anfangen konnte, das System zu benutzen. *

	1	2	3	4	5	
trifft gar nicht zu	○	○	○	○	○	trifft voll zu

A.3 TECHNICAL DETAILS

A.3.1 *Hardware*

All processes used for the tutoring spotter system, including the face-tracking and the object-tracking systems, run on two Dell Latitude D630 Laptops (Intel Core 2 Duo T7500, 2,2 GHz, 4GB RAM).
A Logitech QuickCam Pro 9000 2.0 Mp webcam was used for the face and object tracking systems.
A Microsof Kinect sensor was used for 3D skeleton-tracking.
The iCub humanoid robot was developed as part of an EU project. Except the built-in PC104 computer, it can use clusters of external computers in an distributed, scalable way, connected by the YARP middleware.
An Apple iSight webcam was used for the Ackachan system.
Two Sony High Definition 1080i HDR-FX1 camcorders and two Sanyo Xacti HD1010 1080i camcordes were used for recording the experiments.

A.3.2 *Software*

Ubuntu Linux 10.04 was used as the operatings system.
For the Kinect sensor, the OpenNI, Nite and Primesense software were used.
For the analysis of the data, Matlab 2008b and SPSS v19 were used.
The annotation tools used were the ELAN, Interact, and Praat.
For video processing, FFMpeg andFinal Cut Pro were used.
Netbeans 6.2.1 was used for development.
For the gaze detection, iceWing (with custom extensions) was used.
For the face detection system, FaceAPI was used.
The iCub robot was running the YARP middleware and software modules from the iCub subversion repository.
For tracking the AR markers, the ARToolkit software was used.

A.3.3 *Statistical tests*

In the work presented in this thesis, the hypotheses where tested with either a Student's t-test or an one-way ANOVA.
The t-test is null if the null hypothesis is supported, it is testing if the test statistic follows a Student's t distribution. This test is used to compare the given data with a normal distribution as well as if two sets of data are significantly different from each other.
The one-way ANOVA (analysis of variance) is comparing the variances of data sets. The one-way ANOVA is a generalisation of the t-test but, in the case of two groups, it is the same as a t-test.

BIBLIOGRAPHY

[1] Salience (neuroscience); salience (language). URL http://en.wikipedia.org/wiki/Salience_(neuroscience);http://en.wikipedia.org/wiki/Salience_(language).

[2] Interact software. URL http://www.mangold-international.com/en/products/interact.html.

[3] R. Achanta and S. Suesstrunk. Saliency detection using maximum symmetric surround. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2653–2656. IEEE, 2010.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE, 2009.

[5] M. Argyle and M. Cook. *Gaze and mutual gaze.* Cambridge University Press, 1976.

[6] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34, 2009. ISSN 1943-0604.

[7] A. Bangor, P.T. Kortum, and J.T. Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[8] S. Baron-Cohen. The eye direction detector (edd) and the shared attention mechanism (sam): Two cases for evolutionary psychology. *Joint attention: Its origins and role in development*, pages 41–59, 1995.

[9] J.B. Bavelas, L. Coates, and T. Johnson. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580, 2002.

[10] H. Benedict. Early lexical development: Comprehension and production. *Journal of Child Language*, 6(02):183–200, 1979.

[11] A.E. Bigelow and S.A.J. Birch. The effects of contingency in previous interactions on infants' preference for social partners. *Infant Behavior and Development*, 22(3):367–382, 1999.

[12] K. Bloom. Evaluation of infant vocal conditioning. *Journal of experimental child psychology*, 27(1):60–70, 1979.

[13] R.J. Brand, D.A. Baldwin, and L.A. Ashburn. Evidence for'motionese': modifications in mothers' infant-directed action. *Developmental Science*, 5(1):72–83, 2002.

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

[15] J. Brooke. Sus: a "quick and dirty" usability scale. 1996. *PW Jordan, B. Thomas, BA Weerdmeester and AL McClelland.*

[16] H. Brugman and A. Russel. Annotating multimedia/multimodal resources with elan. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2065–2068. Citeseer, 2004.

[17] G. Butterworth. Origins of mind in perception and action. *Joint attention: Its origins and role in development*, pages 29–40, 1995.

[18] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, et al. Integration of action and language knowledge: A roadmap for developmental robotics. *Autonomous Mental Development, IEEE Transactions on*, 1(99), 2009. ISSN 1943-0604.

[19] M. Castrillón, O. Déniz, and M. Hernández. The encara system for face detection and normalization. *Pattern Recognition and Image Analysis*, pages 176–183, 2003.

[20] B. Chanda and D.D. Majumder. *Digital image processing and analysis*. PHI Learning Pvt. Ltd., 2004. ISBN 8120316185.

[21] H.H. Clark and S.E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

[22] J. Cohn and B. Beebe. Sampling interval affects time-series regression estimates of mother-infant influence. *Infant Behavior and Development*, Abstracts Issue, 13:317, 1990.

[23] G. Csibra. Recognizing communicative intentions in infancy. *Mind & Language*, 25(2):141–168, 2010. ISSN 1468-0017.

[24] G. Csibra and G. Gergely. Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance*, 21, 2005.

[25] G. Csibra and G. Gergely. Natural pedagogy. *Trends in Cognitive Sciences*, 13(4):148–153, 2009. ISSN 1364-6613.

[26] K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents–implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, 2002.

[27] A. Eliëns. *Object-oriented software development*. Addison Wesley, 1995.

[28] B. Estigarribia and E.V. Clark. Getting and maintaining attention in talk to young children. *Journal of child language*, 34(04):799–814, 2007. ISSN 1469-7602.

[29] V. Evans and M. Green. *Cognitive linguistics*. Edinburgh University Press Edinburgh, 2006.

[30] T. Falck-Ytter, G. Gredebäck, and C. Von Hofsten. Infants predict other people's action goals. *Nature neuroscience*, 9(7):878–879, 2006.

[31] T. Farroni, M.H. Johnson, M. Brockbank, and F. Simion. Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition*, 7(6):705–718, 2000.

[32] I. Fasel, G.O. Deak, J. Triesch, and J. Movellan. Combining embodied models and empirical research for understanding the development of shared attention. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 21–27, 2002.

[33] I. Fasel, N. Butko, and J. Movellan. Modeling the embodiment of early social development and social interaction: Learning about human faces during the first six minutes of life. In *Society for Research in Child Development Biennial Meeting*, 2007.

[34] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59 (2):167–181, 2004.

[35] K.; Foth K. Fischer, K.; Lohan. Levels of embodiment: Linguistic analyses of factors influencing hri. In *In Proceedings of HRI'12*, 2012.

[36] P. Fitzpatrick, G. Metta, P. Fitzpatrick, and G. Metta. Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 361(1811):2165–2185, 2003.

[37] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action-initial steps towards artificial cognition. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 3, pages 3140–3145. IEEE, 2003.

[38] J.R. Flanagan and R.S. Johansson. Action plans used in action observation. *Nature*, 424(6950):769–771, 2003.

[39] A. Fogel and A. Garvey. Alive communication. *Infant Behavior and Development*, 30(2):251–257, 2007. ISSN 0163-6383.

[40] K. Foth, I. Schröder, and W. Menzel. A transformation-based parsing technique with anytime properties. In *In Proceedings of the 4th International Workshop on Parsing Technologies*. Citeseer, 2000.

[41] G. Gergely and J.S. Watson. The social biofeedback theory of parental affect-mirroring: The development of emotional self-awareness and self-control in inf. *International Journal of Psycho-Analysis*, 77:1181–1212, 1996. ISSN 0020-7578.

[42] G. Gergely and J.S. Watson. Early socio-emotional development: Contingency perception and the social-biofeedback model. *Early social cognition: Understanding others in the first months of life*, pages 101–136, 1999.

[43] L.J. Gogate, L.E. Bahrick, and J.D. Watson. A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71(4):878–894, 2000. ISSN 1467-8624.

[44] L.J. Gogate, L.H. Bolzani, and E.A. Betancourt. Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word–object relations. *Infancy*, 9(3):259–288, 2006. ISSN 1532-7078.

[45] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007. ISSN 1049-5258.

[46] J.A. Hawkins. *A performance theory of order and constituency*, volume 73. Cambridge Univ Pr, 1994.

[47] J.S. Herberg. Audience-contingent variation in action demonstrations for humans and computers. *Cognitive Science: A Multidisciplinary Journal*, 32(6):1003–1020, 2008.

[48] K. Hirsh-Pasek and R.M. Golinkoff. *The origins of grammar: Evidence from early language comprehension*. The MIT Press, 1999.

[49] I. Ibarretxe-Antuñano. Linguistic typology in motion events: Path and manner. *Anuario del seminario de Filología Vasca 'Julio de Urquijo'-International Journal of Basque linguistics and Philology*.

[50] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine*

Intelligence, IEEE Transactions on, 20(11):1254–1259, 1998. ISSN 0162-8828.

[51] W. James. *The principles of psychology, Vol I.* Henry Holt and Co, 1890.

[52] R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan. Eye–hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917, 2001.

[53] K. Kaye. *The mental and social life of babies: How parents create persons*. University of Chicago Press, 1982. ISBN 0226428478.

[54] H. Keller, A. Lohaus, S. Völker, M. Cappenberg, and A. Chasiotis. Temporal contingency as an independent component of parenting behavior. *Child Development*, 70(2):474–485, 1999.

[55] T.A. Kindermann. Natural peer groups as contexts for individual development: The case of children's motivation in school. *Developmental psychology*, 29(6):970, 1993. ISSN 1939-0599.

[56] M. Knoll and L. Scharrer. Acoustic and affective comparisons of natural and imaginary infant-, foreigner-and adult-directed speech. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[57] K.L. Koay, K. Dautenhahn, S.N. Woods, and M.L. Walters. Empirical results from using a comfort level device in human-robot interaction studies. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 194–201. ACM, 2006.

[58] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4): 219–27, 1985.

[59] P.K. Kuhl. Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120, 2007.

[60] J. Lee, J.F. Kiser, A.F. Bobick, and A.L. Thomaz. Vision-based contingency detection. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 297–304. ACM, 2011.

[61] M.T. Legerstee. *Infants' sense of people: precursors to a theory of mind*. Cambridge Univ Pr, 2005. ISBN 0521818486.

[62] Frank Loemker. icewing–a graphical plugin shell, 2005. URL http://icewing.sourceforge.net.

[63] K. Lohan, K. Rohlfing, K. Pitsch, J. Saunders, H. Lehmann, C. Nehaniv, K. Fischer, and B. Wrede. Tutor spotter: Proposing a feature set and evaluating system. *International Journal of Social Robotics*, 2012.

[64] K.S. Lohan, A.L. Vollmer, J. Fritsch, K. Rohlfing, and B. Wrede. Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations? In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. IEEE, 2009.

[65] K.S. Lohan, A.L. Vollmer, J. Fritsch, K. Rohlfing, and B. Wrede. Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations? *IEEE International Workshop on Social Signal Processing*, 2009.

[66] K.S. Lohan, S. Gieselmann, A.L. Vollmer, K Rohlfing, and B. Wrede. Does embodiment effect tutoring behavior? 2010.

[67] K.S. Lohan, K. Pitsch, K. Rohlfing, K. Fischer, J. Saunders, H. Lehmann, C. Nehaniv, and B. Wrede. Contingency allows the robot to spot the tutor and to learn from interaction. In *ICDL-EpiRob 2011*, 2011.

[68] M. Lohse, M. Hanheide, K. Pitsch, K.J. Rohlfing, and G. Sagerer. Improving hri design by applying systemic interaction analysis (sina). *Interaction Studies*, 10(3):298–323, 2009. ISSN 1572-0373.

[69] M. Lopes, A. Bernardino, J. Santos-Victor, K. Rosander, and C. von Hofsten. Biomimetic eye-neck coordination. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–8. IEEE, 2009.

[70] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, volume 81, pages 674–679, 1981.

[71] G. Markova and M. Legerstee. Contingency, imitation, and affect sharing: Foundations of infants' social awareness. *Developmental psychology*, 42(1):132, 2006.

[72] D.J. Matatyaho and L.J. Gogate. Type of maternal object motion during synchronous naming predicts preverbal infants' learning of word–object relations. *Infancy*, 13(2):172–184, 2008. ISSN 1532-7078.

[73] J.S. McCartney and R. Panneton. Four-month-olds' discrimination of voice changes in multimodal displays as a function of discrimination protocol. *Infancy*, 7(2):163–182, 2005.

[74] G. Metta, P. Fitzpatrick, and L. Natale. Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, 3(1): 43–48, 2006.

[75] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori. The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 50–56. ACM, 2008.

[76] T. Minato, M. Shimada, S. Itakura, K. Lee, and H. Ishiguro. Does gaze reveal the human likeness of an android? In *Development and Learning, 2005. Proceedings. The 4th International Conference on*, pages 106–111. Ieee, 2005.

[77] T. Minato, M. Shimada, S. Itakura, K. Lee, and H. Ishiguro. Evaluating the human likeness of an android by comparing gaze behaviors elicited by the android and a person. *Advanced robotics: the international journal of the Robotics Society of Japan*, 20 (10):1147, 2006.

[78] C.J. Mondloch, T.L. Lewis, D.R. Budreau, D. Maurer, J.L. Dannemiller, B.R. Stephens, and K.A. Kleiner-Gathercoal. Face perception during early infancy. *Psychological Science*, 10(5):419, 1999. ISSN 0956-7976.

[79] C.E. Moore and P.J. Dunham. *Joint attention: Its origins and role in development.* Lawrence Erlbaum Associates, Inc, 1995.

[80] J.R. Movellan. An infomax controller for real time detection of social contingency. In *Development and Learning, 2005. Proceedings. The 4th International Conference on*, pages 19–24, 2005.

[81] D. Muir and K. Lee. The still-face effect: Methodological issues and new applications. *Infancy*, 4(4):483–491, 2003. ISSN 1525-0008.

[82] Y. Nagai and K.J. Rohlfing. Can motionese tell infants and robots' what to imitate? In *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*, pages 299–306, 2007.

[83] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[84] Y. Nagai, C. Muhl, and K.J. Rohlfing. Toward designing a robot that learns actions from parental demonstrations. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3545–3550, 2008.

[85] A. Needham, J.F. Cantlon, and S.M. Ormsbee Holley. Infants' use of category knowledge and object attributes when segregating objects at 8.5 months of age. *Cognitive psychology*, 53(4):345–360, 2006.

[86] C.L. Nehaniv and K. Dautenhahn. Like me?-measures of correspondence and imitation. *Cybernetics and Systems*, 32(1):11–51, 2001. ISSN 0196-9722.

[87] M. Okanda and S. Itakura. Development of contingency: How infants become sensitive to contingency? Kyoto, Japan, 2006. Proc. of the XVth Biennial International Conference on Infant Studies.

[88] O. Pascalis and D.J. Kelly. The origins of face processing in humans: Phylogeny and ontogeny. *Perspectives on Psychological Science*, 4(2):200, 2009. ISSN 1745-6916.

[89] U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini. An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1668–1674, 2010.

[90] K. Pitsch and B. Koch. How infants perceive the toy robot pleo: An exploratory case study on infant-robot-interaction. 2010.

[91] K. Pitsch, H. Kuzuoka, Y. Suzuki, L. Sussenbach, P. Luff, and C. Heath. "the first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in hri. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 985–991. IEEE, 2009.

[92] K. Pitsch, A.L. Vollmer, J. Fritsch, B. Wrede, K. Rohlfing, and G. Sagerer. On the loop of action modification and the recipient's gaze in adult-child interaction. In *Gesture and Speech in Interaction*, Poznan, Poland, 24/09/2009 2009.

[93] S.M. Pruden, K. Hirsh-Pasek, and R.M. Golinkoff. Current events: How infants parse the world and events for language. *Understanding events: from perception to action*, 4:160, 2008.

[94] C.T. Ramey and L.L. Ourth. Delayed reinforcement and vocalization rates of infants. *Child Development*, 42(1):291–297, 1971.

[95] D. Regan and K.I. Beverley. Looming detectors in the human visual pathway. *Vision Research*, 18(4):415–421, 1978. ISSN 0042-6989.

[96] A. Riegler. The role of anticipation in cognition. In *AIP Conference Proceedings*, pages 534–544. IOP INSTITUTE OF PHYSICS PUBLISHING LTD, 2001.

[97] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.

[98] K.J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006. ISSN 0169-1864.

[99] C.K. Rovee and D.T. Rovee. Conjugate reinforcement of infant exploratory behavior* 1. *Journal of Experimental Child Psychology*, 8(1):33–39, 1969.

[100] C. Rovee-Collier. Learning and memory in infancy. *J. D. Osofsky (Ed.), Handbook of infant development*, 2:139–168, 1987.

[101] H. Sacks, E.A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.

[102] M. Scaife and J.S. Bruner. The capacity for joint visual attention in the infant. *Nature*, 1975.

[103] E.A. Schegloff. *Sequence organization in interaction: A primer in conversation analysis I.* Cambridge Univ Pr, 2007. ISBN 0521532795.

[104] A. Senju and G. Csibra. Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9):668–671, 2008. ISSN 0960-9822.

[105] H.J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[106] A. Slater, P.C. Quinn, D.J. Kelly, K. Lee, C.A. Longmore, P.R. McDonald, and O. Pascalis. The shaping of the face space in early infancy: Becoming a native face processor. *Child Development Perspectives*, 4(3):205–211, 2010. ISSN 1750-8606.

[107] A. Stefanowitsch and A. Rohde. The goal bias in the encoding of motion events. *Motivation in Grammar*, pages 249–268, 2004.

[108] D.N. Stern. A micro-analysis of mother-infant interaction. behavior regulating social contact between a mother and her 3 1/2 month-old twins. *Journal of the American Academy of Child Psychiatry*, 10(3):501–517, 1971.

[109] T. Striano, A. Henning, and D. Stahl. Sensitivity to social contingencies between 1 and 3 months of age. *Developmental Science*, 8 (6):509–518, 2005. ISSN 1467-7687.

[110] H. Sumioka, Y. Yoshikawa, and M. Asada. Development of joint attention related actions based on reproducing interaction contingency. In *7th IEEE International Conference on Development and Learning, 2008. ICDL 2008*, pages 256–261, 2008.

[111] H. Sumioka, Y. Yoshikawa, and M. Asada. Reproducing interaction contingency toward open-ended development of social actions: Case study on joint attention. *Autonomous Mental Development, IEEE Transactions on*, 2(1):40–50, 2010. ISSN 1943-0604.

[112] L. Talmy. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3:57–149, 1985.

[113] L. Talmy. Path to realization: A typology of event conflation. In *Proceedings of the seventeenth annual meeting of the Berkeley Linguistics Society*, volume 17, pages 480–519. Berkeley: Berkeley Linguistic Society, 1991.

[114] L. Talmy. *Toward a cognitive semantics, Vol. 1: Concept structuring systems.* the MIT Press, 2000.

[115] F. Tanaka, A. Cicourel, and J.R. Movellan. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 104(46):17954, 2007.

[116] P. Ten Have. *Doing conversation analysis*. Sage Publications Ltd, 2007.

[117] M. Tomasello. Joint attention as social cognition. *Joint attention: Its origins and role in development*, pages 103–130, 1995.

[118] M. Tomasello and M.J. Farrar. Joint attention and early language. *Child development*, 57(6):1454–1463, 1986. ISSN 0009-3920.

[119] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(05):675–691, 2005. ISSN 0140-525X.

[120] E. Tronick. The structure of face-to-face interaction and its developmental functions. *Sign Language Studies*, 1978.

[121] J.K. Tsotsos. *A computational perspective on visual attention*. MIT Press, 2011.

[122] L.A. Van Egeren, M.S. Barratt, and M.A. Roach. Mother–infant responsiveness: Timing, mutual regulation, and interactional context. *Developmental psychology*, 37(5):684, 2001.

[123] T.N. Vikram, K.S. Lohan, M. Tscherepanow, K. Rohlfing, and B. Wrede. Can state-of-the-art saliency systems model infant gazing behavior in tutoring situations? 2011.

[124] T.N. Vikram, M. Tscherepanow, and B. Wrede. A visual saliency map based on random sub-window means. In *proceedings of Iberian Conference on Pattern Recognition and Image Analysis*, pages 33–40, 2011.

[125] T.N. Vikram, M. Tscherepanow, and B. Wrede. A random center surround bottom up visual attention model useful for salient region detection. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 166–173. IEEE, 2011.

[126] A.L. Vollmer, K.S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. Rohlfing, and B. Wrede. People modify their tutoring behavior in robot-directed interaction for action learning. In *International Conference on Development and Learning*, volume 8, Shanghai, China, 04/06/2009 2009. IEEE, IEEE.

[127] A.L. Vollmer, K.S. Lohan, J. Fritsch, B. Wrede, and K. Rohlfing. Which motionese parameters change with children's age? In *Cognitive Development Society, VI Biennial Meeting*, San Antonio, Texas, USA, 2009.

[128] A.L. Vollmer, K. Pitsch, K.S. Lohan, J. Fritsch, K. Rohlfing, and B. Wrede. Developing feedback: How children of different age contribute to an interaction with adults. In *International Conference on Development and Learning*, 2010.

[129] J.S. Watson. Smiling, cooing, and 'the game'. *Merrill-Palmer Quarterly: Journal of Developmental Psychology*, 18(4), 1972.

[130] J.S. Watson. Contingency perception in early social development. *Social perception in infants*, pages 157–176, 1985.

[131] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-l1 optical flow. In *Proceedings of the British machine vision conference*, 2009.

[132] R. Williams. *Keywords: A vocabulary of culture and society.* Oxford University Press, USA, 1985.

[133] A.L. Woodward. Infants' understanding of the actions involved in joint attention. *Joint attention: Communication and other minds: Issues in philosophy and psychology*, pages 110–128, 2005.

[134] B. Wrede, K. Rohlfing, M. Hanheide, and G. Sagerer. Towards learning by interacting. *Creating Brain-Like Intelligence*, pages 139–150, 2009.

[135] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka. Precision timing in human-robot interaction: coordination of head movement and utterance. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 131–140. ACM, 2008.

[136] C. Yu, M. Scheutz, and P. Schermerhorn. Investigating multimodal real-time patterns of joint attention in an hri word

learning task. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 309–316. ACM, 2010.

[137] P. Zukow-Goldring and M.A. Arbib. Affordances, effectivities, and assisted imitation: Caregivers and the directing of attention. *Neurocomputing*, 70(13-15):2181–2193, 2007. ISSN 0925-2312.