In: Proceedings IEEE Ro-Man 2014, 36-42. [PRE-PRINT]
The final version is available at: http://rehabilitationrobotics.net/ro-man14/

# When a robot orients visitors to an exhibit. Referential practices and interactional dynamics in real world HRI*

Karola Pitsch, Sebastian Wrede

***Abstract*— A basic task for robots interacting with humans consists in guiding their focus of attention. Existing guidelines for a robot's multimodal deixis are primarily focused on the speaker (talk-gesture-coordination, handshape). Conducting a field trial with a museum guide robot, we tested these individualistic referential strategies in the dynamic conditions of real-world HRI and found that their success ranges between 27% and 95%. Qualitative video-based micro-analysis revealed that the users experienced problems when they were not facing the robot at the moment of the deictic gesture. Also the importance of the robot's head orientation became evident. Implications are drawn as design guidelines for an *interactional* account of modeling referential strategies for HRI.**

## I. Introduction

In a range of robot applications and scenarios, a basic task consists for the robot in guiding the human user's focus of attention. For example, a museum guide robot needs to orient visitors to a particular exhibit when providing information about it. To guide the user's attention and to establish co-orientation, the robot has at its disposal a range of communicational resources depending on its own embodiment, such as e.g. talk, head orientation/gaze, gestures etc. While such orienting behavior is currently used in a range of settings, rarely the concrete design of the robot's referential practices is detailed nor their success or failure evaluated. Only recently, a small number of studies begin to explore 'robot deixis' investigating the user's interpretation of different hand shapes or (combinations of) modalities in lab experiments [14, 13, 5]: the robot produces a deictic reference and the user's perception of the target is evaluated through questionnaires. While these studies provide important information for the choice and design of modalities, little is know about the success/failure of such strategies in situated real-time human-robot-interaction (HRI).

Investigating video recordings of visitors interacting with a museum guide robot (small-size humanoid NAO) in an arts museum, we observe that visitors, who attempt to follow the robot's explanations, not always manage to successfully orient to the corresponding exhibit. Visitors happen to orient to one painting while the robot is offering information about a different exhibit or they visibly search for the corresponding referent. Such observations point to the relevance of gaining a better understanding of the *interactional* dimension of a robot's referential practices in human-robot-interaction.


Figure 1. NAO as museum guide. Multimodal deixis and user orientation.

In this paper, we present analysis of the ways in which visitors to a museum interpret a guide robot's referential practices and react to them in real-time. We investigate:

1. How successful are basic referential strategies when being deployed within the dynamics of real-time interaction in the wild? What are the conditions for their success/failure?

2. Which implications can we draw for designing a robot's referential practices taking into consideration the dynamic process between presenter and recipient?

Initial results of a combined quantitative and qualitative evaluation are presented. They constitute the basis for design considerations of robot deixis with the long-term goal of providing building blocks and interactional models for technical systems to engage in sequential action with humans.

## II. Referential Practice & Multimodal Deixis

When designing referential practices for technical systems a range of issues need to be considered. Considerable effort has been placed, mainly in the field of robotics, on developing ways for technical systems to *recognize* a human's reference to objects while in the area of embodied agents a strong focus has been on modeling its *production* [e.g. 1, 9]. For a guide robot's referential practices, a set of aspects needs consideration:

*Choice of modalities*: Psycholinguistic accounts have explored either the interplay of talk and gesture or the role of gaze in referential practices, which has been at the basis of modeling deictic procedures in virtual agents. Following a 'trade off hypothesis', gesture has been considered as a fallback strategy if verbal referencing becomes too cumbersome or the distance to the target increases [e.g. 19, 12, 2] – yet, the effectiveness of such strategies has rarely been tested in systematic studies. In a different vein, e.g. [7]

*The authors acknowlegde the financial support from the Volkswagen Foundation (Dilthey Fellowship 'Interaction & Space', K. Pitsch), CITEC (project: Interactional Coordination & Incrementality in HRI) and CoR-Lab.

Karola Pitsch is with the Cognitive Interaction Technology Centre of Excellence (CITEC), Applied Linguistics & HRI group, Bielefeld University, Germany, karola.pitsch@uni-bielefeld.de (author of this text).

Sebastian Wrede is with the Cognitive Interaction Technology Centre of Excellence (CITEC), Cognitive Systems Engineering group, Bielefeld University, Germany, sebastian.wrede@uni-bielefeld.de (project collaborator, responsible for robot system during study).

have modeled the role of gaze for establishing 'joint attention' to an object in both robot systems and virtual agents. Most recently, a small number of HRI studies have begun to explore the users' perception of different combinations of modalities. [16] show that, for a robot, a combined 'head and arm' movement is more successful in indicating a location than single modalities or a cross-body gesture. [14] investigated the user's perception of a robot's deictic gestures and found that the accuracy of pointing gestures (80-90% under different conditions (e.g. normal, distant, clustered objects, noise)) ranges between those where the agent touches the object and sweeping/grouping gestures, and is thus relevant for the design of a guide robot's conduct.

*Timing in the intra-personal coordination of modalities*: Once we assume 'multimodal packages' of talk, gesture and gaze [3], the speaker has to intra-personally coordinate the different modalities. The general principle 'gesture precedes the lexical affiliate' [15] has been confirmed in a range of further studies. [6] measured the gestural phrase for deictic gestures of human narrators and found that the gesture starts 1000 ms before the lexical affiliate and ends 366 ms afterwards. The authors found also that they gaze to the target in about 80% of the cases. Based on these observations, they developed a model for pointing gestures in HRI (including: 'gesture onset precedes the lexical affiliate', 'gaze directed to the target') and found that deictic gestures predicted information recall in a narration scenario.

*Orientation to the recipient*: In addition to the previous individualistic accounts, also the orientation to the recipient plays a role. For the example of a robot giving route directions, [8] designed a robot's pauses between sentences based on previously measuring the time that a listener needs to understand and process a robot's sentence in a similar situation. In the experiment, the best ratings (in questionnaires) were indeed achieved with a robot using gestures and (fixed) listener-modeled pause duration even though their length exceeded the common pause timing.

*Interaction between speaker and recipient*: From a Conversation Analytic point of view the question arises how the speaker's multimodal deictic reference is co-produced together with the recipient. Sequential interactional structures involving repair of referential practices are shown [e.g. 4].

Thus, in social interaction referential practices are part and parcel of complex multimodal interactional dynamics. However, at the present state, we have no empirical information about how the existing *individualistic* approaches to modeling referential practices perform under the condition of real-time interaction nor do *interactional* models for robotic reference production seem to exist to our knowledge (but see [17] for reference resolution). Therefore, we undertake a first step exploring how users interpret a guide robot's (individualistically designed) referential practices under the condition of real-time interaction in the real world.

## III. Design of the Robot's Referential Practices

A humanoid NAO robot was deployed as guide in an arts museum. It was positioned in the corner of a room and set up to get in contact with visitors, to provide information about paintings and artists, and to finally close the encounter.

### A. Interactional conditions for referring actions

The robot's explanation lasted for 2 minutes and was designed to cover three structurally different cases for the referential actions. In addition to the opening and closing of the encounter, the explanation was structured in four parts (Fig. 2: tier 'Topics/Activities') during which the robot gave information about painting 3 (Case A = Ref-1), all artists in the room (Ref-2), some more details of painting 3 (Case B = Ref-3) and painting 6 (Case C = Ref-4). For each topic, the robot produced a deictic reference in its first utterance and then provided more information (Fig. 2: 'Deixis/Content').

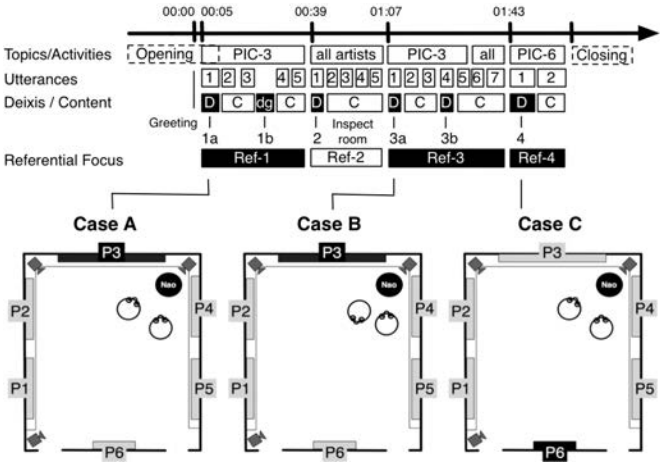



Figure 2. Structure of robot's explanation, interactional conditions during referential actions and room layout (robot stands between painting 3 and 4).

**Case A – Referring presumably attentive recipients to a nearby location close to their line of sight**: After getting in contact with the visitors, NAO refers to painting 3 (P3) situated behind the robot (Ref-1). Its talk is structured in five utterances with the first containing the deictic reference (see fig. 2). As Case A occurs directly after the opening of the encounter, visitors were expected to focus at the robot.

**Case B – Orienting recipients with presumably diverse states of participation**: After explaining P3, the robot attempts to invite the visitors to inspect the other paintings in the room (Ref-2). This creates a situation, in which visitors are expected to be oriented to different parts of the room. Thus, the robot's next referring action – i.e. Case B (Ref-3, to specific features of P3) – will have to deal with variability in the users' state of participation and conduct.

**Case C – Referring presumably attentive recipients to a distant location at the opposite side of the room**: The robot refers to painting 6 (P6) at the opposite side of the room (Ref-4). As the robot's previous explanation was focused on P3, visitors were expected to be oriented towards the robot.

### B. Robot's intra-personal coordination of modalities

For designing the robot's referential practice, we assumed a multimodal perspective in which talk, gesture and – partly – head orientation are considered as *one* 'multimodal package' [3]. In each case, the robot's utterance contains, at its beginning, the deictic expression "this"/"here"/"over there" coupled with a pointing gesture and is followed either by the referent (e.g. "this painting") or by a localization and the referent (e.g. "here at the left hand side the yellow person's

leg"). Timing of the different resources follows the 'gesture precedes the lexical affiliate'-principle [15, 14]. In our case, the gesture phase extends from (i) the onset about 1 to 2 seconds before the lexical affiliate, over (ii) the peak held (at least) during the verbal deictic expression, before it is (iii) retracted. The original design of talk-gesture-coordination shows variance in the trial due to limited computing power.[1]

For the coordination of the robot's 'talk & gesture'-packages with its head orientation, the autonomous system was faced with two competing demands: (i) to guide – in concert with talk and gesture – the recipient's focus of attention; (ii) to keep the contact with the recipients. Due to limited processing power at the moment of the study (see section 4), we had to use a very basic design, which resulted in the principle that the robot directed its head to the nearest visitor at the end of each utterance. While this is not ideal from an interactional point of view, this design choice reflects an additional issue. Due to the material design of current humanoid robots, the robot's head orientation needs (iii) to allow the system to orient itself in the environment and is thus not generally 'free' for the design of the interaction. Only for the more complex interactional situation in Case B (different visitor orientation) also the robot's head orientation was included in the deictic reference.

**Case A**: Case A consists of the basic referential design.

| | | Deix | Ref | | |
|---|---|---|---|---|---|
| darf ich ihnen etwas | über | dieses | bild | erzählen? | (1.0) |
| may i you something | about | this | painting | explain? | |
| onset | | peak | | retract to home | |

Figure 3. D-1a: Robot's gesture with gaze directed to visitors.

**Case B**: As the situation including visitors with potentially diverging states of participation was expected to be interactionally more demanding, the robot's gaze was included in the multimodal package of the referring action and also directed to the target location (@O). The referential structure was designed to explore a stepwise reference resolution process when referring to specific *features* of an object from an initial verbally rather vague description to a precise depiction at the end of the utterance.

[1] Although rarely reported in HRI studies, to precisely synchronize different communicational resources in an autonomous system currently constitutes a separate research challenge [e.g. 13].

| | Deix | Loc | | |
|---|---|---|---|---|
| sie sehen | hier | bei den umrissen | dokupil arbeitet | schnell und skizzenhaft |
| you see | here | at the countours | dokupil works | fast and sketchy |
| onset | | peak | retract to | arm up/down |
| | @O | | | |

| | Deix | Loc | Ref | |
|---|---|---|---|---|
| so wie | hier | auf der linken seite | das bein der | gelben figur |
| just like | here | on the left hand side | the leg of the | yellow person |
| onset | peak | | retract to | home |
| | @O | | | |

Figure 4. D-3a and D-3b: Robot's gestures incl. gaze aligned with gesture.

**Case C**: As the target was located at a distant location at the opposite side of the room, the robot's gesture was made more prominent through the extended duration of its peak.

| | | Deix | | Ref | |
|---|---|---|---|---|---|
| | haben sie | dort vorne | schon | das loch im bild | von w.d. gesehen? |
| | did you | overthere | already | the hole in the painting by | w.d. see? |
| onset | peak | | | | retract |

Figure 5. D-4: Robot's gesture and gaze directed to visitors.

## IV. Robot System

As the robotic platform should provide intuitive access for lay-users and be robust enough to be deployed in the real world, a humanoid NAO robot (Aldebaran, version 3+, 52 cm high) was used. It was positioned on a small table in the corner of the exhibition room and set up to offer information to visitors using talk, gesture and head orientation.

The system was configured to run autonomously. Yet, to circumvent the platform's hardware limitations at the time of the study, the robot's perception was realized through a Vicon infrared tracking system, processed in a dedicated person tracking module and integrated with the robot system using a robotics middleware [20]. The visitors wore marked hats, so that their position in space and head orientation could be detected and classified (directed to robot vs. elsewhere; proximity to robot in three zones). Information about the visitors' position in space and head orientation was used to adjust the robot's conduct to the visitor(s) at three levels: (1) for opening a focused encounter, (2) reacting on the visitors' general loss of interest, and (3) directing its head to the nearest visitor at the end of each utterance.[2]

The robot's multimodal utterances consisted of preconfigured, synchronized speech-gesture behaviors (for Case B: speech-gesture-gaze behaviors), which occurred in a fixed order during the robot's explanation. These different behaviors (multimodal utterances, gaze strategies) of the robot were activated through a coordination module following a dual dynamics-inspired arbitration scheme.

[2] With a more advanced system the design of issues (3) needs refining.

## V. Study and Data

To investigate the users' reactions to the robot's conduct, we conducted a field trial at the Bielefeld arts museum [10].

### A. Study

The NAO robot was placed in the corner of a regular exhibition room (5x6 m) and, due to its small size, positioned on a small table. Ordinary visitors to the museum were asked, when entering the adjacent room, if they were willing to participate in a study and, if so, to wear hats equipped with markers. They were informed that they would be video-recorded, but were not given any information about the nature of the study nor the function of the hats/markers or how to handle the robot. They could ask any questions afterwards and have their recording deleted if they felt uncomfortable with it. Also they could disengage from the robot and walk away at any time, and there was always the possibility of other visitors entering or leaving the room.

The study took place on 7 days (6 hours each day). Each HRI-trial lasted for 2 minutes and was recorded with 3 HD video cameras. We obtained recordings of the visitors' talk, gestures, head orientation/gaze, spatial conduct, and facial expressions. The data from the infrared cameras (Vicon) used for the robot's perception was also stored for offline analysis.

### B. Data

During the experiment, 260 HRI-trials with visitors of different group sizes were recorded. For analysis presented in this paper, a sub-corpus of 64 visitors taking part in 38 trials was used: From the original 260 episodes we discarded all trials (although relevant for other issues) in which users familiar with the system or large user groups came along. Also, those episodes were disregarded, in which the system's performance showed unforeseen behavior (e.g. long pauses between utterances, highly unsynchronized talk-gesture) or the recording quality was problematic (e.g. visitors blocking the camera). Only visitors who wore the marked hats and remained until the end of an episode were considered.

## VI. Analytical Method

The success/failure of the robot's referential practices is evaluated with a combined qualitative-quantitative approach.

In a first step, the visitors' reactions to the robot's deictic reference are quantified based on manual annotation (Elan) of the visitor's videotaped conduct. The following features – derived from qualitative data analysis – were annotated: (i) Visitor's focus of attention (to robot [@R], to other visitor [@V], to which painting [@P1, … @P6],) or as being in motion [≈≈]. (ii) Referential structure, i.e. the stretches of the robot's talk during which the referential focus (Ref-1, -2, -3, -4) established by the robot's deixis is valid (Fig. 2). Some visitors may additionally comment verbally on their ability to follow the robot's reference, which is not considered here.

In a second step, we aim at gaining a better understanding of the interactional micro-processes and reasons why, in some cases, the visitors have difficulties in orienting to the painting indicated by the robot. We use a qualitative method that provides insights into the sequential structure of the interaction and which is based on Conversation Analysis (CA) and its multimodal extensions [18]. This allows us to investigate the interrelationship between the robot's and the visitor's actions and how they respond to each other on the structural level. Important is the aim to reconstruct the participant's view ("member's perspective"), i.e. we investigate the user's understanding of the robot's actions and to which extent they treat them as meaningful relevant actions at particular moments in time. Here, case analyses are undertaken and consist of manual analysis, i.e. repeated inspection of video-data and transcribing/annotating the interaction to uncover the timing and relationship of the actions. The goal is to find the structural organization and how one action makes another one contingently relevant.

## VII. Quantitative Evaluation: Success and Failure

To investigate the success/failure of the robot's referring actions quantitative analysis was undertaken. Based on manual annotations of the video-recorded HRI trials, we counted whether a visitors' head (indicating focus of attention) orients to the referent signaled out by the robot during the corresponding referential timespan. Analysis reveals that visitors experience problems in correctly following the robot's deictic reference when they might not be oriented to the robot at the moment of its deictic production. The basic case of referring to an object close to the recipient's line of sight (case A) is successful in 89%, to a distant location (case C) in 79%. Visitors who are not necessarily oriented to the robot (case B) manage to follow the robot's orientation initially only in 26.5%; after the second deictic gesture another 9.3% manage to orient correctly, and only with the final precise verbal description of the relevant feature a success rate of 95.3% is achieved.

TABLE I. Referring actions: Success and Failure

| | **Case A** | **Case B** | **Case C** |
|---|---|---|---|
| **Location of Referent** | Behind R | Behind R | Opposite side |
| **Visitors' orientation** | To robot | Variable | To robot |
| **Success/Failure** | 89% (57/64) | 26.5 % (17/64)<br>9.3 % (6 /64)<br>59.3 % (38/64)<br>----------<br>95.3% (61/64) | 79.6% (51/64) |

From these findings the question arises how and under which conditions users are able to follow the robot's orienting hints in the concrete interaction.

## VIII. Qualitative Evaluation: Understanding the Interactional Dimension of Referential Practices

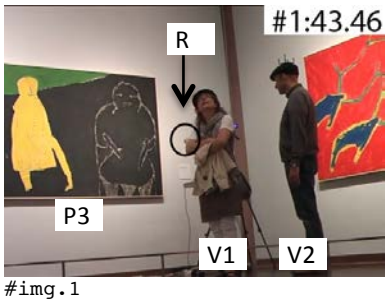
Given the visitors' varying success in following the robot's deictic reference, we aim at understanding the interactional reasons for these problems. To reveal the conditions for successful referential acts, video-based micro-analysis of two problematic cases (B and C) is presented.

### A. Dealing with varying states of participation: Securing vs. orienting attention

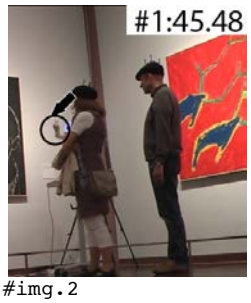
In authentic situations of HRI (as opposed to precisely designed laboratory experiments), the users' state of participation and focus of attention is not always predictable. This becomes particularly visible in our data in case C where the users initially inspect different paintings in the room. Given

the low success rate of 26.5% for the robot's first attempt to orient the user (which amounts to 95% at the end of the explanation), this provides a good starting point for analysis. Here, we will provide detailed analysis of one case (VP 222).

(1) *Deictic reference secures the visitor's attention*: In this fragment, two visitors are oriented to different features of the room when the robot refers to the painting P3 and suggests: "you see here (.) at the contours dokupil's way of working is fast and sketchy" (underlined: deictic gesture). While, at this moment, the male visitor (V2) is gazing to the robot and thus able to see its gesture, the female visitor (V1) inspects the ceiling (#1). She reacts to the robot's conduct (i.e. a combination of deictic reference + address term + sound of robot's arm movement) by shifting her focus of attention to it (#2). As this re-orientation takes time, the robot's deictic gesture is already finished and its arm engaged in an iconic up-down motion (parallel to saying "dokupil's way of working is fast and sketchy") once she looks at the robot. Thus, for V2 the robot's reference to the painting functions as a device to *secure her attention*, but then no further orientational hints to the referent are available for her.



#img.1



#img.2

```
01 R-ver:  sie sehen |hier, (.) bei den|
           you see    here      at  the
   R-gest:           |@P3 .............
   V1-gaz: @ceiling ...................

02 R-ver:  umrissen;|(0.6)|dokupil |arbeitet
           contours        dokupil  works
   R-gest: .........|
   V1-gaz: .........|≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈|@R
                    |#1              |#2

03 R-ver:  schnell und |skizzen|haft;|3.0|
           fast    and  sketchy
   R-gest:             |hand-retr    |
   V1-gaz:             |searching    |@R
                                     |#3 #4 #5 #6
```

(2) *Searching for the referent and checking with the robot*: As V2 does not find any information about the referent, she scans the room (#3), re-orients to the robot (#4) but again finds no hint to the referent, and continuous to search (#5, #6) while the robot's explanation continuous.

#img.3

#img.4

#img.5

#img.6

(3) *Subsequent referential act*: As the robot's explanation proceeds, it produces a second deictic reference: "just like here on the left hand side". V1 reacts immediately: after "just like here on the" she is re-oriented to the robot (#7, #8) which displays a deictic gesture. Ultimately, V1 follows its direction (#9) and once the referent "yellow person's leg" is named, she is able to identify it and points it out to V2 (#10).

#img.7

#img.8

#img.9

#img.10

```
06 R-ver:  so wie    |hier auf der |linken
           just like here on  the  left hand
   R-gest:          |@P1 ..................
   V1-gaz: searching|≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈|@R.....
                    |#7              |#8

07 R-ver:  |sei|te (.) |das |BEIN
            side        the  LEG
   R-gest: |@P1-retract     |home
   V1-gaz:  ...|≈≈≈≈≈≈≈|@P1 ......
                        |#9

08 R-ver:  der    gelben figur;|(4.5)
           of the yellow person
   V1-gaz: ........................
   V1-act:                      |point@P1
                                |#10
```

Upshot: In real-world interactions, the visitors' state of participation is likely to be variable and their focus of attention is often not oriented to the robot when it produces a deictic reference. In such cases, the robot's deixis rather functions as *attention getter*, but then further orientational hints are required. To deal with such situations, the system would need to observe and understand that the initial deictic reference has not been successful in referring the visitor to the target. Thus, it would need to adjust the progression of its explanation and e.g. offer a repeated deictic reference.

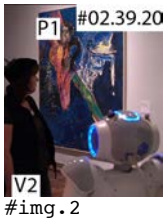
### *B. Interplay of communicational resources: Gaze & gesture*

When performing a deictic reference, the robot's head orientation generally fulfills several tasks: to guide (with talk and gesture) the recipient's focus of attention; to keep the contact with the recipients; to allow the system to orient itself in the environment (here: realized through external vision). Existing individualistic models of deictic reference production have mostly focused on talk and gesture and only rarely included the speaker's gaze. Once we attempt to use these models for real-time HRI, the question arises which impact the robot's gaze assumes. This is explored in case C (here: VP043) where the robot attempts to orient visitors to the opposite side of the room and uses a gaze strategy according to which its head is oriented to the nearest visitor.

(1) *Relevance of robot's head orientation*: At about 1'45'' min. in the robot's explanation, two female visitors are positioned vis-à-vis the robot. After having concluded its utterance, the robot attempts to adjust its face to the nearest

visitor, which results – due to insecurities of perception – in a series of head movements, towards V2 (#1, #2) and back to the original position facing towards image P1 (#3). This draws the visitors' attention to the robot who experience the system as being dynamic and the robot's head orientation therefore as a potentially relevant communicational means.

#img.1



#img.2

#img.3

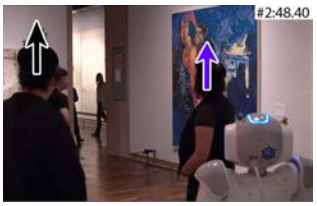
(2) *Different orientation of head and gesture:* When the robot refers to the new painting ("did you already see the painting by walter dahn over there?"), its extended arm points to P6 while its head is (in the robot's concept) oriented to V2 (#3). The users react to this double orientation: (a) V1 firstly turns to the painting (P1) located behind V2 (#4) following the robot's head orientation; then she rotates further to P6 (#5) following the robot's pointing gesture. (b) V2 turns in a whole body motion firstly to P6 (#4), afterwards to P1 (#5). This way, both V1 and V2 display their orientation to the robot's diverging orientation in head pose and gesture (#5).

```
01 R-ver:              |haben sie dort vorne
                        did   you over there
   R-gaz:  @P1|@V2     |@P1.............
   R-ges:     |@P6-on  |@P6-peak.........
           #1 |#2      |#3

02 R-ver:  schon  das loch im     bild     von
           already the hole in the painting by
   R-gaz:  ................................
   R-ges:  ................................

03 R-ver:  walter dahn gesehen? |(1.5) |(1.4)
           walter dahn see?
   R-gaz:  ........................ |
   R-ges: ........................ |P4-re
   V1-gaz:                        |@P6   |@P1
   V2-gaz:                        |@P1   |@P6
                                  |#4    |#5
```

#img.4



#img.5

(3) *Searching for the referent and checking with the robot*: In this situation of diverging referents, V1 re-orients to the robot (#6a). However, its arm is already retracted and does not provide any information to disambiguate the referent (#6b). Its head is still directed to V2/P1, which V1 interprets as a deictic reference to P1 by turning her head to P1 (#7).

```
04 R-ver:  |auch das ist ein stil|mittel      der
            also this is a   stylistic device of
   R-gest: |home ...
   V1-gaz: |@R...........        |@P1
   V2-gaz: |............         |
           |#6a+6b               |#7
```

#img.6a


#img.6b

```
05 R-ver:  MÜHLheimer freiheit
           <name of group of artists>
   V1-gaz: ........................
   V2-gaz: ........................
```

#img.7

Upshot: This fragment shows that visitors orient to a robot's diverging referential hints and subsequently follow both the direction indicated by gesture and by head orientation. If they are in doubt about the referent, they appear to return back to the robot and follow the orientational cues provided at that moment in time. As a similar conduct appeared in the first fragment (VP222), there seems to be a reoccurring practice of visitors to check back in case of doubt and both expect and provide occasion for the robot to deliver additional orientational hints. Thus, a robot would need to adjust its conduct to the visitors' hesitation and engage in a repair sequence providing information tailored to the user.

## IX. Summary and Implications

To investigate how users interpret a robot's referential strategies, a museum guide robot was placed in three structurally different situations when referring to an exhibit. The robot's deictic conduct was based on the current state of the art in HRI and ECAs using speaker-centered strategies. Opposed to the existing laboratory studies, these individualistic strategies were tested in the dynamic conditions of spontaneous real-world HRI. Given the discrepancy between the robot's *individualistic* strategies and the *interactional* requirements, we expected – in the light of interactional accounts in HHI – some visitors to fail when attempting to orient to the painting indicated by the robot. Indeed, quantitative evaluation confirmed that users experience problems, and in particular when they were not oriented to the robot at the moment when the deictic gesture was produced. Qualitative sequential micro-analysis of video-data provided insights into the nature of difficulties:

(1) When visitors were not oriented to the robot at the moment when it produced the multimodal deictic reference, the deixis served to *attract* the visitor's attention. However, when the visitor then looked at the robot, there was no visible orientational hint available any more as the robot's gesture was already retracted to home position.

(2) With regard to the multiplicity of communicational resources it turned out that, if gesture and head orientation do

not point in the same direction (e.g. because the robot attempts to look at visitors or uses its camera to orient itself in the environment), visitors get confused about the target.

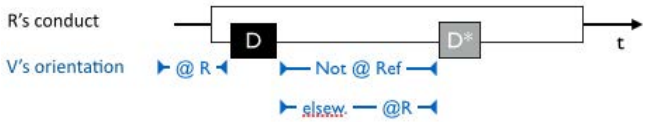
In several fragments, we observed a recurring practice: In case of doubt, users re-oriented to the robot and then followed the orientational hints available at that moment. Thus, the visitors' conduct offers the possibility (i.e. provides a structural provision) for the robot to 'repair' the problem and to provide additional information in a subsequent step.

From these observations a set of implications can be drawn.

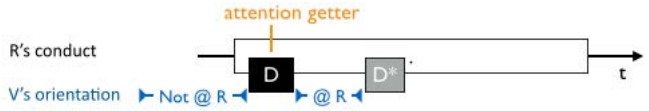
(A) For the *design of the robot's referential conduct*, two different, potentially combinable strategies can be used. The *first* strategy is based on rendering the robot's actions *more explicit*, i.e. longer extension of gestures, verbally more explicit description of the referent and its location. A *second* strategy attempts to enable the robot system to engage in *interactional coordination* and sequential organization. To do so, the robot would need to monitor the visitors' reactions to its deictic references (e.g. head orientation), and to interpret these, at particular moments, as success/failure and in case of failure provide a repair action. The following three micro-models would be a relevant starting point based on the robot's permanent monitoring of the user's focus of attention:

(1) If the user is oriented to the robot before/during the deictic reference (D), afterwards she should look at the target.


(2) If the user is oriented to the robot before/during the deictic reference (D) and afterwards she is not oriented to the referent (i.e. to the robot or elsewhere), the system needs to initiate a repair sequence and suggest anew the deictic reference (D*).




(3) If the user is not oriented to the robot before/during the deictic reference (D), it needs to first attract the visitor's attention e.g. with the deictic reference. Once the user's attention is secured, then the robot should repeat the deictic reference to the target.




This way, a small set of building blocks for situated interactional coordination will be provided that should allow an autonomous system to react appropriately without neglecting the flexibility and contingency of human interactional conduct [see also 11].

(B) To deal with such new sequential structures some *technical requirements* arise: the robotic architecture and the dialog system need to be based on incremental processing.

## X. Future Work

Future work will consist in exploring these design considerations in use with an autonomous robot system using its internal perception and an incremental architecture.

## Acknowledgment

The authors thank J.-C. Seele, L. Süssenbach, R. Gehle, L. Rix, L. Schröder and A. Amrhein for helping setting up the robot, conducting the study and tedious annotation work.

## References

[1] Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In HRI 2006 (pp. 297-304).

[2] de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. Topics in Cognitive Science, 232-248.

[3] Goodwin, C. (2000). Action and embodiment within situated human interaction. Journal of Pragmatics, 32(10), 1489-1522.

[4] Goodwin, C. (2003). Pointing as situated practice. In Pointing: Where language, culture and cognition meet (pp. 217-241). Mahwah, NJ: Lawrence Erlbaum.

[5] Hato, Y., Satake, S., Kanda, T., Imai, M., & Hagita, N. (2010). Pointing to space: Modeling of deictic interaction referring to regions. In HRI 2010 (pp. 301-308).

[6] Huang, C. M., & Mutlu, B. (2012). Modeling and evaluating narrative gestures for humanlike robots. In RSS 2013.

[7] Kaplan, F., & Hafner, V. (2004). The challenges of joint attention. In EpiRob 2004 (pp. 67-74).

[8] Okuno, Y., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2009). Providing route directions: Design of robot's utterance, gesture, and timing. In HRI 2009 (pp. 53-60).

[9] Pfeiffer, T. (2011). Understanding multimodal deixis with gaze and gesture in conversational interfaces. Aachen: Shaker.

[10] Pitsch, K., Gehle, R., & Wrede, S. (2013). Addressing multiple participants: A museum robot's gaze shapes visitor participation. In ICSR 2013 (pp. 587-588).

[11] Pitsch, K., Kuzuoka, H., Suzuki, Y., Süssenbach, L., Luff, P., & Heath, C. (2009). "The first five seconds". Contigent stepwise entry into an interaction as a means to secure sustained engagement. In RO-MAN 2009 (pp. 985-991).

[12] Piwek, P., Beun, R., & Cremers, A. (2008). 'Proximal' and 'distal' in language and cognition: Evidence from deictic demonstratives in dutch. Journal of Pragmatics, 40(4), 694-718.

[13] Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., & Joublin, F. (2012). Generation and evaluation of communicative robot gesture. International Journal of Social Robotics, 4(2), 201-217.

[14] Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In HRI 2014.

[15] Schegloff, E. A. (1984). On some gestures' relation to talk. In J. Heritage & J. M. Atkinson (Eds.), Structures of social action (pp. 266-296). Cambridge: Cambridge University Press.

[16] St Clair, A., Mead, R., & Mataric, M. J. (2011). Investigating the effects of visual saliency on deictic gesture production by a humanoid robot. In RO-MAN 2011 IEEE (pp. 210-216).

[17] Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2007). Natural deictic communication with humanoid robots. In Intelligent robots and systems, 2007. IROS 2007. IEEE/RSJ international conference on (pp. 1441-1448).

[18] J. Sidnell & T. Stivers (Ed.). The Handbook of Conversation Analysis. (2013). The handbook of conversation analysis. Chichester, West Sussex, UK: Wiley-Blackwel.

[19] Van Der Sluis, I., & Krahmer, E. (2001). Generating referring expressions in a multimodal context an empirically oriented approach. Language and Computers, 37(1), 158-176.

[20] Wienke, J., & Wrede, S. (2011). A middleware for collaborative research in experimental robotics. In System integration (SII), 2011 IEEE/SICE international symposium on (pp. 1183-1190).