Deixis: How to Determine Demonstrated Objects Using a Pointing Cone

Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, Ipke Wachsmuth

Collaborative Research Centre SFB 360
Faculty of Technology and Faculty of Linguistics
University of Bielefeld, 33594 Bielefeld, Germany
{akranste, tpfeiffe, ipke}@techfak.uni-bielefeld.de,
{andy.luecking, hannes.rieser}@uni-bielefeld.de

Abstract. We present an collaborative approach towards a detailed understanding of the usage of pointing gestures accompanying referring expressions. This effort is undertaken in the context of human-machine interaction integrating empirical studies, theory of grammar and logics, and simulation techniques. In particular, we attempt to measure the precision of the focussed area of a pointing gesture, the so-called pointing cone. The pointing cone serves as a central concept in a formal account of multi-modal integration at the linguistic speech-gesture interface as well as in a computational model of processing multi-modal deictic expressions.

1 Introduction

Research in cognitive science shows that deixis, pointing or demonstration is at the heart of reference. On the other side, the robust grounding of reference in situated human-machine communication is an open issue until now. In this paper we concentrate on pointing gestures in deictic expressions. Following McNeill (1992), we distinguish between abstract pointings and pointings into concrete domains. Here we focus on pointings into concrete domains co-occurring with verbal expressions.

In our research on human computer interfaces for natural interaction in Virtual Reality (VR) we employ an anthropomorphic agent called Max able to produce synchronised output involving synthetic speech, facial display and hand gestures (Kopp and Wachsmuth, 2004). Doing so, we focus on scenarios in the construction task domain, where a kit consisting of generic parts is used to construct models of mechanical objects and devices. A typical setting consists of a human user instructing a VR system represented by Max in aggregating composite objects. Speech and gesture are used to specify tasks and select relevant referents. To improve the communicative abilities of Max, he will be equipped with a natural pointing behaviour meeting the requirements of deictic believability (Lester *et al.*, 1999).

A central problem we are faced with is the vagueness of demonstration. The question is how to determine the focus of a pointing gesture. To deal with that, we establish in the course of a parameterisation of demonstration (Section 2) the concept of a pointing cone. For our ongoing empirical studies we developed novel empirical methods using tracking technology and VR simulations to collect and evaluate analytical data (Section 3). In Section 4 a multi-modal linguistic interface is described integrating the content of the verbal expression with the content of the demonstration determined *via* the pointing cone. The application of the pointing cone concept in computational models for (1) reference resolution and (2) the generation of multi-modal referring expressions embedded in our agent Max is outlined in Section 5. Finally, in Section 6 we discuss the trade-offs of our approach.

2 The Parameters of Demonstration

If we want to consider the multiple dimensions of deixis more systematically, then we must account for various aspects:

(a) Language is in many cases tied to the gesture channel via deixis. Acts of demonstration have their own structural characteristics. Furthermore, co-occurrence of verbal expressions and demonstration is neatly organised, it harmonises with grammatical features. Gestural and verbal information differ in content. This results from different production procedures and the alignment of different sensory input channels. The interaction of the differing information can only be described via a multi-modal syntax-semantics interface.

(b) We concentrate on two referential functions of pointing, *i.e. object-pointing* and *region-pointing*. If an act of pointing uniquely singles out an object, it is said to have *object-pointing* function; if the gesture fails to do so it is assigned *region-pointing* function. As shown in earlier studies (Lücking *et al.*, 2004), classifying referential functions needs clear-cut criteria for the function distinction.

(c) Pointing gestures are inherently imprecise, varying with the distance between pointing agent and referent. Pointing singles out a spatial area, but not necessarily a single entity in the world. To determine the set of entities delimited by a pointing gesture, we have to analyse which parameters influence the topology of the pointing area. As a first approximation we can model a cone representing the resolution of the pointing gesture. Empirical observations indicate that the concept of the pointing cone can be divided into two topologically different cones for object- and for region-pointing, with the former having a narrower angle than the latter.

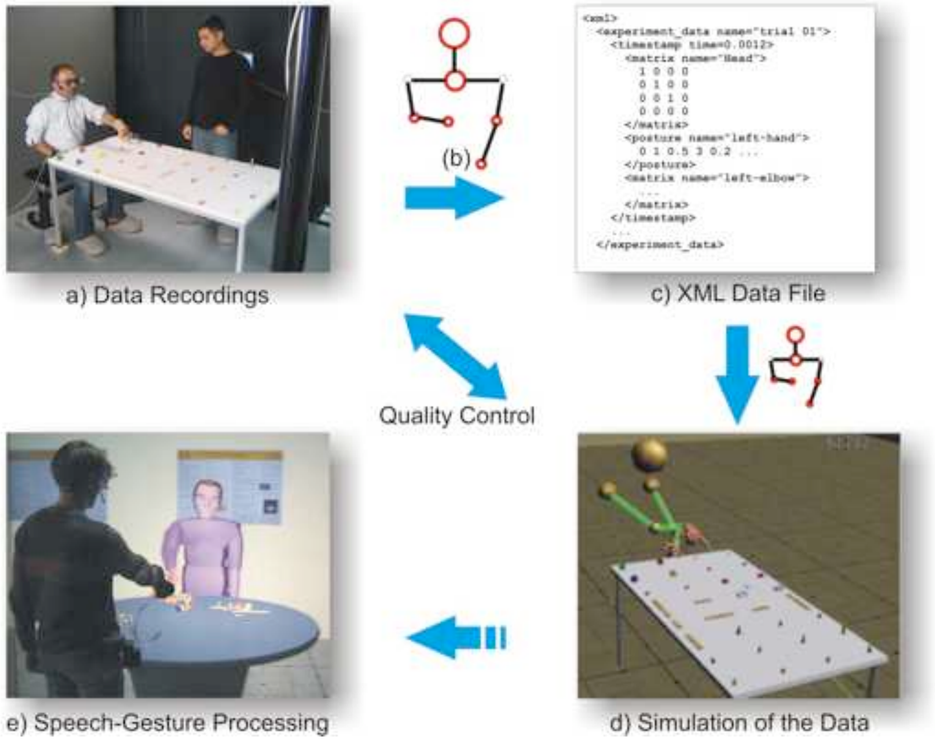
(d) Pointing gestures and speech that constitute a multi-modal utterance are time-shared. One point of interest, then, is whether there is a constant relationship in time between the verbal and the gestural channel. Our investigation of temporal *intra*-move relations is motivated by the synchrony rules stated in (McNeill, 1992). Since the so-called "stroke" is the meaningful phase of a gesture, from a semantic point of view the synchronisation of the pointing stroke and its affiliated speech matters most.

(e) With respect to dialogue, a further point of interest is whether pointings affect discourse structure. To assess those *inter*-move relations, the coordination of the gesture phases of the dialogue participants in successive turns has to be analysed. For instance, there is a tight coupling of the retraction phase of one agent and the subsequent preparation phase of the other suggesting that the retraction phases may contribute to a turn-taking signal.

To sum up, elaborating a theory of demonstration means at least dealing with the following issues: (a) the multi-modal integration of expression content and demonstration content, (b) assigning referential functions to pointing, (c) the pointing region singled out by a demonstration ("pointing cone"), (d) *intra*-move synchronisation, and (e) *inter*-move synchronisation.

3 Empirical Studies on the Pointing Cone

To address the issues named in the preceeding section we started to conduct several empirical studies in a setting where two subjects engaged in simple object identification games. One subject has the role of the "description-giver". She has to choose freely among the parts of a toy airplane lying on a table equally distributed, the pointing domain (Fig. 1a), and to refer to them. The other subject, in the role of the "object-identifier", has to resolve the description-givers reference act and to give feedback. Thus, reference has to be negotiated and established using a special kind of dialogue game.



```
<xml>
  <experiment_data name="trial 01">
    <timestamp time=0.0012>
      <matrix name="Head">
        1 0 0 0
        0 1 0 0
        0 0 1 0
        0 0 0 0
      </matrix>
      <posture name="left-hand">
        0 1 0.5 3 0.2 ...
      </posture>
      <matrix name="left-elbow">
        ...
      </matrix>
    </timestamp>
    ...
  </experiment_data>
```

Fig. 1. The description-giver is tracked using optical markers and data gloves (a). The data is integrated in a geometrical user model (b) and written to an XML file (c). For simulation the data is fed back into the model and visualised using VR techniques (d). The findings are transferred to improve the speech-gesture processing capabilities of the agent Max (e)

In a first study described in (Lücking *et al.*, 2004) the object identification games were recorded using two digital cameras, each capturing a different view of the scene. The annotations of the video data comprise speech, gesture phases, and the structure of the dialogue games in terms of dialogue moves. This study yields useful results concerning the temporal relations of pointing and speech both within a single dialogue move and between the moves of the dialogue participants. However, concerning the topology of the pointing cone no reliable results could be obtained based only the recorded video data.

3.1 Tracker-based Data Recording

To obtain more exact data concerning the pointing behaviour we use a marker-based optical tracking system for the body of the description-giver and data gloves for the fine-grained hand postures. The optical tracking system uses eight infrared cameras arranged in a cube around the setting to track optical markers each with a unique 3-dimensional configuration. A software module integrates the gathered information providing absolute coordinates and orientations. We track head and back of the description-giver to serve as reference points. Arms are tracked by two markers each, one for the elbow and one for the back of the hand. The hands are tracked using CyberGloves® measuring flexion and abduction of the fingers directly.

The information provided by both tracking systems (Fig. 1a) is integrated in a graph-based geometrical model of the user's posture (Fig. 1b). This is done in real-time using the VR frameworks Avango (Tramberend, 1999) and PrOSA (Latoschik, 2001). Special recording modules are attached to the geometric user model to make the recorded data available for annotation and stochastic analysis (Fig. 1c).

To test the experimental setting we run a preliminary study in November 2004 in which our primary concerns were the question of data reliability and the development of methods for analysing the data. The following section describes a simulative approach to support raters with visualisations of the collected data.

3.2 Simulation-based Data Evaluation

For the simulation we use VR techniques to feed the gathered tracking data (Fig. 1c) back into the geometric user model, forming now the basis of a graphical simulation of the experiment (Fig. 1d). This simulation is run in a CAVE-like environment, where the human rater is able to walk freely and inspect the gestures from every possible perspective. While doing so, the simulation can be run back and forth in time and thus, e.g., the exact time-spans of the strokes can be collected. To further assist the rater, additional features can be visualised, e.g., the pointing beam or its intersection with the table. For the visualisation of the subject we use a simple graphical model (Fig. 1d) providing only relevant information.

For a location independent annotation we created a desktop-based visualisation system where the rater can move a virtual camera into every perspective possible and generate videos to facilitate the rating and annotation process when the graphic machines for the real-time rendering are not available. Using the annotation software, these videos can be shown side-a-side in sync with the real videos and provide additional perspectives, e.g., seeing through the eyes of the instruction-giver.

3.3 Computation of Pointing Beam and Pointing Cone

The principal aim of collecting analytical data was to fix the topology of the pointing cone and to measure its size.

A pointing beam is defined by its origin and its direction, the pointing cone in addition by its apex angle. Therefore, to grasp the spatial constraints of pointing, one has to identify the anatomical anchoring of origin and direction in the demonstrating hand and to calculate the apex angle of the pointing cone.

There are four different anatomical parts (the three phalanxes of the index finger and the back of the hand) at disposition for the anchoring. To discriminate between them a hypothetical pointing beam is generated for each of them, see Fig. 2. We will choose the anchoring resulting in the least mean orthogonal distance over all successful demonstrations between the hypothetical pointing beam and the respective referent.

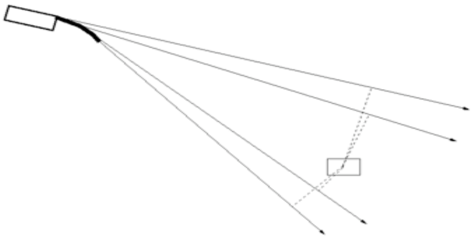


Fig. 2. Four hypothetical pointing beams anchored in different anatomical parts of the hand

Given the anchoring thus obtained, the calculation of the apex angle of the pointing cone can be done as follows: For each recorded demonstration the differing angle between the pointing beam and a beam with the same origin but directed to the nearest neighbour has to be computed. The computed angles decrease with the increasing distance between the demonstrating hand and the referent analogously to the perceived decreasing distance between the objects, see Fig. 3.

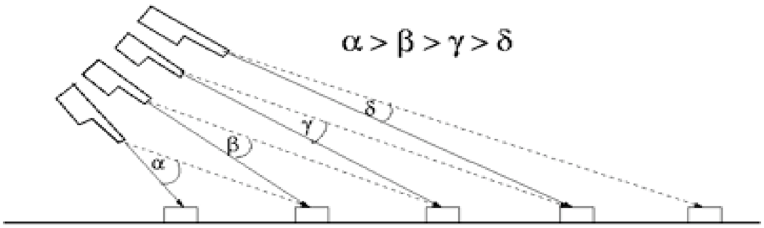


Fig. 3. The angles between the beams to the referent and the next neighbour decrease with the increasing distance to the referent. The dashed arrows represent the beams to the next neighbour

We pursue two strategies for the calculation of the apex angle. In one experimental setting the description-givers are allowed to use both, speech and gesture, to indicate the referent. Analysing this data, we have to search for the differing angle correlating with a shift to more discriminating verbally descriptions. This angle indicates the borderline of the resolution of pointing the description-givers manifests. In the other experimental setting the description-givers are bounded to gestures only. In this data we have to search for the differing angle correlating with the distance where the number of failing references exceeds the number of successful references. This angle indicates the borderline in the object density where the object-identifier cannot identify the referent by pointing alone.

We assume that these two borderlines will be nearly the same, with the former being a little bit broader than the latter due to the demonstrating agent's intention to ensure that the addressee is able to resolve the reference act. The corresponding angles define the half apex angle of the pointing cone of object-pointing.

A first assessment of the apex angle of this pointing cone using a similar calculation based on the video data recorded in our first studies resulted in a half apex angle between 6 and 12 degrees, see (Kühnlein and Stegmann, 2003) and (Kranstedt *et al.*, 2004). However, these results can be only taken as a rough indication.

To establish the apex angle of the pointing cone of region-pointing we have to investigate the complex demonstrations including verbal expressions referring to objects in the distal region. We hope that we can determine the contrast set from which the referent is distinguished by analysing the attributes the description-giver uses to generate the definite description. The location of the objects in the contrast set gives a first impression of the region covered by region-pointing.

In the next section, we introduce a formal attempt to integrate gestural deixis, in particular the pointing stroke, in linguistic descriptions, aiming at a theoretical model of deixis in reference (Rieser, 2004).

4 A Multi-modal Linguistic Interface

4.1 Complex Demonstrations: Object and Restrictor Demonstration

Objects originating from pointing plus definite descriptions are called complex demonstrations ("CDs"). The pointing stroke is represented as "↘" indicating the start of the stroke in the signal and hence its scope. (1) presents a well-formed CD "↘this/that yellow bolt" embedded into a directive as against (1') which we consider as being non-well-formed.

(1) Grasp ↘this/that yellow bolt. (1') *Grasp this/that yellow bolt.

A unified account of CDs will opt for a compositional semantics to capture the information coming from the verbal and the visual channel. CDs are considered as definite descriptions to which demonstrations add content either by specifying an object independently of the definite description or by narrowing down the description's restrictor. We call the first use "object demonstration" and the second one "restrictor demonstration".

Hypothetically then, demonstrations (a) act like verbal elements in providing content, (b) interact with verbal elements in a compositional way, (c) may exhibit forward or backward dynamics depending on the position of ↘.

4.2 Interpretation of CDs

The central problem is how to interpret demonstrations. This question is different from the one concerning the ↘'s function tied to its position in the string. We base the discussion on the following examples showing different empirically found ↘ positions and turn first to "object demonstration":

(2) Grasp ↘ this/that yellow bolt. (3) Grasp this/that ↘yellow bolt.

(4) Grasp this/that yellow ↘bolt. (5) Grasp this/that yellow bolt↘.

Our initial representation for the speech-act frame of the demonstration-free expression is

(6) $\lambda P\ \lambda u(P\ \lambda v\ F_{dir}\ (grasp(u, v)))$.

Here "F_{dir}" indicates directive illocutionary force; "P" abstracts over the semantics of the object-NP/definite description, and "(grasp(u, v))" presents the proposition commanded. The ↘ provides new information. If the ↘ is independent from the reference of the definite description the only way to express that is by extending (6) with "v = y":

(7) $\lambda P\ \lambda u\ \lambda y(P\ \lambda v\ F_{dir}\ (grasp(u, v)\ \wedge\ (v = y)))$.

The idea tied to (7) is that the reference of *v* and the reference of *y* must be identical, regardless of the way in which it is given. Intuitively, the reference of *v* is given by the definite description "ιz(yellowbolt(z))" and the reference of *y* by ↘. The values of both information contents are independent of each other. In the restrictor demonstration case the ↘ contributes a new property narrowing down the linguistically expressed one. The bracketing we assume for (3) in this case is roughly

(8) [[grasp] [this/that [↘yellow bolt]]].

As a consequence, the format of the description must change. This job can be easily done by (9):

(9) $\lambda D \lambda F \lambda P.P(\iota z(F(z)\ \wedge\ D(z)))$

The demonstration ↘ in (3) will then be represented simply by

(10) $\lambda y(y \in D)$,

where *D* intuitively indicates the demonstrated subset of the domain. Under functional application this winds up to

(11) $\iota z(yellowbolt(z) \wedge z \in D)$.

Intuitively, (11), the completed description, then indicates "the demonstrated yellow bolt" or "the yellow-bolt-within-D".

4.3 Multi-modal Meaning as an Interface of Verbal and Gestural Meaning

Even if we assume compositionality between gestural and verbal content, we must admit that the information integrated comes from different channels and that pointing is not verbal in itself, *i.e.* cannot be part of the linguistic grammar's lexicon. The representation problem for compositionality becomes clear, if we consider formula (12)

(12) $\lambda Q\lambda P\ \lambda u\ (P(Q(\lambda y\ \lambda v F_{dir}\ (grasp(u, v)\ \wedge\ (v = y))))) \ \lambda P.P(a)$ /*[grasp + ⇲]

Evidently, (12) does more than a transitive verb representation for "grasp" should do. It has an extra slot Q designed to absorb the additional object, *i.e.* the demonstration $\lambda P.P(a)$. We must regard (12) as a formula belonging to a truly multi-modal domain, where, however, the channel-specific properties have been abstracted away from. This solution only makes sense, however, if we maintain that demonstration contributes to the semantics of the definite description used.

This idea is illustrated in greater detail in Fig. 4. The interface construction shown there for (12) presupposes two things: The lexicon for the interface contains expressions where meanings of demonstrations can be plugged into; demonstrations have to be represented in the interface as well.

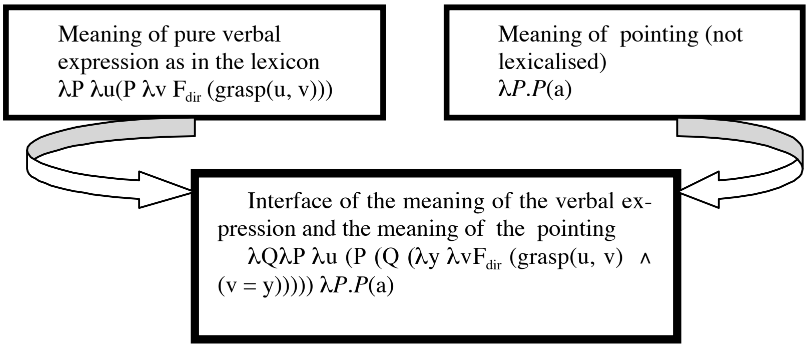

Meaning of pure verbal expression as in the lexicon
$\lambda P\ \lambda u(P\ \lambda v\ F_{dir}\ (grasp(u, v)))$
Meaning of pointing (not lexicalised)
$\lambda P.P(a)$
Interface of the meaning of the verbal expression and the meaning of the pointing
$\lambda Q\lambda P\ \lambda u\ (P\ (Q\ (\lambda y\ \lambda v F_{dir}\ (grasp(u, v)\ \wedge\ (v = y))))) \ \lambda P.P(a)$

Fig. 4. Multi-modal interface: meanings from the verbal and the gestural channel integrated via translation of ⇲

4.4 Underspecified Syntax and Semantics for Expressions Containing ⇲

The varying position of ⇲ can be captured in an underspecification model. The model coming nearest our descriptive interests is the *Logical Description Grammars* (LDGs) account of Muskens (2001).

A simplified graphical representation of inputs (1) and (3) is given in Fig. 5. '+' and '–' indicate components which can substitute ('+') or need to be substituted ('–'). Models for the descriptions in Fig. 5 are derived pairing off + and – -nodes in a one-to-one fashion and identifying the nodes thus paired. Words can come with several lexicalisations as can ⇲-s.

The *logical description of the input* has to provide the linear precedence regularities for our example "Grasp the yellow bolt!" The *description of the input* must fix the underspecification range of the ⇲. It has to come after the imperative verb. The *lexical descriptions for words* will also have to contain the type-logical formulas for compositional semantics as specified in (7) or (9).

Based on the syntax given in Fig. 5 and the type-logical formulas for compositional semantics specified in (12), we can now provide an interpretation for the speech act represented in

(13) F_{dir} (grasp(you, ιz(yellowbolt(z))) ∧ ιz(yellowbolt(z)) = a).

A full interpretation of (13) has to specify its having been performed, its success, the commitments it expresses and its satisfaction in the context of utterance *i*.

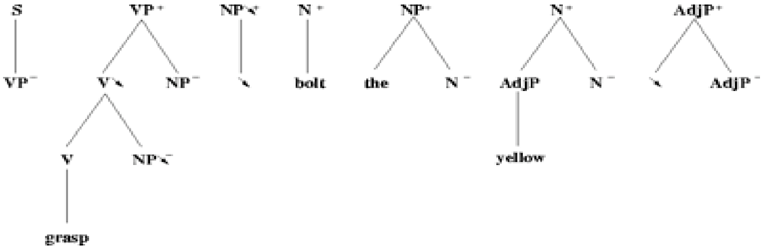


Fig. 5. Graphical representation of an input example in LDG

5 The Pointing Cone in Speech-Gesture Processing

In this section we discuss the relevance of pointing and pointing cones from the HCI perspective. The first part highlights the computational advantages for the Reference Resolution Engine (RRE, Pfeiffer and Latoschik, 2004) from the view of speech and gesture understanding. In the second part it is demonstrated how the cone can be used on the production side to decide whether object- or region-pointing is appropriate for a specific deictic referring expression and how it influences content selection.

5.1 Reference Resolution

The task of the RRE is to interpret complex demonstrations (CDs) according to the current world model represented in heterogeneous knowledge bases (KB) for symbolic information such as type, colour or function and for geometrical information. This is done using a fuzzy logic-based constraint satisfaction approach. A technical query to the RRE matching the CD “the ⇲yellow bolt” would be formulated like this:

```
(inst ?x OBJECT)   (pointed-to instruction-giver ?x time-1)
                   (has-colour ?x YELLOW time-1)

(inst ?y TYPE)     (is-a ?y BOLT time-2)

(has-type ?x ?y time-2)
```

The RRE solves this type of queries and returns a list of possible interpretations ordered by likelihood. Time is an important factor as in our dynamic scenes the constraints can only be computed on demand. But especially geometric constraints formulated verbally, e.g., by “to the left of the block” are computationally demanding. These constraints are highly ambiguous and the fuzziness keeps adding up when several constraints are spanning over a set of variables. To improve performance the RRE uses a hierarchical ordering of the constraints to reduce the search space:

- Constraints on single variables are preferred on those over multiple variables, e.g., (has-colour ?x yellow t_1) is evaluated before (is-left-of ?x ?y t_2)
- Constraints over fast accessible properties are preferred, e.g., (has-colour ?x yellow t_1) is evaluated before (has-size ?x big t_2) as the latter is context dependent.
- Hard constraints evaluating to true or false are preferred, such as constraints over names or types. They are looked up in the symbolic KB. In contrast, constraints over geometric properties are generally soft and less restrictive.

The pointing cone is represented in the same KB as the geometrical aspects of the world model, so the variables can be resolved directly with optimised intersection algorithms. With an accurate direct representation of the pointing cone, the RRE bypasses the described problems with constraints extracted from speech. The geometrical context of a CD can be computed less costly, and thereby faster, while yielding more precise results. So to say, pointing focuses attention.

5.2 Generation of Deictic Expressions

While much work concerning the generation of verbal referring expressions has been published in the last 15 years, work on the generation of multi-modal referring expressions is rare. Most approaches use idealised pointing in addition or instead of verbal referring expressions; see e.g. (Classen, 1992), (Reithinger, 1992), (André *et al.*, 1999) and (Lester *et al.*, 1999). Only Krahmer and Sluis (2003) account for vague pointing and distinguish the three types *precise*, *imprecise*, and *very imprecise* pointing.

We propose an approach, for details *cf* (Kranstedt and Wachsmuth, 2005), which integrates an evaluation of the discriminating power of pointing using the concept of pointing cones with a content selection algorithm for definite descriptions founded on the incremental algorithm published by (Dale and Reiter, 1995).

Based on our empirical observations, we use the pointing cone to define the focus of a planned pointing gesture and distinguish the two referential functions object-pointing and region-pointing discussed above. As a first step, disambiguation of the referent by object-pointing is checked. Doing so, a pointing cone with an apex angle of 12 degree anchored in an approximated hand-position and directed to the referent is generated. If only the intended referent is found inside this cone, we can refer by conducting object-pointing without an additional description of the object uttered verbally. If object-pointing does not yield a referent, region-pointing is used to focus the attention of the addressee to a certain area making the set of objects inside this area salient. This set of salient objects is determined by the pointing cone of region-pointing characterized by a wider apex angle than the cone of object-pointing. In our current implementation we chose heuristically the value 25 degrees.

The objects inside this cone have to be distinguished by additional properties. For determining them we use an adapted version of the incremental algorithm of Dale and Reiter (1995), which exploits domain-specific knowledge about typical properties to achieve a determined sequence in property evaluation and to avoid backtracking. This approach computes in linear time and the results fit well with the empirical findings. In our construction domain typically the property hierarchy, type, colour, relative size related to form, is used. The algorithm is adapted as much as relational properties are considered.

The results of the content selection algorithm are represented by a list of attribute-value-pairs, which are fed into a surface realisation module generating a syntactically correct noun phrase. This noun phrase is combined with a gesture specification and both are inserted into a surface description of a complete multi-modal utterance. Based on these descriptions, an utterance generator synthesizes continuous speech and gesture in a synchronised manner to be uttered by Max (Kopp and Wachsmuth, 2004).

6 Conclusion

The collaborative research presented in this paper scrutinised the issue of pointing in complex demonstrations. This issue was approached from interlocked perspectives, spanning the complete cycle of speech-gesture processing.

A genuine effort has been started in collecting multi-resolutional empirical data on deictic reference ranging from the high levels of speech acts down to the details of finger movements. The analysis of data on complex descriptions led to the notion of *pointing cone* fusing the parameters relevant for the discriminating power of pointing. A detailed procedure has been worked out to assess the geometrical properties of the pointing cone using tracking technology for measuring the pointing behaviours of subjects. Based on the described methods, the results of the studies will ultimately allow the fixation of a set of parameters relevant for the computation of the pointing cone's size and form. Furthermore, the sophisticated simulation of the collected data enriches the traditional video-based annotation approach; a technique that can easily be transferred to other topics of investigation.

The empirically justified concept of pointing cone enables an integrative approach to object- and region-pointing as part of complex demonstrations in concrete dialogue situations. In result, complex demonstration, and pointing as part of it, can be modelled in a more natural manner than in previous approaches. In utterance generation the pointing cone covers the object(s) to be made salient to the addressee. These objects constitute the contrast set for content-selection in planning a definite description. This idea in turn is taken up by the reference resolution procedure where the area of the cone is used to narrow down the search space. Finally, as has been shown with the multi-modal linguistic interface, the concept of the pointing cone enters into formal definitions of performance, success, commitments and satisfaction of speech acts containing complex demonstrations in an utterance's context.

Acknowledgement

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre SFB 360.

References

André, E., Rist, T., and Müller, J. (1999). Employing AI Methods to Control the Behavior of Animated Interface Agents. *Applied Artificial Intelligence*, 13:415-448.

Claassen, W. (1992). Generating Referring Expressions in a Multimodal Environment. In Dale, R. et al. (eds.), *Aspects of Automated Natural Language Generation*. Springer, pp. 247-262.

Dale, R. (1989). Cooking up Referring Expressions. In *Proceedings of the 27th Annual Meeting of the ACL*. Vancouver, pp. 68-75.

Dale, R. and Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 18:233-263.

Krahmer, E. and Sluis, I. (2003). A New Model for the Generation of Multimodal Referring Expressions. In *Proceedings European Workshop on Natural Language Generation (ENLG 2003)*. Budapest, pp. 47-54.

Kopp, S., and Wachsmuth, I. (2004). Synthesizing Multimodal Utterances for Conversational Agents. *Comp. Anim. Virtual Worlds,* 15:39-52.

Kranstedt, A., Kühnlein, P., and Wachsmuth, I. (2004). Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach. In Camurri, A., and Volpe, G. (eds.), *Gesture-based Communication in Human-Computer Interaction*. Springer, LNAI 2915, pp. 112-123.

Kranstedt, A. and Wachsmuth, I. (2005). Incremental Generation of Multimodal Deixis Referring to Objects. In *Proceedings European Workshop on Natural Language Generation (ENLG2005)*. Aberdeen, UK. To appear.

Latoschik, M. E. (2001). A General Framework for Multimodal Interaction in Virtual Reality Systems: PrOSA. In *Proceedings of the Workshop The Future of VR and AR Interfaces - Multimodal, Humanoid, Adaptive and Intelligent*. IEEE Virtual Reality 2001, Yokohama, pp. 21-25.

Lester, J., Voerman, J., Towns, S., and Callaway, C. (1999). Deictic Believability: Coordinating Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents. *Applied Artificial Intelligence*, 13(4-5):383-414.

Lücking, A., Rieser, H., and Stegmann, J. (2004). Statistical Support for the Study of Structures in Multimodal Dialogue: Inter-rater Agreement and Synchronisation. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog '04)*. Univ. Pompeu Fabra, Barcelona, pp. 56-64.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Pfeiffer, T., and Latoschik, M.E. (2004). Resolving Object References in Multimodal Dialogues for Immersive Virtual Environments. In *Proceedings of the IEEE Virtual Reality 2004*. Chicago, pp. 35-42.

Reithinger, N. (1992). The Performance of an Incremental Generation Component for Multimodal Dialog Contributions. In Dale, R. et al. (eds.), *Aspects of Automated Natural Language Generation*. Springer, pp. 263-276.

Rieser, H. (2004). Pointing in Dialogue. In *Catalog '04*. Op. cit., pp. 93-101.

Tramberend, H. (1999). Avocado: A Distributed Virtual Reality Framework. In *Proceedings of IEEE Virtual Reality 1999*, pp. 14-21.