Deictic Object Reference in Task-Oriented Dialogue

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

Abstract. This chapter presents a 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 take steps to classify the role of pointing in deictic expressions and to model the focussed area of pointing gestures, the so-called pointing cone. This 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 computational models of processing multi-modal deictic expressions.

1. Introduction

Deixis, especially deictic expressions referring to objects, play a prominent role in the research undertaken in the course of the Collaborative Research Centre SFB 360. This research focuses on scenarios in the construction task domain. A typical setting has two interlocutors communicating in face-to-face manner about the construction of mechanical objects and devices using a kit consisting of generic parts. In the investigated dialogues both participants typically use deictic expressions consisting of speech and gesture to specify tasks and select relevant objects.

This setting is also applied in the development of human computer interfaces for natural interaction in Virtual Reality (VR). Doing so, we employ an anthropomorphic virtual agent called Max who is able on the one hand to interpret simple multi-modal input by a human instructor and on the other hand to produce synchronised output involving synthetic speech, facial display, and hand gestures (Kopp and Wachsmuth 2004). To improve the communicative abilities of Max, he needs to be equipped with the competence to understand and produce multi-modal deictic expressions in a natural manner.

This chapter describes (1) a genuine effort in collecting multi-resolutional empirical data on human pointing behaviour, (2) formal considerations concerning the interrelation between pointing and referring expressions in dialogue, and (3) the application of the results in the course of reference resolution and utterance generation for the agent Max.

There is little doubt in the cognitive science literature that pointing is tied up with reference in various ways. Since Peirce at least, this has been the philosophers' concern when discussing reference and ostension. Its systematic investigation was considerably pushed ahead by McNeill's (1992, 2000) and Kendon's (1981, 2004) work on gesture. Especially McNeill's thesis that gesture and speech form an "idea unit" spread and has been reconstructed in cognitive psychology paradigms (de Ruiter 2000; Krauss, Chen, and Gottesman 2003). Moreover, the tight relation between motor skills and grasp of reference is investigated in developmental psychology. The index finger's prominent role for the evolution of species is a topic in anthropology and biology (Butterworth 2003). Concerning the ontogeny of pointing, there is a social and cultural-specific reinforcement of the infant coupling index-finger extension with the use of syllabic sounds (Masataka 2003). Clark's (1996) interactionist approach treats pointing as information on a concurrent dialogue track, and pointing and placing as attention getters in his recent article (Clark 2003).

The following quotation from Lyons (1977: 654), early as it is, subsumes much of the linguists' wisdom concerning the field of deixis and reference:

> *When we identify an object by pointing to it (and this notion, as we have seen, underlies the term 'deixis' and Peirce's term 'index' [...]), we do so by drawing the attention of the addressee to some spatiotemporal region in which the object is located. But the addressee must know that his attention is being drawn to some object rather than to the spatiotemporal region.*

Pointing, then is related to objects indicated and regions occupied. Lyons also emphasises that certain kinds of expressions are closely linked to pointing or demonstration (Lyons 1977: 657):

> *[...] definite referring noun-phrases, as they have been analysed in this section, always contain a deictic element. It follows that reference by means of definite descriptions depends ultimately upon deixis, just as much as does reference by means of demonstratives and [...] personal pronouns.*

However, it is not discussed in the literature how exactly pointing and verbal expressions are related compositionally. This is our main focus of interest here. Pursuing it, we follow a line of thought associated with Peirce, Wittgenstein and Quine, who favour the idea of gestures being part of more complex signs. Transferring this idea to deictic expressions we shall henceforth call complex signs composed of a pointing gesture and a referring expression *complex demonstration*. In other words, complex demonstrations are definite descriptions to which pointings add content, either by specifying an object independently of the definite description (Lyons' *attention being drawn to some object*) or by narrowing down the description's restrictor (Lyons' *spatiotemporal region*). In what follows, we refer to these two possibilities as the respective functions of demonstration, *object-pointing* and *region-pointing*, see (Rieser 2004).

If we take the stance that pointing provides a contribution to the semantic content of deictic expressions the question concerning the interface between the verbal and the gestural part of the expression arises. How can the interrelation between the two modalities be described and treated in computational models for speech-gesture processing? A central problem we are faced with in this context is the vagueness of demonstration, i.e. the question how to determine the focus of a pointing gesture. To deal with that, we establish the concept of *pointing cone* in the course of a parameterisation of demonstration (Section 2). In Section 3 we investigate the role of pointing gestures and their timing relations to speech on the one hand and evaluate analytical data concerning the focus of pointing gestures (modelled as pointing cone) that were collected using tracking technology and VR simulations on the other hand. In Section 4 a multi-modal linguistic interface is conceived which integrates the content of the verbal expression with the content of the demonstration determined via the pointing cone. The application of the pointing cone concept to computational models for reference resolution and for the generation of multi-modal referring expressions is described in Section 5. Finally, in Section 6 we discuss the trade-offs of our approach.

2. The parameters of demonstration

In accordance with Kita (2002) we conceive of pointing as a communicative body movement that directs the attention of its addressee to a certain direction, location, or object. In the following we concentrate on hand pointing with extended index finger into concrete domains. In the context of multi-modal deictic expressions pointing or demonstration serves to indicate what the referent of the co-uttered verbal expression might be (Kendon 2004). If we want to consider the multiple dimensions of this kind 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 (McNeill 1992). Finally, since demonstration is tied to reference, it interacts with semantic and pragmatic information in an intricate way. Gestural and verbal information also 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-semantic interface.

(b) Besides the referential functions of pointing discussed in literature (see e.g. (Kita, 2002) and (Kendon, 2004)), which draw on the relationship between gesture form and its function, we concentrate on two referential functions of pointing into concrete domains depending on the spatial relationship between demonstrating hand and referent. If an act of pointing uniquely singles out an object, it is said to have *object-pointing* function; if the gesture refers only with additional restricting material it is assigned *region-pointing* function. As we will see (Section 3.1), 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.

It has to be stressed, however, that a cone is an idealisation of the pointing area. First of all, we have to consider that depth recognition in vision is more difficult than recognition of width. Furthermore, the focus of a pointing gesture is influenced by additional parameters, which we can divide in perceivable parameters on the one hand (like spatial configuration of demonstrating agent, addressee, and referents, as well as the clustering of the entities under demonstration) and dialogue parameters on the other.

(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. Investigating 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 on 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 pointing

As mentioned in the introduction, reference is one of the key concepts for every theory of meaning. Reference and denotation guarantee the *aboutness* of language – the property of being about something in the world. It is well explored how we refer with words (see e.g. (Lyons 1977: ch. 15), (Levelt 1989: 129-134) or (Chierchia and McConnell-Ginet 2000: chs. 2 and 6)). Similarily, there is a bulk of research on the usage of co-verbal gesture (see e.g. the functions of gestures and their synchronisation with speech in narrations (McNeill 1992)).

However, there is only little work dedicated to demonstration as a device for referring to objects in multi-modal deixis. The empirical studies reported in (Piwek and Beun 2001) and (Piwek, Beun, and Cremers 1995) show that there is a different deictical treatment (*high vs low* deixis) of objects distinguished by their degree of salience (*givenness* and *noteworthiness*) in Dutch cooperative dialogues. Beun and Cremers (2001) proved for task-oriented dialogue that focusing the attention by pointing reduces the effort needed to refer to objects as well as to identify them. Van der Sluis and Krahmer (2004) observe a dependence of the length of the verbal part of the expression on the distance between demonstrator and object demonstrated.

Although the above-mentioned studies support the assumption that pointing carries some part of the meaning of multi-modal deixis, a lot of questions concerning the details of the interface between the modalities in such expressions are still open. In 2001 we started our empirical work with explorative studies on these matters. The setting and the design of those studies were chosen to investigate temporal as well as spatial relations that tie together gesture and speech. On the one hand, we wanted to look whether the synchronisation between the modalities as found in narratives (McNeill 1992) can be replicated in task-oriented dialogues. On the other hand, we wanted to get some insight into how the spatial properties of density and distance constrain the use of pointing gestures.

In the ongoing section we start with a brief sketch of the setting used for the studies and continue with a description of their results. Then we propose new methodologies to elucidate the pointing region as represented by the pointing cone and finally discuss current results.

3.1. Simple object identification games

We conduct our empirical studies in a setting where two subjects are engaged in simple object identification games (Fig. 1), which restrict the *instructor-constructor* scenario investigated in the SFB 360 to the problem of referring. One subject (instructor) has the role of the "description-giver". She has to choose freely among the parts of a toy airplane spread on a table, the pointing domain, and to refer to them. The other subject (constructor), in the role of the "object-identifier", has to resolve the description-giver's reference act and to give feedback. Thus, reference has to be negotiated and established using a special kind of dialogue game (Mann 1988).

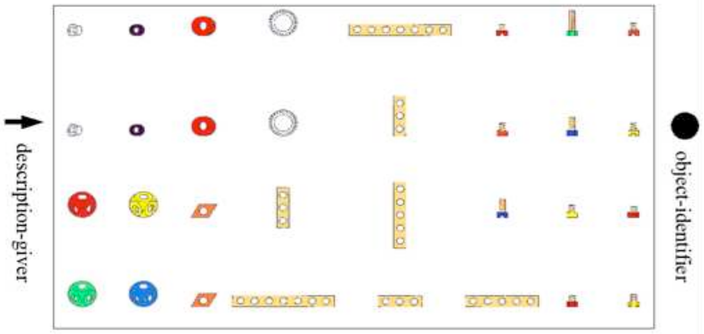


Figure 1. Simple object identification games in settings with objects arranged in a shape-cluster

3.2. Explorative studies on demonstration in dense domains

In the first explorative studies described in (Kühnlein and Stegmann 2003) and (Lücking, Rieser, and Stegmann 2004) the object identification games were recorded using two digital cameras, each capturing a different view of the scene. One camera recorded a total view seen from one side orthogonally to the table, the other gave an approximate perspective of the description-giver's.

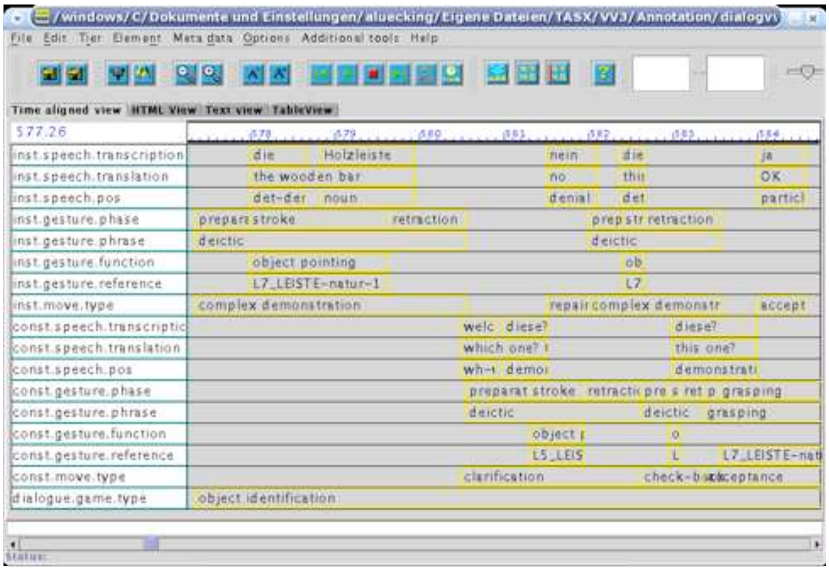
The objects of the pointing domain were laid out equi-distantly, that is, the distance between their centres was the same for all objects lying side by side. Their positions on the table top fit in a regular coordinate system and were not changed over the time of the study (Fig. 1). This move not only allowed us to determine the density holding among the objects but also provided us with a simple notion of distance, namely in terms of object rows, which can easily be converted into a linear measure.

Positioning of objects was clustered in two ways: according to colour and according to shape (Fig. 1). The different distributions of objects should prevent subjects' pointing behaviour from being influenced by certain prevalent traits. The two clusters together with a change of the subjects' roles yielded four sub-settings for each single execution of the experiment.

The subjects were not forced to use pointing gestures. Contrary to our assumption that this move assures natural referring behaviour a lot of subjects avoided pointing. This problem has to be solved in future studies by giving more precise instructions.

From seven explorative studies conducted only two involve the use of demonstration. Because of the role change, the results given below are based on four subjects acting as description-givers. They produced a total of 139 referring acts.

In order to get results concerning the relations between gesture and speech in dialogue, we applied descriptive and analytical statistical methods to the time-based annotation stamps of suitable dialogue data.


/windows/C/Dokumente und Einstellungen/aluecking/Eigene Dateien/TASX/VV3/Annotation/dialogv
File Edit Tier Element Meta data Options Additional tools Help
Time aligned view
HTML View
Text view
TableView
Status:

577.26	578 579 580 581 582 583 584
inst.speech.transcription	die Holzleiste nein die ja
inst.speech.translation	the wooden bar no thi OK
inst.speech.pos	det-der noun denial det particl
inst.gesture.phase	preparz stroke retraction prep str retraction
inst.gesture.phrase	deictic deictic
inst.gesture.function	object pointing ob
inst.gesture.reference	L7_LEISTE-natur-1 L7
inst.move.type	complex demonstration repair complex demonstr accept
const.speech.transcriptic	welc diese? diese?
const.speech.translation	which one? t this one?
const.speech.pos	wh-ı demoı demonstrati
const.gesture.phase	preparat stroke retractic pre s ret p grasping
const.gesture.phrase	deictic deictic grasping
const.gesture.function	object ŗ o
const.gesture.reference	L5_LEIS L L7_LEISTE-nat
const.move.type	clarification check-backceptance
dialogue.game.type	object identification

Figure 2. Annotation of a complex dialogue game. A screenshot from a TASX annotation session that exemplifies the annotation scheme applied in score format, see example for transcription of speech parts. Taken from (Lücking, Rieser, and Stegmann 2004)

3.2.1. Annotation

The analysis of our corpus of digital video data is based on an annotation with the TASX-Annotator software package (Milde and Gut 2001; http://medien.informatik.fh-fulda.de/tasxforce). It allows an XML-based bottom up approach. Since the annotation data is stored in XML format, the extraction of the relevant information for purposes of statistical analysis can be realized via XSLT script processing straightforwardly. Details connected with the empirical setting and different annotation approaches are laid out in (Kühnlein and Stegmann 2003).

As illustrated in Fig. 2, the set of annotation tiers includes a transcription of the agent's speech at word level (`speech.transcription`) and a classification of the dialogue move pursued (`move.type`). The annotation of deictic gestures follows in essence the framework established in (McNeill 1992). A gesture token has three phases: wrt pointing gestures, the maximally extended and meaningful part of the gesture is called *stroke,* respectively *grasping* if an agent grasps an object. Stroke or grasping is preceded by the *preparation* phase, that is, the movement of the arm and (typically) the index finger out of the rest position into the stroke or grasping position. Finally, in the *retraction* phase the pointer's arm is moved back to rest position. The distinction between object- and region-pointing is captured on the `gesture.function` tier. The discriminating criterion was whether the annotator could resolve the description-givers pointing gesture to a single object.

All tiers are specified for the description-giver and the object-identifier; the respective tier names have an `inst.` or `const.` prefix, see Fig. 2. So, for example, there is a tier labelled `inst.speech.translation` containing the utterance of the description-giver, and one labelled `const.speech.translation`, for recording the utterance of the object-identifier (the naming of the prefixes is due to the subjects' role names in the "standard scenario" of the SFB 360.)

To get a better grip on the kind of data we are concerned with, the speech portions of the sample dialogue from Fig. 2 were extracted and are reproduced below.

(1) Inst: The wooden bar
[pointing to object1]
(2a) Const: Which one?
(2b) This one?
[pointing to object2]

(3a) Inst: No.
(3b) This one.
[pointing to object1]
(4) Const: This one?
[pointing to object1 and grasping it]
(5) Inst: O.K.

We have the dialogue move of a `complex demonstration` of the description-giver in (1) here, followed by a `clarification` move involving a pointing of the object-identifier (2a, 2b). The description-giver produces a `repair` (3a), followed by a new `complex demonstration` move (3b) to the object she had introduced. Then we have a new `check-back` from the object-identifier (4) coming with a pointing and a grasping gesture as well as an acceptance move by the description-giver (5). The whole game is classified as an `object identification game`. The following events from different agents' turns overlap: (2b) and ((3a) and (3b)); (3b) and (4).

3.2.2. Results

Rather than being mere emphasis markers, gestures contribute to the content of communicative acts. This can be substantiated by findings related to the semantic, the pragmatic, and the discourse level summarised in the following.

I. Gestures Save Words. The total amount of 139 referring acts adds up out of 65 referential NPs escorted by a pointing gesture (hereafter CDs, for *complex demonstrations*) and 74 NPs without pointing (DDs, short for *definite descriptions*). We (Lücking, Rieser, and Stegmann 2004) found strong evidence for the semantic contribution of pointings in comparing the number of words used in CDs with that in DDs by means of a t-test. It results in a (highly) significant difference ($t = 6.22$, $p \cong 0$, at the risk level $\alpha = 0.05$), *cf.* Fig. 3a. This result can be couched into the slogan “Gestures save words!”. Thus, gestures contribute content that otherwise would have to be cast into clumsy verbal descriptions, making communicative acts more efficient.

II. Gestures as Guiding Devices. A related cognitive hypothesis was that the time the object-identifier needs to interpret the description-giver's reference (hereafter called *reaction time*) is less after a CD than after a DD. The pointing gesture can be seen as guiding the object-identifier's eyes towards the intended object – or at least towards a narrow region where the object is located – and thus as shortening the object-identifier's search effort. To assess this point, we calculated 48 (39 CDs and 9 DDs, taken from two description-givers) differences between the start time of the object-identifier's move and the end time of the description-giver's referring act. A subsequent t-test applied to the two resulting sets of time stamps did not come out with a significant difference (t = -1.4, p = 0.166, α = 0.05) but there seems to be a tendency for shorter reaction times after CDs, cf. Fig. 3b (Lücking, Rieser, and Stegmann 2004).

What might have prevented a significant outcome was the fact that some objects are unique and therefore more salient, e.g., there is only one yellow cube (as opposed to several yellow bolts), so that the object-identifier could quickly spot such objects when directed with appropriate DDs only. In addition, the object-identifier may have used the description-giver's gaze as a guiding device, especially with toy airplane parts that lie very close to the description-giver (Kühnlein and Stegmann 2003). Nonetheless, the small difference found in reaction times might become significant in larger samples.

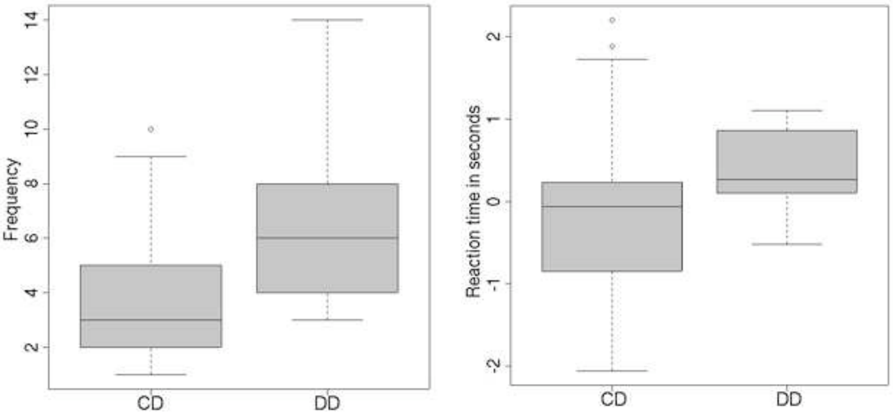


Figure 3. Boxplots displaying a) the number of words in CDs and in DDs, b) object-identifiers' reaction times (in seconds) following instruction-givers' CDs or DDs. The horizontal lines delimit the range of measurement values; run-away tokens (that are values that fall out of the range given by 1.5-times inter-quartile distance added to the quartiles) are indicated with a small circle. The boxes span the 0.25 and the 0.75 quantile and show the median. Taken from (Lücking, Rieser, and Stegmann 2004)

III. *Intra*-move Temporal Relations. At the beginning of this paper, a distinction was made between *intra*- and *inter*-move synchronisation at the dialogue level. As regards *intra*-move synchronisation we accounted for the temporal relations holding between gesture phases and escorting utterances. Above all, we focused on two synchronisation effects, namely *anticipation* and *semantic synchrony* (McNeill 1992: 25-26, 131). The semantic synchrony rule states that gesture and speech present one and the same meaning at the same time (McNeill's "idea unit"). Anticipation refers to the temporal location of the preparation phase in relation to the onset of the stroke's co-expressive portion of the utterance. This rule states that the preparation phase precedes the linguistic affiliate of the stroke. Table 1 summarises the descriptive statistics (N = 25). The different rows were calculated as follows: (P) preparation$_{\text{start}}$ -speech$_{\text{start}}$, (R) speech$_{\text{end}}$ -retraction$_{\text{start}}$, and (S) stroke$_{\text{start}}$ -speech$_{\text{start}}$. Note, that we take the verbal affiliate to be the complete denoting linguistic expression, i.e. a possibly complex noun phrase.

Row P gives the values for the start of the preparation phase relative to the onset of the first word of the noun phrase. For each speech-gesture ensemble, the time stamp associated with the beginning of the first word of the utterance was subtracted from the time stamp for the start of the respective gesture's preparation phase. Hence, negative values in row P indicate that the start of the preparation phase precedes the verbal affiliate as is to be expected in the light of McNeill's anticipation rule. Contrary to (McNeill 1992: 25, 131), we found that the utterance usually starts a little before the initiation of the gesture (compare the positive mean value in Table 1. This seems to contradict anticipation, given the way we operationalised McNeill's concept of the idea unit.

Table 1. Temporal *intra*-move synchronisation values: The *minimum* (the smallest measurement value), the *maximum* (the largest measurement value), the *arithmetic mean*, the *standard deviation*, the *first quartile* (or 0.25 quantile, the value that divides the data ordered according to size such that 25% of the measurement values lie below this value), and the third quartile (0.75 quantile, 75% of the measurement values lie below this value)

	Min.	**1st Qu.**	**Mean**	**3rd Qu.**	**Max.**	**Std.Dev.**
P	–0.8	–0.2	0.3104	0.48	4.68	1.0692
R	–0.86	0.0	0.564	1.06	3.38	0.89
S	–0.02	0.48	1.033	1.24	5.54	1.128

Similarly (compare the mean value in row R), the stroke ends (or the retraction starts) normally around 0.5 seconds before the end of the affiliate. Together with an average start of the stroke around 1 second after the onset of the utterance (mean for row S) this shows, that the prototypical stroke does not cross utterance boundaries (Lücking, Rieser, and Stegmann 2004). This is as to be expected in the light of McNeill's semantic synchrony rule. Note, however, that some extreme tokens (compare respective min. and max. values in Table 1) were observed that seem to contradict the McNeill regularities, cf. (Kühnlein and Stegmann 2003).

IV. *Inter*-move Temporal Relation. Concerning *inter*-move synchronisation, one point of interest was the alignment of the end of description-giver's preparation phase with object-identifier's retraction phase. A look into the dialogue video data reveals that two different cases have to be distinguished here. If the object referred to lies within object-identifier's reach, his initiation seems to regularly overlap with the description-giver's retraction. If the object referred to lies at the opposite side of the table, that is, out of his reach, the object-identifier first has to move around the table which delays initiation of his gesture. The temporal differences between the two gesture phases (preparation$_{OI}$ – retraction$_{DG,}$ where the indices stand for the respective roles) were grouped accordingly into a within-reach case and an out-of-reach case. The outcomes are given in Table 2.

Table 2. Inter-move synchronisation of preparation and retraction

	Min.	1st Qu.	Mean	3rd Qu.	Max.	Std. Dev.
within-reach	–2.06	–0.96	–0.4984	–0.06	2.2.6	0.89
out-of-reach	–1.36	0.4	1.54	1.7	8.76	2.19

If the object in question is within object-identifier's reach his initiation of grabbing it overlaps with the retraction of the description-giver by an average amount of time of half a second – compare the mean value in Table 2 (note also that the third quantile still yields a negative result!). This indicates that the description-giver's retraction phase might contribute to a turn-taking signal. Not surprisingly, there is no such overlap if the object is out of object-identifier's immediate reach (Lücking, Rieser, and Stegmann 2004).

V. Partitioning of the Pointing Domain. Moving from semantic and temporal to pragmatic issues, we also tried to find out whether there are contextual conditions constraining the use of gestures. This was defined in terms of frequencies of DDs *vs* CDs utilised to refer to objects in different rows of the pointing domain – that is, basically, wrt their distance as seen from the instructor.

What is at stake here is whether the asymmetry that seems to be revealed in the bare data – compare Table 3 and the plot depiction in Fig. 4 – could be statistically validated.

Table 3. Descriptive values for referring to objects in different rows of the domain

Row	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
CDs	3	6	10	10	10	11	7	8
DDs	10	11	7	9	6	6	7	18
Total	13	17	17	19	16	17	14	26

Roughly three regions emerge (Kühnlein and Stegmann 2003; Kranstedt, Kühnlein, and Wachsmuth 2004): the first two rows constitute an area which is nearest to the description-giver, called the *proximal* region. In opposition, rows seven and eight form the *distal* region, the area that is farthest away from the description-giver. The remaining 4 rows in the middle of the pointing domain are the *mid-range* region. Note, that this partitioning corresponds to the ratings of gesture function, cf. finding VI below.

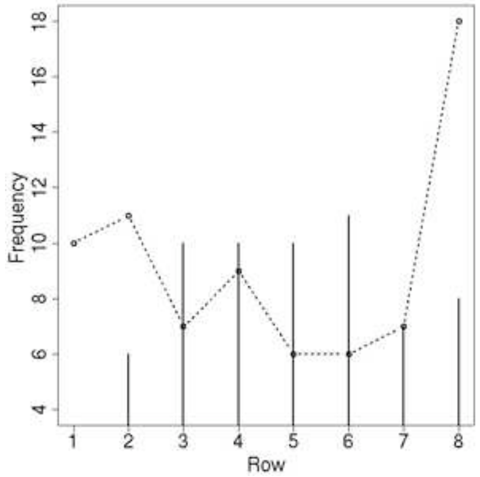


Figure 4. Plot for the modes of reference modelled by the eight rows of the reference domain; the bars depict the frequency distribution of CDs over the rows, the dashed line that of DDs. Taken from (Lücking, Rieser, and Stegmann 2004)

While the decrease of CD's and the increase of DD's in the distal region correspond with intuition, the results concerning the proximal reason are surprising. Maybe, one reason could be that some of the subjects use gaze and head movements accompanied by a DD to guide the attention of the addressee to objects in the proximal region. Though, to capture this in the video data is difficult. This phenomenon of head or gaze pointing and possible other reasons for the observed decrease of CD's has to be addressed in further investigations.

However, the relative distance of the object in question to the description-giver seems to be a contextual factor for the choice of the mode of reference to that object (Lücking, Rieser, and Stegmann 2004).

VI. Object-Pointing vs. Region-Pointing. As introduced in parameter (b) above, we assume that pointing gestures serve one of two semantic functions: they uniquely pick out an object (*object-pointing*) or merely narrow down the region in which the intended object lies (*region-pointing*). In order to illustrate this distinction, an occurrence of each gesture function is shown in Fig 5. The extension of pointing gestures is modelled with a pointing cone. Fig. 5b depicts a case of region pointing, where several objects are located in the conic section of the pointing cone and the tabletop. There, the extension of the index finger does not meet the object in question. Against this, in object pointing the object is unequivocally singled out, i.e. it is the only object within the conic section (Fig. 5a).

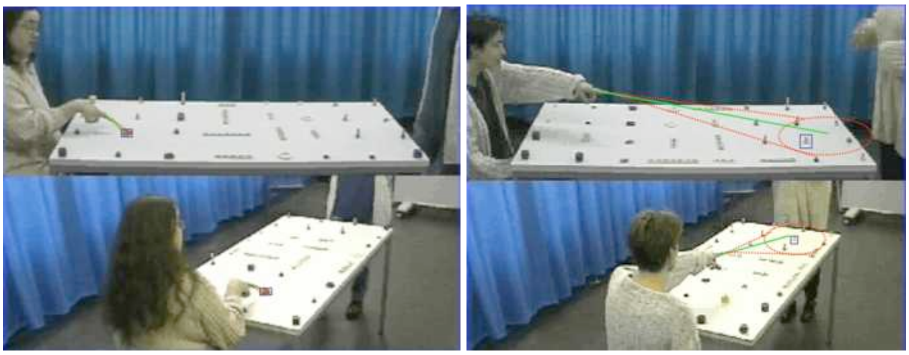


Figure 5. The two kinds of pointing found in the data, a) object-pointing, b) region-pointing. The prolongation of the index finger is indicated with a line, the pointing cone is indicated using dotted lines, and the box frames the intended object. Taken from (Lücking, Rieser, and Stegmann 2004)

From a semantic point of view, object pointings behave very much like referring expressions, whereas region-pointing tokens may be said to be predicative or relational in nature. The difference in meaning between those functions is formally explicated in the linguistic interface described in Section 4.

In the course of proving whether the dialogue scheme used is reliable in terms of *inter*-rater agreement, the distinction between the two gesture functions turned out to be problematic in some ways: Although there is a strong consensus concerning the classification of pointings in regions very near and very far from the description-giver, there is a broad region in the middle where the raters differ in their estimation, cf. Table 4. We see three kinds of reasons for the disagreement. Above all, the two-dimensional video-data lack the necessary depth of focus to admit the classification. Furthermore, the rating criterion is probably not well-defined, so that the raters used varied interpretations (for example, one rater might be content with exactly one object lying in the projected pointing cone to vote for object pointing, while the other raise the bar in requiring the prolongated pointing finger (the "pointing beam") to meet the object). At last, it is feasible that the theoretically motivated function-distinction has no clear-cut realisation in the empirical realm of the real world.

Table 4. Gesture function ratings. The region of disagreement is highlighted

Row		***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
Rater 1	object-pointing	2	4	8	6	7	1	0	0
	region-pointing	0	1	2	1	3	9	7	5
Rater 2	object-pointing	2	4	6	2	2	0	0	1
	region-pointing	0	1	4	5	8	10	7	4

VII. Distance-dependence of Gesture *vs* Speech Portions. The following two assumptions are corroborated: Firstly, there is a division of labour between gesture and speech in referring to objects; secondly, pointings loose resolution capacity in greater distances. Hence it follows that description-givers have to put the larger identifying burden into the verbal expression the farer away the intended object is in order to perform successful deictic acts. Indeed, in (van der Sluis and Krahmer 2004) the dependence of the distance of the object in question on the informational share that has to be provided via each channel could be proved. To verify this dependence in our study, we can make use of the pre-structuring of the pointing domain into rows.

The obvious statistical computation is to compare the number of words used in CDs to refer to objects in the different regions (reminder: distal, mid-range, and proximal). Therefore, an analysis of variance (ANOVA) was carried out on the number of words modelled by regions. Although there is a minor difference in the bare data, cf. Fig. 6, the ANOVA did not yield a significant outcome (F = 0.53, p = 0.6).

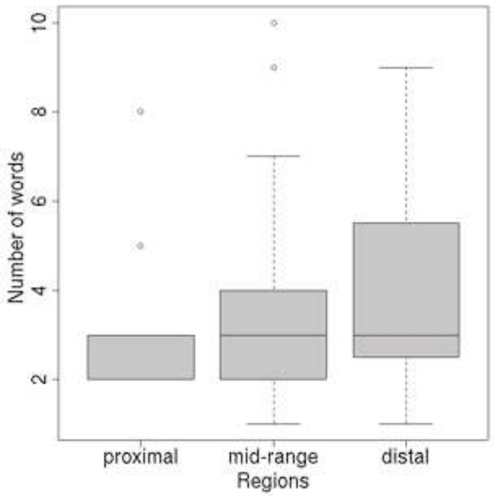


Figure 6. Boxplot displaying the number of words used to refer to objects in the different regions. Though there is a decrease in the inter-quartile distance from the proximal to the distal region, the median remains all about the same

This unexpected result can be explained by two facts: firstly, the sample is clearly too small to render such small differences in means significant. Secondly, a look in the videos reveals that the subjects make use of overspecification: they provide more information than necessary to identify the object referred to, and thus – superficially – violate rules of parsimony and economy. This in turn might be an artefact of the setting. The simplicity and repetition of the identification task tempted subjects to use recurrent patterns of simple NPs, mostly composed of a determiner followed by an adjective and the head noun. On the other hand, the description-giver is anxious for securing object-identifier's comprehension, so that the latter is able to successfully and smoothly resolve the former's referential behaviour.

3.2.3. Discussion

As has been shown above, our experimental setting provides us with rich empirical evidence to support our parameterisation of demonstration presented in section 2. Our findings that gestures save words (I) and the tendency for shorter reaction times after CDs (II) further emphasise the need for a multi-modal linguistic interface (parameter (a)). This view is also empirically supported by the findings of Piwek and Beun (2001) and Beun and Cremers (2001).

The question of the temporal relations subsumed by the parameters (d) and (e) are captured by the findings III and IV. It has to be noted that in our task-oriented setting we find higher temporal variability than in narrative dialogues (McNeill 1992). This imposes greater restrictions especially onto the speech-gesture resolution module which has to be sufficiently general in order to process all occurrences of the relatively loose temporal relations of multi-modal deixis.

The partitioning of the pointing domain according to the distribution of CDs and DDs presented in V (proximal/mid-range/distal) provided us with a useful spatial categorisation, which is picked up in the description of our findings regarding the spatial constraints of demonstration. The distinction between the two referential functions object- and region-pointing, as proposed in parameter (b), are backed by this partitioning (VI). Together they provide the descriptive framework to describe our findings on the distance dependence of gesture and speech (VII). Dealing with this interrelationship is necessary for both sides of speech-gesture processing, speech-gesture generation and speech-gesture recognition. The tendency we find in our experiments accords with the findings of van der Sluis and Krahmer (2004).

All issues touching upon the distance of referents are affected by the pointing cone, which is bound up with the vagueness of pointing. In this context, the cone also can be seen as a device to capture the focusing power of pointings in the sense of (Piwek, Beun, and Cremers 1995) and (Beun and Cremers 2001). Assessing the pointing cone (parameter (c)) and its three-dimensional topology is essential for our theoretical and computational models of the interface between gesture and speech in deictic expressions. However, the two-dimensional video data do not afford accurate statements about the spatial area singled out by a pointing gesture. Especially the position and orientation of the demonstrating hand and the stretched index finger wrt the table and the objects lying on it, which are necessary for the computation of the size and form of the pointing cone, can only be estimated inexactly.

In sum, the empirical results in this study address the parameters (a), (b), (d), and (e). First approximations of the pointing cone, parameter (c), give some clues but the empirical method used does not provide means to really grasp the pointing cone's topology. Hence, the pointing cone needs to be assessed in more precision, in particular, to account for possibly different cones associated with object-pointing and region-pointing.

3.3. Assessing the pointing cone

The inappropriate results concerning the topology of the pointing cone are a consequence of the methods used for data collection and analysis. The two perspectives provided by the video recordings lead to too many ambiguities in the ratings, which have become evident in our inter-rater agreement tests. Therefore, methods are needed which grasp the topology of the pointing cone in its three-dimensionality and provide exact spatial data concerning the pointing behaviour.

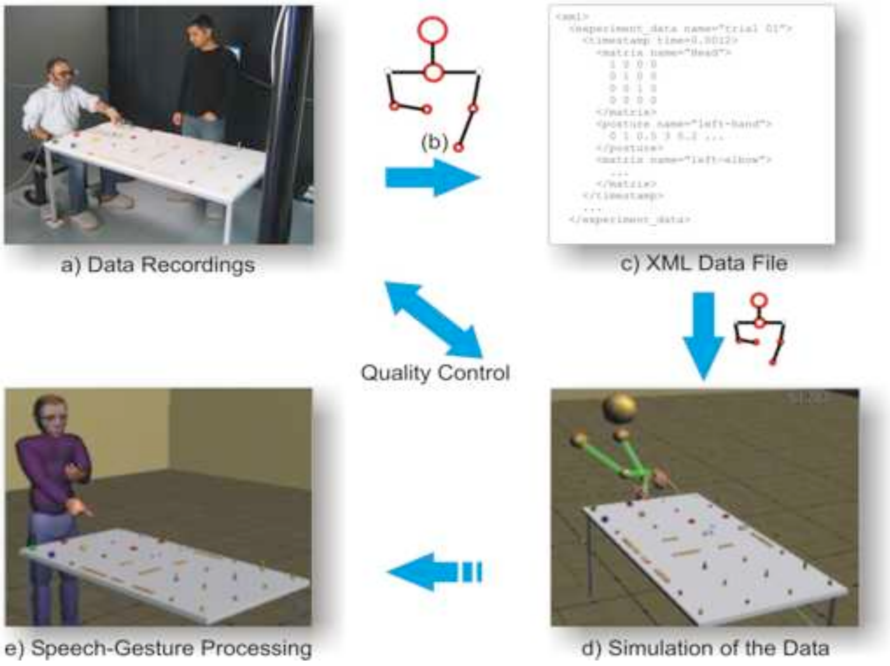
In addition, we search for methods to visualize pointing-beam, pointing cone, and the intersection of them with the pointing domain to support analysing the data.

3.3.1. Tracker-based experiments

In our search for such methods we settled on a tracker based solution, see also (Kranstedt et al. 2005). It uses a marker-based optical tracking system to obtain adequate analytical data for the body of the subject. Additional data for the fine-grained hand postures is collected using data gloves (Fig. 7a). 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 information gathered providing their absolute coordinates and orientations. We track head and back of the description-giver to serve as reference points. With two markers each, one for the elbow, and one for the back of the hand the arms are tracked. The hands are tracked using CyberGloves® measuring flexion and abduction of the fingers directly. We do not specially track the object-identifier, as the relevant information, especially the identification of the demonstrated object, can easily be extracted from the recorded videos.

3.3.2. Representing the data

The information provided by the tracking systems (Fig. 7a) is fed into our VR application based on the VR framework Avango (Tramberend 2001), which extends the common scenegraph representation of the visual world.



```
<xml>
  <experiment_data name="trial 01">
    <timestamp time=0.3012>
      <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 5.2 ...
      </posture>
      <matrix name="left-elbow">
        ...
      </matrix>
    </timestamp>
    ...
  </experiment_data>
```

Figure 7. 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 enhance the speech-gesture processing models (e). Taken from (Kranstedt et al. 2005)

A scenegraph consists of nodes connected by arcs defining an ownership relation. The nodes are separated into grouping nodes and leaf nodes. Every node is the target of an ownership relation, then called a “child”, but only grouping nodes can also be a source or “parent”. In addition to this basic distinction, the nodes in a scenegraph can have different types: geometry nodes, material nodes, etc., are used to define visual appearance. A single visual object may be the product of a combination of several such nodes interacting, separately defining one or more shapes, colours, or textures of the object. The position of an object in the world is determined by the multiplication of matrices defined in transformation nodes along a chain from the root node of the scenegraph to the object's geometry nodes. A special feature of the Avango VR framework is the datagraph, which is defined orthogonally to the scenegraph. It does not operate on the nodes in the scenegraph, but on subcomponents of them, the fields. Each node in the scenegraph can exhibit a set of fields defining its data interface. Examples of such fields are the matrices of the transforming nodes. The datagraph connects these fields with a dataflow relation, defining that the data from the parent field is propagated to the child field. Every time such propagation results in the change of a child field, a special trigger function is called in the scenegraph node owning the field. The node can then operate on the new data, change its state, and eventually provide results in some of its fields, which may induce the next propagation.

A group node acting as root of a subgraph represents the description-giver. This type of node does not have a graphical representation. It is a special kind of group node, a transformation group node, which is not only grouping its siblings but also defines a transformation to position them in space. The matrices of the transformation nodes in this subgraph are connected to *actuator* nodes representing the different tracking devices. These actuator nodes are defined in the PrOSA (Patterns On Sequences of Attributes, (Latoschik 2001a)) framework, a set of data processing nodes specialised for operating on timed sequences of values utilising the data-processing facilities of Avango. The subgraph representing the description-giver is updated according to the posture of the tracked user using field connections from the actuator nodes providing a coherent geometric user model (Fig. 7b). For recording the tracked data this user model is written to an XML file and can later be used for annotation or stochastic analysis (Fig. 7c).

3.3.3. Simulation-based data evaluation

To support data evaluation we developed tools to feed the gathered tracking data (Fig. 7c) back into the geometric user model, which is now the basis of a graphical simulation of the experiment in VR (Fig. 7d). 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. 7d) providing only relevant information. We preferred this in contrast to our anthropomorphic agent (Fig. 7e), as the visualisation of information not backed by the recordings, such as the direction of the eye gaze, could mislead raters.

For a location independent annotation we created a desktop-based visualisation system where the rater can move a virtual camera into every desired perspective 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., looking through the eyes of the description-giver.

3.3.4. Computation of pointing beam and pointing cone

The pointing beam is defined by its origin and its direction, the pointing cone in addition by its apex angle. To grasp the spatial constraints of pointing, one has to specify

a) the anatomical anchoring of origin and direction in the demonstrating hand and
b) the apex angle.

We can calculate these parameters under the following assumptions:

(i) We know the exact position and orientation of the demonstrating hand and the extended index finger (provided by the tracking data).
(ii) We know the intended referent (identified in the dialogue annotation).
(iii) We have a statistically relevant amount of demonstrations to each object and each region in the pointing domain.

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. 8. 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.

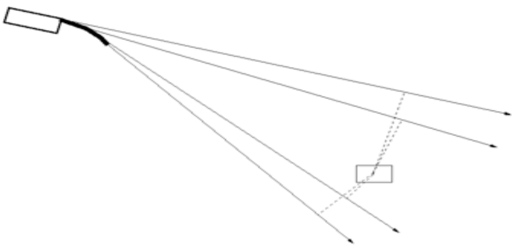


Figure 8. Four hypothetical pointing beams anchored in different anatomical parts of the hand. Taken from (Kranstedt et al. 2005)

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. 9.

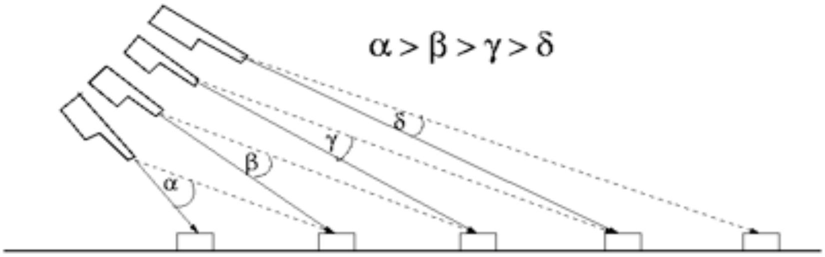


Figure 9. The angles between the beams to the referent and the next neighbour decreases with the distance to the referent (the dashed arrows represent the beams to the next neighbour). Despite similar distance to the referent, the beam to the object behind the referent results in a smaller angle than the beam to the object in front of the referent. This is because of the greater distance of the former one to the demonstrating hand. Taken from (Kranstedt et al. 2005)

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 the first substantial increase of the verbal expressions describing the referent. 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 referential 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, Kühnlein, and Wachsmuth 2004). However, for this assessment a fixed hand position heuristically determined over all demonstrations was assumed and only a small number of annotated data was used. So, these results should be 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. The idea is to 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. The angle between the pointing beam and a beam touching the most distant object defines then in a first approximation the half apex angle of the pointing cone of region-pointing.

3.3.5. Discussion

The method proposed was tested in a first study in November 2004. There, our primary concerns were the question of data reliability and the development of methods for the analysis. The main study was conducted in September 2005. Video and tracking data from 60 subjects consisting of 30 description-givers and 30 object-identifiers were collected. At the time of writing this text the analysis of the data is under preparation. The results seem promising, so that we will discuss our experience and highlight some interesting advantages of this approach.

The tracker-based recordings supplement the video recordings by providing 3D coordinates of the markers on the body of the description-giver specifying a full posture for every frame of the video. This data is more extensive and more precise than that gathered annotating the videos. Its collection can be automated to disencumber the manual annotation significantly speeding up the overall analysis. As the posture of the description-giver is known for every frame, extensive data for a statistical analysis is available, a precondition for gathering the anchoring of pointing beam and pointing cone and the topology of the cone.

The visual simulation of the gathered data provides us with a qualitative feedback of the tracker recordings. This proved to be useful, especially when running on-line. This way important preparations of the experimental setting, such as adjusting the illumination, avoiding occlusions or positioning and calibrating the trackers are easily done before the experiment, improving the quality of the data to be recorded. After the experiment, the simulation is used to review the data and identify problems, so that incomplete or defective recordings are recognised and separated as early as possible. This applicability renders the simulation a perfect tool for the quality assurance of the recorded tracking data. Furthermore, the simulation can be used to facilitate the annotation of the video recordings by providing a dynamic perspective on the setting. It is also possible to add a virtual pointing beam or pointing cone to the simulation. The intersection of the pointing beam and the table top can then be interpreted as an approximation of the location pointed to and the intersection of the cone with the table top as the area covered by the pointing gesture.

On the other hand the tracker-based recordings are a compromise where we preserve the natural dialogue setting only to some extent, e.g., as the subjects are not used to wear trackers, which are physically attached to their bodies. In one trial this showed up in an extreme fashion when a subject used her hands with an outstretched index finger in a tool-like manner without relaxation. To compensate for such effects an interactive preparation phase has to be introduced where subjects can familiarise themselves with the new environment. Still we believe this method to be less obtrusive than any modification concentrating on the index finger or the gesturing arm alone, as it involves the whole body of the description-giver without putting too much emphasis on a specific aspect, e.g., pointing gestures as such.

Overall, we are aware that the combination of optical markers and data gloves is more invasive than relying on video cameras alone. But at the time being they seem to be our most powerful empirical tool for a deeper investigation of the pointing cone's topology.

4. A multi-modal linguistic interface

In this 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 that captures the object-/region-pointing distinction.

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 "⇲", mimicking the index finger in stroke position. ⇲ is concatenated with the verbal expression, indicating the start of the stroke in the signal and hence its functional role. In this respect, ⇲ is treated like a normal linguistic constituent. Its insertion can be directly derived from the annotated data. (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, the ⇲ being absent in the CD.

(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. Abstracting from other less well understood uses such as abstract pointings, CDs are considered as definite descriptions to which demonstrations add content either by specifying an object independently of the definite description, thus acting as a definite description in itself, or by narrowing down the description's restrictor. We call the first use "object demonstration", pointing to an object, and the second one "restrictor demonstration", a semantic classification of pointing to a region. Graspings are the clearest cases of object demonstration.

Before we show how to represent demonstrations with descriptions in one logical form, we specify our main hypotheses concerning their integration. These are related to content under compositionality, i.e. their roles in building up referential content for the embedded proposition, and the scope of the gesture. Hypothetically then, demonstrations (a) act much 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 ⇲ (see examples (2) to (5) below), (d) involve, empirically speaking, a continuous impact over a time interval, comparable to intonation contours, and (e) can be described using discrete entities like the ↘.

4.2. Interpretation of complex demonstrations

The central problem is of course how to interpret demonstrations. This question is different from the one concerning the ↘'s function tied to its position in the string. We base our discussion of these matters 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) $\lambda N\ \lambda u(N\ \lambda v\ F_{dir}\ (grasp(u,v)))$.

Here "F_{dir}" indicates directive illocutionary force; "N" abstracts over the semantics of the object-NP/definite description "this/that yellow bolt", i.e. "$\lambda Z.Z(\iota z(yellowbolt(z)))$", and "(grasp(u,v))" presents the proposition commanded. The ↘ provides additional information. If the ↘ is independent from the reference of the definite description the only way to express that is by somehow extending (6) with "v = y":

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

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. This property of independence will be reconstructed in the interface for multi-modal semantics.

Object demonstration and restrictor demonstration are similar insofar as information is added. In the object demonstration case, this is captured by a conjunct with identity statement; 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]]].

Here, the demonstration contributes to the content of the N'-construction "yellow bolt". As a consequence, the format of the description must change. This job can be easily done with

(9) $\lambda R \lambda W \lambda K.K(\iota z(W(z) \wedge R(z)))$.

Here, K abstracts over the semantics of the directive, W is the predicative delivered by the noun, and R is the additional restrictor.

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 as given by the pointing cone. We use the $\in$-notation here in order to point to the information from the other channel. Under functional application this winds up to

(11) $\lambda K.K\ (\iota z(\text{yellowbolt}(z) \wedge z \in D))$.

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

We started from the hypothesis that verbal descriptions and gestural demonstrations yield complex demonstrations, the demonstrations either independently identifying an object or contributing an area demonstrated, extending an underspecified definite description.

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 deeper reason, however, is that integrating values for pointings-at would make the lexicon infinite, since infinitely many objects can be pointed at.

The representation problem for compositionality becomes clear, if we consider the formulas used for the imperative "grasp", i.e. the different forms (12), (13), and (14), stated below.

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

(12) is the demonstration-free expression of the imperative form corresponding to the semantic information in a lexical entry for "grasp something". (13) already specifies an identity condition and says that one of the arguments to "grasp", *v*, has to be identical to some other, *y*, the latter being reserved for the pointing, but it does not yet contain a device which can guarantee compositionality of definite description and pointing information. In other words, there is no way of putting a value for *y* into the formula. This is achieved using (14). Evidently, and that's the important issue here, (14) does more than a transitive verb representation for "grasp" in the lexicon should do. It has an extra slot Q designed to absorb the additional object *a*, tied to the demonstration $\lambda P.P$(a). Given the infinity argument above, we must regard (14) as a formula in the model-bound interface of speech and gesture, i.e. as belonging to a truly multi-modal domain, where, however, the channel-specific properties have been abstracted away from. That is, in the semantic information coded in the interface you do not see any more where it originates from. This solution only makes sense, however, if we maintain that demonstration contributes to the semantics of the definite description used.

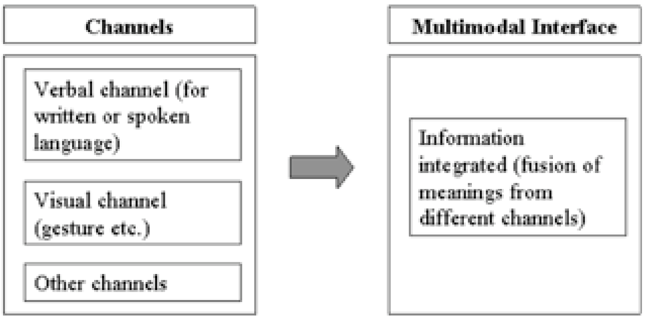
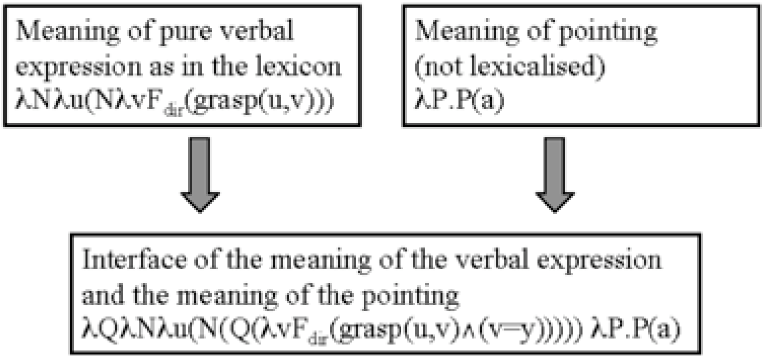

Channels
Verbal channel (for written or spoken language)
Visual channel (gesture etc.)
Other channels
Multimodal Interface
Information integrated (fusion of meanings from different channels)

Figure 10. Information from different channels mapped onto the multi-modal interface

The general idea is shown in a rough picture in Fig. 10 and illustrated in greater detail in Fig. 11. The interface construction shown there for (12) to (14) 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. The number of demonstrations is determined by the intended model, see section 4.5.2.

Syntax and semantics have to be mapped onto one another in a systematic way. Now, the position of ⇲ varies as examples (2) to (5) above show, in other words, the ⇲ might go here or there. We can capture this feature in an underspecification model, which implies that we generally deal with descriptions instead of structures. The underspecification model coming nearest our descriptive interests is the *Logical Description Grammars* (LDGs) account of Muskens (2001), which has evolved from Lexicalised Tree Adjoining Grammar (LTAG), D-Tree Grammar, type logics and Dynamic Semantics. The intuitive idea behind LDGs is that, based on general axioms capturing the structure of trees, one works with a *logical description of the input*, capturing linear precedence phenomena, and *lexical descriptions for words* and *elementary trees*. A *parsing-as-deduction* method is applied yielding semantically interpreted structures.



Meaning of pure verbal expression as in the lexicon
$\lambda N\lambda u(N\lambda vF_{dir}(grasp(u,v)))$

Meaning of pointing (not lexicalised)
$\lambda P.P(a)$

Interface of the meaning of the verbal expression and the meaning of the pointing
$\lambda Q\lambda N\lambda u(N(Q(\lambda vF_{dir}(grasp(u,v)\wedge(v=y))))) \lambda P.P(a)$

Figure 11. Multi-Modal interface: meanings from the verbal and the gestural channel integrated via translation of ⇲

4.4. Underspecified syntax and semantics for expressions containing ⇲

A simplified graphical representation of inputs (1) and (3) is given in Fig. 12. ‘+’ and ‘–’ indicate components which can substitute (‘+’) or need to be substituted (‘–’). Models for the descriptions in Fig. 12 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. (a) specifies the elementary tree for the imperative construction. VP^{-}marks the place where a tree tagged VP$^+$ can be substituted. (b) indicates how the demands of the multi-modal interface have to be fulfilled: V needs an NP⇲-sister whose tag ⇲ says that only stroke-information can be substituted resulting in a constituent V⇲ taking then a normal NP as an argument. (c) introduces a referring stroke. (d) is the lexical entry for “bolt”. (e) describes an NP-tree anchored with “the”. The insertion of “yellow” is brought about using (f). Finally, (g) is used for ⇲-insertion before an AdjP. NP⇲$^+$ is needed to build up (14) and, similarly, AdjP⇲$^+$ for getting at (9).

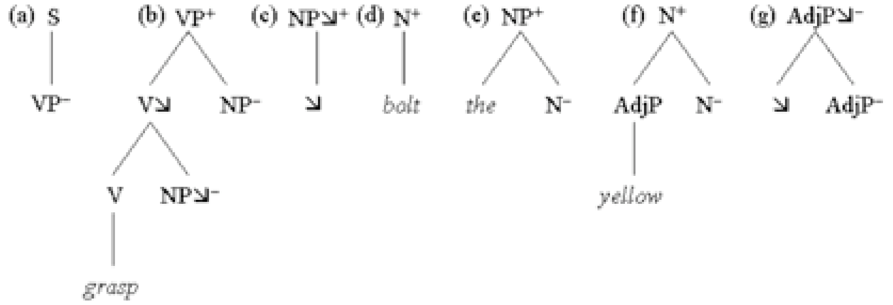


Figure 12. LTAG representation of the syntax interface for pointing and speech

The *logical description of the input* has to provide the linear precedence regularities for our example “Grasp this yellow bolt!”

The *description of the input* must fix the underspecification range of the ⇲. It has to come after the imperative verb, but that is all we need to state; in other words, an underspecified description is at the heart of all the models depicted in (2) to (5). The *lexical descriptions for words* will also have to contain the type-logical formulas for compositional semantics as specified in (7) or (9). From the descriptions of the *elementary trees* we will get the basics for the “pairing-off” mechanism. Fig. 13 shows the derived tree for the directive “Grasp ⇲ this yellow bolt!” with semantic tagging using the LTAG in Fig. 12.

4.5. On the question of structures anchoring multi-modal meanings

We now want to seriously consider the problem of providing some meaning for formulas of the sort

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

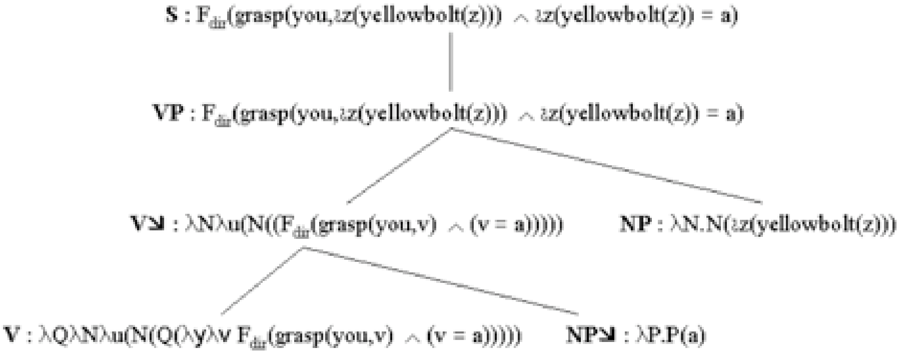

S : F_dir(grasp(you,ɩz(yellowbolt(z))) ∧ ɩz(yellowbolt(z)) = a)
VP : F_dir(grasp(you,ɩz(yellowbolt(z))) ∧ ɩz(yellowbolt(z)) = a)
V↘ : λNλu(N((F_dir(grasp(you,v) ∧ (v = a)))))
NP : λN.N(ɩz(yellowbolt(z)))
V : λQλNλu(N(Q(λyλv F_dir(grasp(you,v) ∧ (v = a)))))
NP↘ : λP.P(a)

Figure 13. Derived tree for the directive “Grasp ↘ this yellow bolt!” with semantic tagging using the LTAG in Fig. 12

paraphrased as the directive speech act “Grasp the yellow bolt demonstrated!” Pursuing this we must discuss the following problems: 1. Which is the structure to be used for speech act interpretation? 2. Which are the conditions of success for speech acts in general and (15) in particular? 3. Which are the conditions of commitment and the satisfaction conditions for speech acts in general and (15) in particular? 4. What is the relation between empirical setting and model structure? To discuss these problems in a very preliminary way, we use Searle and Vanderveken’s Illocutionary Logic (*IL*) (here (Searle and Vanderveken 1989)), which allows us to touch upon some points of interest.

4.5.1. The Structure Used for Speech Act Interpretation

Formula (15) describes an elementary illocutionary act with the directive illocutionary force as indicated by F_{dir}.

Hence, we will concentrate on how elementary (i.e. atomic in the strict sense) directives are treated in IL. In IL one uses the notion of *context of utterance* in order to specify the semantic and pragmatic conditions of illocutionary acts such as these. For building up contexts of utterance, we need four sets, I_1, I_2, I_3, I_4 for, respectively, possible speakers, hearers, times and places of utterance. In addition, we postulate a set W of *possible worlds of utterance*.

The set I of all possible contexts of utterance is a proper subset of the Cartesian product of the sets introduced individually: $I \subset I_1 \times I_2 \times I_3 \times I_4 \times W$.

As a consequence, every context of utterance $i \in I$ has five constituents, the so-called coordinates of the context: speaker a_i, hearer b_i, time t_i, location l_i and the world w_i. A context i is identified with the 5-tuple $< a_i, b_i, t_i, l_i, w_i>$. There is a linear ordering $\prec$ on I_3 (times). Possible worlds are taken to be primitive; as usual in modal logics, we need a designated world w_0, for the actual world. In addition, the set W comes with a binary relation R of accessibility, which we need in order to express different styles of possibility and necessity, mental states and future or past courses of events.

So far, we have provided an answer to our first question concerning the structure to be used for speech act interpretation. We now turn to the conditions of commitment and satisfaction for speech acts as mentioned in the second question.

What do success, commitment and satisfaction conditions, respectively, amount to for example (15)? First we investigate success, i.e. successful performance. To discuss this question, we need a couple of notions from general modal logics and from IL: The notion of possibility, $\Diamond$, is used as in normal systems of modal logics, *Des* is a modal operator indicating desire, and $\Pi_!$ serves as a modal operator for the directive illocutionary point used to model the semantics and pragmatics of requests. *U(w)* is the domain of objects associated with some world $w \in W$; in addition, domains for all possible worlds can be defined.

An elementary illocutionary act of the form $\mathrm{F}_{\mathrm{dir}}$ (grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) = a) is performed in the context of utterance i iff the speaker = description-giver a_i succeeds in the context of utterance i to

- express the illocutionary point $\Pi_!$ (request) on P = (grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) = a),
- issue the commanded proposition P, i.e. issue the relevant locutionary act,
- presuppose that it is possible ($\Diamond$) for the addressee to grasp the yellow bolt demonstrated, i.e. $\Diamond$(grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) = a), where you = addressee = object-identifier, and
- express a desire (*Des*) concerning the intended act, i.e. Des(grasp(object-identifier, ιz(yb(z))) ∧ ιz(yb(z)) = a).

These conditions provide only success requirements for the illocutionary act. We now turn to the description giver's commitments. Since the description-giver produces an utterance of (15) in w_i, we may assume that he is committed to the conditions of $\mathrm{F}_{\mathrm{dir}}$ (grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) = a), i.e., presuppositions, mental states and the like, for example, he must believe that the object to be grasped exists, that the addressee has not grasped it so far and he must sincerely intend that it should be grasped.

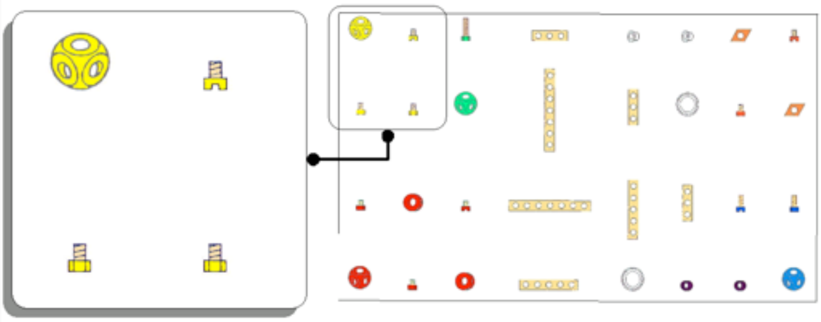


Figure 14. Experimental domain used as a sub-domain of the intended model for speech act interpretation

Finally, we turn to the notion of satisfaction for elementary speech acts of the form *F(P)*: An illocutionary act of the form *F(P)* is *satisfied* in a context of utterance *i* iff P(w_i) = 1 and is *not satisfied* otherwise in *i*. For F_{dir} (grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) = a) this means that it is satisfied in *i*, iff (grasp(you, ιz(yb(z))) ∧ ιz(yb(z)) (w_i) = 1, i.e. iff the object-identifier grasps the demonstrated yellow bolt in w_i.

4.5.2. Logics and reality: experimental setting and model structure

Normally, if one has to set up models for speech act representations such as in (15) one is hard pressed for providing intuitive model descriptions, especially, if problems of reference are at stake and the models should in a way imitate natural referring conditions. We are better off in this respect: As the empirical data show, we have all the information necessary in order to add substance to the formal model described in the previous passages: Both agents in our object-identification dialogue are possible speakers and hearers, hence I_1 = I_2 = {description-giver and object-identifier}, I_3 and I_4 get a natural interpretation as being related to the time and the place of the experiment, respectively. It is perhaps more difficult to decide on the *possible worlds of utterance*. The most suitable choice seems to be to map occurrences of speech act tokens onto contexts *i*. Our agents reside in the actual world, i.e. in the experimental setting. Hence, we have, paralleling speech act occurrences, contexts *i* of the following sort, distinguishable by the values of t_i: <*description-giver*$_i$, *object-identifier*$_i$, t_i, l_i, w_0>. We can exactly specify, what the relevant part of $U(w_0)$, the set of objects that can be pointed at, is. It is shown in Fig. 14 and is identical to one of the settings used in the experimental studies described in section 3 above. Using Fig. 14 as depiction of our relevant sub-domain, we notice three yellow bolts in the left corner. This means that *wrt* this model the satisfaction of (15) fails, since the definite description ιz(yb(z)) cannot be satisfied. As a consequence, the object-identifier might try to check-back saying "Which one do you mean?" Indeed, some such reaction is frequently found in our corpus. Notice that restrictor-demonstration has more chances of success, if D in ιz(yellowbolt(z) ∧ z ∈ D) can be instantiated to contain one of the yellow bolts, still, there are various options for a proper choice of D, see Fig. 15.

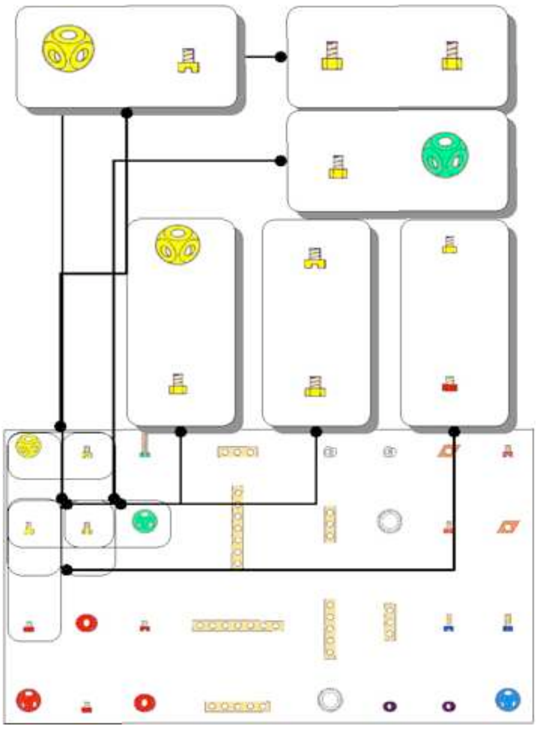
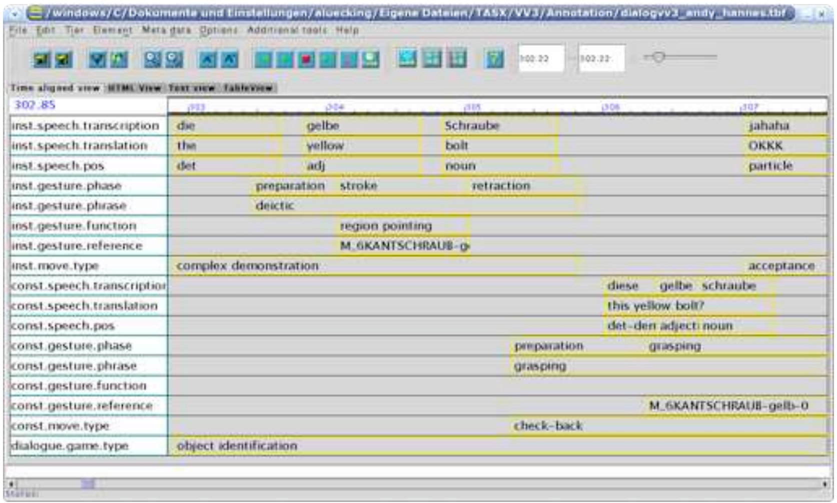


Figure 15. Some pair-subsets of the spotted sub-domain. Note that some pairs constitute models for successful CDs, while others do not

4.6. Modelling pointing effects in current theory of dialogue

So far, we have not developed a systematic description of the role demonstration can play in natural dialogue. This will be the aim of this section after a brief recapitulation of what we have got up to now. First, we showed how multi-modal content, speech and gesture, can be integrated into a theory distinguishing between object-pointing and region-pointing. The theory maps multi-modal objects onto a speech act representation containing complex demonstrations, i.e. definite descriptions accompanied by demonstrations. This step is based on examples from an annotated corpus of object identification games. Secondly, we specified conditions of success, commitment and satisfaction for speech acts using “Grasp ↘ this yellow bolt!” as an example. In this context we also discussed the relation between empirical setting and model structure, showing that the empirical setting can be used as an intended model. Thirdly, using statistical methods, we extracted a potential regularity concerning turn-taking and demonstration from our corpus data, namely, that the description-giver’s retraction phase might contribute to a turn-taking signal. Considering all that, we have already gone some way towards the description of dialogue.

In order to estimate what is still missing, we turn to a fairly easy example from the corpus: Figure 16 shows the transcript of a complete sub-dialogue, the wording of which is given in (16). Figure 17 offers snapshots of the description-giver’s and the object-identifier’s actions.


/windows/C/Dokumente und Einstellungen/aluecking/Eigene Dateien/TASX/VV3/Annotation/dialogvv3_andy_hannes.tbf
File Edit Tier Element Meta data Options Additional tools Help
302.22
302.22
Time aligned view
HTML View
Text view
TableView
Status:

302.85	303	304	305	306	307
inst.speech.transcription	die	gelbe	Schraube		jahaha
inst.speech.translation	the	yellow	bolt		OKKK
inst.speech.pos	det	adj	noun		particle
inst.gesture.phase	preparation	stroke	retraction		
inst.gesture.phrase	deictic				
inst.gesture.function		region pointing			
inst.gesture.reference		M_6KANTSCHRAUB-g			
inst.move.type	complex demonstration				acceptance
const.speech.transcriptio				diese gelbe schraube	
const.speech.translation				this yellow bolt?	
const.speech.pos				det-den adject noun	
const.gesture.phase			preparation	grasping	
const.gesture.phrase			grasping		
const.gesture.function					
const.gesture.reference				M_6KANTSCHRAUB-gelb-0	
const.move.type			check-back		
dialogue.game.type	object identification				

Figure 16. Complete sub-dialogue from the object-identification corpus

(16) a. Description-giver: The yellow ↘bolt! [demonstrates yellow bolt]
b. Object-identifier: This ↘yellow bolt? [grasps indicated yellow bolt]
c. Description giver: OKOKOK.

We have a directive in (16a), a clarification question in (16b) and an acceptance move in (16c). The directive and the clarification question are elliptical, lacking appropriate finite verbs. We can substitute 'grasp' and 'should I grasp' respectively. Taking the previous sections 4.1 to 4.5 as background, we are now able to develop the relevant intuitions: The description-giver issues a command. Its grammar and multi-modal semantics is as shown in Section 4.5 (15). The command is successfully performed but not satisfied. It would be satisfied, so we may assume, if the object-identifier simply took the yellow bolt with some sort of assertion or without a comment and the description-giver accepted the dialogue move. Why does the situation arise? Looking at the intended model for the satisfaction of the command, i.e. the table plus objects depicted in Fig. 17, we see, why it is not satisfied.

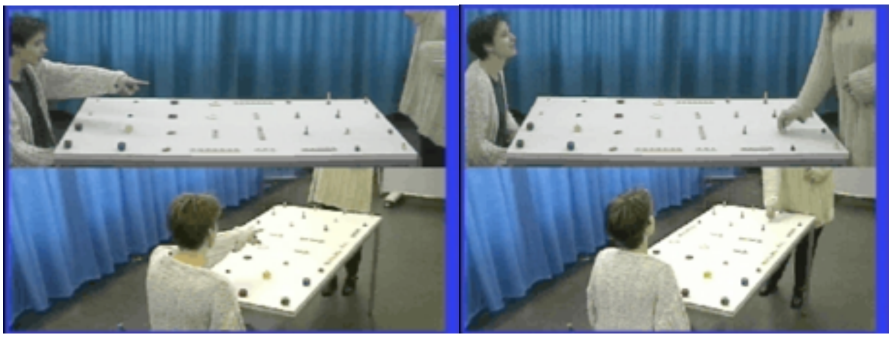


Figure 17. Description giver's and object-identifier's actions

As Fig. 18 clearly indicates, neither conceptualising the pointing as object-pointing nor as region-pointing will yield a uniquely referring definite description. We can even assume that the description-giver had the right intention to refer to the bolt which the object-identifier finally grasped, thus emphasising that success and commitment conditions were indeed met, but the pointing resolution does not suffice for attaining satisfaction, since it is defined for <*description-giver*$_i$, *object-identifier*$_i$, t_i, l_i, w_0>, i.e. it also depends on the object-identifier. This explains, why we have a clarification question of the object-identifier's ✋*This yellow bolt?* Observe that the reference of the grasping act provides no problem, since grasping can be conceived as borderline case of pointing. The clarification question thus functions as a means to achieve alignment between description-giver and object-identifier. In terms of the concept of pointing cone, the sequence of command and clarification question can be explained as follows: The semantics of the pointing cone taken as a "Platonic entity" may be OK, that is, it may single out a sub-domain which can be fused with the definite description in the multi-modal interface as discussed in Section 4.2, but its pragmatics is obviously not, the main problem being that the gauging of the pointing cone by the object-identifier does not yield an applicable description. More generally, in dialogues involving pointing the alignment of the pointing cone and its projection by the addressee of the pointing act have to be considered. In informal terms, the clarification question can be paraphrased as: Does the object grasped meet your referring intention? The description-giver's accept shows that it does.

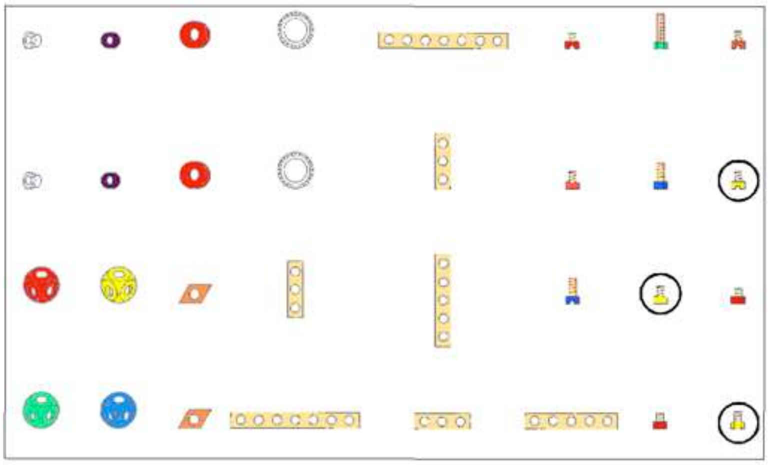


Figure 18. Intended model for satisfaction of the elliptical directive *The yellow*⭸*bolt!* There are three yellow bolts at the right border, which explains that neither object-pointing nor region-pointing can be satisfied in conjunction with the definite description

A final observation coming from the transcript in Fig. 16 is that description-giver's retraction phase and object-identifier's preparation phase overlap. If we want to use this trait in our theorising, we have to introduce special annotation devices indicating the full structure of the demonstration. So, let us use ↓ for the preparation phase of a demonstration, ⭸ for its stroke as before, usurping it now also for grasping, and ↑ for its retraction phase. In order to distinguish contributions of various agents, we decorate the arrows with agents' labels like ↓$_{\text{description-giver}}$, ⇲$_{\text{description-giver}}$, ↑$_{\text{description-giver}}$ etc. Using these means, we get the following annotation for the turns of (16):

(17) a. [$_{\text{NP}}$ [$_{\text{DET}}$ The] [$_{\text{N'}}$ ↓$_{\text{description-giver}}$ [$_{\text{ADJ}}$ yellow] ⇲$_{\text{description-giver}}$ [$_{\text{N'}}$bolt]]
↑$_{\text{description-giver}}$].
b. [$_{\text{NP}}$ ↓$_{\text{object-identifier}}$ [$_{\text{DEM}}$ This] [$_{\text{N'}}$ ⇲$_{\text{object-identifier}}$ [$_{\text{ADJ}}$ yellow] [$_{\text{N'}}$bolt]]].
c. OKOKOK.

After these preliminaries, we look at the structure of the three-turn dialogue. Here we must integrate different traditions of dialogue description: The basic idea of agents cooperating and coordinating in dialogue comes from Clark (1996) and, more recently, from Pickering and Garrod (2004), the proposal that newly attached turns are bound to old content on the basis of discourse relations has been developed in dialogue game theory (Levin and Moore 1977), RST (Mann and Thompson 1987), and SDRT (Asher and Lascarides 2003); finally surface orientedness as a program for dialogue description goes back to a proposal of Poesio and Traum (1997).

Now we determine the discourse relations involved in (16). (16a) and (16b) are related by the fact that (16c) is a clarification question following up a command. The command cannot be satisfied, since the object identifier is not able to spot the object indicated. The object-identifier's question is such that if it is answered by the description-giver, he knows whether the command is satisfied or not. We suggest a binary relation ICSP(α, β) (called 'indirect command satisfaction pair') to capture that α is a command and β a question. The answer to β will as a rule indicate whether the command is already satisfied by the addressee's action or whether he has to initiate a new action to finally carry out the description-giver's request. In other words, the question is closely tied to the satisfaction conditions of the command. More precisely, it is a question solely concerned with establishing the satisfaction of the command. As a consequence, it must be followed by an answer. Fulfilling this need, (16b) and (16c) compose a question answer pair, QAP, a relation as proposed in Asher and Lascarides (2003: 313). The description-giver's accept is also only concerned with the satisfaction problem.

We forgo specifying the formal details here, they are straightforward, anyway. The structure of the whole dialogue is thus simply as depicted in Fig. 19, in addition satisfying the constraint that $\uparrow_{\text{description-giver}} \circ \downarrow_{\text{object-identifier}}$, i.e. $\uparrow_{\text{description-giver}}$ and $\downarrow_{\text{object-identifier}}$ overlap.

α: The $\downarrow_{\text{description-giver}}$ yellow $\searrow_{\text{description-giver}}$ bolt! $\uparrow_{\text{description-giver}}$

ICSP

β: $\downarrow_{\text{object-identifier}}$ This $\downarrow_{\text{object-identifier}}$ yellow bolt?

QAP

γ: OKOKOK

Constraint: $\uparrow_{\text{description-giver}} \circ \downarrow_{\text{object-identifier}}$

Figure 19. Dialogue structure for example (16) according to SDRT

5. Processing deictic expressions

In this section we discuss the relevance of pointing in complex demonstrations from the perspective of human-computer interaction. The scenario under discussion consists of task-oriented dialogues, which pertain to the cooperative assembly of virtual aggregates, viz. toy airplanes. These dialogues take place in face-to-face manner in immersive virtual reality, realised in the three-side CAVE-like installation mentioned in Section 3. The system is represented by a human-sized virtual agent called Max, who is able on the one hand to interpret simple multi-modal (speech and gesture) input by a human instructor and on the other hand to produce synchronised output involving synthetic speech, facial display and gesture (Kopp and Wachsmuth 2004). As illustrated in Fig. 20, Max and the human dialogue partner are located at a virtual table with toy parts and communicate about how to assembly them.

In this setting demonstration games can be realised to focus on the understanding and generation of complex demonstrations. In analogy to the empirical setting described in Section 3 these demonstration games follow the tradition of minimal dialogue games as, e.g., proposed in (Mann 1988). However, we reduce the interaction to two turns. This enables us to directly compare the empirically recorded data with the results of speech-gesture processing, since our HCI interface already provides a framework for handling these basic interactions.

The narrow description of speech-gesture processing is split into two subsections. In the first one below the role of the pointing cone for speech-gesture understanding is highlighted. Special attention is given to its relevance for the computation of reference in the Reference Resolution Engine

(Pfeiffer and Latoschik 2004). The second subsection describes the algorithm for generating deictic expressions, especially how demonstrating by object- respectively region-pointing interacts with content selection for the verbal part of the expression.

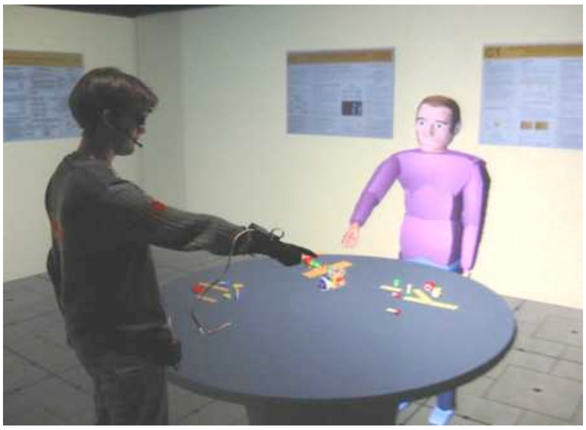


Figure 20. Interacting with the human-sized agent Max in an immersive VR-scenario concerning the assembly of toy airplanes. Taken from (Kranstedt and Wachsmuth 2005)

5.1. A framework for speech-gesture understanding gesture recognition

In 3.3 we have seen how the information of the trackers is made accessible by the actuator nodes of the PrOSA framework to the VR application. For recognising gestures, the fields exported by the actuators are connected to specialised detector nets, subgraphs of evaluation nodes designed to classify certain postures or trajectories. For instance, there are detector nets to detect an extended index finger called "`right-hand-index-posture`" or an extended arm called "`right-arm-extended`". Their results are provided in timed sequence fields, e.g., as collection of Boolean values identifying whether at a certain point in time the index finger was extended or not. High-level concepts such as "`right-is-pointing`" can then be identified combining the results of existing detector nets. Note that this is only a didactic example, the composition of detector nets used in the current system is far more complex. A more detailed description can be found in (Latoschik 2001b).

5.1.1. The role of the pointing cone in early gesture processing

The dynamic environment of a VR setting imposes some difficulties for modelling the pragmatic effect of pointing gestures, that is for identifying the intended objects or regions. At the time the system has finally reached the conclusion that the spatial area of the pointing gesture is important and the objects enclosed in the pointing cone are relevant, they might already have changed their positions or appearances. Their positions at the production time of the gesture are needed, but to gather tracking information about all objects in the environment during the full course of interaction is almost impossibile. Instead we follow a proactive approach. After a gesture has been classified as a pointing gesture, additional nets take care of evaluating the corresponding pointing cones, collecting all enclosed objects in a special structure called space-map. These space-maps are then used by the following processes for the semantic interpretation of the pointing gesture. In this early processing steps, elaborated models of the pointing cone(s) help to sustain a low memory profile while maintaining the descriptiveness of the gesture. This is accomplished taking a highly localised snapshot of the gesture's visual context.

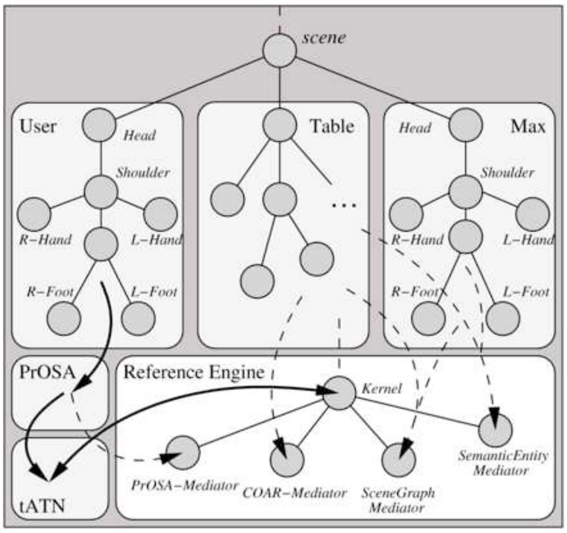


Figure 21. The framework for speech and gesture understanding. Taken from (Pfeiffer and Latoschik 2004)

5.1.2. Speech and gesture integration

For the understanding of multi-modal instructions and direct manipulative actions in the VR system, a tATN is used, an ATN specialised for synchronising multi-modal inputs (Latoschik 2003). It operates on a set of states and defines conditions for state transitions. The actual state thereby represents the context of the utterance in which the conditions will be processed. In the extension, states are anchored in time by an additional timestamp `reach`. Possible conditions classify words, access PrOSA sequence fields for the gestural content or test the application's context. An important part of the context is the world model representing the visual objects. A module called reference-resolution engine (RRE) enables the tATN to verify the validity of the object descriptions specified so far, finding the matching objects in the world model. The set of possible interpretations of an object description delivered by the RRE will incrementally be restricted during the further processing of the utterance by the tATN. If the parsing process has been successful, these sets are used to finally fill in the action descriptions used for initiating the execution of the instruction. It is the RRE where the content of the pointing is finally integrated with content from other modalities and where the cone representations find their application.

5.1.3. The relevance of the pointing cone for the reference resolution

The task of the RRE is to interpret complex demonstrations (CDs) according to the current world model represented in heterogeneous knowledge bases (see Fig. 21) for symbolic information such as type, colour or function (*SemanticEntity Mediator*, *COAR Mediator*) and for geometrical information (*SceneGraph Mediator*, *PrOSA Mediator*). This is done using a fuzzy logic-based constraint satisfaction approach.

When incrementally parsing a multi-modal utterance such as (1), "*Grasp ↘this/that yellow bolt*", the tATN tries to find objects in the world satisfying the complex demonstration. For this the tATN communicates with the RRE using a constraint query language. A query corresponding to the example (1) would be formulated like this:

```
(inst ?x) (pointed-to instruction-giver ?x time-1)
          (has-colour ?x YELLOW time-1)
          (has-type ?x BOLT time-2)
```

To process this query the RRE has to gather the knowledge of several heterogeneous knowledge bases. The *PrOSA Mediator* is used to evaluate the `pointed-to` constraint. The `has-colour` constraint requires the *SemanticEntity Mediator* and for the `has-type` constraint the knowledge of the *COAR Mediator* is used. The RRE integrates the responses from each mediator and tries to satisfy `(inst ?x)`. This could be a single object in the case of an object demonstration or in the case of a restrictor demonstration a set of possible objects, the subdomain of the world defined by the query (and initially by the CD). In both cases the RRE provides additional information about the saliency of the match(es) and the contributions of the single constraints to the overall saliency.

In our dynamic scenes the constraints can only be computed on demand, so fast evaluating constraints are necessary to meet the requirements of real-time interaction. Unfortunately, especially geometric constraints formulated verbally, e.g., by "to the left of the block" are computationally demanding: Even single constraints are highly ambiguous and fuzziness keeps adding up when several constraints are spanning over a set of variables. To improve performance the RRE uses therefore a hierarchical ordering of constraints to reduce the search space as soon as possible:

– Constraints on single variables are preferred on those over tuples of variables, e.g., (has-colour ?x yellow t_1) is evaluated before (is-left-of ?x ?y t_2)
– Constraints on 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. Typical examples are constraints over names or types, which can be solved by looking them up in the symbolic KB. In contrast, constraints over geometric properties are generally soft and less restrictive.

The pointing cone is directly 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 faster, while yielding more precise results. So to speak, pointing focuses attention.

5.1.4. Differentiating object-pointing and region-pointing

Per default the `pointed-to` constraint discriminates between object-pointing and region-pointing based on the distances of the objects. This behaviour can be overwritten by explicitly specifying the intended interpretation using the parameters `'object-cone` or `'region-cone`. As in `(pointed-to instruction-giver ?x time-1 'object-cone)` where object-pointing, and therefore a more narrow cone, is forced.

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 multimodal referring expressions is rare. Most approaches use idealised pointing in addition to or instead of verbal referring expressions, see e.g. (Classen 1992; Reithinger 1992 and Lester et al. 1999). In contrast, only Krahmer and van der Sluis (2003) account for vague pointing, and distinguish the three types *precise*, *imprecise*, and *very imprecise* pointing.

We propose an approach (Kranstedt and Wachsmuth 2005) which integrates an evaluation of the discriminatory power of pointing with a content selection algorithm founded on the incremental algorithm published by Dale and Reiter (1995). Based on empirical observation and theoretical consideration, we use the pointing cone to model the discriminatory power of a planned pointing gesture and to distinguish its two referential functions, object-pointing and region-pointing discussed above. Fig. 22 presents the algorithm, Fig. 23 depicts an example which will be explained in detail further on in this section.

Using terminology proposed by Dale and Reiter (1995), we define the *context set* C to be the set of entities (physical objects in our scenario) that the hearer is currently assumed to be attending to. These can be seen as similar to the entities in the focus spaces of the discourse focus stack in the theory of discourse structure proposed by Grosz and Sidner (1986). We also define the set of *distractors* D to be the set of entities the *referent* r has to be distinguished from by a set of *restricting properties* R each composed of an attribute-value pair. At the beginning of the content selection process the distractor set D will be the context set C, at the end D will only contain r if content selection has been successful.

To achieve linear compute time Dale and Reiter (1995) propose a determined sequence of property evaluation and dispense with backtracking. This leads to overspecification, but they can show that the generation results fit well with the empirical findings if the sequence of properties is chosen accurately wrt the specific domain. As described in Section 3, overspecification is also often found in our data. Therefore, the content selection algorithm gets a sorted list of properties in addition to the referent and the context set as input. Concerning the order of properties, in our corpus we typically observe the hierarchy type, colour, size and relative location in the verbal part of the deictic utterances. In addition we consider absolute location to be expressed by pointing.

As a first step in the proposed algorithm for deictic expressions (see Fig. 22, 1.), disambiguation of the referent by object-pointing is checked if the referent is visible to both participants. Using the PrOSA tools mentioned above, this is achieved generating a pointing cone with an apex angle of 20 degree anchored in an approximated hand-position and directed to the referent. If only the intended referent is found inside this cone referring is done by object-pointing. 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 in this area salient. The distractor set D is narrowed down to this set of objects. In both cases the property *location* with the value *pointingAt* indicating a pointing gesture is added to R.

For determining the other properties we use a simplified version of the incremental algorithm of Dale and Reiter (1995), which tests every property in P wrt its discriminatory power (Fig. 22, 2.). Our algorithm is simplified in as much as in our current implementation the *findBestValue* function defined by Dale and Reiter is replaced by the simpler *getValue* function. The task of *findBestValue* is to search for the most specific value of an attribute that both, discriminates the referent r from more elements in D than the next general one does, and is known to the addressee. Only for the special case *type* we realise this search of the appropriate vaue on a specialisation hierarchy ("screw" instead of "pan head slotted screw" is used). We operate in a highly simplified domain with objects characterised by properties having only a few and well distinguished values. Thus, for the other prperties like colour we do not need such a sophisiticated approach.

However, extending the basic algorithm by Dale and Reiter we also account for relationally expressed properties often found in our corpus. To evaluate these properties we use a function named *getRelationalValue*. This function needs a partial order for each property; in the current system this is only implemented for size and relative position. In the case of size we relate the property to the shape of the objects under discussion. Shape is a special property often used if the type of an object is unknown but is difficult to handle in generation. Therefore, we currently only account for it by evaluating size. The shape of some of the objects in our domain is characterised by

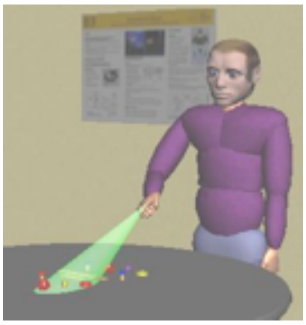
```
contentSelectRE(referent r, properties P, context set C)
      restricting properties R ← {}
      distractors D ← C
      α ← objectPointingConeApexAngle
      β ← regionPointingConeApexAngle
1.    if    reachable?(r)
            then  R ← {(location,pointingAt)}
                  ($\vec{h}$,$\vec{r}$) ← generatePointingBeam(r)
                  if    getPointingMap(($\vec{h}$,$\vec{r}$),C,α) = {r}
                        then return R ∪ {type,getValue(r,type)}
                        else D ← getPointingMap(($\vec{h}$,$\vec{r}$),C,β)
2.    for each p ∈ P
            if    relationalProperty?(p)
                  then value v ← getRelationalValue(r,p,D)
                  else value v ← getValue(r,p)
            if    v ≠ null ∧ rulesOut(p,v,D) ≠ {}
                  then  R ← R ∪ {(p,v)}
                        D ← D \ rulesOut(p,v,D)
            if    D = {r}
                  if    (type,x) R for some x
                        then return R
                        else return R ∪ {type,getValue(r,type)}
      return failure
--------------------------------------------------------------------------------
getPointingMap(($\vec{h}$,$\vec{r}$),C,α)
      pointing map M ← {}
      for each o ∈ C
            $\vec{x}$ ← getPosition(0,$\vec{h}$)
            β ← getAngle($\vec{x}$,$\vec{r}$)
            if    β ≤ α
                  then insert(o,M,α)
      return M

rulesOut(p,v,D)
      return {x|x ∈ D ∧ getValue(x,p) ≠ v}

getRelationalValue(r,p,D)
      if    min{v|v=getValue(x,p) ∧ x ∈ D} = getValue(r,p)
            then return minValue(p)
      if    max{v|v=getValue(x,p) ∧ x ∈ D} = getValue(r,p)
            then return maxValue(p)
      return null
```

Figure 22. The content selection algorithm. It gets the referent, the set of properties holding true for this referent, and the set of objects in the domain under discussion and returns a list of property value pairs. The first part realises the evaluation of pointing using the pointing cone. `generatePointingBeam` generates the pointing beam defined by two vectors, the origin and the direction. `getPointingMap` returns all objects inside the pointing cone defined by the beam and the apex angle. The second part is an adapted version of the incremental algorithm proposed by Dale and Reiter (1995)

one or two designated dimensions. For these objects size is substituted by, e.g., length respectively thickness (“long screw” is used instead of “big screw”). In the case of relative location we also use substitution. The relative location is evaluated along the axes defining the subjective coordinate systems of the dialogue participants (left-right, ahead-behind, and top-down). E.g., *getRelationalValue* returns “left” if the referent r is the left most located object in D.

“Meinst Du die
lange Leiste?”

(Do you mean
the long bar?)



```
<definition>
   <parameter name="NP"/>
   <parameter name="Object"/>
   <utterance>
      <specification>
         Meinst Du <time id="t1"/>$NP? <time id="t2"/>
      </specification>
      <behaviorspec id="gesture_0">
         <gesture>
           <affiliate onset="t1" end="t2"/>
           <function name="refer_to_loc">
              <argument name="refloc" value="$Object"/>
           </function>
         </gesture>
      </behaviorspace>
   </utterance>
</definition>
```

Figure 23. A parameterised utterance specification expressed in MURML (Kranstedt, Kopp, and Wachsmuth 2002). The picture illustrates the resulting animation (German speech) including the visualised pointing cone

The content selection for the example depicted in Fig. 23 can be described as follows: The starting point is a query concerning the reference to a specific object named *five-hole-bar-0*, the intended referent r. As mentioned before, first the pointing cone for object-pointing is evaluated (see Fig. 22, 1.). In this case, more than one object is inside the cone and region-pointing is evaluated next. The cone is visualised in Fig. 23. As a result, the set of distractors D for property evaluation in part two of the algorithm is narrowed down to the two bars *five-hole-bar*-0 and *three-hole-bar-0* and some other objects. The property location with the value *pointingTo* indicating a pointing gesture is added to R. The second part starts with testing the property *type*. The type *five-hole-bar* is too specific, so the super-type *bar* is chosen. It rules out all objects except the two bars (now D = {*five-hole-bar-0, three-hole-bar-0*}), and *type* with the value *bar* is added to R. Next, the property *colour* is tested; it has no discriminatory power concerning the two bars. But the following relational property size discriminates the two objects. The shape of the bars is characterised by one designated dimension, length. For these objects *size* is substituted by *length*. In our case r has the maximum length of all objects in D, the property *length* with the value *long* is added to R. Now D contains only r, the algorithm finishes and returns R = {(*location*, *pointingAt*), (*type*, *bar*), (*length*, *long*)}.

The results of the content selection algorithm represented as a list of attribute-value-pairs 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 template of a multi-modal utterance fetched from a database. The resulting description represents the locutionary act of one single communicative act (that is a multi-modal extension of speech act). As far as communicative acts are concerned, currently instances of the general types *query*, *request*, and *inform* can be expressed.

In the utterance descriptions cross-modal synchrony is established by appending the gesture stroke to the affiliated word or subphrase in the co-expressive speech. Based on these descriptions, an utterance generator synthesises continuous speech and gesture in a synchronised manner (Kopp and Wachsmuth 2004). To replicate the empirical findings an offset of 0.2 seconds between the beginning of the gesture stroke and the affiliate is implicitly added during realisation. In our example (Fig. 23), based on R = {(*location*, *pointingAt*), (*type*, *bar*), (*length*, *long*)} a pointing gesture directed to r is specified, the noun phrase “die lange Leiste” (the long bar) is built, and both are inserted into the utterance template. The complete utterance is synthesised and uttered by the agent Max.

First evaluations of the generation results support the assumption that different apex angles for the pointing cones of region-pointing and object-pointing in settings with high object density are needed. In our VR-setting 40 degrees for region-pointing seems to be a good initial choice to get robust distinctions and natural expressions. However, this has to be investigated in more detail empirically. The concept of the pointing cone based on a set of parameters guarantees that the cone’s form and size can be adjusted as further findings become available.

6. Conclusion

The collaborative research presented in this chapter raised the issue of pointing in complex demonstrations. We approached this issue from interlocked perspectives including empirical research, theoretical modelling and speech-gesture processing in human-computer interaction (see Fig. 24).

Complex demonstrations comprise two fundamental kinds of referring to objects, indicating via pointing and describing using a definite description. The meaning of this kind of utterances is seen as a composition of the meaning of the gesture and the meaning of the verbal expression while the gesture and the definite description are often underspecified by their own. Therefore, we differentiate two referential functions of pointing, object-pointing, referring successfully on its own, and region-pointing, successfully referring only in combination with a description. To model the distance dependent decreasing precision of pointing we introduced the concept of a pointing cone. The pointing cone captures the geometrical aspects of pointing and is used as an interface between the spatial context of pointing and its referential semantics.

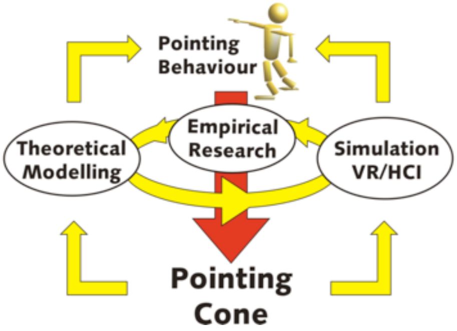


Figure 24. The pointing cone as the central concept is theoretically grounded and empirically measured wrt the needs in speech-gesture processing. Inversely, it constitutes a central building block in the formal construction of the meaning of complex demonstrations and it is essential for setting up efficient methods of processing complex demonstrations in human-machine interaction

In our studies, a genuine effort was undertaken in collecting multi-resolutional empirical data on deictic reference ranging from the high levels of speech acts down to the delicate movements of the fingers. We worked out a detailed procedure to assess the geometrical properties of pointing using tracking technology for measuring the set of parameters relevant for computation of the pointing cone's size and form.

The results concerning the sub-domain determined by the base of the pointing cone serve as a basis for getting at the "pure semantics of pointing". According to the semiotics tradition, the pointing gesture itself can be conceived of as a sign with its own syntax, semantics and pragmatics. Following this lead, we may assume that the pointing gesture in itself is able to determine an extension, much like a proper name or relations as interpreted in logical semantics with respect to a model. As a consequence, the described experimental settings serve as a basis for the construction of realistic models lacking for example in the philosophical literature on demonstration.

Applying the concept of a pointing cone to human-computer interaction it is shown that in reference resolution the cone not only accounts for expressing the extension of pointing. Its topology is also used for generating snapshots of the visual context associated with a gesture in early processing steps. These snapshots allow a low memory profile and help to unfold the restrictive power of pointing by narrowing down the search space and hence speed up the computation of reference.

In utterance generation, we use the empirically determined size of the pointing cone to estimate the borderline of the discriminative power of object-pointing in a planned utterance. If object-pointing does not yield a referent, region-pointing is used to draw the attention of the addressee to a spatial area. The objects inside this area constitute the contrast set for a content-selection based on an adapted version of the incremental algorithm by Dale and Reiter (1995).

It has to be emphasised that the pointing cone as described in this contribution is an idealised concept. Observations from our empirical data indicate that several context dependent parameters influence the focus of a pointing gesture and therefore interact with the geometrical concept of a pointing cone. Especially the focus of region-pointing is influenced by additional spatial constraints on the one hand and the dialogue history on the other. For instance, it seems plausible that region-pointing singles out a whole object cluster even if the corresponding pointing cone does not cover the whole cluster. Or it may be clear to the interlocutors that a pointing gesture singles out a specific set of objects, even if the cone covers additional objects because they just talked about this set.

Extending our approach to incorporate dialogue semantics and pragmatics, a first step can be taken in the following way. Instead of using a model for success and satisfaction of directives along the lines of Searle and Vanderveken (1989) which now has contexts of utterance $i \in I$ with five constituents, speaker a_i, hearer b_i, time t_i, location l_i and the world w_i, we can also take account of the description giver's position at the table, positions of trunk, head, hand, index finger, the apex angle etcetera. We can then let the interpretation of the gesture depend on these fine-grained parameters and say that, relative to such and such parameters, the demonstration's extension will be such and such. This will be a refinement in comparison with the pure semantics approach moving the whole issue into the direction of "classical" pragmatics but still relying on an objective ontology.

As far as we can tell from experiments it could well be that real object-identifiers lack the full interpretive power of both, pure semantics and classical pragmatics. A case in point is the little multi-modal dialogue analysed, where we have a clarification question and the referent of the preceding multi-modal reference act is determined by agents' coordination. This moves us more into the direction of speaker's meaning which relies on the speaker's individual possibilities given the situation at hand. Classical paradigms, situated in a Platonic realm, will not always do justice to speakers' worldly reactions.

References

Asher, N., and A. Lascarides
2003 *Logics of Conversation.* Cambridge UK: Cambridge University Press.

Beun, R.-J., A. and Cremers
2001 Multimodal reference to objects: An empirical approach. In *Proceedings of Cooperative Multimodal Communication*, 64-86. Berlin/Heidelberg/New York: Springer.

Butterworth, G.
2003 Pointing is the royal road to language for babies. In *Pointing: Where Language, Culture and Cognition Meet,* S. Kita (ed.), 9-35. Mahwah, NJ: Erlbaum.

Chierchia, G., and S. McConnell-Ginet
2000 *Meaning and Grammar – An Introduction to Semantics.* 2nd edition. Cambridge, MA: MIT press.

Claassen, W.
1992 Generating referring expressions in a multimodal environment. In *Aspects of Automated Natural Language Generation*, R. Dale, E. Hovy, D. Rosner, and O. Stock (eds.). Berlin/Heidelberg/New York: Springer.

Clark, H. H.
1996 *Using Language*. Cambridge, UK: Cambridge University Press.
2003 Pointing and placing. In *Pointing: Where Language, Culture and Cognition Meet,* S. Kita (ed.), 243-269. Mahwah, NJ: Erlbaum.

Dale, R., and E. Reiter

1995 Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 18: 233- 263.

de Ruiter, J. P.

2000 The production of gesture and speech. In *Language and Gesture,* McNeill D. (ed.), 284-312. Cambridge, UK: Cambridge University Press.

Grosz, B., and C. Sidner

1986 Attention, intention, and the structure of discourse. *Computational Linguistics* 12: 175-206.

Kendon, A.

1981 *Nonverbal Communication, Interaction, and Gesture*. The Hague: Mouton.

2004 *Gesture. Visible Action as Utterance*. Cambridge, UK: Cambridge University Press.

Kopp, S., and I. Wachsmuth

2004 Synthesizing multimodal utterances for conversational agents. *Comp. Anim. Virtual Worlds.* 15: 39-52.

Krahmer, E., and I. van der Sluis

2003 A new model for the generation of multimodal referring expressions. In *Proceedings European workshop on Natural Language Generation (ENLG 2003)*.

Kranstedt, A., S. Kopp, and I. Wachsmuth

2002 MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop Embodied Conversational Agents – let's specify and evaluate them!*

Kranstedt, A., P. Kühnlein, and I. Wachsmuth

2004 Deixis in multimodal human computer interaction: An interdisciplinary approach. In *Gesture-based communication in human-computer interaction*, A. Camurri and G. Volpe (eds.), 112-123. Berlin/Heidelberg/New York: Springer.

Kranstedt, A., A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth

2005 Deixis: How to determine demonstrated objects using a pointing cone. In *Proceedings of the 6th International Workshop on Gesture in Human-Computer Interaction and Simulation.* Berlin/Heidelberg/New York: Springer. To appear.

Kranstedt, A., and I. Wachsmuth

2005 Incremental generation of multimodal deixis referring to objects. In *Proceedings of the European Workshop on Natural Language Generation (ENLG2005)*, 75-82.

Krauss, R. M., Y. Chen, and R. F. Gottesman

2000 Lexical gestures and lexical access: A process model. In *Language and Gesture*, D. McNeill (ed.), 261-284. Cambridge, UK: Cambridge University Press.

Kühnlein, P., and J. Stegmann

2003 Empirical issues in deictic gesture: Referring to objects in simple identification tasks. Technical report 2003/3, SFB 360, University of Bielefeld.

Latoschik, M. E.

2001a A general framework for multimodal interaction in virtual reality systems: PrOSA. In *The Future of VR and AR Interfaces - Multimodal, Humanoid, Adaptive and Intelligent. Proceedings of the workshop at IEEE Virtual Reality 2001*, W. Broll and L. Schäfer, (eds), 21-25.

2001b *Multimodale Interaktion in Virtueller Realität am Beispiel der virtuellen Konstruktion.* (infix DISKI Vol. 251). Berlin: Akademische Verlagsgesellschaft Aka GmbH.

2003 Designing transition networks for multimodal VR-interactions using a markup language. In *Proceedings of the IEEE fourth International Conference on Multimodal Interfaces, ICMI 2002*, 411-416.

Lester, J., J. Voerman, S. Towns, and C. Callaway

1999 Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence* 13 (4-5): 383- 414.

Levelt, W. J. M.

1989 *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levin, J. A., and J. A. Moore

1977 Dialogue games: Meta-communication structures for natural language interaction. In *ISI/RR-77-53*. Information Sciences Institute, Univ. of Southern California.

Lücking, A., H. Rieser, and J. Stegmann

2004 Statistical support for the study of structures in multimodal dialogue: Interrater agreement and synchronisation. In. *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog '04)*, 56-64.

Lyons, J.

1977 *Semantics.* Volume 2. Cambridge, UK: Cambridge University Press.

Mann, B., and S. A. Thompson

1987 Rhetorical structure theory: A framework for the analysis of texts. In *International Pragmatics Association Papers in Pragmatics* 1: 79-105.

Mann, W.C.

1988 Dialogue games: Conventions of human interaction. *Argumentation* 2: 512-532.

Masataka, N.

2003 From index-finger extension to index-finger pointing: Ontogenesis of pointing in preverbal infants. In *Pointing: Where Language, Culture and Cognition Meet,* S. Kita (ed.), 69-85.

McNeill, D.
1992 *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.
2000 Catchments and contexts: Non-modular factors in speech and gesture production. In *Language and Gesture*, D. McNeill (ed.), 312-329. Cambridge, UK: Cambridge University Press.
2003 Pointing and morality in Chicago. In *Pointing: Where Language, Culture and Cognition Meet,* S. Kita (ed.), 293-306.

Milde, J.-T., and U. Gut
2001 The TASX-environment: An XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on linguistic databases*.

Pfeiffer, T., and M. E. Latoschik
2004 Resolving object references in multimodal dialogues for immersive virtual environments. In *Proceedings of the IEEE Virtual Reality 2004*, 35-42.

Pickering, M. J., and S. Garrod
2004 Towards a mechanistic psychology of dialogue. In *Behavioral and Brain Sciences* 27: 169-226.

Piwek, P., and R. J. Beun
2001 Multimodal referential acts in a dialogue game: From empirical investigations to algorithms. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*, 127-131.

Piwek, P., R. J. Beun, and A. Cremers
1995 Demonstratives in dutch cooperative task dialogues. IPO Manuscript 1134, Eindhoven University of Technology.

Poesio, M., and D. Traum
1997 Conversational actions and discourse situations. In *Computational Intelligence* 13 (3): 1-44.

Reithinger, N.
1992 The performance of an incremental generation component for multimodal dialog contributions. In *Aspects of Automated Natural Language Generation*, R. Dale, E. Hovy, D. Rosner, and O. Stock (eds.). Berlin/Heidelberg/New York: Springer.

Rieser, H.
2004 Pointing in dialogue. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog '04)*, 93-101

Searle, J. R., and D. Vanderveken
1989 *Foundations of Illocutionary Logic*. Cambridge, UK: Cambridge University Press.

Tramberend, H.
2001 Avango: A distributed virtual reality framework. In *Proceedings of Afrigraph '01.*

van der Sluis, I., and E. Krahmer
2004 The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the 8th International Conference on Spoken Language Processing*.