Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach

Alfred Kranstedt[1], Peter Kühnlein[2], and Ipke Wachsmuth[1]

[1] Artificial Intelligence Group, Faculty of Technology, University of Bielefeld, D-33594 Bielefeld, Germany
{akranste,ipke}@techfak.uni-bielefeld.de
[2] Linguistics Dept., Faculty of Linguistics & Literary Science, University of Bielefeld, D-33594 Bielefeld, Germany
p@uni-bielefeld.de

Abstract. Focusing on deixis in human computer interaction this paper presents interdisciplinary work on the use of co-verbal gesture . Empirical investigations, theoretical modeling, and computational simulations with an anthropomorphic agent are based upon comparable settings and common representations. Findings pertain to the coordination of verbal and gestural constituents in deictic utterances. We discovered high variability in the temporal synchronization of such constituents in task-oriented dialogue, and a theoretical treatment thereof is presented. With respect to simulation we exemplarily show how the influence of situational characteristics on the choice of verbal and nonverbal constituents can be accounted for. In particular, this depends on spatio-temporal relations between speaker and the objects they refer to in dialogue.

1 Introduction: Interdisciplinary gesture research

The formation of new fields of applications, e.g. in virtual reality, raises new demands on how to interact with computational systems. In response, there are numerous attempts at utilizing natural multi-modal communication skills humans employ in face-to-face conversation. This includes the reception and generation of synchronized verbal and nonverbal utterances. The development of computational models requires detailed knowledge about the involved mechanisms, in particular the deep coordination of speech and gestures and their formation throughout the production of utterances. Specific investigations induced by the needs of simulation work can help close knowledge gaps.

In this paper, we present our research into the use of multi-modal deixis in a restricted domain of assembly tasks. From the viewpoint of linguistics there is quite a remarkable extent of theoretical discussion concerning the way deixis works. The status of gestures is not unequivocally accounted for by the different theories in the philosophy of language community, cf. Section 2. The works of Kaplan [5] and his successors serve as a model for linguistic reference by

demonstratives that in our view can plausibly be accommodated to cover pointing. There is no comparably easy way to translate other manners of gesturing into linguistic theories without begging certain relevant questions. On the other hand, simulative work dedicated to deixis offers an opportunity to develop exemplary computational production models of utterances most often expressed in more than one modality and structured in direct relation to the spatio-temporal environment.

The interdependence between the empirical, theoretical and simulative parts and their methodical connection described in Section 3 is based on a common setting called pointing games. We describe empirical studies realized in these setting concerning (1) the temporal synchronization between speech and pointing gestures; (2) the influence of the spatio-temporal restrictions of the environment, in particular the density of the objects referred to, on deictic behavior. The transfer between empirical, theoretical, and simulative work is fostered by the creation of translations between the annotations we use in our empirical studies and a common representation language for utterance descriptions, MURML[9].

In Section 4 we outline some findings and their implications on theoretical modeling of the phenomenon deixis. An extension of our theoretical framework is proposed to deal with the high temporal variability in synchronization we found in the task-oriented dialogues we investigated. Moreover, it is shown how the empirical results guide the enhancement of the virtual anthropomorphic agent MAX. MAX was developed as a mediator in an immersive 3D virtual environment for the simulation of assembly tasks [25, 11]. He is able to grasp simple multimodal instructions which may include deictic and iconic gestures, and to produce smooth co-verbal gestures in synchrony with synthetic speech from descriptions of their surface form in real time [7, 8]. For deixis these descriptions can be generated from abstract prototypical templates in a situated planning process.

2 Related Work

There is hardly any research dedicated to the phenomenon of deixis with the goal of computational simulation that implicates more than simple pointing. As a subcase of co-verbal gesture, pointing is usually treated as a putatively simple case. The concurrence of deictic gestures and corresponding verbal expressions, however, is not explicitly acknowledged.

The empirical data available in this area of research is also quite sparse. There is noteworthy work by Paul Piwek from ITRI in Brighton and co-workers, cf. [21, 22], that aims in the same direction as our work does. A problem is that their findings seem to be specific for Dutch.

There is not much work comparable to our approach in the field of the theoretical investigations reported, either. Philosophically, the status of gestures is a matter of controversy. Researchers in the line of Kaplan [5], take the stance that (1) gestures are logically irrelevant for determining the meaning of demonstrative expressions and (2) the speaker's intentions already suffice to fix meanings. In contrast, we adopt in our theoretical work what we call a *neo-Peirce-Quine-*

Wittgenstein view. According to this position, gestures are part of more complex signs and have to be treated on a par with speech. Few other researchers in the linguistics community adopt this view when it comes to formal modeling. E.g., in the STAGING project a related integrative account is pursued, utilizing attribute-value grammar and unification [20]. However, an elaborate theory of integration is still missing.

In simulative work on utterance generation, the production of speech has gained most attention, in recent work replenished, modified or partly substituted by gesture. Current approaches, e.g. [19, 2], enrich speech with nonverbal behaviors fetched from a database. But investigations on human-human communication reveal an intricate temporal synchronization between speech and gesture related to semantic and pragmatic synchronization.

In recent psycholinguistic work several models of speech-gesture production have been proposed to approach this problem. While *RP-models*[3] [4, 3, 10] extend the speech-production model of Levelt [12] and suggest a parallel production process in specialized modules, McNeill [16, 17] emphasizes the inseparable connection of the modalities. If we want to follow the computationally more manageable RP-approach the examplary treatment of deixis can help to realize the cross-modal interactions on the different levels of the production process which are inexplicit in the suggested models. In section 4 we describe how the environmental restrictions on successful pointing influence the conceptualization of the co-articulated speech.

3 Methodological Issues

3.1 The Pointing Game Scenario

We investigate deixis in a reduced setting of interaction between two agents we call pointing games. Pointing games inherit their names from the dialogue games as proposed by [13, 14]. We start with the minimal form of these games consisting only of two turns. The underlying idea of pointing games is to integrate signal interpretation and the generation of an answer in one unique setting. This gives the chance to investigate deixis in dialogue imagined as part of an interactive process of fixing reference.

Pointing games are embedded in instructor-constructor settings involving the assembly of *Baufix*[4]-aggregates, e.g. toy airplanes, investigated in the Collaborative Research Center "Situated Artificial Communicators" (SFB 360).

In parallel to the empirically investigated human-human setting we build a human-computer interaction scenario embedded in immersive virtual reality (VR), realized in a three side cave-like installation (see Fig. 1). MAX and the user are located at a table with Baufix parts and communicate about how to assemble them.

[3] **R**epresentations and **P**rocesses, models related to the information processing approach.

[4] Baufix is the trade name of a children's construction toy used in our scenario.

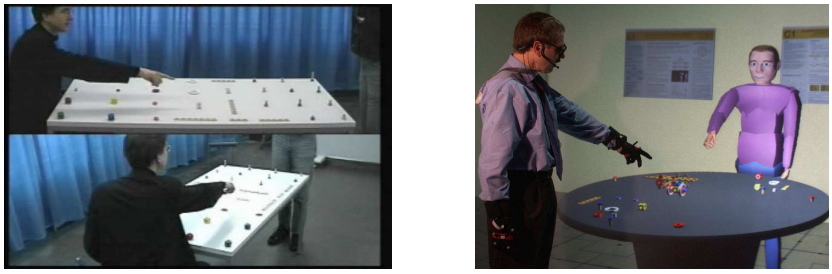

Fig. 1. The pointing game scenario: Comparable settings for empirical studies and VR

Intended empirical studies realized in this VR-setting are aimed at evaluating MAX capacity to interpret and generate situated co-verbal gesture and testing for the naturalness and acceptability of the simulation.

3.2 From Data to Qualitative Descriptions: Empirical Settings and Preparation of the Collected Data

We conducted a number of studies to obtain data concerning the timing relations and spatial conditions for successfully received pointing. A central question was whether in construction dialogues the temporal relations are the same as reported in the literature for narratives and related gestures. It was not clear if the timing would be the same, given the fact that the task was completely different and gestures serve a different purpose during pointing. The spatial conditions we investigated concern how the perceived density of objects influenced the pointing behavior of subjects.

For our studies we used a carefully selected setting. Two subjects, one called instructor, the other constructor, were to cooperate in identifying the atomic parts of a toy airplane distributed on a table. Instructor had to tell constructor via some multi-modal utterance which of the parts s/he had in mind, and constructor in turn had to briefly lift the object to indicate that s/he had identified it and then put it back in place.

The physical density of the objects on the table (though not the *perceived* density) was constant all over the area and did not change over time. The gestures performed to identify objects were very likely to be simple pointing gestures with the fingers. The dialogue patterns were expected to be simple, as were the sentential constructions.

The studies were video-graphed and annotation was done using the TASX annotator, cf. [18], a tool that allows to annotate of sound and video data without prescribed categories. We can thus use the inventory of MURML [9], a symbolic description of the surface shape of utterances developed for the simulation of speech and gesture with MAX. We devised an XSLT script as an output filter for TASX that produces complete MURML descriptions. The XSLT script serves the purpose of transforming the TASX output by reduction of information, extracting qualitative description from quantitative data, and ordering them in a

hierarchical structure. This step opens the path to the inverse of the generation of quantitative data by the empirical part of the project in that the behavior of MAX can be studied using qualitative judgments in turn.

3.3 Utterance Form Descriptions as a Link between Analysis of Coverbal Gestures and their Simulative Synthesis

Organizing utterance generation as a process on several levels, a qualitative description of their overt shape can be an important link between the mental planning process and the physical realization of utterances. A notation system of such descriptions, MURML, was developed as starting point for MAX's generator modules which form a hierarchical system of controllers computing the upper-limb movements and feed the text-to-speech system [7, 8].

We adopt the empirical assumption [16] that continuous speech and gesture are co-produced in successive units each expressing a single idea unit. The correspondence between gesture and speech at the surface realization is commonly assumed to depend on a correspondence between certain units on different levels of the hierarchical structure of both modalities [6, 16]. As introduced in [9], we define *chunks* of speech-gesture production to consist of an intonational phrase in overt speech and a co-expressive gesture phrase (see Fig. 2). Within each chunk, the gesture stroke corresponds to the focused constituent in the intonational phrase (the *affiliate*) that carries the nuclear accent. Complex utterances with multiple gestures are conceived as being divided in several chunks.

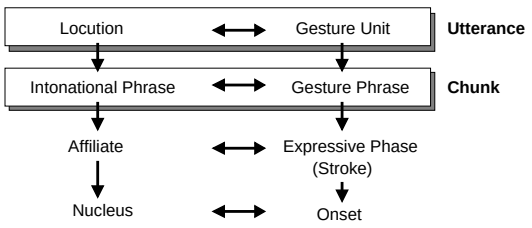

Fig. 2. Units of speech-gesture correspondence (taken from [9])

MURML utterance specifications are hierarchically organized in an XML-notation [9]. They start from the textual output augmented with certain points in time expressing the correspondence between speech and the subsequently defined gesture by specifying the affiliate's onset and end. Gestures can either be stated by specifying a communicative function or be explicitly described symbolically in terms of their spatiotemporal features. We use an inventory derived from the sign language notation system HamNoSys [23]. The optional parametrization of all features give us the possibility to define prototypical templates for frequently used utterances and instantiate them during the generation process, adapting the situational requirements. This will be exemplified in Section 4.3.

4 Results and Discussion

4.1 Empirical Findings

The results of the described studies were in part unexpected. E.g., pointings occurred that had their onset *after* the associated expression. In other cases the pointing took place after all the linguistic material was uttered. This is evidence that the timing relations in construction dialogues are more varied than timing relations during narratives, with no such events reported there [16].

"Good"[5] cases are those where the stroke of the gesture lies before or on the noun phrase that is its affiliate. These cases are good, because the gesture can be understood as being a normal modifier for the noun phrase. Compare the following linguistic constructions:

(1) Take the yellow bolt.
$V \quad Det\ Adj \quad N$

(2) Take the ↘ bolt.
$V \quad Det - \quad N$

Yellow is an adjective that modifies the noun *bolt*. This modifier can be seen as an operator that takes the following phrase as an argument, hence binds to the right (cf. section 4.2). The symbol "↘" is used to represent the stroke of a pointing gesture, and is intended to indicate that the stroke occurs at the time between uttering "the" and "bolt" in the present example. So, when the gesture is a semiotic object on a par with linguistic signs, the gestures stroke and the modifying adjective are of the same category here.

Pointings with strokes *after* the linguistic material are called "bad", because the model of the modifying adjective that is supposed to be of the same category breaks down. An utterance like (3) simply is not well-formed in English[6], but pointings like (4) are perfectly possible. Cf. section 4.2 for discussion.

(3) * Take the bolt yellow.
$V \quad Det\ N \quad Adj$

(4) Take the bolt ↘.
$V \quad Det\ N \quad -$

Concerning the spatial conditions, we observed a two-way interaction between perceived density and complexity of linguistic material that can be measured counting the frequency of pointing behavior, cf. Fig. 3. Whenever the perceived density is very low, instead of pointing with fingers, subjects seem to use gaze direction as a pointing device. When there is a mid-density of objects, they use noun phrases with low complexity and frequently point with their fingers. High density surprisingly leads to a slight decrease in pointing and an increase of the complexity of noun phrases.

The borderline between the far and the mid area is an indicator of the resolvability of human pointing. On the evidence of these data, we could determine the size of a *pointing cone* as being of $\approx 8°$ around the axis of the pointing finger. The

[5] "Good" in the sense that they are simple to treat with our theoretical apparatus.

[6] As one of the reviewers pointed out, adjectives in post-position are perfectly acceptable in, e.g., French. Our claim that the meaning of pointing gestures can be assimilated to linguistic meanings is, however, not touched by that fact. Our model is in this regard parametried for English (and German) grammar.

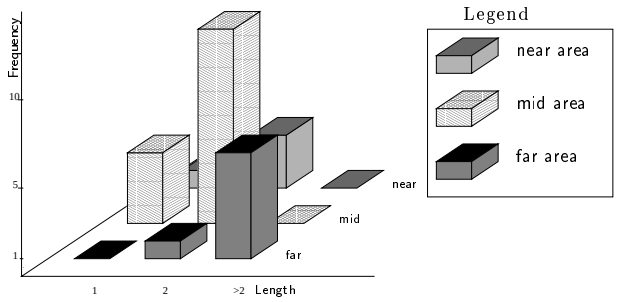

Fig. 3. The two-way interaction between perceived density and complexity of the linguistic material. "Frequency" denotes the frequency of successful gestures, "Length" the length of their affiliate in speech

behavioral clues we used were the conditions under which instructors were able to point out objects to constructors successfully. The results were correlated with the perceived object density, the distance between the intended object and the closest one relative to the distance of the finger root to the object and the angle between object plane and pointing ray. We hypothesize that a similar pointing cone can be found for the case where eye gaze is used for pointing[7].

4.2 Theoretical Modeling: What about the Bad Cases?

Theoretical modeling at the moment is restricted to deal with the temporal phenomena observed, and does not comprise the spatial structure of the domain. Doing that step will involve using structured models for the semantics, planned at a later time.

Our current interest is to find a sound and up-to-date explanation of the interface between syntax and semantics of speech-gesture complexes. Facing the problem that we want to obtain computer-implementable results we strive for mathematically tractable methods. We adopted a constraint-based version of HPSG, which had to be enriched in a multitude of ways. The typed feature structure descriptions that form the representations of the HPSG analyses spell out under which conditions a given (multi-dimensional) sign is well-formed. We set up a type-logical interpretation for the syntax, instead of the flat semantics currently favored.

From a logical point of view the results reported in section 4.1 mean that the gesture corresponds to a polymorphic operator. The standard case for operators in linguistics is that they are taken to bind to the right, which means that subsequent material can be in the scope of the operator. It is hence straightforward to give a semantics for a pointing gesture in the "good" cases. It is not even difficult in principle to define an operator as binding to the left and then use it as an interpretation for a gesture in the "bad" cases. But it is unsatisfactory to have a multiplicity of interpretations in contrast to uniform interpretations.

[7] This will be verified in future studies using eye tracker technology.

For a type-logical treatment this multiplicity means that there have to be multiple solutions at first, and some subsequent filter. Things here are not even easy for the "good" cases, as different realizations of gesture positions are possible. Accordingly, for the "good" cases like (1), there could be six possible interpretations for the pointing gesture, four of which would be different in logical type. The various interpretations that are possible at each of the positions of the gesture imply that the corresponding logical operator is not only polymorphic, but also polysemous.

For the "bad" cases (where $\searrow$ follows the relevant linguistic material in temporal order) we have all the interpretations that are possible for the "good" cases, except that the binding is in the other direction. An example that is close enough to the "bad" case (4) could be (5):

(5)
Take	the	bolt	that	lies	rightmost
V	Det	N	RC		

It is obvious that post-modification as exemplified in (5) is possible. (For a discussion of the use of post-modification in Dutch and English in the context of task oriented dialogues and pointing see [21, 22].) And it is obvious that this parallels the "bad" cases again, cf. (4). Note that mixed cases again add one dimension of complexity. Let this suffice as an indication of how the semantics of the gesture is treated and how the results of the studies influence and inform the semantic representation. Analogously, if a syntactic representation of speech plus gesture is desired, it has to respect the complex data that were found in the studies.

In our project—and in contrast to [20]—we developed a syntactic apparatus, based on the suggestions by [24], which uses constraints in order to define well-formed multi-modal expressions. The interface we got thus far contains a lexical component with definitions of the logical types of the linguistic and gestural material as discussed above. To be clear, "lexical component" here means that the entries for the lemmas contain rules for the uses of expressions (for the gestures, e.g.) and only in cases of rigid designators it contains also the values for reference. Following logical tradition, we view pronouns as carrying values only if they were uttered on a certain occasion. Analogously, pointing gestures are in a sense lexicalized, but this does by far not mean that their reference is fixed in the lexicon. Rather, we have a multiplicity of rules for uses of pointing gestures in order to capture the polysemy and polymorphy discussed above.

The calculation of utterance meanings then is rather straightforward. The semantic composition follows the syntactic analysis just as usual. Here it proves useful to have a type-logical apparatus, as the calculations done within this framework have been especially well-studied.

4.3 Simulative Synthesis

The empirical results with respect to the "bad cases" of synchronization and the high failure rate in human-human communication give us an idea of the difficulties and the limits of getting machines to understand multimodal utterances.

"Do you mean the long bar?"

```
<definition>
  <parameter name="NP" default="this one"/>
  <parameter name="Object"/>
  <utterance><specification>
      Do you mean <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"/>
        <argument name="frame_of_reference" value="world"/>
      </function>
    </gesture></behaviorspec></utterance>
</definition>
```

Fig. 4. Sample abstract MURML specification of a deictic utterance fetched from a database and the simulative realization. The variable "Object" is instantiated with the identifier of the intended object and in a further step by its coordinates. The variable "NP" in the speech specification is substituted in the planning process by an object description that allows the user to identify the intended object in the scene. This depends on the discriminating power of the pointing gesture. For a more detailed description see the text

But we can learn from the way humans handle misunderstandings, namely, they clarify them in dialogue. Analogously, we reduce the analytic process w.r.t. synchronization on the frequent "good cases" and relocate gaps of understanding in the interaction. In most cases, these apply to a further enquiry. Furthermore, there are no cues which suggest that the "bad cases" are an indispensable part of human communication. This allows reducing the range of our agents' communicational behavior to the "good cases" and automatizing them in the generation process. Thus, we have defined an implicit parameter representing the offset between the beginning of the gesture stroke and the affiliate in the co-expressive speech that has the default value 0.2 seconds, but can optionally be redefined in MURML.

The described empirical investigations suggest a "zoning" of the pointing area depending on the perceived object density. We can make these results fruitful for utterance planning using the pointing cone as a central concept. This approach has the advantage of offering a distance-invariant description of the observed phenomena. The *near* area then is that area where all pointing is unambiguous, meaning only one object is in the pointing cone, the perceived object density is low. The *far* area is the area where an unambiguous pointing gesture is impossible, the perceived object density being high.

This view can be used for the assignment of information to the modalities and the adaption of the overt shape of the planned utterance to conditions resulting from the environment. A reasonable heuristics in utterance planning should be the minimization of the extent and complexity of the target utterance. Deixis is a very good example of utterance types we could build up from prototypical

templates representing the simplest form to utter, here a simple pointing gesture connected to a short unspecific noun phrase in speech like "this one".

Beginning with a prototypical template fetched from an utterance database we can describe the realization in two planning steps. In the example illustrated in Fig. 4 the MURML description of the template scripted on the left contains an abstract function for a pointing gesture and a parametrized speech specification. The first step is the assignment to the modalities beginning with a check if a pointing gesture with only the desired object in the pointing cone is possible. To check this, an approximate hand position on the line from a point between the shoulders to the intended object is anticipated. If more than one object is detected in the estimated pointing cone, an adequate verbal restrictor is chosen discriminating the objects in the pointing cone by a comparative analysis of the object knowledge in the following order: *color, general type* (typically descriptions in natural speech), *shape, size*, and *location*. For a discussion of order of adjectives and an overview of the literature for German see [26]). The first discriminating attribute, preferably the color, is used to specify the reference-object. In our example there are two bars with the same color, so the first discriminating attribute is size relative to the shape, that is, length.

The second step contains the parallel realization in the involved modalities. A text production module replaces the variables in the speech specification with a syntactical correct combination of the chosen verbal restrictors. The function "refer_to_loc" in the behavior specification is substituted by a shape description containing parametrized hand/arm configurations and movement constraints in terms of the three features hand shape, hand orientation, and hand position. In a further step this feature descriptions must be adapted to the spatial requirements, in the case of a pointing gesture the direction and distance of the referred object. Finally, cross-modal correspondence is established by appending the coverbal behavior to its respective chunk. Its affiliate in speech receives a pitch accent. The resulting utterance plan feeds the text-to-speech system and the motor planner that generates a hierarchical system of movement control primitives (for details see [8]).

5 Conclusions and Perspective

It was explained how the interdisciplinary approach taken in our project furthers the understanding of the functioning of deictic gestures and leads way to their natural simulation. The empirical data have guided theoretical modeling in that they made salient the relevant cases concerning timing. They also revealed findings to support simulations for natural human-machine interactions.

In future work we will intensify the role of the virtual agent MAX in the empirical investigations. As an example, MAX can be programmed to perform the gesture-to-speech synchronization at different points in time for uttering one and the same sentence in multiple ways. When Max interacts with subjects under varied conditions, their reactions are expected to show whether the changes in MAX's behaviour are comparable to effects in human-human interactions.

We are currently exploring dialogue models that go beyond two-turn sequences of speech acts based on the *dialogue games theory* as proposed by [13] and extended by [15]. The syntax-semantics interface developed in the theoretical part of the project is well suited for an extension in that line.

Acknowledgment

This research is partly supported by the Deutsche Forschungsgesellschaft (DFG) in the Collaborative Research Center "Situated Artificial Communicators"(SFB 360).

References

1. Niels Ole Bernsen and Oliviero Stock, editors. *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue* — IPNMD 2001, *Verona, Italy.* ITC-irst, December 14–15, 2001.
2. J. Cassell, H. Vilhjalmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In Eugene Fiume, editor, *Proceedings of SIGGRAPH 2001*, pages 477–486. ACM Press/ACM SIGGRAPH, 2001.
3. Justine Cassell and Scott Prevost. Distribution of Semantic Features Across Speech and Gesture by Humans and Machines. In *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, 1996.
4. Jan Peter deRuiter. The production of gesture and speech. In David McNeill, editor, *Language and gesture*, chapter 14, pages 284–311. Cambridge University Press, 2000.
5. David Kaplan. On the Logic of Demonstratives. *Journ. Phil. Logic*, 8:81–98, 1979.
6. Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. The Hague, Mouton, 1980.
7. S. Kopp and I. Wachsmuth. A Knowledge-based Approach for Lifelike Gesture Animation. In W. Horn, editor, *ECAI 2000 - Proceedings of the 14th European Conference on Artificial Intelligence*, pages 663–667, Amsterdam, 2000. IOS Press.
8. S. Kopp and I. Wachsmuth. Model-based Animation of Coverbal Gesture. In *Proceedings of Computer Animations 2002*, pages 252–257. IEEE Press, 2002.
9. A. Kranstedt, S. Kopp, and I. Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents — let's specify and evaluate them*, Bologna, Italy, July 2002.
10. R. Krauss, Y. Chen, and R. Gottesman. Lexical gestures and lexical access: a process model. In D. McNeill, editor, *Language and gesture*, chapter 13, pages 261–283. Cambridge University Press, 2000.
11. Marc Erich Latoschik. A Gesture Processing Framework for Multimodal Interaction in Virtual Reality. In A. Chalmers and V. Lalioti, editors, *Afrigraph 2001, 1st International Conference on Computer Graphics, Virtual Reality and Visualization in Africa, 5 - 7 November 2001*, pages 95–100, New York, NY 10036, 2001. ACM SIGGRAPH.
12. W. J. Levelt. *Speaking.* MIT Press, Cambridge, Massachusetts, 1989.

13. James A. Levin and James A. Moore. Dialogue Games: Metacommunication Structures for Natural Language Interaction. *Cognitive Science*, 1(4):395–420, 1978.
14. William C. Mann. Dialogue Games: Conventions of Human Interaction. *Argumentation*, 2:512–32, 1988.
15. William C. Mann. Dialogue macrogame theory. http://www-rcf.usc.edu/ billmann/dialogue/dmt-paper1.htm, 2002. Revised version of a paper presented at SIGdial, Philadelphia, Pennsylvania USA, July 2002.
16. David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
17. David McNeill. Models of speaking (to their amazement) meet speech-synchronized gestures. Obtained from the cogprints archives: http://cogprints.soton.ac.uk/, 1998.
18. Jan-Torsten Milde and Ulrike Gut. The TASX-environment: an XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on linguistic databases, Philadelphia*, 2001.
19. T. Noma and N. Badler. A Virtual Human Presenter. In *Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent*, pages 45–51, 1997.
20. Patrizia Paggio and Bart Jongejan. Multimodal communication in the virtual farm of the STAGING project. In: [1], pages 41–45, 2001.
21. P. Piwek, R.J. Beun, and A. Cremers. Demonstratives in Dutch Cooperative Task Dialogues. IPO Manuscript 1134, Eindhoven University of Technology, 1995.
22. P. Piwek and R. J. Beun. Multimodal Referential Acts in a Dialogue Game: From Empirical Investigation to Algorithm. In: [1], pages 127–131, 2001.
23. Siegmund Prillwitz. *HamNoSys. Version 2. Hamburger Notationssystem für Gebärdensprachen. Eine Einführung*. SIGNUM-Verlag, 1989.
24. Ivan Sag and Thomas Wasow. *Syntactic Theory — A Formal Introduction*. CSLI, 1999.
25. T. Sowa, S. Kopp, and M.E. Latoschik. A Communicative Mediator in a Virtual Environment: Processing of Multimodal Input and Output. In: [1], pages 71–74, 2001.
26. Petra Weiß and Stefan Barattelli. Das Benennen von Objekten. In Th. Herrman and J. Grabowski, editors, *Enzyklopädie der Psychologie: Themenbereich C, Theorie und Forschung*, volume III of *Sprache*. Hogrefe, 2003.