Situated Generation of Multimodal Deixis in Task-Oriented Dialogue

Alfred Kranstedt, Ipke Wachsmuth
Artificial Intelligence Group, Faculty of Technology
University of Bielefeld, D-33594 Bielefeld
{akranste, ipke}@techfak.uni-bielefeld.de

June 18, 2004

1 Introduction

This poster describes ongoing work concerning the generation of multimodal utterances, animated and visualized with the anthropomorphic agent Max. Max is a conversational agent that collaborates in cooperative construction tasks taking place in immersive virtual reality, realized in a three-side CAVE-like installation. Max is able to produce synchronized output involving synthetic speech, facial display, and gesture from descriptions of their surface form [Kopp and Wachsmuth, 2004]. Focusing on deixis here it is shown how the influence of situational characteristics in face-to-face conversation can be accounted for in the automatic generation of such descriptions in multimodal dialogue.

2 Context-dependent conceptualization of deictic utterances

The task-oriented dialogues in our setting pertain to the cooperative assembly of virtual aggregates, e.g., toy aeroplanes. These face-to-face dialogues are characterized by an extensive use of nonverbal modalities both for conveying information and for structuring the interaction. Therefore, speech and gesture production cannot be treated as separated; for discussion, see e.g. [McNeill, 2000]. Additionally, the perceived environment, in particular the spatial relations between speaker, listener, and the objects they refer to, constrains the construction of multimodal utterances. Empirical investigations focusing on the internal structure of multimodal deixis reveal a relationship between the perceivable spatial density of the objects communicated about and the number and complexity of the verbal constituents in the occurring deictic utterances [Kranstedt et al., 2004].

To account for these findings, we propose analogously to Levelt's speech production model a generation of multimodal deixis in three steps, conceptualization, formulation, and articulation [Levelt, 1989]. Conceptualization contains the search of an appropriate combination of communicable object attributes capable to identify the referent in an unambiguous manner. This includes the decision about the modalities in which to utter them. We name these attributes *restrictors* because of their capacity to confine the set of potential referents. Pointing is seen as the most appropriate way to refer to objects visually accessible for both communicators. It directs the attention of the addressee to a spatial region, as an initial step to make the intended object salient.

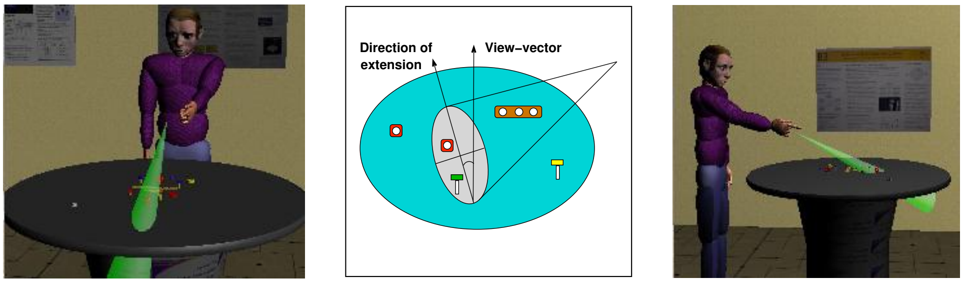
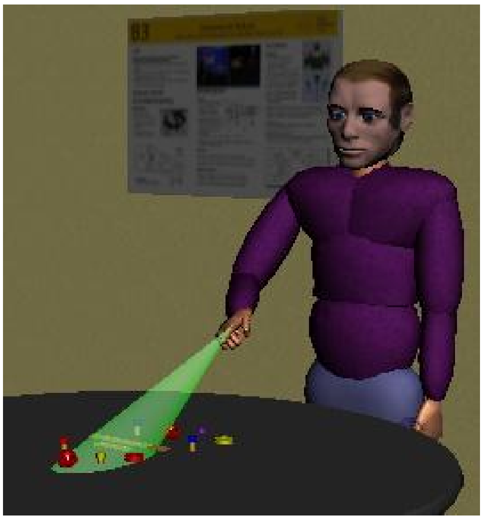


Figure 1: With respect to the visual restrictions the pointing cone is extended in the direction that results from projection of the view-vector of the addressee into the orthogonal cutting plane of the cone; the cut becomes an ellipse.

Based on our empirical results mentioned above [Kranstedt et al., 2004], a pointing cone is modeled to represent the resolvableness of pointing gestures from the perspective of the addressee. Objects in this cone are not distinguishable from each other by the addressee on the base of a single pointing gesture. In its morphology the pointing cone is adapted to the specific restrictions of the display technology we use; see Fig. 1. In the depth the resolvableness is remarkably worse than in the breadth. Such a pointing cone models the first restrictor evaluated during conceptualization. If there is more than one object in the pointing cone, additional attributes of the intended object (type, color, size, and relative position) are recursively evaluated, confining the set of potential referents step by step (for discussion of the naming of objects in German, especially choice and ordering of describing attributes, see [Weiß and Baratelli, 2003]).

Using a simple grammar, the verbal restrictors are syntactically formulated and inserted in an utterance template fetched from a database. The resulting overt form of the utterance is denoted using an XML-based description language for synchronized speech and gesture output and feed into the utterance generators of Max. Based on an incremental model, continuous speech and gesture are co-produced in successive ”chunks”, whereby each chunk is a synchronized pair of an intonation phrase and a co-expressive gesture phrase (for detail cf. [Kopp and Wachsmuth, 2004]).

“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”/>
          <argument name=”frame_of_reference” value=”world”/>
        </function>
      </gesture>
    </behaviorspec>
  </utterance>
</definition>
```

Figure 2: A parameterized utterance specification and the resulting animation (German speech) with the visualized pointing cone

3 Preliminary results and further steps

Currently Max can express simple speech acts of the type *ask*, *actionRequest*, and *confirm*, each including a deictic reference. This is to be extended in ongoing work. Actually only the object attribute position is nonverbally indicated using pointing gestures. One next step will be to use more complex gestures in referential expressions, e.g., iconic gestures to specify form and size features of objects. This will enable Max to utter expressions like

"Give me a bar with this length" to be co-uttered with a two-hand gesture specifying the length of the intended object.

In future work the simple grammar now used will be replaced by a more advanced formalism based on Lexicalized Tree Adjoining Grammar (LTAG). Furthermore, on the way toward more interactivity, verbal and nonverbal feedback signals will be accounted for during production. This gives the chance to monitor if the expressed utterances appear to be understood, and if not, hold and adapt the ongoing turn as necessary.

Acknowledgement

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

References

[Kopp and Wachsmuth, 2004] Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Comp. Anim. Virtual Worlds*, 15:39–52.

[Kranstedt et al., 2004] Kranstedt, A., Kühnlein, P., and Wachsmuth, I. (2004). Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction*, LNAI 2915, pages 436–447. Springer.

[Levelt, 1989] Levelt, W. (1989). *Speaking.* MIT Press, Cambridge, Massachusetts.

[McNeill, 2000] McNeill, D. (2000). *Language and Gesture.* Cambridge University Press, Cambridge, UK.

[Weiß and Baratelli, 2003] Weiß, P. and Baratelli, S. (2003). Das Benennen von Objekten. In Herrman, T. and Grabowski, J., editors, *Enzyklopädie der Psychologie: Themenbereich C, Theorie und Forschung*, volume III, Sprache. Hogrefe.