



# Situated Generation of Multimodal Deixis in Task-Oriented Dialogue



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

## Scenario

- CAVE-like virtual environment
- Cooperative construction tasks
- Anthropomorphic agent (**Max**), able to produce synchronized multimodal utterances (Kopp & Wachsmuth, 2004) including
  - synthetic speech
  - facial display (visems, emotions)
  - gesture (generated from descriptions of their surface form)
- Task oriented face-to-face dialogue, characterized by
  - an extensive use of nonverbal modalities
    - ⇒ Speech and gesture production cannot be treated as separated (McNeill, 2000)
  - a strong influence of the perceived environment: Relationship between perceived spatial object density and number and complexity of verbal constituents in occurring deictic utterances (Kranstedt et al., 2004)

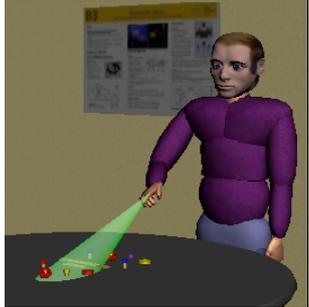


**„Meinst Du die lange Leiste?“**  
**(Do you mean the long bar?)**

```

<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"/>
        </function>
      </gesture>
    </behaviorspec>
  </utterance>
</definition>

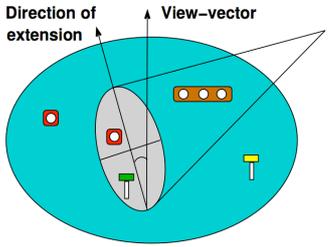
```



## Context-dependent conceptualisation of deictic utterances

**Pointing cone**

- Represents the resolvableness of pointing gestures from the perspective of the addressee
- Stretched in the depth to adapt to the specific restrictions of the display technology
- Objects inside are not distinguishable from each other based only on pointing



- Production model in three steps, see (Levelt, 1989), extended to gesture (de Ruiter, 2000): Conceptualisation, formulation, and articulation
- Conceptualisation includes the search for appropriate object attributes (*restrictors*)
- Pointing is seen as most appropriate way to refer
  - ⇒ Pointing Cone models the first restrictor
- Recursive evaluation of additional restrictors (type, colour, size, and relative position), for ordering cf. (Weiß & Baratelli, 2003)

## Speech-gesture realisation

- Library of parameterized utterance descriptions including speech, gesture and facial expressions
- MURML (XML-based utterance specification language): Describes surface form of synchronized gesture and speech (Kranstedt et al., 2002)
- Instantiation of utterance descriptions using the set of selected restrictors (syntactically correct formulated)
- Successive production of *chunks*, each consisting of an intonation phrase and a co-expressive gesture phrase
- Synchrony within a chunk between the affiliated subphrase and the gesture stroke is accomplished by the gesture adapting to the timing of running speech
- Building appropriate animations using a kinematic figure model and a text-to-speech system (TXT2PHO, MBROLA)

## Further steps

- Enlargement of the speech act repertoire (*ask*, *actionRequest*, and *confirm*)
- Integration of iconic gestures referring to objects size and form attributes
- Integration of an advanced grammar formalism (LTAG)
- Consideration of feedback signals during production

## References

De Ruiter, J. (2000): The production of gesture and speech. In McNeill, D. (ed): *Language and Gesture*, pages 284-311. Cambridge University Press.

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

Kranstedt, A., Kühnlein, P., & Wachsmuth, I. (2004): Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach. In Camurri, A., & Volpe, G., (eds): *Gesture-Based Communication in Human-Computer Interaction*, LNAI 2915, Springer, 436-447.

Kranstedt, A., Kopp, S., Wachsmuth, I. (2002): MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. Workshop proceedings *Embodied Conversational Agents - let's specify and evaluate them*, AAMAS02.

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

McNeill, D. (ed) (2000): *Language and Gesture*. Cambridge University Press.

Weiß, P., & Baratelli, S. (2003): Das Benennen von Objekten. In Herrman, T., & Grabowski, J., (eds): *Enzyklopädie der Psychologie: Themenbereich C, Theorie und Forschung*, Volume III, Sprache. Hogrefe.

