Speech-Gesture Alignment

Hannes Rieser, Stefan Kopp & Ipke Wachsmuth

The project Speech-Gesture Alignment focuses on the investigation of iconic and referential gesture concurrent with speech in route description dialogues. Our research is in many ways inspired by work inaugurated by David McNeill, Justine Cassell and their co-workers combining several methodologies:

(a) experiments in the psycho-linguistic tradition using VR-technology,
(b) classical corpus annotation integrating VR-simulation as an ancillary device,
(c) statistical investigation,
(d) semantical and pragmatical description of structures and, finally,
(e) simulation of speech and iconic gesture production in the embodied agent Max.

Above all, (a) to (d) are concerned with speech and gesture-structures, information about the morphology of gesture tokens, and timing facts. They serve as a basis for describing the function of gestures, their overall meaning, the set-up of multimodal content and, most importantly, to inform the simulation. Simulation of speech-accompanying iconic gesture (e) is our main field of application. It is on the one hand theory-bound and serves as a test-bed and falsification device for (a) to (d) on the other hand. The semantical and pragmatical description of structures (d) proceeds from linguistic structure and gesture morphology, maps gesture morphology onto depictional content and finally into meanings suited to interact with verbal meanings. Aligning interactions will be achieved via a common representation or via context modeling techniques. Here the relevant background methodology is dialogue-oriented dynamic semantics, underspecification theory and Neo-Gricean pragmatics.

At this workshop we present

- our Empirical Study,
  - our Annotation, Preliminary Gesture Typology and Reconstruction of Gesture Meaning,
  - and our approach to Simulating Speech and Gesture Production.

Empirical Study

Kirsten Bergmann, Stefan Kopp, Andy Lücking, Hannes Rieser

In our empirical study we follow the hypothesis that the similarity between a gesture's form and its referent entity is not the only driving force behind a depictive gesture. This is based on experience from a previous study on iconic gestures in direction-giving, which yielded only few, weakly significant correlations between the visuo-spatial features of a referent object and the morphological features of a gesture (Kopp et al., to appear). Thus we now aim at identifying (1) which features of a referent people select under which circumstances, and (2) what strategies they follow in representing them, and (3) how a particular gesture is shaped by that strategy and the selected set of features. In our experimental setup two interlocutors engage in a spatial communication task combining direction-giving and sight descriptions. This scenario is well-suited for systematically studying aspects of natural speech and gesture use requiring to communicate information about the shape of objects and the spatial relations between them. The stimulus is a model of a town presented in a Virtual Reality environment, affording better determination and experimental control of the content of messages. Another novel feature is that gaze and hand/arm movement data of the interlocutors have been captured using a motion tracking system (cf. Kranstedt et al., 2006). The movement tracking data supplements the annotation of video recordings in order to facilitate the extraction of statistically significant behavior patterns, e.g. the association between the features of gestures and referents will be analysed by correlation techniques. Further exploration of the data will rely on heuristic procedures provided by factor analysis and cluster analysis. The annotation itself will be evaluated by reliability assessment.

Annotation, Preliminary Gesture Typology and the Reconstruction of Gesture Meaning

Hannes Rieser, Stefan Kopp, Andy Lücking & Kirsten Bergmann

This part of the talk hooks up to our description of the experimental setting, especially the VR stimulus, the raw data generated and the on-going “rough” annotation of the corpus. Initially, working hypotheses concerning the notion of meaning used for the speech gesture interface are given. In essence, we use customary concepts like reference and denotation and assume properties of meaning such as compositionality and vagueness. Indeed, the very assumption of a speech-gesture interface is founded on compositionality.

Based on the annotation we point out that gesture accompanies all levels of the content communicated from partial word meanings to dialogue structures and the depiction of situations. Annotation also provides the criteria and categories of our incipient typology. Criteria are for example handedness and the number of objects depicted. Categories comprise deictic reference and single semantic features. The ensuing section discusses reconstruction principles for a gesture speech interface. Various cases exemplified by empirical data are dealt with: Gesture features replicate features of word meaning, when the gesture features overlap with the set of features of the word meaning. Gesture can also add information to underspecified word meaning, if so, we can unify both meanings. Often, gesture meaning cannot be amalgamated with some particular word meaning, since it works like the full meaning of a constituent, say an adjective phrase or a clause modifying a noun. In rare cases, gesture contributes full propositional content.

Gesture can also accompany self- or other-repair, resulting in a revision of a proposition. Finally, gesture meaning sometimes captures objects and relations of dynamically developed situations.

All these observations demonstrate that the speech gesture interface works from the word level up to the description of situations.

Simulating speech and gesture production

Stefan Kopp & Kirsten Bergmann

In the computational modeling of multimodal communicative behavior the generation of combined language and iconic gesture poses big challenges for semantic integration and meaning-form mapping. We will discuss existing systems as well as current empirical evidence in order to develop a simulation account of human-like speech and gesture production. So far, only the NUMACK system (Kopp et al., 2004) has applied systematic meaning-form mappings in the formation of iconic gestures. It operates on an unified logic-based representation and taking similarity between a gesture's form and its referent's form as primordial in depictive gesture. This approach proved sufficient to generate a set of multimodal direction-giving instructions, but it could not account for the whole range of multimodal performances one can find in spontaneous instructions (Kopp et al., to appear).A crucial issue to consider is the coexpressivity of the modalities. Recent empirical findings indicate that the interplay between speech and gesture in forming multimodal utterances is complex and depending on a number of contextual aspects (Bergmann & Kopp, 2006). From a psycholinguistic perspective particularly the model proposed by Kita & Özyürek (2003), representing the Interface Hypothesis, seems to be able to account for the data. Based on this model we will draw conclusions as to how the design of computational models needs to go beyond the uni-directional three-staged process usually adopted in current systems. The resulting model will be implemented and tested in a simulation prototype based on our virtual human Max (Kopp & Wachsmuth, 2004). Building on previous work, it will turn active representations into multimodal behavior, and it will be extended to the distribution of information across modalities and gestural depiction strategies.