Effective and Spurious Ambiguities due to some Co-verbal Gestures in Multimodal Dialogue

Frédéric Landragin[1]

[1] CNRS, Lattice Laboratory, 1 rue Maurice Arnoux,
92120 Montrouge, France
Frederic.Landragin@ens.fr

Abstract. This paper deals with some ambiguous situations linked to the use of co-verbal gestures with a touch screen. With the aim to clarify the process for the interpretation of multimodal utterances, we study the potential polysemy of a curve-like gesture produced together with a "put-that-there" verbal utterance. We emphasize the role of the task model when identifying the polysemy and determining if it is effective (real ambiguity that can lead to a question from the system) or spurious (artifact ambiguity that the task model must resolve).

Keywords: Multimodal interfaces, co-verbal gesture, deictic gesture, verbal semantics, multimodal fusion, multimodal ambiguity, reference to objects.

When designing human-machine dialogue systems, one step consists of testing the behavior of subjects face to a fake system (Wizard of Oz paradigm), and to save the interaction traces, with the aim to study them and to exploit the relevant phenomena. The experiments may involve the use of video cameras, or may limit themselves to the traces that are captured by the interaction devices (microphone, mouse, touch screen). When the dialogue between the system and its user implies a visual scene, and particularly when a touch screen is used, one may observe essentially deictic co-verbal gestures. The user speaks about and points out the objects and the places that are visible on the screen, following the classical "put-that-there" paradigm [2].

In some cases, one curved gesture is observed instead of two pointing ones. What we want to emphasize here is that an advanced study of such cases is important and relevant for the design of multimodal systems. As a starting point, we consider the situations from Fig. 1, where the gesture that is produced together with the classical "put-that-there" utterance has the form of a curve, and leads to a potential ambiguity: does the curve illustrate the way to "put", i.e. the move trajectory, or not?

The illustrating property of the gesture is explicit in the third situation, because of the arrow form. It is not really explicit but likely in the fourth situation where the position and orientation of the target object (compare to the other objects) leads us to interpret the gesture as a pointing and rotating one. The illustrating hypothesis is reinforced in the second situation by the presence of cumbersome squares between the departure and the arrival. In the first situation, the gesture may be interpreted as the only combination of two deictic aims, and therefore the hypothesis of the illustrating property may be ignored.

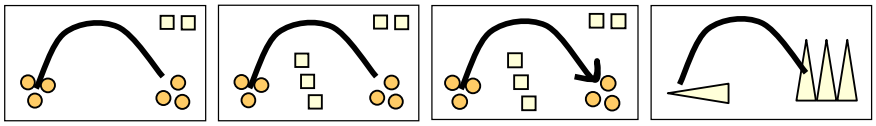

Fig. 1. Some examples that follow the classical "put-that-there" paradigm, with various interpretations considering the visual context and the form of the gesture.

The simulation described by [3] led to a corpus where we can find some examples of illustrative gestures together with "put-that-there"-like utterances. In some cases, the authors (or corpus annotators) have interpreted the curved form as the mandatory transition between the object pointing and the destination pointing. In these cases, only the two curve extremities are exploited when interpreting. When unifying (or fusing) the gesture with "that", a point (x_1, y_1) is exploiting for determining the object (or the group of objects in its immediate proximity). When fusing the gesture with "there", the point (x_2, y_2) is exploiting for determining the exact target position. [4] proposed a formalization of the Gestalt proximity and similarity criteria in this aim.

The curved gesture may also be exploiting during the multimodal fusion: the curved form can be unified to "put" in order to make precise the route of the object move. In this case, the gesture trajectory has to be analyzed from a temporal point of view (regularity, significant stop points, etc.) as well as from a structural point of view (analysis of the form itself, see [1]). A lot of algorithms from signal processing are then required. The interest of this hypothesis relies in the abilities of the algorithm for the resolution of the references to actions. If the application or task model performs the two primitives "move (object, place)" and "move (object, place, route)", an ambiguity appears (effective or spurious considering the importance of routes in the task). If only the first primitive works, i.e. if a move is like a disappearance followed by an apparition, no ambiguity appears – except if the system is able to split one "move (object, place)" into several "move (object, $place_1$)", "move (object, $place_2$)", etc. Moreover, if the task model makes a distinction between the "move" action and a "put" action that implies a new object, then the presence of a gesture presupposes that the object exists and that the verbal "put" unifies with one of the "move" primitives and not with the "put" one. With "put that there", the presence of "that" is sufficient, but not with "put the circles there" or "put circles there".

References

1. Bellalem, N., Romary, L.: Structural Analysis of Co-verbal Deictic Gesture in Multimodal Dialogue Systems. In: Progress in Gestural Interaction, Proceedings of Gesture Workshop, pp. 141–153. Springer, Heidelberg (1996)
2. Bolt, R.A.: Put-That-There: Voice and Gesture at the Graphics Interface. In: Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques. Seattle (1980)
3. De Angeli, A., Romary, L., Wolff, F.: Ecological Interfaces: Extending the Pointing Paradigm by Visual Context. In: Proceedings of the 2nd Conference on Modeling and Using Context. Trento, 91-104 (1999)
4. Landragin, F.: Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems. Signal Processing 86(12), Elsevier, 3578-3595 (2006)