Reference interpretation in a multimodal environment combining speech and gesture

Nadia Bellalem, Laurent Romary

Reference interpretation in a multimodal environment combining speech and gesture

Nadia Bellalem & Laurent Romary
CRIN-CNRS & INRIA Lorraine
Bâtiment LORIA
BP 239 F-54506
Vandœuvre-lès-Nancy
nbell@loria.fr, romary@loria.fr

Summary: The study presented in this paper is dedicated to the integration of pointing gestures within a task oriented man-machine dialogue system. From our point of view, it is natural to see this problem as a sub-part of the more general study of reference in a dialogue since we took the option to limit the analysis of gestures to those explicitly associated with speech and more specifically for singling out objects in the task. Even under these hypotheses, understanding the global message resulting from the combinaison of speech and gesture implies that a precise analysis of the gestural trajectory should be done. This analysis may be split up into two main steps. The first, which we may call structural, is centred upon the shape of the gestural trajectory to mark the meaningful parts to which a specific designation role will be given. The second step has to do with the contextual interpretation of the gesture for which one has to take into account a) the features of the application and the way it is visualized and b) the oral dialogue and more specifically the instructional content of the referring expression accompanying the gesture.

1 Introduction

Using speech as the sole means of communication between a human and a machine may seem inappropriate when the problem at hand is to manage and pilot an application with a strong visual component (e.g. presented on a graphical display). Indeed, the task constraints may force the user/speaker to create utterances containing complex referential expressions to isolate a specific object among others with very similar characteristics, in particular their categories and thus the noun used to refer to them. Nevertheless, considerable progress has been made in unravelling the mechanisms underlying the interpretation of spatial prepositions such as on *the right of, above, behind* etc. [Pribbenow, 93], [Schang, 94].
The other possibility at hand to refer to objects in the task to be managed is to use, in close connection with specific linguistic expressions bearing a deictic value, a designation gesture which has the strong advantage of providing a quick and direct access to visual objects [Wahlster, 91]. As a matter of fact, this is very similar to what happens during human to human communication when reference is made to the surrounding environment.
These two ways of referring to objects in a man-machine dialogue are to be seen as complementary, and neither of them must be left aside when designing dialogue systems as it has been proved in Wizard of Oz simulations [e.g. Mignot, 93] that both are to be observed. In this framework, our study will focus upon the specific problem of designating gestures, seeing them as participating in the overall communicative framework within man-machine dialogue.
To this end, we will present the main linguistic elements which may be interpreted in combination with gesture, as well as the main steps leading to the understanding of the corresponding trajectories in dialogue.

2 Linguistic expressions and designation

A classical view of natural language reference to objects in discourse could lead to the following classification of French linguistic markers which are presenting a deictic spatial value [Cosnier, 82]: demonstrative adjectives (*ce, cet, cette, ces*), demonstrative pronouns (*ça, celui-ci, celui-là*), "pure" deictics such as *ici* (here) and *là* (there). These markers take part in the construction of demonstrative Noun Phrases.
It may appear strange to limit ourselves to such restricted set when intuition seems to indicate that gesture may appear together with many other indicators (e.g. definite description, pronouns, etc.). For

more details relating to this issue, cf. [Corblin, 87] and [Kleiber, 92].

Within a natural language dialogue context, it is necessary to define precisely the respective roles of both the designation gesture and the referential expression which supports it up. The designation gesture may be roughly characterized as a hand movement whose role is to direct the user's gaze towards a specific area in the shared visual space.
The aim is thus to focalize the addressee's attention upon a given region in order to isolate either certain objects which are being presented, or simply a locus in the scene. As a counterpart, the verbal channel provides a frame for the interpretation of reference. In particular, it provide specific categorial information which a gesture cannot provide, as well as some constraints regarding the number of objects belonging to the sub-space marked out by the gesture. To give a more precise account of this contrastive analysis of the pair [referential expression+gesture], we may briefly look at the specific case of demonstrative NPs of the type *ce N* (this N). In previous works [Gaiffe, 94], following some orientations given by [Corblin, 87] and others it has been shown, that such an expression carried an intra categorial contrast within a set. This means that an element which can be attached to the category 'N' is to be distinguished from other elements of the same category, because of its specific focal situation. In the specific case of a graphical presentation of a task, this principle may be seen as equivalent to a sifting of the presentation space so that only objects of the type 'N' are considered, any other referent being made opaque at this step of the interpretation. Considering this, the analysis of the designation trajectory may be carried out on the basis of the level of granularity (i.e. the scale) providing by this N-filtering, without taking into account finer grained levels of precision. The final interpretation thus corresponds to the *more focal element*, taking into account the different *singularities* (we will explain this notion in section 3.1.1) encountered within the gestural trajectory.

3 Analysing the designation gesture

To illustrate our views, we will consider an application to interior furnishing which seems to correspond clearly to the kind of tasks where a gestural component could be beneficially introduced. In this context, we will try to make clear the different steps leading to a proper understanding of the oral + gestural message. The task consists in manipulating graphical objects which are viewed in a scene. The verbal utterances are essentially positioning statements containing the following elements:

- action
- set of objects (possibly one)

More specifically, we will centre our analysis upon the example shown in figure one corresponding to the utterance "put this armchair here".

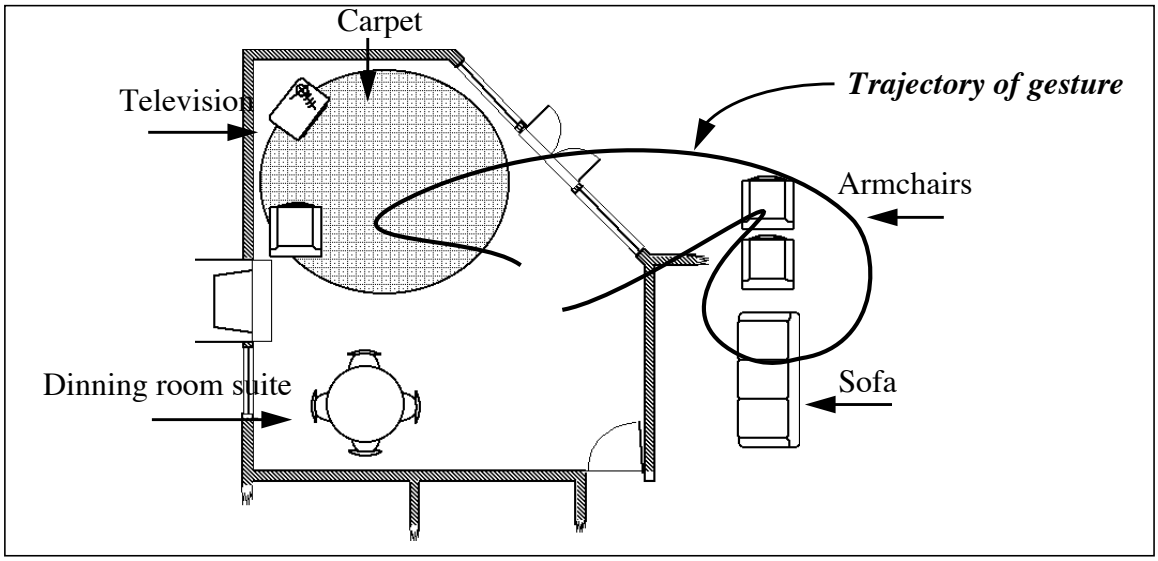

Figure 1: Designation gesture for an application to interior furnishing.

3.1 A structural analysis of gesture

The first step in the recognition and understanding process for a referring gesture consists in analysing the gestural signal independently from the scene upon which it has occurred, in order to detect the different components of meaning that it may contain. As a matter of fact, it should be noticed that the

gestural signal coming from the pointing device is a stream of data corresponding to a complete hand motion, that is, comprising a motion towards the designation area, the actual designation associated with the verbal message and finally a motion away from the designated area. It is clear that only the central portion is of importance to us since it embodies most of the meaning of the gesture. The main problem is to be able to differentiate their position from the rest of the signal. To this end, we will put forward the idea that the designation gesture aims at singling out a sub-space of the overall space which has specific contrasting features. We thus propose a model for these contrasting features, which we name "singularities", providing us with, on the one hand, a way to take into account different patterns as potential designations, and, on the other hand, a solution for the resolution of possible ambiguities.

3.1.1 Defining a singularity

The meaningful part of a gesture is characterized by two aspects: from the point of view of the surface expression, the gesture will have a particular shape which we will call singularity and from the point of view of meaning, the gesture shape may be associated with semantic component of an act of designation.

A singularity is to be observed relative to an overall property of the trajectory (e.g. curvature) and relative to a specific segment of the trajectory which is considered as stable for this property. It is thus a local event. A complete analysis of a gesture should be carried out on all the possible properties and for the whole signal; the global meaning of the gesture is obtained by associating and combining the different singularities which have been detected. Indeed, the notion of singularity provides the means to relate the structure and the meaning of the gesture.

The main idea is to build a symbolic representation from the original structure of the trajectory within which singularities will be identified, with the underlying hypothesis that no singularity is accidental but results from the user's intention to mean something (e.g. to designate something). When really accidental singularities are encountered, they will be automatically eliminated at the interpretation stage, either because no linguistic expression may be associated to them or because they do not correspond to any object which could be contrasted. From a general perspective, three kinds of singularities may be distinguished:

a) punctual singularities such as abrupt changes of directions corresponding to important variations in the curvature. These may correspond to the designation of the reduced areas where the curvature reaches its maximum;

b) simple trajectory singularities such as a circling trajectory with straight segments at its ends;

c) repetitive trajectory singularities which may be exemplified by the zigzag used to designate a whole region. They usually correspond to a combination of punctual singularities and trajectory segments which form a regular pattern.

3.1.2 Modelling the gestural trajectory

In order to be able to observe the different singularities, it is mandatory to achieve a proper modelling of the gestural trajectory. To do so, we have studied several kinds of properties, among which we may mention [Bellalem, 93]:

- the motion speed along the trajectory,
- the different crossing points,
- curvature,
- device specific features (e.g. mouse).

Figure 2 illustrates the structural analysis of the trajectory presented in our example.

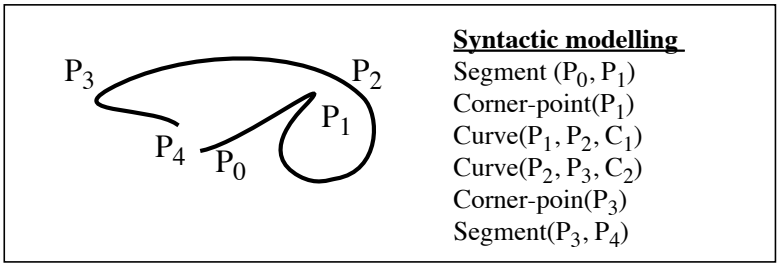

Figure 2: Modelling the gestural trajectory

3.1.3 Studying the meaningful elements

The second step in the structural analysis consists in recognizing, within the trajectory representation, the different singularities and to build up the global meaning represented by a proper association of these singularities. In the case of our example, this leads to two possible hypotheses corresponding to two different presences of the punctual and trajectory singularities observed in the gesture. The first one has preserved the information carried by the punctual singularities, whereas the second one stresses upon the partial circling (figure 3). The ambiguity will only be resolved at the interpretation stage.

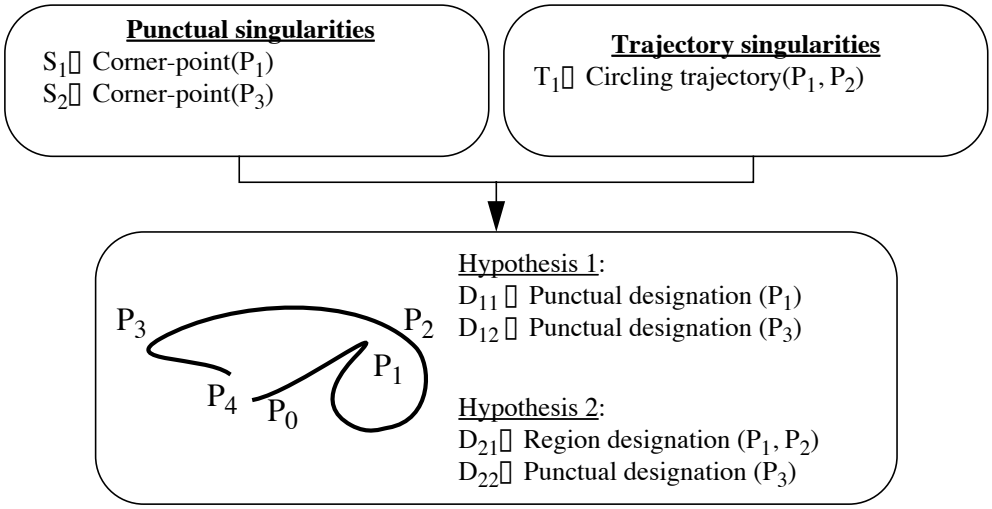

Figure 3: Meaning elements

3.2 Towards a contextual interpretation of gesture

This interpretation process in context poses the problem of obtaining a model of the visual space in order to take a proper account not only of the spatial layout of the different objects, but also of the different (spatial or functional) relations between the objects, since these may evolve depending on the kind of visualization which is being chosen (2D,3D). The final aim is to build a representation which is as close as possible to that which is considered by the human operator when uttering a referring expression. Among the different spatial relations that have to be considered, we may quote proximity and superposition, to which we may add some specific compositional relations (e.g. the fact that the living room is composed of tables, chairs etc.).

Moreover, it is important to be able to deal with ambiguities resulting either from a possible imprecision in the gestural trajectory, from different possible spatial or functional configurations between objects or even from a difficulty in contrasting different candidates. The solution we have adopted is to manage a set of hypotheses which are ordered according to the likelihood of being designated in the context of a given demonstrative natural language expression.

At this stage, we may mention the possible role of the dialogue context in the interpretation process. Until now, we have made the implicit assumption that the interpretation was made relatively to the complete graphical space presented to the user. However, it appeared to us that it is often relevant to constrain the interpretation of a multimodal referring expression to particular subspaces where objects may be contrasted. Such sub-spaces will typically correspond to stable working zones at a given step in the dialogue process and which may detected by means of different clues: referential expressions from preceding utterances, specific markers of intentional break (e.g. "Now, we turn to the living room"), markers of continuity, overall structure of the task (the decomposition of a flat into rooms etc.).

In our example, we suppose that such phenomena have been taken into account in describing the object structure given to the interpretation process. In the example, the utterance "place this armchair here", which contains both a demonstrative NP and a "pure" deictic, allows us to infer that the associated gesture is made of maximally two designating portions. The two hypotheses that we have put forward are thus valid since they both correspond to two such designations. The integration of these within the object universe leads to the following result for the first hypothesis (cf. figure 4):

- the first designation leads to the selection of {armchair1},
- the second designation isolate the object {carpet},

For the second hypothesis, we have the following interpretation:

- the first designation selects the objects {armchair2, armchair1, sofa} in the order given by their likelihood of being designated by the gesture (depending upon different numerical criteria),
- the second designation gives the same result as for hypothesis 1.

The study of the demonstrative NP indicates that the candidate to be considered for the first designation is unique and is of the category "armchair". There again, both hypotheses remain valid since hypothesis 1 proposes armchair1 and hypothesis 2 armchair2. Consequently, we have to look again at the trajectory to differentiate the two. Indeed, we observe that hypothesis two may be favoured, considering that the corner point P1 may be integrated in the circling hypothesis since it is its beginning element whereas the circling shape cannot be explained and is thus not relevant if we consider that P1 is directly pointing at a localized area.

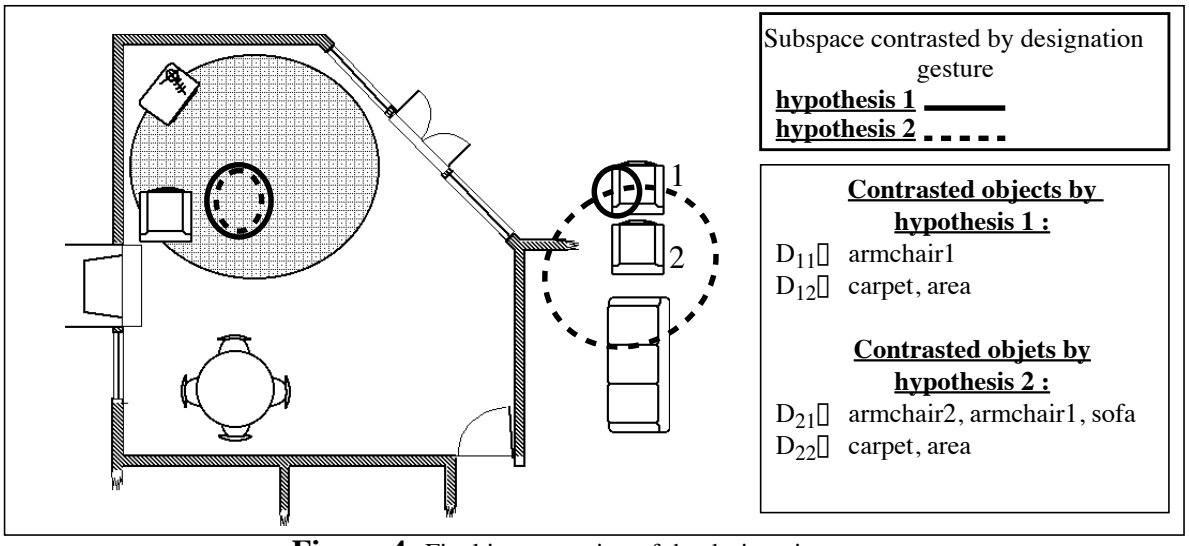

Figure 4: Final interpretation of the designation gesture

4 Conclusion

The notion of singularity that we have introduced corresponds to a specific conception of gesture interpretation closely related to the dialogue system in which it is to be integrated. Indeed, it allows us to identify a close interaction between the gesture and the task as it is visualized on a graphical space. In addition, it provides a plausible account of linguistic information and of the overall process of dialogue management.

References

Bellalem N. & Romary L., Le dialogue homme-machine multimodal, vers la compréhension du geste de désignation, *Actes des 2èmes journées internationales, L'interface des mondes réels et virtuels*, Montpellier 22-26 Mars 1993, p. 217-228

Caelen J., Garcin P., Wret J., Reynier E., Interaction multimodale autour de l'application ICPDraw, *Actes du Workshop IHM'91*, Dourdan, December 1991

Corblin F, 1987, *Indéfini, défini et démonstratif*, Droz, Genève-Paris.

Cosnier J., Berrendoner A., Orecchioni C.,1982, Communications et langages gestuels, *Les voies du langage, Communications verbales gestuelles et animales*, Edition Dunod-Bordas.

Gaiffe B., Reboul A., Romary L., 1994, Références et gestion du dialogue, *Actes de TALN'94*, Marseille

Kleiber G., Y a-t-il un il ostensif ?, *Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas*, Universidade de Santiago de Compostela, 1989, Publicadas por Ramón Lorenzo, A Coruña, 1992

Mignot C., Valot C., Carbonell N.,1993 , An experimental study of future "natural" multimodal human-computer interaction, *INTERCHI'93*, Amsterdam.

Pribbenow S., 1993, *Computing the meaning of localization expressions involving prepositions: the role of concepts and spatial context, Mouton de Gruyter*, Cornelia Zelinsky-Wibbelt (Editor), pages 441-470

Romary L., 1993, "Mets ça ici" où quand "ici" dépend de "ça". L'interprétation de "ici" dans des énoncés de positionnement, *Workshop Caen*.

Schang D. and Romary L., Frames, 1994, a unified model for the representation of reference and space in a Man-Machine Dialogue, *3rd International Conference on Spoken Language Processing*, Yokohama .

Wahlster W., 1991, User and discourse models for multimodal communication, *Intelligent user interface*, ACM Press Services, J. Sullivan and S. Tymler (Eds), Adisson-Wesley .