Constraints on the Use of Language, Gesture and Speech for Multimodal Dialogues

Bertrand Gaiffe
CRIN-CNRS & INRIA Loraine,
Bâtiment Loria, B.P. 239
54506 Vandœuvre Lès Nancy
gaiffe@loria.fr

Laurent Romary
CRIN-CNRS & INRIA Loraine,
Bâtiment Loria, B.P. 239
54506 Vandœuvre Lès Nancy
romary@loria.fr

1. Introduction

In the domain of natural language understanding and more precisely man-machine dialogue design, there are usually two trends of research which seem to be rather differentiated. On the one hand, many studies have tackled the problem of interpreting spatial references expressed in verbal utterances, focusing in particular on the different geometric or functionnal constraints which are bound to the existance of a 'source' (or site) element in relation to which a 'target' is being situated. Such studies are usually based upon fine grained linguistic descriptions for different languages (Vandeloise, 1986). On the other hand, the problem raised by the integration of a gestural mode within classical NL interfaces has yielded some specific research about the association of demonstrative or deictic Nps together with designations, as initited by Bolt some two decades ago (cf. Thorisson et alii, 1992; Bellalem and Romary, 1995). Our aim in this paper is to show that the different phenomena described in the context of spatial reference or multimodal interaction should not necessarily be considered as two independant issues, but should rather be analysed in a unified way to account for the fact that they are both based on linguistic and perceptual data.

As a matter of fact, if we consider a situation of man-machine dialogue where the user is presented with a graphical representation of his task, it is clear that, given a certain informational content he wants to convey, he will essentially choose a referring mode which seems most relevant in the current communicative situation. For example, if we consider a graphical situation such as that described in figure 1.1, he may either use *the black triangle*, *this triangle* (+ pointing gesture), *the leftmost triangle* to refer to the left most object, and it would be quite annoying to consider these different expressions as corresponding to uncomparable referring modes[1].

Figure 1.1

In this context, we will try to show how language, gesture and perception can be seen in a uniform way from the perspective of referential analysis, even if doing so we will have to look at the specific constraints which underly the speaker's choice of a given expression. To this end, we will first quickly situate the relative importance of speech and gesture in man-machine communication. Then, we will concentrate upon the specific effects resulting from the combination of verbal, gestural and perceptual information, showing that on the one hand the three provide structural constraints to the objects which are being referred to and on the other hand that any referring operation, whatever its origin, has to be interpreted within a localized frame, with some consequences upon dialogue management.

2. Several means to make a referring act

When designating a given object within a visual environment, it seems at first sight that Natural Language provides uncomparable means to do so as opposed to gesture. Beyond the different determiners which are present in most natural languages either explicitely or implicitely (indefinite, definite or demonstrative), nominal categories allows one to set the proper level of granularity corresponding to the intended object. Indeed, in a situation where a gesture would be ambiguous and point to the overall scene (a set of geometrical shapes), a specific

[1] In particular, gricean maxims as well as relevance theory (Sperber and Wilson, 1986) would tend towards an analysis which compare the different referring expressions in terms of cognitive cost.

object (a triangle) or any of its part (a segment, a point etc.), the sole phrase *the triangle* may directly designate what is being intended.

Another important aspect is that pointing gestures[2], when used in the general framework of an oral dialogue, can seldom appear in isolation, wheras a definite description such as *the blue triangle* can clearly be expressed independantly of any gesture. The reason for this is, as we said, that the intrisic ambiguïty of gesture implies that it should be complemented with a categorizing expression, but also because a gesture cannot express very easily an action to be performed upon the designated object and has thus to be also complemented by a predicative utterance. In this latter case, it is hard to imagine that any combination will be possible between linguistic chunks and gestural acts. In particular, gesture can hardly fill a role which is mandatory for a given predicate, since it would lead to odd utterances such as *?give me the color of [pointing].*

3. Reference and contrast

The schematic algorithm used in most dialogue systems in order to deal with referential NPs (in the case they get their reference within a context which is visually presented) can be expressed as:

a) get all the indices from the expression;

b) deduce from these indices some constraints which must be true in the visual representation;

c) filter the referent(s) thanks to these constraints.

In such a framework, what would be expected as the system's perceptuel abilities boils down to an ability to build the set of objects appearing on the screen. Such an approach would compute the "correct" referents in such examples as:

Figure 2.1.a: the leftmost triangles

or

Figure 2.2.a: these triangles

[2] We do not consider here other types of gestures, either ergative, epistemic or mimetic as described in (Cadoz, 1992).

by means of formulae (at step b) such as:

- *triangle(R) & (card(R) >= 2) & R included R' & R at the left as compared to (R' - R)* in what concerns 2.1.a and
- *triangleR) & (card(R) >= 2) & (made sallient (by gesture) R)* in what concerns 2.2.a.

Although these two referential expressions, if ever used, are unambiguous, we would certainly prefer such examples as:

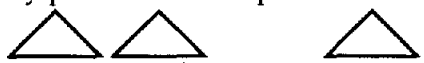


Figure 2.1.b: the leftmost triangles

and

Figure 2.2.b: these triangles

What seems to lack in examples 2.1.a and 2.2.a is a visual contrast that pre-exists to the refering expression. The refering expression itself is not sufficient to establish such a contrast. If the user intends to refer to the first two triangles of figure 2.1.a, he would probably prefer an expression such as *the two leftmost triangles* or *these triangles* together with a "peripheral" designation as we will describe it latter. Our claim about the necessity of a perceivable discrimination seems in accordance with what Robert Dale (Reiter and Dale 1992, Dale 1995) observes about referential expression generation. Just as we do, he argues that the relevance of a refering expression does not only rely on its ability to filter a unique referent but also on its ability to establish a contrast in a contextual set of objects.

Examples 2.1.b and 2.2.b rely on an already accessible discrimination based upon spatial cohesion. In such a case, the definite referring expression (*the leftmost triangle*) directly maps the spatial discrimination. Such examples as:

Figure 2.3.a: the leftmost triangles

and

Figure 2.3.b: these triangles

show that any percepually based discrimination (we could as well have a contrast in size, texture or whatsoever) is sufficient to justify *the leftmost triangles* (*these triangles* + gesture respectively)[3].

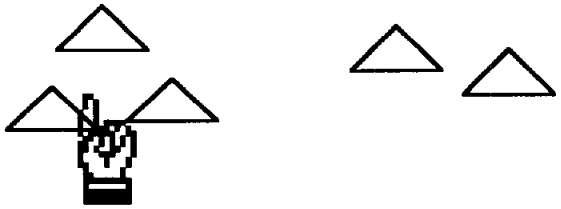
If a dialogue system has to understand such expressions as those we mentioned so far, he therefore should perceive its environment on a more "user compatible" basis. We suggest at least that perceptual contrasts should be taken into account in order to structure the set of visible objects.

When no contrast pre-exists which would directly support an intended reference, we mentioned the possibility to build a group on the basis of individuals by such an explicit expression as "the two leftmost triangles". A corresponding solution in terms of demonstrative use would be something like *this triangle and this one* (or *these two triangles*) together with two pointing gestures. Another solution consists in building the contrast by means of a "peripheral designation" which justifies our claim about considering perception and designation on a unified constrastive basis. In order to argue that claim, we will now re-consider gestures almost independantly from the referential expressions they accompany.

4. Gesture and contrast

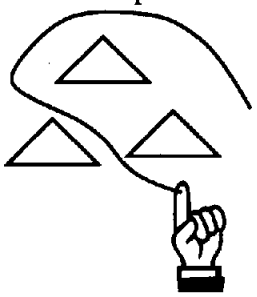
Our analysis of demonstrative and definite NPs (when referring within a perceptual environment) relies on perceptually founded contrasts. The required precision of a designation gesture therefore depends upon these perceptive constrasts. In such an example as:

the gesture only has to separate the two triangles along the horizontal, since the perceptive contrast relies upon a separation of the objects on that direction. No strict inclusion of the pointing into the left triangle is required. The situation depicted below relies on the same kind of horizontal discrimination:



the only difference with the preceeding example is that we refer to a cohesive group. The horizontal discrimination identifies here two groups from which gesture only has to select one. However, such situations as:

do not provide any perceptual grouping of something which would correspond to "the upper and the righmost triangles in the left group of three". If the user intends to refer to these two triangles, he has to build a discrimination into the group. A possible gesture to do that is depicted bellow:



Such "peripheral" designations take up for the absence of a shared perceptual feature (such colour), as it both gather up the two objects and put them into focus. The analysis of the whole intervention (the gesture plus the NP "these triangles") is then of the same kind as its equivalent in 2.3.b. As such, we clearly see here that gesture, instead of just being another mode of communication, pertains to the same domain as perceptual information.

Our analysis so far can thus be summarized as follows:

- a contrast based on the category has to match a perceptive contrast in the case of simple definite Nps, thus meaning that perceivable triangles should be considered when analyzing *the triangle(s)*
- a contrast based on saliance has to match a perceptive contrast in the case of demonstrative Nps. As we only considered in this paper demonstrative plus gestures,

[3] In some cases, the speaker has the possibility to elicit the contrastive feature. E.g. *The black triangles* (fig.2.3.b)

the required saliance is yielded by gesture itself

- a spatial contrast has to match a perceptive contrast (not necesserally spatial) in the case of spatial definite NPs.

The remaining problem is now to limit the context in which we consider these contrasts. There are expressions in dialogue corpora that cannot be properly understood if we do not take into account focusing phenomena as well as attentional contexts and visual capabilities. Moreover, as we will justify, in such reduced contexts, functionality associated to the objects considered may introduce specific orientations.

5. Localizing spatial references

Having shown that any spatial —in the broad sense we want to advocate — reference is based upon a structural organisation of a set of elements, we will now see how this very set plays a real role of contextualizing the referring process, with some consequences upon dialogue management. Indeed, all our examples so far were simple enough to imply that there was only one context in which to find the intended referent. On the contrary, if we consider a more complex situation taken from a Wizard of Oz simulation in the domain of interior furnishing (Dauchy et alii, 1993; Mignot et alii 1993), we will see that our analysis should actually be drawn a step further. Figure 4.1 examplifies a typical situation that was presented to the user during the experiment, with an empty drawing room to be furnished using the presented elements.

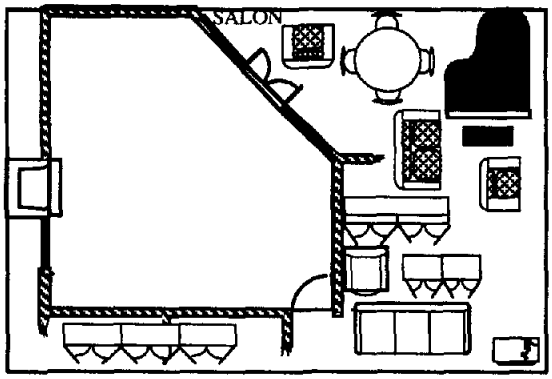


Figure 4.1 : Initial scene for the interior furnishing scenario.

Following the observation that there should be a prior structure shared by both the speaker and the hearer for a spatial reference to be understood, we can quickly see that this structure can only be inferred within a localized context which first limits its extension, but also subsumes its general characteristics such as the categories of objects, their perceptual or functional properties etc. Paradoxically, we could say that it is difficult to contrast objects having little or nothing in common as there would be no reason for a speaker to compare them in any way. Besides, such contexts seem to have a certain amount of stability during a dialogue, as can be seen in the following example associated with figure 4.2:

U1 : *turn the sofa round*

U2 : *move up a bit the armchair on the right*

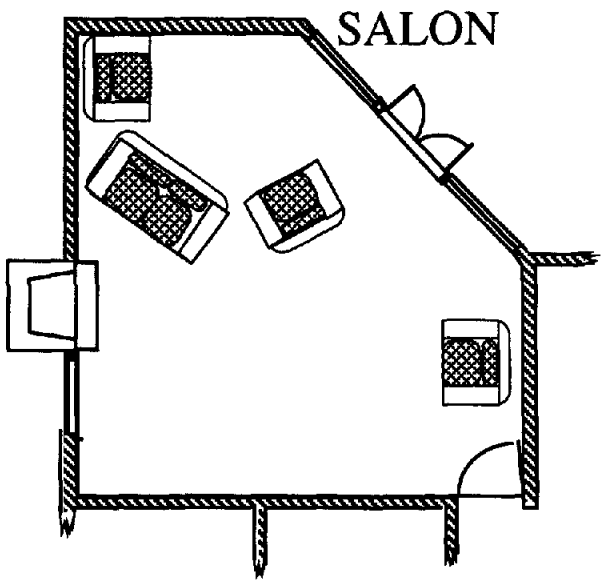


Figure 4.2 : Referential contextualisation.

Here, it appears that the spatial reference in the second utterance is not computed globally on the visualized scene but upon a sub-space resulting from the interpretation of the first utterance and thus centered on the sofa. Such a sub-space is characterized by its spatial inclusion within that of the drawing room, but also by the different characteristics (especially functionnal ones) of the objects it contains.

At this stage we can thus caracterize a spatial referring operation as a double system of vertical and horizontal relationships within a context which encompasses the object which is being referred to, but also the set of alternatives which being stated either explicitaly or implicitaly during the current

referring act or the rest of the dialogue. Figure 4.3 summarizes these different constraints for a reference to object O1 within a context C, the alternatives being reduced here to a single object O2.

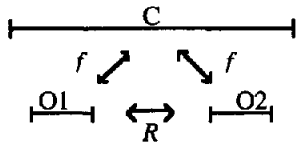


Figure 4.3 : localizing links

In this schema, f is the functionnal link (FL) which unites the different objects to their current context[4], and which has to be shared by the whole set of alternatives. R is the contrasting relation which allows the interpreter to isolate the referent of the expression from the other members of the set of alternatives. The very existance of both C and f are supposed to be implied by the nature of the referring expression. Similarly, the relation R is constrained by the type of the expression and can be further specified as follows:

- definite description (*the triangle*): Cat(O1) ≠ Cat(O2) (intercategorial contrast)
- demonstrative NP (*this triangle*): Cat(O1) = Cat(O2) (intra-categorial contrast)
- spatial definite description: Pos(O1) ∩ Pos(O2) = Ø (localization contrast)

In this last case, the referential expression is usually explicit (e.g. *the leftmost armchair*) about the actual contrasting relation and thus makes the presence of alternatives all the more obvious. For the two other cases, it is usually through the following utterances that, as we have seen, we can justify the presence of the set of alternatives. As a matter of fact, one consequence of constraining a referential operation to a localized frame is that these have a certain amount of stability from utterance to utterance.

6. Conclusions

In this paper, we argue on the one hand for a unified account of gesture and perception, and on the other hand for a matching between contrastive conditions required by referential expressions and a pre-existing perceptual contrast. We show that such a contrast has to be localized and to exhibit traces of these contexts through dialogue structures as well as specific orientation properties related to such perceptual contexts.

[4] There can be of course many different contexts projected upon a given object, as this depends upon the intention that the speaker wants to convey about it.

7. References

Allen James F. and Raymond C. Perrault 1980, Analyzing intention in utterances, *Artificial Intelligence, 15*, p. 143-178.

Bellalem Nadia and Laurent Romary 1995, Reference interpretation in a multimodal environment combining speech and gesture, Actes First IMMI Workshop, Edinburgh.

Cadoz Claude 1992, Le geste canal de communication homme/machine - la communication instrumentale, *Technique et Science Informatique, 13*, 1, p. 31-61.

Dale Robert and Ahud Reiter 1992, Computational Interpretations of Gricean Maxims in the Generation of Referring Expressions, Actes Coling-92.

Dale Robert 1995, Generating One-Anaphoric Expressions: Where Does the Decision Lie?, Actes Working Papers of PACLING-II, Brisbane, Australia, p. 49-58.

Dauchy P., C. Mignot and C. Valot 1993, Joint speech and gesture analysis : some experimental results, Actes Eurospeech 93, p. 1315-1318.

Mignot C., C. Valot et N. Carbonell 1993, An Experimental Study of Future ``Natural'' Multimodal Human-Computer Interaction, Actes INTERCHI'93 1993 Conference on Human Factors in Computing Science INTERACT'93 and CHI'93, Amsterdam (The Netherlands).

Schang D. et L. Romary 1994, Framing the world, towards a localised spatial reasoning, Actes 3rd International Conference on the Cognitive Science of Natural Language Processing (CSNLP-94), Dublin.

Sperber Dan et Deidre Wilson 1986, *Relevance, communication and cognition*, Basil Blackwell, Oxford.

Thorisson K., D. Koons et R. Bolt 1992, Multi-modal natural dialogue, Actes CHI'92, USA, p. 653-654. (Annelies).

Vandeloise Claude 1986, *L'espace en français*, Editions du Seuil, Paris.