Modeling Context for Referring in Multimodal Dialogue Systems

Frédéric Landragin

DRAFT VERSION

Abstract. The way we see the objects around us determines speech and gestures we use to refer to them. The gestures we produce structure our visual perception. The words we use have an influence on the way we see. In this manner, visual perception, language and gesture present multiple interactions between each other. The problem is global and has to be tackled as a whole in order to understand the complexity of reference phenomena and to deduce a formal model. This model may be useful for any kind of man-machine dialogue system that focuses on deep comprehension. We show how a referring act takes place into a contextual subset of objects, called 'reference domain,' and we present the 'multimodal reference domain' model that can be exploited in a dialogue system when interpreting.

1 Introduction

The understanding performance of natural language dialogue systems more and more relies on their pragmatic abilities. Indeed, modeling the context and modeling the interpretation process are particularly complex aspects of pragmatics for multimodal dialogue systems. For systems where a user interacts with a computer through a visual scene on a screen, the combination of visual perception, gesture and language involves interactions between the visual context, the linguistic context and the task context. There has already been several proposals related to the representation of the linguistic and the task contexts, considering components such as dialogue history, salience, focus of attention, focus spaces, topics, frames, plans and so on. Still, less attention has been put on how to deal with the visual context in such a framework: some works focus on structuring the visual scene into perceptual groups [24], others focus on the management of a visual focus of attention and on the relations between this notion and salience [1]. What we want to do here is to integrate all these perceptual, linguistic and cognitive aspects (see Figure 1), for the interpretation of reference to objects phenomena. To us, this has to be done by using an unified framework, in order to compare and to merge the various information from the various contextual aspects into homogeneous structures.

It is with this aim that we have been developed since several years the 'multimodal reference domain' model. As opposed to approaches like the DRT (Discourse Representation Theory) [11], this model has been built with multimodal

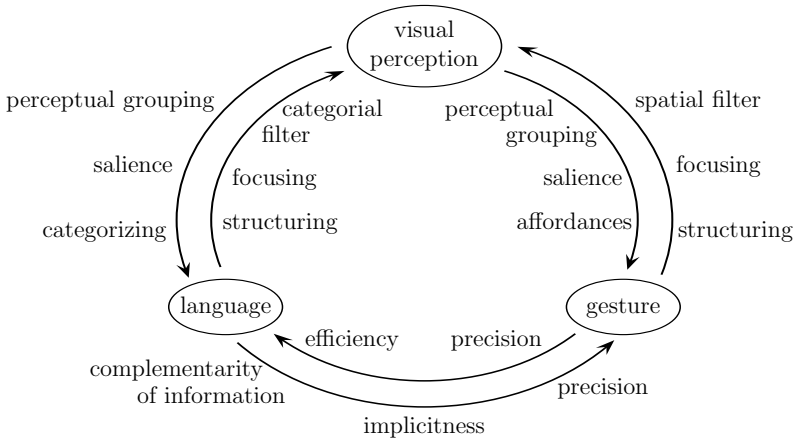


Fig. 1. Some interactions between visual perception, gesture, and language.

concerns from the early first phases of the design. As opposed to approaches based on domains of quantification [4], it takes into account the previous utterances when delimitating the context. Reference domains are then linked to each others. With these two strong points, the reference domain model appears to be useful when designing a multimodal dialogue system. The visual context as well as the linguistic context (dialogue history) can be represented by sets of reference domains, which can be easily compared. In this paper we want to show that multimodal dialogue systems need to take into account the visual and linguistic contexts in a same manner, in order to manage in a proper way all contextual information. We first present the main principles of our model. In the next section, we describe in details how we translate the visual context into visual reference domains. We then describe how a multimodal utterance (a verbal referring expression together with a pointing gesture) from the user can be interpreted with the help of reference domains.

2 Reference domains

The basic idea of the 'multimodal reference domain' model is that when we interpret a multimodal referring expression, we take into account not the complete context (for instance all objects that are present in the communicative situation), but only a reduced part of it (for instance objects that are in the focus of attention of the participants). This part constitutes a 'reference domain.' Reference domains can come from visual perception, language or gesture, or can be linked to the dialogue history or the task constraints. Visual domains may come from perceptual grouping, for example to model focus spaces [1]. Some domains may come from the user's gesture, others from the task constraints. All of them are structured in the same way (see Figure 2). They include a grouping factor ('being in the same referring expression,' 'being in the same perceptual

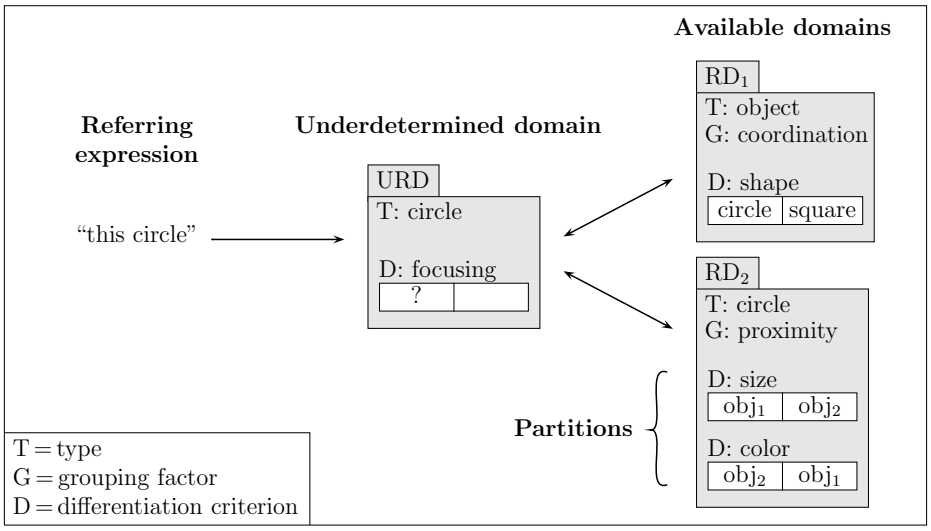


Fig. 2. Interpretation of a referring expression using reference domains.

group'), and one or more partitions of elements. A partition gives information about possible decompositions of the domain [22]. Each partition is characterized by a differentiation criterion, which represents a particular point of view on the domain and therefore predicts a particular referential access to its elements ('red' compared to 'not-red,' 'focused' compared to 'not-focused'). With these formal aspects, reference domains consist of a way to represent data structures maintained in a dialogue system, with a cognitive inspiration.

One important point of the model is the creation of a new reference domain. The linguistic and contextual clues are sometimes not sufficient for the delimitation of such a domain. For this reason, we propose to manage underdetermined reference domains, as it is done in [22] and then in [16], and as it is showed in Figure 2. The linguistic and gestural information allow to build an underdetermined domain that groups all constraints. In Figure 2, the referring expression that is currently treated is "this circle." Such a demonstrative nominal phrase implies that a particular circle is focused. This interpretation constraint can be translated into an underdetermined reference domain, that consists of a partition where one element is focused. "This circle" is making a contrast between a particular circle and other circles. So the reference domain in which the interpretation occurs must include only circles. This is the role of the 'type' attribute.

Then, the reference resolution process consists of the unification of this underdetermined domain with the domains that appear in the context. In Figure 2, two reference domains are available in the context, RD_1 that comes from the dialogue history, and RD_2 that comes from the visual context. More precisely, RD_1 was built on at a previous stage of the interaction, when interpreting a referring expression such as "a circle and a square." RD_2 groups two objects because

of their proximity. The domain with the best unification result is kept for the referent identification.

In the next sections we will first focus on perceptual phenomena that are at the early beginning of reference, including salience and grouping aspects. We then focus on referring phenomena, including multimodal aspects, and we conclude on the algorithm for multimodal reference resolution based on the management of reference domains. Such an algorithm has been developed in the framework of several European project: IST MIAMM (see http://www.miamm.org) and IST OZONE (see http://www.extra.research.philips.com/euprojects/ozone). We don't want here to describe the implementation of this algorithm, because it requires the presentation of a lot of technical problems that are not of importance here (such as the calculus of salience scores, or algorithms for the recognition of gesture trajectories). We want to focus on the exploitation of contextual and communicative clues for the interpretation of multimodal referring expressions, in order to emphasize the way the context can be modeled with reference domains.

3 Perceptual phenomena

3.1 Focusing (salience)

Since we consider that salience is at the origin of referring phenomena, we want here to clarify the way to take visual salience into account in multimodal dialogue systems. In the absence of information provided either by the dialogue history or the task history, an object can be considered as salient when it attracts the user's visual attention more than the other objects. Several classifications of the underlying characteristics that may make an object be perceived as salient have been proposed. For instance, Edmonds [5] has provided some specific criteria in direction-giving dialogues when the objects are not mutually known by the instructor and learner. However, such classifications are by far too dependent upon the task to be achieved (for example there is one specific classification for each type of object) and narrows down on the notion of salience to specific aspects. Merging them and adding to them the major results of pictural arts studies (Itten [9], Kandinsky, etc.) may lead us to contemplate a more generic model which in turn could be implemented for an application-driven system.

First, a salience model requires a user model of perception. Indeed, visual salience depends on visual familiarity. Some objects can be familiar to all users. It is the case for human beings: when a picture includes a human (or when a virtual environment contains an avatar), he will be salient and the user's gaze will be first attracted by his eyes, and then his mouth and nose, as well as his hands, when a specific effort has been made to simulate natural gestural behavior. For other objects, familiarity depends on the user. When a photographer enters a room, the pictures on the walls might be more salient than the computer on the table; whereas it might be the opposite for a computer scientist. Everyone acquires his own sensitivities, for example his own capacity in distinguishing colors. The choice of the right color term can show these sensitivities. Somebody

may prefer to name 'red' a color that somebody else is used to naming 'pink.' No need to be color-blind for that.

Second, a salience model needs a task model. Visual salience depends on intentionality. When you invite colleagues in your office, you search chairs in your visual space, and so chairs are more salient than the other furniture.

Third, visual salience depends on the physical characteristics of the objects. Following the Gestalt theory, the most salient form is the 'good form,' i.e., the simplest one, the one requiring the minimum of sensorial information to be treated. This principle has been first illustrated by Wertheimer [25] for the determination of contours, but it is also suitable for the organization of forms into a hierarchy. Nevertheless, when the same form appears several times in the scene, one of the instances can be significantly more salient than the others. The salience of an object then depends on a possible peculiarity of this object, which the others do not have, such as a property or a particular disposition within the scene. Basically, those peculiarities can be summarized as follows:

1. classification of the properties that can make an object salient in a particular visual context:
   (a) category (in a scene with one square and four triangles, the square is salient),
   (b) functionality, luminosity (in a room with five computers, with one of them being switched on: this one is salient),
   (c) physical characteristics: size, geometry, material, color, texture, etc. (in a scene with one little triangle and four big triangles, the little one is salient, etc.),
   (d) orientation, incongruity, enigmatic aspect, dynamics (object moving on the screen)...
2. salience due to the spatial disposition of the objects: in a room containing several chairs, a chair which is very near the participant may be more salient than the distant ones, and an isolated chair may be more salient than the others if these ones are grouped.

When no salient object can be identified by means of the previous methods, visual salience also depends on the structure of the scene, i.e., the frame, the positions of the strong points in it, and the guiding lines that may restrain the gaze movements. The strong points are classically the intersections of the horizontal and vertical lines at the 1/3–2/3 of the rectangular frame. If the perspective is emphasized, vanishing points can also be considered as strong points. If the scene presents a symmetry or balance which hinges upon a particular place, this very place becomes a strong point. As a whole, the objects that are situated at strong points are usually good candidates for being salient. If they can be identified (from continuities in the disposition of the objects), the guiding lines go from salient objects to salient objects. Salience can thus be propagated.

The four stages that we have identified in this section correspond to the four stages of the algorithm we propose to automatically detect salient objects in a visual context. If a given stage cannot lead to significant results, the next stage is considered. Each result must be associated with a confidence rate (for example

the number of characteristics that distinguish the salient object from the others). When no result is found, the whole visual context has to be taken into account.

3.2 Grouping

Following the Gestalt theory [25], the major principles to group objects are proximity, similarity and good continuation. From the list of visible objects and their coordinates, algorithms can build groups, which allows the system to have an idea of the user's global perception of the scene. An example of such algorithm is given by Thórisson [24]. The notion of salience can be extended from an object to a group. When the user sees a scene for the first time, one group may attract his attention more than the others and may be perceived first. According to our definition, this group will be salient. Based on proximity and similarity, the algorithm of Thórisson produces groups ordered according to goodness, and therefore according to salience.

Grouping on the sole basis of the proximity principle amounts to the computation of distances between objects. Applying a classic algorithm of automatic classification, we obtain a hierarchy of partitions of the objects in groups, each group being characterized by a compactness score (see Figure 3-B). When a 2-D display of a 3-D scene is made, for example with a virtual environment displayed on a screen, grouping can be done in 3-D, or in 2-D with the coordinates of the projections of the objects. Strictly following the Gestalt theory, this second solution is in line with the application of proximity principle at the retina level. An experiment of Rock and Brosgole [21] shows however that users restore the third dimension, and that grouping is done at a later level than the early processing of retina information. Rock and Brosgole introduce the notion of phenomenal proximity, and the relevance of grouping objects in the underlying 3-D representation.

Grouping by taking into account the good continuation principle can be done by means of a recursive processing: groups are built from each single object and are extended to their nearest proximity, and so on until the whole space has been covered. Continuities are identified by doing linear regressions. Grouping with one Gestalt criterion or another leads us to different results (Figure 3). Moreover, only considering the proximity criterion produces various results depending on the compactness level at which the hierarchy is read. We cannot consider priorities between the criteria (as we did with salience criteria), because we do not know when it is better to consider groups with a high compactness or groups with a linear global shape. For the moment, we have to manage several results. Each of them must be associated with a confidence rate, for example the compactness.

Visual reference domains can be built on by using these focusing and grouping methods. The existence of a strong visual reference domain relies on the demarcation of a group in the dendrograms. The grouping factor of the domain will be the combination of criteria (for instance, proximity plus continuity) used when grouping. When a salient object is present in the group, a partition is created where this salient object is focused. The differentiation criterion of this partition is labeled as 'visual salience.'

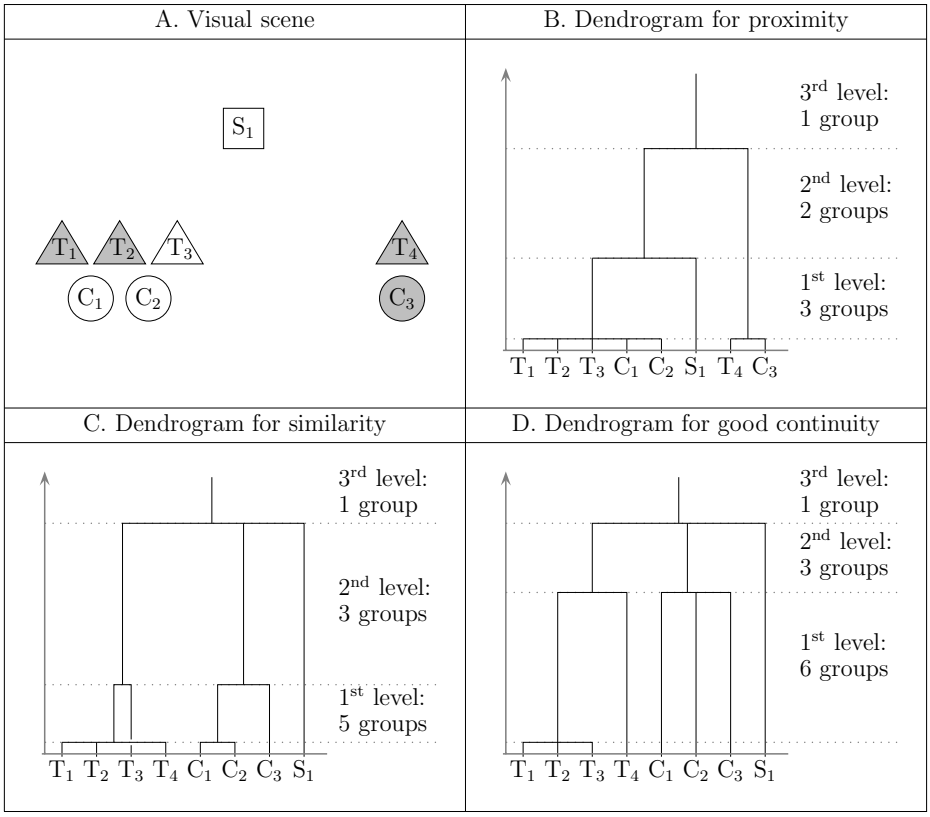


Fig. 3. Grouping objects using a dendrogram for each grouping factor.

4 Multimodal referring phenomena

4.1 Referential gestures

Cosnier and Vaysse [3] propose a synthesis of different classifications of conversational gestures, taking into account the one of Efron [6], which was the first to focus on the referential aspect of gesture, and that of McNeill [17], which does so in a more thorough manner. How the fact of communicating with a machine incites the user to restrict his gestures on his own, especially when the support of the communication is a touch screen? Even if the machine as an interlocutor is symbolized by a human-like avatar, a user does not talk to it as he would to an actual human being [10]. Likewise, we suppose the user will produce neither synchronization nor expressive gestures because he knows that the machine will not perceive or be sensitive to them. As a general rule, we suppose that the user will produce only informative gestures, as opposed to gestures that facilitate the speech process, such as 'beats' and 'cohesives' [17]. For the moment, we focus our work on the design of systems with a touch screen. See the work of Bolt [2]

for the origin, and for instance the work of Wolff *et al.* [26] for a more recent work. In such an interaction mode, the user may be conscious that touching the screen must be informative. Even when not explicitly prohibited from doing so, he will not produce gestures that do not convey meaning. He will also leave out gestures which require anything beyond 2-D, in particular 'emblems' [6] and a lot of 'iconic' and 'metaphoric' gestures [17]. Of the remaining gesture types, we are left with deictic, some iconic and some metaphoric gestures. We note here that these gestures are all referential, which emphasizes on the problem of reference.

The most frequent referential gesture in communication with a touch-screen is the deictic one [26]. What are its functions and the condition of its production, in term of effort (or cost)? As demonstratives or indexicals in language, deictic gesture is an index, i.e., an arbitrary sign that has to be learned and whose main function is to attract the interlocutor's attention to a particular object. A deictic gesture is produced to bring new information by making an object salient which is not already so [15]. Moreover, deictic gestures, as iconic and metaphoric ones, are often produced when a verbal distinguishing description is too long or too complicated, in comparison with an equivalent multimodal expression (a simple description associated with a simple gesture). A distinguishing description has a high cost when it is difficult to specify the object through its role or its properties in the context. It is the case for example when other objects have the same properties: the user has to identify another criteria to extract the referent from the context. He can use a description of its position in the scene, that leads to long expressions like "the object just under the big one at the right corner." Deictic gesture has a cost as well. It depends on the size of the target object and, in 3D-environments, its distance from the participant. Fitts' Law [7], a score that can be computed from these two parameters, is an indicator of the effort in pointing. Another indicator is given by the disposition of the objects in the scene. If the target object belongs to a perceptual group, it is more difficult to point out it than if it is isolated from the other objects. A score can also be computed to quantify the aggregation of the perceptual group. If several Gestalt criteria are simultaneously verified, this score will be high. Then, a gesture whose intention is to extract an object from this group will have a high cost, proportional to the difficulty of breaking the group. On the contrary, a gesture whose intention is to point the whole group will have a low cost.

As a pointing gesture on a single object can be extended to a group, it seems, from the system point of view, that several interpretations are often possible. What are the possible forms of a deictic gesture, and what are the possible interpretations that can be done considering the visual context? On a touch screen, deictic gestures can take several forms: dots ('pointing'), lines, opened or closed curves, 'scribbling.' Trajectories can pass between objects, in order to separate some of them (generally by surrounding them) from the other ones ('circling'), or pass on the target objects ('targeting'). Pointing, scribbling, circling and targeting were the four categories of trajectories extracted from the corpus study by Wolff *et al.* [26]. This study leads to strategy ambiguity (individual reference opposed to group reference), as we already discuss, and to

form ambiguity and also to scope ambiguity. There is a form ambiguity when the same trajectory, for example an unfinished circling curve, can be interpreted as a circling or as a targeting, as shown on the first scene of Figure 4 (the gesture can target the triangles, can surround two circles, or, following a mixed strategy, can point out all of them). There is a scope ambiguity when the number of referents can be larger than the number of target objects, as shown on the second scene of Figure 4 (the gesture can target two or three triangles).

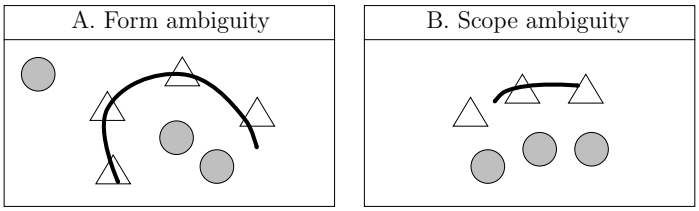


Fig. 4. Form and scope ambiguity.

These possible ambiguities emphasize an additional problem, that the target objects (the referents of the gesture) are not always the referents of the multimodal expression. In the next section we explore the links between speech and gesture and we characterize the links between the referents of the gesture and the referents of the multimodal expression. We then deduce a list of clues that the system may exploit to interpret the reference.

4.2 Gesture referent and multimodal referent

We have seen that the verbal referring expression guides the interpretation of gesture. This can be illustrated by considering the possible expressions "these triangles" and "these circles" in the first scene of Figure 4, and by considering "these two objects" and "these three objects" in the second. In these expressions, only one word, the category in the first case and the numeral in the second, is sufficient to interpret the gesture and then to identify the referents. The demonstrative indicates the presence of a gesture in the referring action, that is if no set of triangles or circles is salient in the dialogue history (possibility of an anaphora). Nevertheless, if the gesture makes one object very salient, a definite article might be used instead of the demonstrative. This situation, more frequent in French than in English, happens in particular during the acquisition of the articles functions by children [12] and can be observed in some spontaneous dialogues (examples can be found in the studied corpus). Another example of the relaxation of linguistic constraints is the use of "him" ("lui" in French) or "he" ("il" in French) with a gesture. In some situations, "il" can be associated with a gesture instead of "lui," which is the usual word to focus on a person [15]. A third example in French is the use of deictic marks. When several objects are placed at different distances, "-ci" in "cet objet-ci" ("this object") and "-là" in

"cet objet-là" ("that object") allow the interlocutor to identify an object closer to or further from him. When a gesture is used together with "-ci" or "-là," the distinction does not operate any more (a lot of examples can be found in the studied corpus).

The referents of some expressions are different from the referents of the associated gesture. It is the case of expressions like "the N_2 preposition this N_1" with a gesture associated with "this N_1." It can be expressions like "the color of this object" (an equivalent of "this color") or spatial expressions like "the form on the left of this object." Their common point is that their interpretation presents two stages, the first (the only one that has an interest here) being the multimodal reference of N_1, and the second being the use of this first identification to resolve the reference of the complete expression, by extracting a characteristic of the referent in the first case, by considering it as a site for the identification of N_2 in the second case.

Fig. 5. Generic interpretation.

One of the classical aspects of reference is the possibility of a specific interpretation and of a generic one. It seems that every multimodal referring expression like "this N" with a gesture, can refer to the specific object that is pointed out, or to all objects of the N category. Sometimes there is a clue that gives greater weight to one interpretation. For example, an unambiguous gesture pointing out only one object will lead to the generic interpretation if it is produced with "these forms," where the plural is the only clue (Figure 5. This interpretation is confirmed by the presence of other objects with the same form, and by the fact that being in a perceptual group these objects need a high cost to be pointed out. On the contrary, the use of a numeral will reject the generic interpretation. When no clue can be found, the task may influence the interpretation (some actions must be executed to specific objects), and, for this reason, we do not settle here. To summarize, we propose the following list of clues:

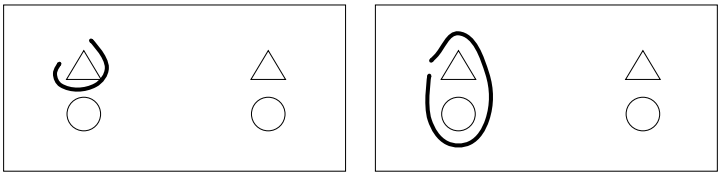
- the components of the nominal phrase: the number (singular or plural, eventually determined by a numeral or a coordination like in "this object and this one" with one circling gesture); the category and the properties (to filter the visible objects and to count the supposed referents);
- the predicate: its aspect and its role considering the task (to reinforce the specific interpretation);

– the visual context: the presence and the relevance of perceptual groups (to interpret a scope ambiguity); the presence of similar objects (to make the generic interpretation possible).

These clues show that the multimodal fusion is a problem that occurs at a semantic level and not at a media level, as it is considered in many works ([2] is a famous example that is still followed).

4.3 Referent and reference domain identification

We show in this section how the reference resolution goes through the identification of the referents and of the context from which these referents are extracted. We first demonstrate the importance of taking this context into account, and, second, we expose the possible links between a gesture trajectory and the context demarcation.



A. "This triangle" B. "The triangle"

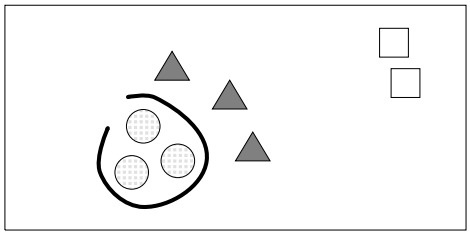
Fig. 6. Referent and domain delimitation.

In the first scene in Figure 6, a triangle is pointed out by an unambiguous gesture associated with a simple demonstrative expression. Supposing that the next reference will be "the circle," it is clear that such a verbal expression will be interpreted without difficulty, designating the circle just under the triangle of the last utterance. Whereas two circles are visible on the scene, the one being in the same visual reference domain than the precedent referent will be clearly identified. This is one role of the proximity criterion of the Gestalt Theory, as we have precised it in section 3.2. If the reference domain is implicit in the first scene of Figure 6, it is explicit in the second scene. In this case, the expression "the triangle" has the role to extract the referent from the domain delimited by the gesture. Thus, Figure 6 shows the two main roles of gesture: delimitating referents or delimitating a domain.

As in Figure 6, we begin to study examples where the gesture is unambiguous, generally when it has a circling form that can not be interpreted as a targeting one. When the set of target objects is identified, it is compared to the linguistic constraints of the referring expression. These constraints are the category and properties filters, and the functionality of the determiner. Following Salmon-Alt [22], the use of a demonstrative implies the focus on some objects in a domain where other objects with the same category are present. This focus is done by

salience, and particularly by the salience due to gesture. The use of a definite article implies an extraction of objects of a given category in a domain where some objects of another category may be present (but not necessarily).

These linguistic constraints give evidence for the role of the gesture. In the second scene of Figure 6, the target objects are not all "triangles." The use of the definite article "the" implies a domain containing triangles and other forms of objects. This domain is clearly the set of target objects. As the expression is singular and as there is one triangle in this domain, the extraction of the referent leads to the unambiguous identification of this triangle. In contrast, the target object in the first scene is a "triangle." As the expression is singular, the multimodal referent may be this target object, and the domain has to be identified. For that, we search a domain containing another triangle. The whole visual context is such a domain. It allows one to interpret the next reference "the other one" as "the other triangle in the domain." There is here a problem: at the beginning of this section, with Figure 6-A, we construct with the proximity criterion the perceptual group at the left of the scene, and we exploit this group, which can be seen as a reference domain, to interpret the next reference "the circle." But this reference domain hypothesis does not fit well with the demonstrative of "this triangle" because it does not contain any other triangle. Our model will handle both hypotheses, to make all interpretations possible. But the reference domain corresponding to the whole visual context will be labeled with a better relevance, and will be tested first in the interpretation process.



"These forms which are the most clear"

Fig. 7. Gesture initiating a domain.

Another example where the gesture is not ambiguous but where the identification of the reference domain is complex is given in Figure 7. The hypothesis of a gesture delimitating the reference domain is impossible, and so the set of target objects may be the multimodal referents. For the identification of the possible reference domains, we must take "the most clear" into account. The hypothesis of the whole visual context is impossible because the three circles are lightly gray whereas the two squares are perfectly white. The proximity criterion gives a solution, by constructing a reference domain including the three circles and

the three triangles. In this domain, the "forms which are the most clear" are the circles indeed.

When the gesture is ambiguous, a way to proceed is to test all the mechanisms seen above. With the example of a pointing gesture that can designate one object or a perceptual group, the use of a definite determiner will give greater weight to the hypothesis of the perceptual group as the reference domain. With the example of a gesture that can target two or three objects, the presence of other objects of the same category will influence the identification of reference domain. Considering the expression "the triangles" with the gesture of Figure 4-B, the hypothesis of the whole visual context will be relevant as reference domain and the referents will be the three triangles. On the other hand, using the demonstrative "these triangles," we restrict the referents to the two triangles under the trajectory, thus leaving the third triangle in the reference domain, and allowing for the demonstrative mechanism to be applied.

5 Conclusion

Reference to objects in multimodal dialogue systems can take several forms which are not linked to particular mechanisms of identification. As opposed to approaches like the one of Kehler [13], we consider that simple algorithms are not sufficient for a multimodal system to identify the referents. As we have seen with our examples, the gesture does not always give the referents, and the components of the verbal expression are not sufficient to distinguish them. But the combination of these ostensive clues with inferred contextual considerations does. Then, the question is: how can we combine all the clues to lead to the right interpretation? To answer this question, we investigate in this paper the 'multimodal reference domain' model, whose aim is to formalize the clues into homogeneous structures (reference domains) and then to combine these clues by comparing and merging reference domains. Our study is based on the concrete examples we found in the corpus of Wolff *et al.* [26] and in linguistic classical works like [18], [20] or [23]. As it is showed with the implementation of reference domains in two multimodal dialogue systems (in MIAMM and OZONE project, see section 2), our model appears to be relevant for different kinds of interaction modalities and for different kinds of applications.

References

1. Beun, R.-J., Cremers, A.H.M.: Object Reference in a Shared Domain of Conversation. Pragmatics and Cognition **6(1/2)** (1998) 121–152
2. Bolt, R.A.: Put-That-There: Voice and Gesture at the Graphics Interface. Computer Graphics **14(3)** (1980) 262–270
3. Cosnier, J., Vaysse, J.: Sémiotique des gestes communicatifs. Nouveaux actes sémiotiques (geste, cognition et communication) **52** (1997) 7–28
4. Dekker, P.: Speaker's Reference, Descriptions and Information Structure. Journal of Semantics **15(4)** (1998) 305–334

5. Edmonds, P.G.: A Computational Model of Collaboration on Reference in Direction-Giving Dialogues. Ms. Thesis, University of Toronto, Canada (1993)
6. Efron, D.: Gesture, Race and Culture. Mouton, The Hague (1972)
7. Fitts, M.: The Information Capacity of the Human Motor System in Controlling Amplitude of Movement. Journal of Experimental Psychology **47** (1954) 381–391
8. Grosz, B.J., Sidner, C.L.: Attention, Intentions and the Structure of Discourse. Computational Linguistics **12(3)** (1986) 175–204
9. Itten, J.: The Art of Color. Reinhold Publishing Corp., New York (1961)
10. Jöhsson, A., Dählback, N.: Talking to a Computer is not like Talking to your Best Friend. In: Proceedings of the Scandinavian Conference on Artificial Intelligence. Tromsø (1988)
11. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer, Dordrecht (1993)
12. Karmiloff-Smith, A.: A Functional Approach to Child Language. Cambridge University Press (1979)
13. Kehler, A.: Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In: Proceedings of the 17th National Conference on Artificial Intelligence. Austin (2000)
14. Kievit, L., Piwek, P., Beun, R.-J., Bunt, H.: Multimodal Cooperative Resolution of Referential Expressions in the DenK System. In: Bunt, H., Beun, R.-J. (eds.): Cooperative Multimodal Communication. Springer, Berlin Heidelberg (2001) 197–214
15. Kleiber, G.: Anaphores et pronoms. Duculot, Louvain-la-Neuve (1994)
16. Landragin, F.: Dialogue homme-machine multimodal. Hermes Science Publishing, Paris (2004)
17. McNeill, D.: Psycholinguistics: A New Approach. Harper and Row, New York (1987)
18. Moeschler, J., Reboul, A.: Dictionnaire encyclopédique de pragmatique. Seuil, Paris (1994)
19. Olson, D.R.: Language and Thought: Aspects of a Cognitive Theory of Semantics. Psychological Review **77** (1970) 257–273
20. Reboul, A., Moeschler, J.: Pragmatique du discours. Armand Colin, Paris (1998)
21. Rock, I., Brosgole, L.: Grouping Based on Phenomenal Proximity. Journal of Experimental Psychology **67** (1964)
22. Salmon-Alt, S.: Reference Resolution within the Framework of Cognitive Grammar. In: Proceedings of the International Colloquium on Cognitive Science. San Sebastian, Spain (2001)
23. Sperber, D., Wilson, D.: Relevance. Communication and Cognition (2nd ed.). Blackwell, Oxford UK Cambridge USA (1995)
24. Thórisson, K.R.: Simulated Perceptual Grouping: An Application to Human-Computer Interaction. In: Proceedings of the 16th Annual Conference of the Cognitive Science Society. Atlanta, Georgia (1994)
25. Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt II. Psychologische Forschung **4** (1923)
26. Wolff, F., De Angeli, A., Romary, L.: Acting on a Visual World: The Role of Perception in Multimodal HCI. In: Proceedings of AAAI'98 Workshop: Representations for Multi-modal Human-Computer Interaction. Madison, Wisconsin (1998)