Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems

Frédéric Landragin

Visual Perception, Language and Gesture: A Model for their Understanding 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 human-machine dialogue system that focuses on deep comprehension. We show how a referring act takes place into a contextual subset of objects. This subset is called 'reference domain' and is implicit. It can be deduced from a lot of clues. Among these clues are those which come from the visual context and those which come from the multimodal utterance. We present the 'multimodal reference domain' model that takes these clues into account and that can be exploited in a multimodal dialogue system when interpreting.

Key words: multimodal communication, visual perception, pointing gesture, natural language processing, reference to objects, salience, interpretation modeling

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 [30], 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 for the interpretation of reference to objects phenomena (see Figure 1 for a personal synthesis of these aspects that will be detailed in the paper). To us, this has to be done by using a unified framework, in order to compare and to merge the various information from the various contextual aspects into homogeneous structures.

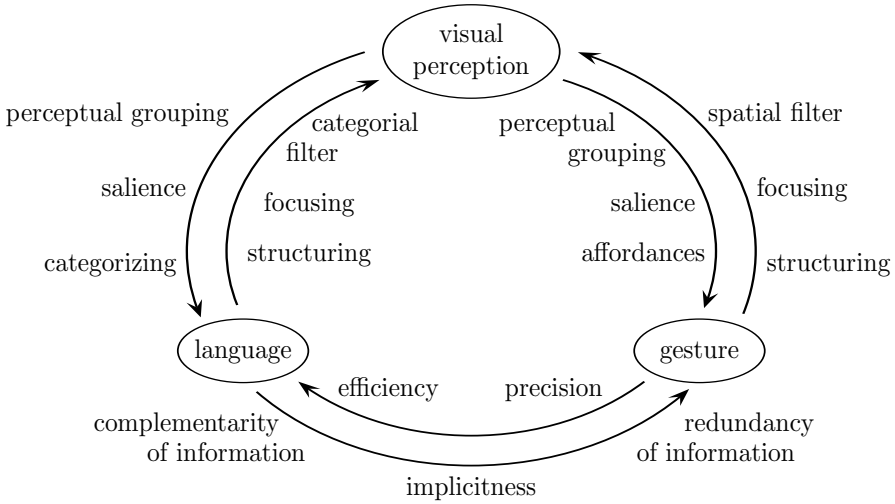


Fig. 1. Some interactions between visual perception, gesture, and language.

It is with this aim that we have been developed since several years the 'multimodal reference domain' model. As opposed to approaches like the Discourse Representation Theory (DRT, [14]), this model has been built with multimodal concerns from the early first phases of the design. As opposed to approaches based on domains of quantification [6], 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 input information. We first present in section 2 the main principles of our model. In the two next sections, we describe in details all phenomena we want our model to take into account. In section 3 we focus on perceptual phenomena and we describe how we translate the visual context into visual reference domains. In section 4 we focus on multimodal referring phenomena and we 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. In section 5 we deduce from all these phenomena a complete model of multimodal reference resolution. We then propose some arguments for the evaluation of this model and we conclude on its strong and weak points.

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 instance 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 group'), and one or more partitions of elements. A partition gives information about possible decompositions of the domain [28]. 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.

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 with linguistic preoccupations in [26] and [28], with multimodal preoccupations in [22], and as it is showed in Figure 2. The linguistic and gestural information allow to build an underdetermined domain that groups all constraints. In the example of the figure, the referring expression that is currently treated is "this circle". Such a demonstrative nominal phrase implies that a particular circle is focused upon. This interpretation constraint can be translated into an underdetermined reference domain, that consists of a partition where one element is focused. In the representation of the underdetermined reference domain (URD), the partition corresponds to the white box with two compartments. Since the differentiation criterion that characterizes the partition is 'focusing', the first compartment is dedicated to the focused element. Since it is the one we are looking for, there is a question mark in it. Moreover, "this circle" is making a contrast between a

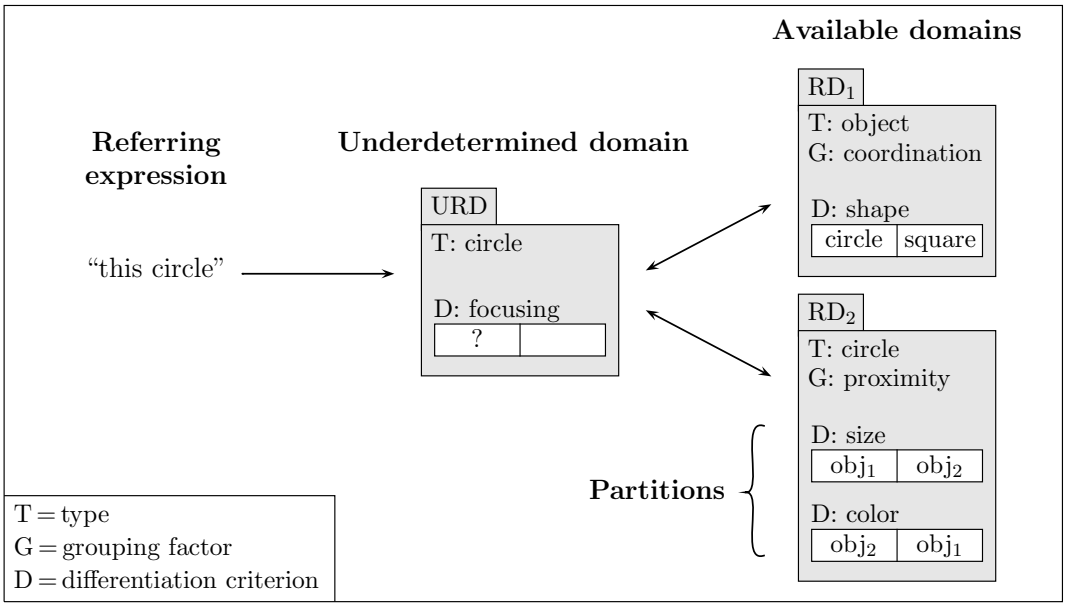


Fig. 2. Interpretation of a referring expression using reference domains.

particular circle and other objects with the 'circle' type (this is one important mechanism of the demonstrative determiner, see [15] and [3]). So the reference domain in which the interpretation occurs must include only circles. This is the role of the 'type' attribute in URD.

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. The important point here is that all input information, i.e., all signals, whatever their nature, visual, gestural, or linguistic, are translated into homogeneous structures, i.e., reference domains, that can be combined, compared, and merged.

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 partly developed in the framework of several European projects: ACTS COVEN (see http://coven.lancs.ac.uk), IST MIAMM (see http://www.miamm.org), and IST OZONE (see http://www.hitech-projects.com/euprojects/ozone). We don't want here to describe the implementation of the algorithm, because it requires the presentation of a lot of technical problems that are not of im-

portance here (such as algorithms for the recognition of gesture trajectories, or the calculus of salience scores). We want to focus on the exploitation of contextual and communicative clues for the interpretation of multimodal referring expressions, in order to emphasize how multimodal interpretation can be made in a systematic way with reference domains.

3 Perceptual phenomena

3.1 Focusing and 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 [7] 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 goal 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 pictorial arts studies (Itten, Kandinsky, etc., see for instance [12]) 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 instance 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 model of the goal to be achieved. 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. In task-oriented human-machine dialogue, the notion of task model is strongly linked to that aspect.

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 [31] 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 [31], 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 [30].

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 [27] 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.

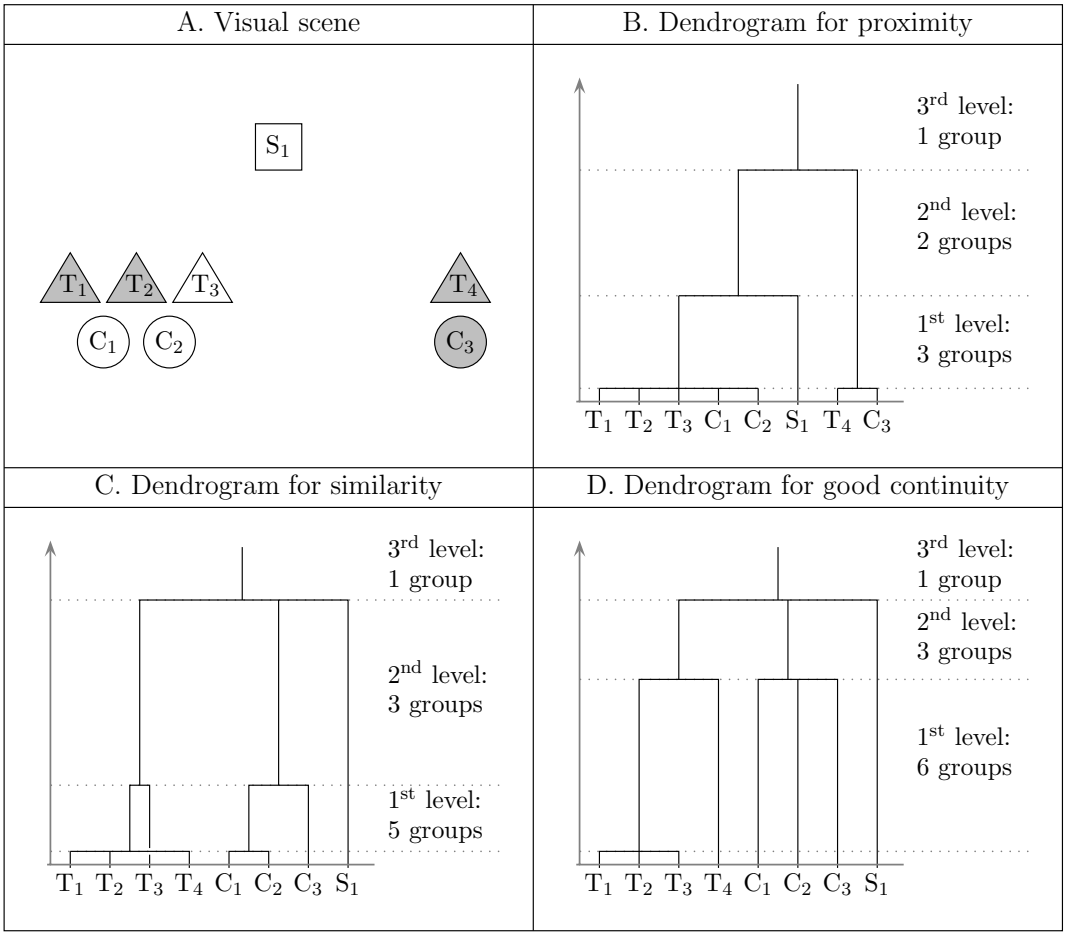


Fig. 3. Grouping objects using a dendrogram for each grouping factor.

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. 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. It was the case with RD_2 from Figure 2, which applied to circles only and was built on using the proximity criterion. 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'. When no salient object intervenes, partitions can be created following the relevant differences between the elements. In the case of RD_2, the size and the color of the elements were relevant as differentiation criteria. In general, the relevant differences are emphasized by the main levels of the den-

drograms (Figure 3). Considering this method, visual reference domains can be numerous. But, as we will see in the next sections, referring phenomena will only exploit some of them.

4 Multimodal referring phenomena

4.1 Referential gestures

Cosnier and Vaysse [5] propose a synthesis of different classifications of conversational gestures, taking into account the one of Efron [8], which was the first to focus on the referential aspect of gesture, and that of McNeill [23], which does so in a more thorough manner. There exists a lot of sorts of conversational gestures, and we can ask whether all of them are suitable for human-machine dialogue. For instance, what is the influence of a touch screen on the user's behavior? Does he restrict his gestures on his own? 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 [13]. 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' [23]. 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.* [32] 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' [8] and a lot of 'iconic' and 'metaphoric' gestures [23]. 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.

In this paper we will not study iconic and metaphoric gestures, because we want to focus on the identification of the referent (as an object which is managed by the application and that the dialogue manager has to identify) of a referring action. Iconic and metaphoric gestures refer to concepts or actions, and they complete verbal information by giving it some additional features. In human-machine dialogue systems, these features may be managed using reference domains, but this leads us too far from our initial problem, and we prefer to concentrate on purely deictic aspects.

Then, the most frequent referential gesture in communication with a touch-

screen is the deictic one [32]. 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 [18]. 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 criterion 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 [9], 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 it out 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.

Since perceptual groups correspond to the visual reference domains that were previously described, this point has an importance. In particular it shows how the interpretation of a gesture relies on the existence of visual reference domains.

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.* [32]. 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).

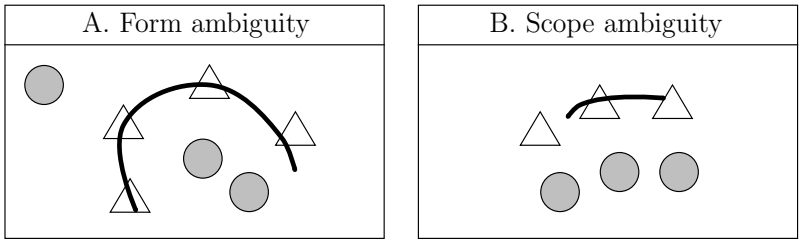


Fig. 4. Form and scope ambiguity.

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 subsection 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 [15] and can be observed in some spontaneous dialogues (examples can be found in the corpus of Wolff *et al.* [32]). 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 [18]. 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 same 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.

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 goal to be achieved may influence the interpretation (some actions must be executed to specific objects), and, for this reason, we do not settle here.

Fig. 5. Generic interpretation.

To summarize, we propose the following list of clues:

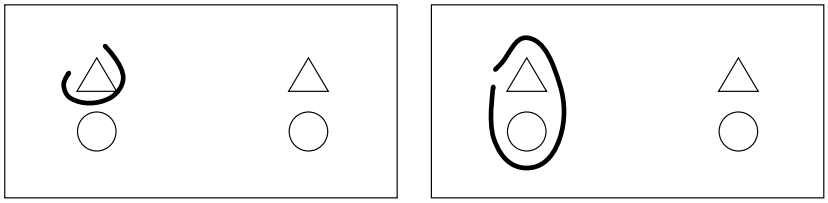
- the components of the nominal phrase: the number (singular or plural, eventually determined by a numeral or a coordination like in "this object and this one" with one circling gesture); the category and the properties (to filter the visible objects and to count the supposed referents);
- the predicate: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

We show in this subsection 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.

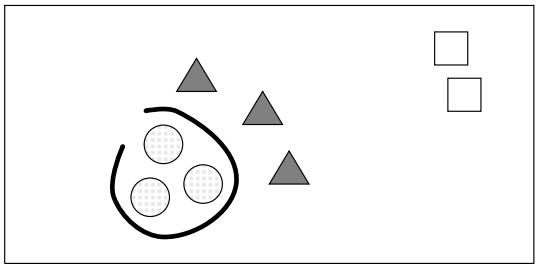
In the first scene in Figure 6, a triangle is pointed out by an unambiguous gesture associated with a simple demonstrative expression. Supposing that the next reference will be "the circle", it is clear that such a verbal expression will be interpreted without difficulty, designating the circle just 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 detailed 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 [28], 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 Multimodal reference resolution

5.1 Approach

At this stage of the paper, we have clarified a lot of understanding processes for multimodal input. In particular we have described how visual, linguistic and gestural signals have to be treated in order to allow a deep comprehension of their working within referring actions. We have proposed a visual processing model based on visual reference domains; we have proposed a natural language processing model based on linguistic reference domains; and we have clarified how the gesture participates to reference resolution. We emphasized for instance the possible roles of a gesture trajectory among the following ones: indicating a referent, delimitating a domain, or initiating a domain. With this basis we need now to determine an algorithm for reference resolution in multimodal dialogue systems. This algorithm will have to take into account all perceptual and multimodal referring phenomena we have talked about in sections 3 and 4, and will be based on rules implying reference domains.

Due to the complexity of perceptual and multimodal referring phenomena, we consider that simple algorithms are not sufficient for a multimodal system to identify the referents. As opposed to approaches like the one of Kehler [16] which is based on a corpus and presents a simple algorithm within the limits of the corpus, we want to go beyond the exploitation of the phenomena we can find in corpora, and to provide an algorithm that takes into account all cases: the ones from the corpora, and others that can be extrapolated from them and from linguistic concerns such as the ones we previously described

and, in particular, the theory of Corblin [3] dealing with determiners. Since no corpus may include all possible multimodal referring phenomena, our claim is that only a systematic approach can apprehend such an issue.

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. In our algorithm, we focus on modeling this combination. The starting point is the access to the referents through reference domains. Such an access is called 'referring mode' and is described in the next subsection. Then, we will propose a systematic algorithm to label any referring action with possible referring modes considering the type of the determiner.

5.2 Referential terms and referring modes

Considering the presence of reference domains, we propose the following list of referring modes, that groups several manners to consider a referent and a reference domain:

- **'new-ref'** mode: introducing a new referent in a linguistic manner;
- **'ext-any-ref'** mode: extracting any element from an activated reference domain;
- **'ext-par-ref'** mode: extracting a particular element from an activated reference domain;
- **'ind-par-ref'** mode: indicating a particular referent that is focused elsewhere;
- **'ind-par-dom'** mode: indicating a particular reference domain whose one element is focused;
- **'gen-ref'** mode: referring to a generic entity, that is not a set of particular referents nor a reference domain.

To a referential term corresponds a preferential referring mode. An indefinite noun phrase is generally used to introduce a new referent; a definite is an indicator to the necessity of extracting a particular referent [3]. A pure demonstrative is generally used together with a pointing gesture that is another cause of focus. But all of these referential terms have other uses. Thus, there is no one-to-one relation between referential terms and modes.

With state of the art studies, the problem is the same. For instance, Reboul [24] proposes to distinguish the following referring modes: direct reference, indirect reference, demonstrative reference, deictic reference and anaphoric reference. Proper names constitute the preferential direct referring mode, and demonstratives the preferential demonstrative referring mode. The problem is that a same noun phrase can be used for several referring modes. For example,

the demonstrative noun phrase "this object" can be used for a demonstrative reference that implies an ostensive gesture, or for an anaphoric reference, the antecedent being a previous noun phrase like "the blue triangle". Indeed, demonstratives just as definites can be used for anaphoric purposes.

In our list, the introduction of a new referent by a multimodal referring action is seen as the linguistic mention of a referent that is focused by the coreferent ostensive gesture ('ind-par-ref' mode). One important point is that we consider that referring directly to a particular object is impossible without an activated domain. Consequently, the direct reference mode of Reboul corresponds here to 'ext-par-ref' mode. Mentional expressions [4] with "first", "second" or "last" also correspond to 'ext-par-ref' mode, the differentiation criterion for the referents identification being the rank. Words like "other" and "next" have particular mechanisms. They refer to not-focused elements of a domain that has just been used, and are then included in 'ext-par-ref' mode. The strong point of our list is that all these phenomena can be modeled using reference domains and labeled with an item of our list.

As we said, there is no one-to-one relation between referential terms and modes. To interpret a referential term, we need to consider the context, which includes perceptual information, previous referring actions (and their results), and various world knowledge. With this linguistic, gestural, and visual information, the context is heterogeneous. Moreover, its scope can be enormous. Considering reference domains, the problem can be apprehended in a more efficient way. Our hypothesis is that each referring action occurs in a reference domain. As we already saw in this paper, this contextual subset is generally implicit and has to be identified by the system. The utterance's components allow to extract the referents from this subset and to prepare the interpretation of a future reference. For instance, the referential term "the red triangle" includes two properties that must be discriminative in a reference domain that must include one or more "not-red triangles". A further referential term like "the other triangles" may be interpreted in the same domain, denoting a continuity in the reference sequence.

5.3 The modes linked to a referential term

The possible modes considering the type of determiner are grouped in the following tables: Figure 8 for indefinite noun phrases (including headless ones), Figure 9 for definite noun phrases, Figure 10 for demonstrative noun phrases, Figure 11 for personal pronouns, and Figure 12 for demonstrative pronouns. Following the works of Corblin [3], Salmon-Alt [28], and also Karmiloff-Smith [15], we give a particular importance on the functions of the determiners. With all the possibilities we present, we show how complex the relation between

Mode	**Mechanism details** including the nature (linguistics or multimodal) of the referential expression		**Examples** with singular, quantifier, plural, numeral adjective
new-ref	The referential expression does not refer but can be the antecedent of a future anaphor. No coreferent pointing gesture.		"Create **a square**", "**some squares**", "**squares**", "**two squares**" (all possibilities, i.e., singular and plural are possible, as well as quantifiers and numeral adjectives).
ext-any-ref	The activated reference domain must be more reduced than the whole ontological class of objects. It can be:	delimited by a coreferent pointing gesture	"Delete **a square**", "**some squares**", "**squares**", "**two squares**" with a gesture delimiting a set of squares (all quantifiers and numeral adjectives are possible).
		delimited by a previous referential term	"Select the squares and the triangles" followed by "delete **a square**", "**some squares**" (interpreted as "delete some of the selected squares"), "**squares**", "**two squares**" (all possibilities).
		imprecise (in which case we consider the whole visual context)	"Delete **two squares**" interpreted as "delete two of the visible squares", and eventually as "delete two of the visually salient squares" (all possibilities).
ind-par-ref	It is the pointing gesture that forces the choice of the referent. This case constitutes a deviance from classical theories like the one of Corblin [3]. Nevertheless, we found it in the corpus of Wolff *et al.* [32].		"Delete **a square**" with a gesture pointing out a particular square, "**some squares**", "**squares**", "**two squares**" with gestures pointing out particular squares (all possibilities).
gen-ref	Reference to a class of objects. No coreferent pointing gesture.		"**A square** has four sides", "**squares** have four sides", "**two triangles** with a common side make a quadrilateral" (quantifiers are impossible).

Fig. 8. Indefinite noun phrases.

terms and modes is. A system that has to interpret spontaneous multimodal expressions must know all this information.

5.4 Reference resolution

In this subsection we exploit the previous tables and we present an algorithm for reference resolution starting with referential terms and using reference domains. From the components of the verbal utterance and from the possible ostensive gesture, we deduce a list of clues that the system may exploit to identify the correct referring mode, the correct reference domain and the correct referent. We start with the determiner and then we detail the role of the predicate and of the other linguistic components. Following Corblin [3], we make the hypothesis that the propositional context (and not only the determiner in the referential term) will favor the specific or the generic interpretation. Indeed, we consider that generic references are not usual in human-computer interaction. Thus, we ignore here the 'gen-ref' mode.

For an indefinite noun phrase, the system may choose between 'new-ref', 'ext-

Mode	**Mechanism details**		**Examples** with singular, plural, numeral adjective
ext-par-ref	The activated domain can be:	delimited by a coreferent pointing gesture	"**The triangle**" with a gesture delimiting a group of geometrical forms including one triangle (all possibilities). This is the case of Figure 6-B.
		delimited by a previous referential term	"Select the blue triangle and the green square" followed by "delete **the triangle**", "select the triangles" followed by "**the two red triangles**", "the red triangle, the green one and the blue one" followed by "**the first**", "the group" followed by "**the triangle**" (all possibilities).
		delimited by a previous focusing on a visual space	After some references to objects at the left of the visual scene, "**the triangle**" can refer to "the triangle on the left" (all possibilities).
		delimited by a precision in the referential term	"**The triangle on the left of the scene**" (all possibilities).
		imprecise (in which case we consider a salient focus space)	"**The triangle**" interpreted as "the salient triangle" (all possibilities).
ind-par-ref	The referent can be:	given by a coreferent pointing gesture	"**The triangles**" with a gesture pointing a group of triangles (all possibilities)
		given by a previous referential term	"Select a red triangle" followed by "**the triangle**", "the triangle and the square" followed by "**the two forms**" (all possibilities).
ind-par-dom	The focused element can be:	given by a coreferent pointing gesture	"**The triangles**" with a gesture pointing out one triangle (in this particular example, a pointed object is extended to a group of similar objects, so only the plural is relevant).
		given by a previous referential term	"The square with circles around" followed by "**the group**" (this is also a particular case).
gen-ref	No coreferent pointing gesture.		"**The triangle** is a simple geometrical form", "**the triangles** have three sides" (numeral adjectives are impossible).

Fig. 9. Definite noun phrases.

any-ref', and 'ind-par-ref' referring modes. The presence of a pointing gesture may help the system: if the gesture delimits a set of objects not reduced to the referent(s), the only possible mode is 'ext-any-ref'; if the gesture is pointing out the referent(s), the only possible mode is 'ind-par-ref'. If no coreferent gesture is produced, there is an ambiguity between 'new-ref' and 'ext-any-ref' modes. The presence of an activated linguistic domain will favor the 'ext-any-ref' mode. In the other case, the predicate will disambiguate: a verb that denotes the introduction of new referent(s) like "add" and "create" will force the 'new-ref' mode. The 'ext-any-ref' mode will be chosen otherwise.

In the 'new-ref' interpretation, the system has to add the new object(s) in the visual domain corresponding to the scene. In this domain, a new partition is created, with a differentiation criterion linked to the predicate. The chosen referent(s) are focused in this partition. In the 'ext-any-ref' interpretation, the considered domain is the activated one and the process is the same. In the

Mode	**Mechanism details**		**Examples** with singular, plural, numeral adjective
ext-par-ref	A coreferent pointing gesture is impossible. Focusing is necessarily due to a previous referential term.		"Select the blue triangle and the green square" followed by "delete **this square**" (all possibilities).
ind-par-ref	The referent can be:	given by a coreferent pointing gesture	"**This triangle**" with a gesture pointing out one triangle. This is the most common multimodal referring expression (all possibilities). This is the case of Figure 6-A and Figure 7.
		given by a previous referential term	"Select the blue triangle" followed by "**this triangle**", "the triangle" followed by "**this form**" (all possibilities).
ind-par-dom	The focused element can be:	given by a coreferent pointing gesture	"**These triangles**" with a gesture pointing out one triangle with a particular aspect (extension to a group of similar objects, so only with a plural).
		given by a previous referential term	"The square with circles around" followed by "**this group**" (the same particular example than in definites).
gen-ref	Three referring modes can be distinguished:	transition from a gestural antecedent to a generic interpretation	"**These forms**" with a gesture pointing out one triangle (numeral adjectives are impossible). This is the case of Figure 5.
		transition from a linguistic antecedent to a generic interpretation	"This strange form" followed by "**these forms**" (only plural with no numeral adjective).
		direct multimodal generic interpretation	"**This form**" with a gesture pointing out one triangle, that can be interpreted as "this type of form", and is by consequence ambiguous with a specific interpretation (only singular).

Fig. 10. Demonstrative noun phrases.

'ind-par-ref' interpretation, the process is the same, except that the choice of referents is not free but constrained by the gestural interpretation.

For a definite noun phrase, the system may choose between 'ext-par-ref', 'ind-par-ref', and 'ind-par-dom' modes. A fine analysis of the referential term and the possible pointing gesture is not sufficient to disambiguate. All hypotheses have then to be kept. In the 'ext-par-ref' interpretation, the system has to extract and to focus the referent from the activated domain. This referent has to be isolated with the category and its modifiers. For the 'ind-par-ref' and 'ind-par-dom' modes, the system has to build a new domain around the focused referent. The differentiation criterion of the new partition in this domain is the referent category.

For a demonstrative, the process is nearly the same than for definites, except that for 'ind-par-ref' and 'ind-par-dom' modes, the differentiation criterion of the new partition is given by the predicate or by the intervention of a pointing gesture. That shows the main difference between definites and demonstratives: the contrast between the referent and the other elements of the reference domain is due to category (and modifiers) for definites, and to focusing for

Mode	Mechanism details		Examples
ind-par-ref	The referent can be:	given by the communicative situation	An object can be so salient in the situation that an expression such as "**it**" is sufficient to refer to it, without any gesture. Another case is when an obvious intention, for instance the movement of a child, is susceptible to capture the interlocutor's attention on him and to make the pronoun sufficient to refer to him (singular and plural are possible).
		given by a coreferent pointing gesture	"**He** has a big head" with a gesture pointing out a man (direct ostension) or his hat (deferred ostension) (singular and plural are possible).
		given by a previous referential term	"Sélectionne le triangle bleu" / "select the blue triangle" followed by "supprime-**le**" / "delete **it**". One other case is when the pronoun refers to another specimen of the referent linked to the antecedent: "j'ai supprimé le triangle mais **il** est revenu" / "I deleted the triangle but **it** appears again" (singular and plural are possible).
ind-par-dom	A coreferent pointing gesture is impossible. The focused element is given by a previous referential term.		"Ajoute un triangle vert" / "add a green triangle" followed by "supprime-**les**" / "delete **them**" (the plural form is necessary to build on the domain).
gen-ref	A coreferent pointing gesture is impossible. This case corresponds to the transition from a linguistic antecedent to a generic interpretation.		"J'ai ajouté un triangle rouge parce qu'**ils** attirent le regard" / "I added a red triangle because **they** are eye-catching" (the plural form is necessary).

Fig. 11. Personal pronouns.

Mode	Mechanism details	Examples
ext-par-ref	In this mode, a coreferent pointing gesture is impossible. The focusing is necessarily due to a previous referential term.	"Le triangle, le carré et le rond" / "the triangle, the square and the circle" followed by the mentional reference "**celui-ci**" / "**this one**" (singular and plural are possible).
ind-par-ref	Demonstrative pronouns combine a demonstrative reference and an anaphor. They are associated to a pointing gesture to refer to a new object with the characteristics of a previous referent. The focusing is then necessarily due to a coreferent gesture.	In "sélectionne ce triangle bleu" / "select this blue triangle" followed by "supprime **celui-ci**" / "delete **this one**", "celui-ci" / "this one" together with a coreferent gesture refers to another blue triangle (singular and plural are possible).
gen-ref	See gen-ref mode for personal pronouns.	"J'ai ajouté un rond vert et un triangle rouge. **Ceux-ci** attirent le regard" / "I added a green circle and a red triangle. **These ones** are eye-catching" (the plural form is necessary).

Fig. 12. Demonstrative pronouns.

demonstratives.

For a personal pronoun, the system may choose between 'ind-par-ref' and 'ind-par-dom' modes. The clue to disambiguate is a change in the use of singular or plural forms: if a transition occurs from a singular to a plural form, then the 'ind-par-dom' is identified. In this case, the system has to build a new domain around the focused element, the differentiation criterion of the new partition being the category. In the other case, the focusing nature does not

change and then no new domain has to be built.

For a demonstrative pronoun, the presence of a pointing gesture forces the 'ind-par-ref' interpretation ('ext-par-ref' interpretation otherwise). In this case, the system has to extract and to focus the referent from the activated domain, the differentiation criterion being the order of mention. For the 'ext-par-ref' interpretation, the system has to build a new domain around the focused element, the new differentiation criterion being the gestural intervention.

All these cases show how the access to the referents depends on the type of utterance, the type of gesture, and the combination of both of them. Since identifying the referring mode is essential for a deep comprehension of multimodality, we claim that simple algorithms are not sufficient to take into account the various possibilities we described. Reference domains are needed to understand multimodal referring actions, and multimodal dialogue systems should be able to manage reference domains.

6 Evaluation and discussion

6.1 Designing multimodal systems

A first method for evaluating an algorithm for multimodal reference resolution consists of designing a multimodal dialogue system and running a set of use tests. We participated to the design of three multimodal dialogue systems, in the framework of the COVEN, MIAMM and OZONE projects (see section 2). Each system involved a set of constraints linked to the application, to the objects that were manipulated, and to the possible communication modalities. In this subsection we present the aspects of our approach that we tested when designing systems or parts of systems.

One of the main aspects of the COVEN (COllaborative Virtual ENvironments) project was to provide a spontaneous multimodal interaction between the user and a 3-D virtual environment. The task consisted of the arrangement of an interior, the objects being chairs, tables and so on. Since these objects were displayed in three dimensions, we had to implement a gesture recognition module dedicated to 3-D problems. We focused on the possible ambiguities due to the third dimension, and we explored the links between speech and gesture in order to resolve these ambiguities. Then our implementation included the management of visual salience and an algorithm for perceptual grouping based on proximity criterion (in the 3-D space and not in the 2-D projection, as discussed in section 3.2). The COVEN system was also our first opportunity to test a speech recognition module, and to confront the possible recognition

errors (for instance with determiners) to the theoretical linguistic constraints and gesture possibilities. In particular, we used to detect inconsistencies between the use of a gesture and the use of a determiner, and then to question the result of the speech recognition module. For this system, only a preliminary form of reference domain was implemented, but a lot of aspects linked to referring modes and referential terms understanding were managed.

The MIAMM (Multidimensional Information Access using Multiple Modalities) framework consisted of a multimodal dialogue system for the access to multimedia information using force feedback gesture devices. It was an opportunity for us to test the relevance of reference domains for haptic human-machine interaction. As it is described in [20], we imagined tactile reference domains. That was a means for proving the interest and the flexibility of our multimodal reference domain model for various interaction paradigms. More precisely, the MIAMM project was an opportunity for us to apply the notions of salience and reference domain to tactile perception. We imagined tactile specific salience factors, and a specific algorithm for building on tactile reference domains. Since they could be translated into the same type of structure, such tactile domains could be combined and merged to linguistic and visual reference domains. We emphasized the strong links between visual and tactile domains, and, as a main result, we concluded that our 'multimodal reference domain' model (with its principle based on the fusion of an underdetermined domain with contextual domains) was relevantly adequate to handle complex multimodal interaction.

Concerning the OZONE (O_3, Offering an Open and Optimal roadmap towards consumer oriented ambient intelligence environment) framework, our purpose was to build on a multimodal dialogue system for the reservation of train tickets. Possible trips were displayed on the screen and the user could point out them and ask questions about them. We exploited the possibilities of the touch screen of a Tablet PC. The architecture and the management of reference domains are described in [21]. Since the interaction was multimodal and involved a 2-D visual scene as a support, we focused our implementation on the management of 2-D gestures, visual salience factors, and visual reference domains. Only few linguistic aspects were taken into account (only the main differences between definites and demonstratives, but not our complete set of referring modes). Then, the OZONE implementation did not reflect the algorithm from section 5. Since the complete algorithm is not implemented in COVEN, MIAMM and OZONE, we consider these design experiences as partial validations, and we complete them with another evaluation. What we can say at this stage is that, with several implementations within several applicative domains, we showed that our model can face to a lot of interaction aspects, i.e., kinds of objects, interaction paradigms and reference behaviors.

A second evaluation method is the classical corpus study. Using the multimodal corpus of Wolff *et al.* [32], we explore by hand how our algorithm works with a set of 98 multimodal referring actions. This set is the same than the one already described and exploited in [19]. We use it because multimodal corpora are very rare, and because this corpus has the advantage to group visual, gestural and acoustic signals that are partially transcribed. More precisely, we transcribed the visual scenes into descriptions involving perceptual groups (and then potential visual reference domains). Concerning the transcription of gesture trajectories, we attribute for each potential referent a numeric score that corresponds to its probability of being pointed out (see [22]). Concerning the linguistic utterances, we transcribed them into texts, and when it was relevant we tried to formalize the content of the dialogue history using linguistic reference domains. In fact only few referring actions exploited the dialogue history, i.e., there were few anaphora in the corpus. Most referring actions were direct multimodal accesses to the referents ('ind-par-ref' or 'ext-par-ref' referring modes). The most complex ones are at the origin of Figures 4 to 7.

More precisely about the corpus, it was collected during a Wizard of Oz simulation directed by Wolff *et al.* [32]. In this technique often used in human-machine dialogue studies, a human (the wizard) plays the role of the computer behind the interface in order to test the efficiency of the planned capacities of a dialogue system before its implementation. Seven students from the University of Nancy participated in the simulation experiment as volunteers. They were French native speakers. Engaging a dialogue in French with the simulated system, they were required to move objects and groups of objects into appropriate boxes. The interaction was based on speech and gesture, mediated by a microphone and an electronic pen for the touch screen. The experimental instructions provided to participants were only related to the task and not to the mode of interaction. In order to assure the spontaneous character of the interaction, users were free to use speech and gesture as they wished. This is important because the goal was to collect the largest possible variety of multimodal referring expressions. To inhibit the only use of unimodal verbal references, the objects to be moved into the boxes were abstract-shaped figures, i.e., having no linguistic term associated with them. No triangles nor circles, but shapes that incited the participants to produce expressions like "this object", "these two figures", "that big form", etc. From the side of the machine, the wizard has the role to determine if a user's utterance is susceptible to be correctly treated by a dialogue system or not. For that, he quickly estimates the quality of the speech delivery, the lexicon, syntax and semantics of the utterance, as well as the possibility to resolve the referring actions. Recognition or syntactic errors are then not really simulated, but unexpected behavior from the user is not answered to.

The result of our corpus study is very simple: all multimodal referring actions are well interpreted when using our algorithm based on reference domains. For most of them, a simple classical algorithm may be sufficient. But for the most complex situations, reasoning with reference domains and referring modes appears as essential for a correct understanding. In some cases, and in particular for the corpus situation that is at the origin of Figure 7, such a reasoning is an efficient way to make an hypothesis on the user's behavior and on its referring intention. To us, such concerns are essential for the design of more comprehensive dialogue systems.

7 Conclusion

Reference to objects in multimodal dialogue systems can take several forms which are not linked to particular mechanisms of identification. The choice of a determiner, of the singular or plural form, of a co-referent pointing gesture, lead to clues that specify some aspects of the interpretation process. In this paper we investigate multimodal human-computer interaction involving visible objects, and we propose 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. We explore the multiple possibilities of referring modes. We show that many ambiguities can occur. We propose a list of disambiguation principles based on the notion of reference domain and of the concrete examples we found in the corpus of Wolff *et al.* [32] and in linguistic classical works like [24], [26] or [29]. The examples we investigate illustrate a number of reference possibilities in terms of anaphor, transition from specific to generic interpretation, associations of referential terms and pointing gestures, etc.

As it is showed with the implementation of reference domains in several multimodal dialogue systems and with a corpus study, our model appears to be relevant for different kinds of interaction modalities and for different kinds of applications. One problem, given our focusing on complex phenomena (for example when the pointed objects are not exactly the referents), is the lack of multimodal corpora suitable for a systematic evaluation. Nevertheless, the phenomena can easily be found in human-human communication, and we need algorithms for a system to understand these phenomena, even if for the moment their evaluation is difficult. As we showed with the complexity of some examples, simple algorithms for multimodal understanding are not sufficient. But algorithms such ours that manage structures like reference domains are useful.

References

[1] R.-J. Beun and A.H.M. Cremers, Object Reference in a Shared Domain of Conversation, *Pragmatics and Cognition* **6(1/2)** (1998) 121–152.

[2] R.A. Bolt, Put-That-There: Voice and Gesture at the Graphics Interface, *Computer Graphics* **14(3)** (1980) 262–270.

[3] F. Corblin, *Indéfini, défini et démonstratif* (Droz, Genève, 1987).

[4] F. Corblin, Mentional References and Familiarity Break, in: *Hommages à Liliane Tasmowski-De Ryck* (Unipress, Padoue, 1999).

[5] J. Cosnier and J. Vaysse, Sémiotique des gestes communicatifs, *Nouveaux actes sémiotiques (geste, cognition et communication)* **52** (1997) 7–28.

[6] P. Dekker, Speaker's Reference, Descriptions and Information Structure, *Journal of Semantics* **15(4)** (1998) 305–334.

[7] P.G. Edmonds, A Computational Model of Collaboration on Reference in Direction-Giving Dialogues (Ms. Thesis, University of Toronto, Canada, 1993).

[8] D. Efron, *Gesture, Race and Culture* (Mouton, The Hague, 1972).

[9] M. Fitts, The Information Capacity of the Human Motor System in Controlling Amplitude of Movement, *Journal of Experimental Psychology* **47** (1954) 381–391.

[10] H.P. Grice, Logic and Conversation, in: P. Cole and J. Morgan, eds., *Syntax and Semantics (vol. 3)* (Academic Press, 1975) 41–58.

[11] B.J. Grosz and C.L. Sidner, Attention, Intentions and the Structure of Discourse, *Computational Linguistics* **12(3)** (1986) 175–204.

[12] J. Itten, *The Art of Colour* (Reinhold Publishing Corp., New York, 1961).

[13] A. Jöhsson and N. Dählback, Talking to a Computer is not like Talking to your Best Friend, in: *Proceedings of the Scandinavian Conference on Artificial Intelligence* (Tromsø, 1988).

[14] H. Kamp and U. Reyle, *From Discourse to Logic* (Kluwer, Dordrecht, 1993).

[15] A. Karmiloff-Smith, *A Functional Approach to Child Language* (Cambridge University Press, 1979).

[16] A. Kehler, Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction, in: *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)* (Austin, 2000).

[17] L. Kievit, P. Piwek, R.-J. Beun and H. Bunt, Multimodal Cooperative Resolution of Referential Expressions in the DenK System, in: H. Bunt and R.-J. Beun, eds., *Cooperative Multimodal Communication* (Springer, Berlin & Heidelberg, 2001) 197–214.

[18] G. Kleiber, *Anaphores et pronoms* (Duculot, Louvain-la-Neuve, 1994).

[19] F. Landragin, A. De Angeli, F. Wolff, P. Lopez and L. Romary, Relevance and Perceptual Constraints in Multimodal Referring Actions, in: K. van Deemter and R. Kibble, eds., *Information Sharing: Reference and Presupposition in Language Generation and Interpretation* (CSLI Publications, Stanford, 2002) 395–413.

[20] F. Landragin, N. Bellalem and L. Romary, Referring to Objects with Spoken and Haptic Modalities, in: *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)* (Pittsburgh, 2002) 99–104.

[21] F. Landragin and L. Romary, Dialogue History Modelling for Multimodal Human-Computer Interaction, in: *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)* (Barcelona, Spain, 2004) 41–48.

[22] F. Landragin, *Dialogue homme-machine multimodal* (Hermès, Paris, 2004).

[23] D. McNeill, *Psycholinguistics: A New Approach* (Harper and Row, New York, 1987).

[24] J. Moeschler and A. Reboul, *Dictionnaire encyclopédique de pragmatique* (Seuil, Paris, 1994).

[25] S.L. Oviatt, Ten Myths of Multimodal Interaction, *Communications of the ACM* **42** (1999) 74–81.

[26] A. Reboul and J. Moeschler, *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours* (Armand Colin, Paris, 1998).

[27] I. Rock and L. Brosgole, Grouping Based on Phenomenal Proximity, *Journal of Experimental Psychology* **67** (1964).

[28] S. Salmon-Alt, Reference Resolution within the Framework of Cognitive Grammar, in: *Proceedings of the International Colloquium on Cognitive Science* (San Sebastian, Spain, 2001).

[29] D. Sperber and D. Wilson, *Relevance. Communication and Cognition (2nd ed.)* (Blackwell, Oxford UK & Cambridge USA, 1995).

[30] K.R. Thórisson, Simulated Perceptual Grouping: An Application to Human-Computer Interaction, in *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (Atlanta, Georgia, 1994).

[31] M. Wertheimer, Untersuchungen zur Lehre von der Gestalt II, *Psychologische Forschung* **4** (1923).

[32] F. Wolff, A. De Angeli and L. Romary, 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).