Refering in a Multimodal Environment : from NL to designation.

Bertrand Gaiffe, Jean-Marie Pierrel, Laurent Romary

Refering in a Multimodal Environment : from NL to designation.

Bertrand Gaiffe - Jean-Marie Pierrel - Laurent Romary

CRIN-CNRS&INRIA, Centre de Recherche en Informatique de Nancy
B.P.239; 54506 Vandœuvre lès Nancy
{gaiffe,romary,jmp}@loria.fr

1. INTRODUCTION

In this article, we intend to put forward several problems related to the design of man-machine multimodal dialogues. We will not take into account every man-machine multimodal dialogue but only those in which the user is in front of a computer system in order to perform a given task, as opposed to - for example - a dialogue dedicated to the interrogation of a database [Carbonell 89]. In such a dialogue, the user handles the objects of the task through an interface which allows him to use all the commands provided by the task. The structure of each interaction may thus be sketched as applying a given predicate to some objects according to some possible parameters (those required by the computer implementation of the commands for instance). Finding the predicate and its associated parameters are of course important in such a framework, but the fact of giving a natural aspect to a dialogue depends a great deal on the way the user is allowed to make references to objects.

Increasing the user's opportunities to refer to objects may be achieved by giving him several modes of communication. This is why we consider multimodal dialogues : the user can make a reference by means of a vocal message or by means of gestures. It is clear that in some cases, a gesture gives a way to avoid long circumlocutions such as "the window at the right of the green icon", an utterance which may be replaced by a mere pointing.

If we consider a multimodal interaction as the sole adjunction of modes with no relations between them we may find ourselves unsatisfied, though. As a matter of fact, it would mean that each reference has to be done by means of only one mode at a time, while the others would be despisedset aside, even if some actual event occurs on them. For example, a multimodal utterance such as "move this window" *together with* a gesture would be forbidden.

Taking such an option, which we might call asynchrounous multimodality (i.e. a multimodality excluding any referential collaboration between modes), would be far

from giving a natural flavor to our dialogues, and this is contrary to what we actually want to be realized here.

This is the reason why we will view multimodality as the need for integrating the referential models of the different modes. This implies that we have to study the ways by which each mode takes its referents. We will first show that treating natural langage designations urges us to manage a historical memory of the discourse referents which is exclusively dedicated to that mode. Following this, we will show that the same kind of treatment is mandatory for the designation mode, leading us to conceive another historical memory. Finally, we will show how these two historical memories have to collaborate in order to achieve a proper treatment of actual multimodal designations such as "this window" together with a gesture.

2. SETTING UP THE REFERENCE PROBLEM

As mentioned in the introduction, we design dialogue systems within the exclusive context of a specific task. This hypothesis implies two types of constraints on our work : on the first hand, references may only be solved among the set of objects belonging to the task universe. Their types, and thus their characteristics, are determined according to the specific relations yielded by the task and, at the same time, depending on the possible actions which may take place in this context. On the second hand, the utterances that the user may express do not cover the whole range of possible NL sentences, since they are essentially based on the following pattern : [action to be performed ; objects operated on].

This pattern underlines the fact that, at the utterance level, predicative and referential elements may be separated in order to implement a specific process dedicated to each aspect. Still, we should not forget that each time an action is performed in the task universe, it induces some changes within the properties and states of the different objects (i.e. referents) in the universe. As a result, the final interpretation should provide us with a combination of both aspects and not with a simple pair of partial referential and predicative instructions. Besides, we will observe that dealing with referents compels us - at least in a close future - to design a specific temporal representation, such as that we would consider when dealing with predicates [Romary 91]. However, if we actually wished to treat the problem of action within the frame of this paper, this would lead us too far away from referents and thus from our main point here.

In spite of this limitation, the problem still remains complex, as the following examples show. Let us suppose several possible sequences of utterances as would be expressed by the user of a dialogue system[1] :

(1) *"Move the green window"*
"Put this window in the background" (without any gesture)

(2) *"Put this window there" (with two gestures)*

(3) *"Move the green window and the blue icon"*
"Put the window in the background"

Through these examples, we observe something that we will call co-references : in the first case as well as in the third case, we have co-references between two NL expressions : "the green window" and "this window" refer to the same object and so do "the green window" and "the window". This type of co-reference involving a previously mentionned object is usually called anaphora[2]. In example 2, we have another type of co-reference, namely a deixis, involving two different modes : NL and designation.

Of course, co-references between the communication modes are specific of multimodal dialogues and should be taken into account with the greatest attention. However, their resolution should not be achieved in a way which would make a correct treatment of anaphoras impossible. The problem is real, since there are linguistic forms admitting the two types of co-references, typically : demonstrative noun phrases. We have these two cases represented in example 1 which correspond to an anaphora as compared to example 2 in which "this window" has to be associated with a designation gesture.

Besides, one should note that the references expressed by each mode may be ambiguous whereas the co-reference resulting from the combination of the modes is not. For example : "this window", associated with a gesture, if analysed from the NL mode only, is ambiguous each time there is more than one window on the screenand

---

[1] All our examples will be about windows, icons and other graphical objects, our aim being to describe a generic task (in a windowing environment) rather then focusing on the constraints induced by a more specific one.

[2] We are conscious of the fact that such a contreversial notion, as regard its actual definition, should be handled with care. Still, we ask the reader to accept the way we will use it from now onwards (cf Reinhart 76; Partee 84; Hinrichs 86; Webber 88; Kleiber 90; Reboul 89).

similarly, the designation gesture may be also ambiguous. For instance, if the gesture points at a character belonging to a text, the objects which may be considered as potential referents are : the character, the word, the line, the paragraph, the whole text or the window. But, the resulting co-reference between the modes is not, in that example, ambiguous at all : there is only one window pointed to by the gesture.

3. CO-REFERENCES VIEWED FROM NL :

A common analysis considers the two kinds of co-references (NL+NL, and NL+gesture) as different and opposable. On the one hand, we find studies pointing at the specific syntactic and discursive articulation of NL-NL co-references (Reinhart 76; Sidner 83), and on the other hand, specific computer architectures are designed for bi-modal references without contemplating a real treatment of anaphoras (Caelen 91). We prefer here to analyse all the co-references involving NL in the same way by focusing on the specific constraints expressed by this rather powerful mode since we showed that there are cases in which we cannot decide a priori between the two types of co-references. This analysis will lead us, in a second section, to an organisation of the referents based on a discourse representation, that is, dedicated to the treatment of anaphoras, but aiming at a more general purpose as far as multimodal references are concerned.

3.1 Pronouns, demonstratives and definite articles :

There has been a lot of research done about the automatic treatment of anaphoras. Most people - Sidner (86) for instance - proposed to consider definite anaphoras (definite pronouns such as "he" or "it" and definite noun phrases such as "the N") in the same way. Such an analysis comes from the fact that most studies have been made in the context of mono informational discourses, that is, discourses in which the only pertinent objects are those mentionned in the preceding utterances. In such discourses, a new object is usually introduced by an indefinite noun phrase (for instance, "I want to organize a meeting") and is refered to afterwards by means of a pronoun, a definite noun phrase or a demonstrative one. This implies that almost all definite or demonstrative noun phrases are anaphoric.

In our context of multimodal command dialogues, the available objects for a definite designation are not limited to those that have been mentionned in previous utterances. Definite descriptions may then be used to refer to objects belonging to an extra-textual reality (objects seen by the two partners, etc...). In particular, definite noun phrases are a very common way to refer to an object represented on the screen of the

computer (for instance : "move the green text window"). We then have an ambiguity between anaphoric definite noun phrases and non anaphoric ones. Similarly, demonstrative noun phrases, when admitting designation gestures, become also ambiguous : they may be anaphoric as in "transform the text window into an icon. Move this icon to the right part of the screen", or they may be co-referential with a gesture ("move this window").

All these possible ambiguities oblige us to analyse separately the three kinds of referential noun phrases we have observed in our analysis of the natural langage mode. We will thus discuss the way pronouns, definite noun phrases and demonstratives access to their referents depending on the actual context given by the previous utterances, the state of the task and the possible gestures of the user.

Pronouns

Pronouns illustrate the prototypic case of an anaphora. This is exemplified by the following sequence of utterances :

(4) *"Move the green window"*
"Put it in the background"

A first observation is that, in its normal use, a pronoun should not lead to a co-reference with a gesture. There may be examples such as : "he is fool" [Kleiber 89] said by someone seing a car charging straight at him, but we can affirm that such examples may be excluded in the applications based on task oriented dialogues such as those we consider here. We will thus consider only anaphoric pronouns.

In such cases, it is often heard that the pronouns works as if the antecedent would replace it in the same linguistic situation. This rewriting approach would transform our example (4 bis) into :

(4 bis) *"Move the green window"*
"Put the green window in the background"

We do not care about whether this rewriting operation should be a syntactic one, that is to say, purely a matter of word or whether it is more a semantic process. The difference would be between rewriting "it" as "the green window" or as (green(?x) and window(?x)). In all cases, "the green window" will be analysed after the substitution as something that looks like (green(?x) $\wedge$ window(?x)).

There are many arguments against this rewriting approach, especially if we take an ambiguous antecedent. Suppose that we have two windows on the screen, and for an unknown reason, the user says :

(5) *"Move the window"*
"Put it in the background"

A rewriting approach could associate the definite noun phrase "the window" to a first window, and the pronoun, after its substitution to the other window. Of course, we know it would be a wrong interpretation[3], proving thus that the pronoun is in that case co-referential. There are examples however for which pronouns are not co-referential. We will give two such examples here. They are actually interesting, although they are not likely to appear in our dialogues but still they will confirm, in a more general frame, our analysis of pronouns.

Let us consider the two following examples :

(6) *"I bought a Toyota because they are reliable and cheap"*

(7) *"Do not buy him this book, he already has it"*

In example (6), the pronoun "they" do not refer to the specific Toyota which is its antecedent and this may leads us to two remarks concerning this example. First, if the antecedent is a definite one, "the last Toyota" for instance, the mechanism put forward above does not work anymore. Instead, we would have - in this situation - a co-reference associating a generic reference to another one all the same generic by means of the pronoun "it". This phenomenon can be explained by the fact that indefinite NPs are very close to generic ones since the only properties given to a referent by means of an expression such as "a Toyota" are those born by a generic Toyota.

The second remark is that, if we produce, in the discursive context, an antecedent allowing a co-reference with "they", it will be the one which will actually be considered as a correct antecedent for the pronoun, as in :

(8) I bought a Toyota to my parents because they ..."

[3]We are not concerned by the way a human-being actually solves pronouns. We are just considering that the result of this interpretation leads to a co-reference.

The conclusion about this example is that the pronoun is co-referent with its antecedent, each time it has the opportunity to do so. If this is not the case, there may be a shift in a generic interpretation.

The second example :

*(7) "Do not buy him this book, he already has **it**"*

is a very interesting one too. Such linguists as A.Reboul [Reboul 88-89, Reboul 89] argue that "this book" is a designation for a first book (the one seen by the speaker), and "it" refers to another book (the one that "he" owns). The other possible analysis that we propose is that because the pronoun is co-referential, the only solution is that "the book" in that example means : the intellectual-production-from-an-author. That possible meaning of "book" seems here the correct interpretation and allows the co-reference we need.

Demonstratives

Demonstrative noun phrases are a real problem in multi-modal dialogues since they are likely to be co-referential with gestures, but they may also be anaphoric[4]. A way to solve this problem, as a first step towards a real modelization of the associated phenomenon, would be to see whether we have a designation or not in the same time period as that in which the oral noun phrase occured.

Unfortunately, when given a dataglove for instance, we may not be sure that there has been a designation gesture. From the speech recognition point of view too, it is not always simple to know where a noun phrase begins and where it ends. And, last problem, and this is the only one which has a theoretical importance, we do not know, as very few studies have been made in real multimodal situations, wether a user synchronizes exactly his gestures and his oral designations.

These arguments do not reject a temporal analysis of co-references ivolving two different modes. It simply means that it may be interesting to find out other constraints which may be added to the temporal ones. We will thus study the way demonstrative noun phrases access to a referent, so that, coming along with this analysis, we will know how a gesture may be associated with such expressions.

---

[4]Especially in french for which definite noun phrases do not have the deicticity its english counterpart provides.

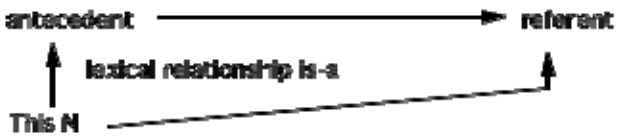
Regarding this problem, it has been observed by linguists such as G. Kleiber (90) that in the case of demonstrative anaphoras, the relation usually called "is_a" has to hold between the antecedent and the anaphoric noun phrase. This can be seen in the following example :

"I saw a car. This vehicle" is authorised, but

*"I saw a vehicle. This car" is not.

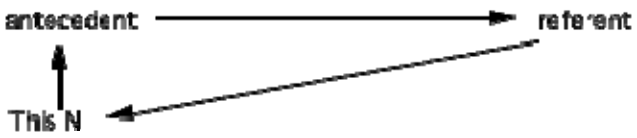
The actual reason for such a behaviour of demonstrative NPs is that a car is a vehicle, but a vehicle is not necessarily a car.

There are two ways of explaining this phenomenon. The first point of view relies on a lexical approach, which says that when uttering "this N2" as an anaphoric expression on "a N1", the N1s have to be N2s from a lexical point of view, which yields, with the terms of our example : cars are vehicles. We could represent this as :



The second point of view is to consider that the is-a relationship is very close to the refering process, since, whenever we refer to something by means of "the car", the associated concept (or referent, or whatever it is) *is-a* car, that is, has all the properties which allows the speaker to call it a car.

When hearing "a car", one builds a referent having the property of being a car, together with possible other properties, as - for example - the fact that it is of a specific type or colour. Following this referential creation, the corresponding car accepts to be called a vehicle. This would lead to the following schema :



These two approaches seem very close to each other. However, the difference between them is obvious if we consider multimodal utterances. As a matter of fact, the computation of the reference associated to a pointing gesture will work in the case of the second hypothesis just because the demonstrative NP will access directly to a referent, whereas the lexical approach would compel us to transcribe a pseudo-linguistic form associated to the gesture. In other words, a lexical approach would oblige to give

an arbitrary name to the designated object before treating the co-reference. Obviously, giving an arbitrary name to an object is dangerous since you may not always find the proper level to which you must ascribe the object. Hence, a car could be qualified as a vehicle, whereas it could no more be referred to by means of "this car" (since we know that a demonstrative NP barely allows a re-classification of a referent).

Definite noun phrases

What we have proved until now is that pronouns and demonstrative noun phrases should be treated in a co-referential way. Definite noun phrases, however, do not work in the same way and there is much linguistic evidence in favour of this opinion. But instead of developping those rather classical arguments, we prefer to tackle here a more pragmatic reason. As we saw, a noun phrase such as "the N" may directly refer to an object in the task (any object in the universe corresponding to the given description), or may otherwise be anaphoric (any object mentionned in the discourse and having the gioven properties). We will first show that in some cases, the ambiguity between an anaphoric interpretation or a non anaphoric one is real : there are cases for which none of them fails to find a referent. This means that we cannot use a hypothesis-test strategy which would successively suppose that "the N" is anaphoric and then that it is not, in case of a failure in the test. As the two hypotheses have to be made in parallel, we will propose a common analysis for all definite noun phrases.

The ambiguities :

Apart from the differet ambiguities that we have mentionned so far concerning the definite NPs, there are other interesting cases that we may present here. Let us compare the following utterances :

(8) *"Move the leftmost window"*

(9) *"Move the green windows"*
"Put the leftmost window in the background"

(10) *"Move the green windows and the blue icons"*
"Put the leftmost window in the background"

(11) *"Move the green windows"*
"Put the leftmost (one) in the background"

The utterance in example 8 is non anaphoric (if we consider it as a first utterance) : the referent is the leftmost among all the windows on the screen. Besides, we think that utterance 9 is of the same kind. In example 10 however, the window

which is to be moved is the leftmost among green windows and so does it work in example 11. As a result, with very similar forms, some examples such as 8 or 9 are non anaphoric, whereas we may consider that examples 10 and 11 are so.

It is very clear that in example 9, a non anaphoric interpretation would not fail. It would correspond to a search for an ?x such that (window(?x) $\wedge$ leftmost(?x)). Neither would an anaphoric interpretation for which the corresponding constraint would be : (window(?x) $\wedge$ leftmost(?x) $\wedge$ green(?x)). The problem is precisely here. Since we can neither make the hypothesis that the definite NP under consideration is anaphoric or not, we observe that we need once again a common analysis for both type of expressions.

The analysis of "the N" :

In the situation mentionned above, the only solution is to consider "the N" as a way to select in a set of elements[5]. If we consider non-anaphoric designations, there is nothing shocking in the fact that "the window" is a way to select in the environment an object with has the property : "window(?x)". Similarly, we see that definite anaphoras are of the same kind. As a matter of fact, they realize a selection among the objects mentioned in the discourse so as to produce the element referred to by the expression. At that point of the demonstration, two problems still have to be solved :

- what criterion must a given set verify in oder to be a proper candidate to the computation of a definite description?

- what are the sets usable in general, since it is very clear that an abitrary set of elements among the elements on the screen is not usable as a basis for the computing of a definite description?

We will discuss the second point in the next paragraph. For the time being, let us look at the "good properties" of a given set for a given definite description.

The first two properties are quite obvious : when one says "the N", there must be a "N" in the set, and there should not be more than one. With these two properties however, we can not always decide between an anaphoric or a non anaphoric interpretation as shown in examples 8 and 9. It is then necessary to state a third property

[5]We will not consider here the problem of associative anaphoras (Kleiber 90), since they are barely used in the type of dialogues we have to deal with. The analysis of "the N" might however be extended to treat them as well.

: there should also be at least one element in the set which is not an "N". If we look again at example 9, we understand all the importance of this constraint :

(9) *"Move the green windows"*
"Put the leftmost window in the background"

It is very clear that all "green windows" are "windows". It would thus not be pertinent for the user to give this property if it was of no immediate use, and this is the very reason why we think that example 9 is not anaphoric.

Of course, we do not say that, given these three properties, a system has to use them in order to accept or reject an utterance. There are actually cases for which redundancy appears in a user's utterance and a dialogue system should by all means try to understand what it expresses, instead of being too normative. However, our point of view is the following : when the system is in front of an ambiguity, it should suppose that the user makes pertinent descriptions. This seems fair an hypothesis, and at least, fairer than supposing that the users systematically expresses himself by means of non pertinent descriptions.

3.2 A word about C-command

At that point, it might be surprising that we did not even mention what c-command, in the more general framework of the government and binding theory [Reinhart 76], has to say concerning the treatment of anaphora. This theory was developed to deal with pronouns and divided them into two classes : reflexives and others. As a matter of fact, we do not encounter reflexives (such as himself, each other ...) in command dialogues and this could be a first (still non-scientific) reason for us not to focus on this therory. The second reason is that government and binding, in the case of non reflexive pronouns, only gives impossibilities of co-referentiation. Even if it may seem interesting to get a synthesis of those impossibilities, there are not of the utmost importance in a context in which utterances are based on a predicate-arguments pattern. This means that we usually observe no subordinated propositions - the more complicated expressions being coordinated propositions. In this syntactic context, based essentially on inter-sentencial relations, government and binding theory does not help us much and the so called binding domain of the pronoun is in our case the whole utterance.

3.3 Discourse referents

In the preceding paragraphs, we only discussed the way pronouns, demonstrative and definite noun phrases take their referents. To this aim, we used, in the case of definite noun phrases, sets of elements. The problem is now to determine which objects are accessible for co-references in the case of pronouns and demonstratives, and which are for solving definite noun phrases.

Our two main information sources for this will be the syntax in some cases, and the structure of the task in some others.

Syntactic elements :

We said that demonstratives and pronouns imply a co-reference with an object refered to by something else. There are cases however in which a pronoun or a demonstrative noun phrase cannot refer to an already mentionned object. For instance after :

(12) "Move the green window and the blue icon"
we cannot have :
?"Put this window in the background"
or
?"Put it in the background"
but we might have :
"Put these objects in the background"
or
"Put them in the background"

This proves, if we consider what we said about pronouns and demonstrative noun phrases, that "the green window and the blue icon" represents an accessible object, whereas the two parts of this object, refered to by "the green window" and "the blue icon", are not accessible objects.

Of course the fact of not being an object has to be understood here as : not being an object as far as the discourse is concerned. The reason for building a unique object out of two sub-objects is purely a syntactic one : the two coordinated elements correspond to the same argument of the predicate and they thus constitute the direct object of 'move'.

Conversely, the object associated with "the green window and the blue icon" may be seen as a set and thus allows such a definite anaphora as :

(13) *"Move the green window and the blue icon"*
"Put the window in the background"

Still, considering what has been said about demonstrative NPs - that are co-referential to an accessible object - it might be argued that in the following example (14), the two sub-objects are referred to by means of demonstratives. However, in this case, the object as a whole is reconstructed by means of a coordination and thus corresponds to a single entity.

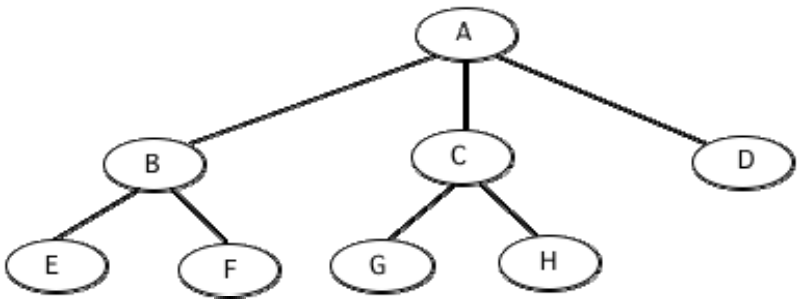
(14) *"Move the green window and the blue icon"*
"Put this window and this icon in the background"

Consequently, there are syntactic clues (simple to detect) which play an important part in the building of discourse referents. But of course, some sets of referents are not that easy to find, in particular, we should not forget that the task is important too as we will see in the end of this section.

Finding the good sets of objects is very important. As we saw in example 11, it allows the refusal of the hypothesis that a demonstrative NP is anaphoric. In such a case the only way to explain what the demonstrative expressed is to suppose that a co-designation with a gesture occured, even if the gesture itself was not very clear (the problem is real with datagloves for instance).

Task as structure :

Grosz (81) observed that a hierarchical task gives very strong contraints on the resolution of anaphoras. The idea was the following, suppose a task described as :



To perform task A, subtasks B, C and D have to be performed in that order and so on for each sub task.

Grosz gave evidence upon the fact that when performing subtask G for instance, the speaker does not make anaphoras on objects defined exclusively in E or F, subtasks of B.

It is clear that the associated notion of focus, which is a psychological one, is very difficult to use in an automatic system. However, in the case of a specific task, this

notion of focus can be retrieved in the way a user concentrate the topic of the dialogue on specific sub-tasks. We thus have a possibility to approximate the more general notion of focus.

We do not intend to discuss much more here the theory of Grosz. It is useful as soon as the task is strongly hierarchical (the task chosen by Grosz was the maintenance of compressors). In our command dialogues however, the task is much less hierarchical and most of all, the structure of the task is not given a priori. We may view each command offered in the system as an elementary piece structured in such a way that the user performs his own task by means of these commands. Suppose for instance that the user wants to put an image into a window containing a text. This is his plan or subtask. He may do it by opening the image in a first window, de-iconify his text window, cut the image and paste it in the text. All the objects mentioned here are likely to appear in anaphoras : the image window, the image itself, the text, the icon, the resulting text with an image and so on. After having finished this subtask, the only objects pertinent for a new subtask are those appearing in anaphoras (typically here, the result).

In our own perspective, we view the task as a means to build some of the sets that we need to interpret definite NPs. For each level of a task such as that we have just defined above, we can consider that the associated predicates build each the set of their arguments; those set being accessible depending on the hierarchy formed by the different sub-tasks. More specifically, when two subtasks are on the same level, the more recent one necessarily masks the older ones, whereas they all keep accessible when they part in a domination relation - i.e. task/sub-task - as defined by Grosz & Sidner (86).

4. CO-REFERENCES VIEWED FROM THE DESIGNATION MODE

4.1 The main characteristics of a designation using a mouse

The designation mode is usually left aside, since it is considered to be so simple that no specific mechanism is needed to analyse its meaning. However, if we only consider what can be expressed by a sole mouse on a windowing environment, we observe that the expressive power of this mode can be close to some parts of natural language. Actually, it is possible to define a specific syntax on the mouse/button events describing their structure along the time axis. For such a syntax, the terminals can be either mouse-clicks (up and down) or elementary trajectories such as linear ones, circular ones or zig-zags for example. As such a syntax is given, it is possible to associate specific meanings to some of the sequences given by the user on the screen.

These meanings are constraints which are used by the dialogue system to filter the different objects described in the universe model or predicates to be applied on those objects. Usually, it is more natural to think that a designation mode will express references rather than predicates. However, in the case of a drag for example (button-down + linear-motion + button-up), we can interpret a sequence of mouse events as expressing both a predicate and some elements on which it can operate (e.g. moving the object). We will mainly focus on the problem of reference since it exemplifies plainly how much a designation should be treated as a mode equivalent to more usual ones such as Natural Language.

4.2 Complex references.

Objects and types - The different objects appearing in the screen to the user cannot be only described by single variables together with their size and position. They may or may not take part in different events (e.g. they may be moved, closed, one may change their color etc…). This means that they shoud be typed objects, eventually related by 'is_a' relation. For example, the general type *framed_object* will cover the range of *windows*, *icons* and all the objects which can be moved directly on the screen. Another type of objects will be *textual_object* which cover the range of paragraph, *word* or *character*. Types are also useful to define the standart relations that can or must be established between two objects. For example, *textual_object*'s must necessarily be included in a *framed_object*, two *framed_object*'s can be related by 'in', 'by' or 'on' relations etc…

Unlike types defined in object oriented languages which are implementation bound types, the types thus defined are related to the behaviour of the objects within the representation space and they may be understood by the user. They must not be brought together with the X-windows data structure for example.

Objects and relations - We have already hinted at relations that may occur between objects appearing on the screen. These relations express the topological link between graphical objects as perceived by the user. Thus they are not necessarily linked to any relation induced by a specific implementation (such as the X window interface for example). The semantics of these relations can be given by means of the way they interact with each other. As it has been defined in [Romary 90], we may introduce two kinds of combination rules which are transitions rules enabling the system to infer new relations from existing ones and compatibility rules which express the fact that some relations cannot appear at the same time.

Focus and reference - Given a set of typed objects and relations between these objects, the reference operation should deal with both types and relations.

To shed further light on this aspect, we will essentially present two complementary examples of reference using the designation mode.

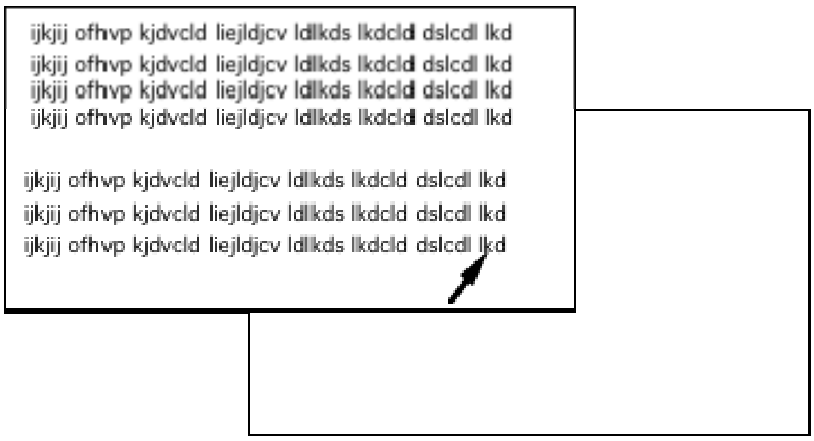


Figure 4.1 : a small window environment

Exemple 1, a simple click - figure 4.1 shows a screen where two windows overlap. In one of these windows there is a text made of paragraphs, words, characters ... We can first consider that the user has clicked on one of the characters. The structural information is rather simple. It may consist of a preparatory phase (motion to the location of the click) the click itself and a conclusion phase (letting the mouse leave the clicked point). Still, the semantics is already ambiguous. It can either be the designation of the locus corresponding to the position of the mouse at the time of the click (e.g. if associated with the utterance "Put the window there") or it can be the designation of any object in which the pointer was at the time of the click (e.g. if associated with the NL tterance "move this X"). If we limit ourselves to the second possibility, we can now see how the click can be interpreted at the referential level. At this level, its meaning is broadly ambiguous since it can refer to the character, the word, the paragraph or one of the windows.

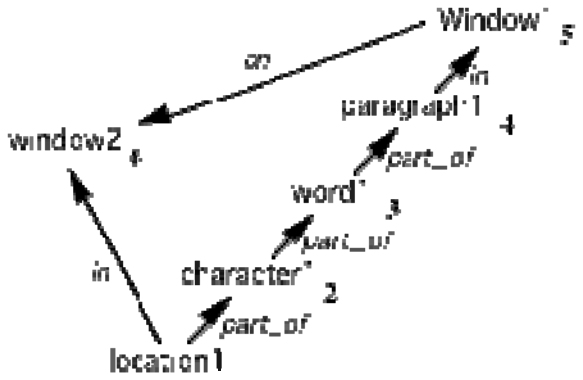


Figure 2.2 : Interpretation of a simple click.

Since these different objects are related to each other by differerent kind of spatial relations, the final result is the complex structure shown in figure 4.2 where the number indicates the order in which the reference candidates will be examined. Here the first element acts as a local focus from which any reference operation will start.

Exemple 2, a circular curve : Let us now suppose that instead of a click, the user has drawn a large circular curve around a paragraph at the top left of the foremost window. We can make a similar analysis concerning the syntax and the semantics of this designation, but its interpretation is far more interesting since it shows some new mechanism that must be taken into account. Figure 4.3 shows the result of the interpretation. Whereas we only had single objects refered to in the previous example, we see that we need here to deal with sets of objects (represented by a circle in the shema). These sets represent all the objects of the same type that can be gathered because they have equally been pointed out by the designation.

Between these sets occur the same kind of relations as those we have presented in example 1.

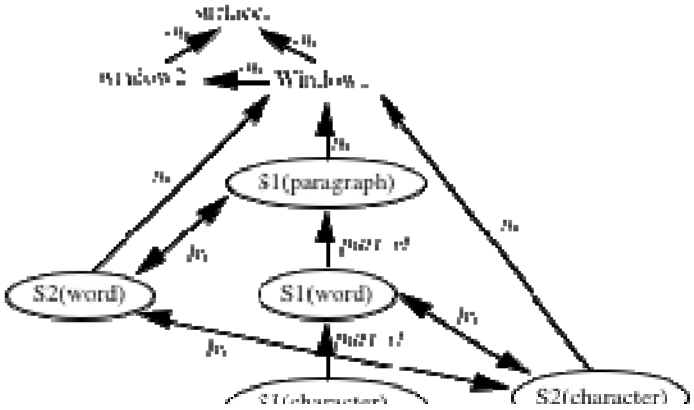


Figure 2.3 : Interpretation of a closed curve designation.

4.3 Using designation in a multimodal environment

The structure thus obtained is included in a local focus.

If a similar referential structure has already been generated from another mode (e.g. natural language) or is to be generated within a short period of time, the two have to be merged in order to form the final referential part of the multimodal utterance.

There may still be some ambiguities depending on the precision with which the user has pointed out the objects he wanted to refer to. For instance, if we take the result of example 2 together with an utterance like "this word", the type *word* generates a set of possible candidates along the interpretation structure. In the case of example one, there is only one candidate left.

5. THE NEED FOR AN HISTORICAL MEMORY

In the preceding sections we studied the ways each mode accesses referents. Still, there is a type of information that we did not take into account at all : co-references are not likely to be effective between elements which are too far away from each other. For instance, the user would not make an anaphora by means of a pronoun to an element he mentioned some ten utterances before.

To treat this kind of phenomenon, the best solution would be to modelize the Short Time Memory (STM) of human beings. Unfortunately, we know very few about the human STM. When psychologists study the STM, they observe that human beings are only able to memorize a few elements in it, but the number of elements we are able to memorize is not the same in the case of numbers, characters, words or sentences. Consequently, tthere is a notion of maximal complexity of the elements which take

place in the STM. Another problem is that elements disappear from the STM after some seconds, but the user may refresh them, and of course we can not modelize the choices he makes for refreshing an element and not another one. So the conclusion is very pessimistic at this time : if we want a computer implementation, we are condemned to a rough approximation of an STM, which we usually call a historical memory.

That historical memory has an arbitrary size in terms of utterances. For instance, we use a historical memory containing all the information extracted from the last 3 or 4 utterances.

Another problem with historical memories is that time is not taken into account. We said for STM that the pieces of information, if not refreshed, are lost. In historical memories, the only time considered lies at the utterance level : a new utterance enters the historical memory, and the oldest one is thrown out. We thus have a stack limited in size.

Two kind of focuses ?

When dealing with multimodal dialogues, the historical memory has to contain pieces of information coming from both modes because these have to collaborate in order to fully treat multimodal references. However, the modes cannot collaborate in any conditions : we know for instance that a co-reference with a gesture is not allowed on an element which has already been interpreted as anaphoric. This means that no double coreference - that is simultaneously - is allowed.

We proved in the preceding sections that each mode needed to collect sets of objects for its own purpose. It is clear however that we cannot mix, in a given set, elements designated by NL and elements designated by a gesture. We cannot give real examples of such things. They simply do not exist. They would look like : "move the green window" together with a pointing to something else than this very window, the intention being to move the green window and the object pointed at.

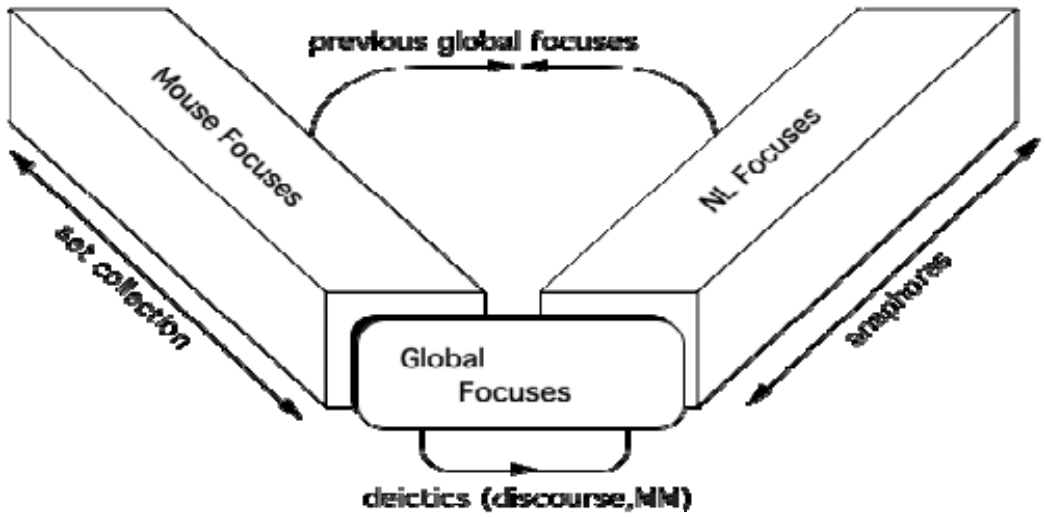
In the same way, it is not allowed, in a collection of objects designated by a gesture, to add an element referred to by means of the sole NL mode. As each mode manages specific structures, the historical memory has to be divided into two parts, one corresponding to each mode.

However, the modes have to communicate with each other if we intend at least to manage multimodal dialogues, and as a result, a part of the two historical memories has to be shared. This sharing is used to solve the current multimodal references (i.e. the

ones of the last utterance). This can be exemplified as follows : suppose that we have to understand the utterance "move this" together with the selection of a set of objects by means of a circling designation. In such a case, we have to treat a co-reference between the set coming from designation and the demonstrative "this", these two elements will be the most recent ones for each mode.

We explained that each mode needs its own historical memory. Some elements are shared, though. In our example, the set of objects belongs to the two historical memories. This double belonging cannot of course disappear after the utterance has been treated.

We thus obtain the following shema which summarizes the different constraints expressed so far concerning a historical memory in a multimodal dialogue system. However, this schema should not be misleading. Far from being the representation of a computer implementation, it represents the actual theoretical stance deduced from the study conducted until now. Each mode dedicated stack of focuses has thus its own structure based on a set representation, as needed for the reference calculus.



6. CONCLUSION

This paper addresses the problem of reference in multimodal dialogues. We propose an architecture based on two historical memories. Each of these historical memories reflects the structure that a mode builds upon the objects which it encountered in the application. This is the reason why we gave a particular attention to the ways each mode refers to objects.

There are open problems however. For instance we did not at all take into account the answers of the system. If the system's answer is not linguistic, for example

highlighting a window, it may affect the focus of the dialogue and thus authorize such an utterance as "this window".

Another issue we did not discuss is the temporal aspects of the reference. It corresponds to the well known problem of the referents that get modified. Of course, this problem is important in command dialogues which aim at working upon objects.

7. REFERENCES

Caelen J. & Coutaz J., "Interaction multimodale homme-machine : quelques problèmes généraux. actes du workshop IHM'91, Dourdan, déc. 91.

Carbonell N. et Pierrel J.M., 1989, "Vers un dialogue naturel homme-machine : apport des études sur les interfaces orales en langue naturelle", *Actes du colloque sur l'ingéniérie des interfaces homme-machine*,Cargèse.

Grosz B. and Sidner C., 1986, "Attention, Intentions and the structure of discourse", *Computational Linguistics*, 12, pp.175-204.

Grosz B., 1981, "Focusing and description in natural language dialogues", In Joshi A., Webber B. and Sag. I. (eds), *Elements of discourse understanding*, pp.48-105, Cambridge University Press.

Hinrichs, E., 1986, "Temporal anaphora in discourses of English", *Linguistics and Philosophy*, 9, pp.63-82.

Kamp H. & Rohrer C. (1983) "Tense in texts", in *Proceedings of the 1981 Linguistics Conference at Konstantz*, Germany, pp.250-269.

Kleiber G., 1981, *Problèmes de référence : descriptions définies et noms propres*, Klincksieck, Paris.

Kleiber G., 1989"Quand "il" n'a pas d'antécédent", revue langage numéro 97 : aux confins de la grammaire, l'anaphore. P. Cadiot et A. Zribi-Hertz (Eds).

Kleiber G., 1990, "Sur l'anaphore associative : article défini et adjectif démonstratif.", *Rivista di linguistica*, vol.2, n°1.

Morin P. et Pierrel J.M., "Partner : un système de dialogue oral homme-machine", *Actes du colloque Cognitiva 87*, Paris 18-22 mai 1987.

Partee B., 1984, "Nominal and temporal anaphora", *Linguistics and Philosophy*, 7, pp.243-286.

Pierrel J.M., 1990, "Vers une meilleure intégration de la parole dans des systèmes de communication homme-machine", *Traitement du signal*.

Reboul A., 1989 "Résolution de l'anaphore pronominale : sémantique ou pragmatique ?", *Cahiers de linguistique française* 10, 77-100.

Reboul A.,1988-89 "Pragmatique de l'anaphore pronominale" *Sigma* 12-13, 197-231.

Reichenbach H., 1947, *Elements of Symbolic Logic* , London, Macmillan.

Reinhart T. , 1976 "The Syntactic Domain of Anaphora", PhD Thesis, MIT.

Romary L. 1990, "Perception, language, raisonnement : une même représentation temporelle", Actes du troisième colloque de l'Arc, Paris, mars 1990.

Romary L., 1991, "Integration of spatial and temporal information produced by a natural language discourse", *in proc. Kmet 91*, Sophia-Antipolis, 22-24 avril 1991.

Roussanaly A., *DIAL : la composante dialogue d'un système de communication orale homme-machine finalisée en langage naturel*, Thèse de doctorat de l'université de Nancy I, 1988.

Sidner C. L., 1986 "Focusing in the comprehension of definite anaphora". in Readings in natural langage processing. Edité par B. Grosz, Jones K.S. , Webber B.L. Morgan Kauffman Publisher.

Sidner C., 1981, "Focusing for interpretation of pronouns", *American Journal of Computational Linguistics*, 4, pp.217-231.

Sidner C., 1983, "Focusing and discourse", *Discourse Processes, 6*, pp.105-142.

Webber B., 1988, "Tense as discourse anaphor", *Computational Linguistics*, 14, pp.61-73.