Generating Referring Expressions in Multimodal Contexts

Susanne Salmon-Alt
Loria – UMR 7503
Campus Scientifique, B.P.239
54506 Vandoeuvre-lès-Nancy, France
Susanne.Alt@loria.fr

Laurent Romary
Loria – UMR 7503
Campus Scientifique, B.P.239
54506 Vandoeuvre-lès-Nancy, France
Laurent.Romary@loria.fr

Abstract

This paper adresses the need of structuring the global context set into subsets or *domains* in order to explain adequately the use of referring expressions in a multimodal corpus. We underline, in particular, the importance of taking into account not only the discourse, but also perception and gestures for the construction of these domains. We propose a unified context model where the context is built up dynamically from different information sources and show that this way of context modelling predicts correctly the use of referring expressions in a corpus of instructional dialogues.

1 Need of a multimodal context model

This paper presents on-going work with the aim to design a tool for the interpretation and generation of referring expressions in multimodal instructional human-machine dialogues. We are considering a context model which has to integrate not only information conveyed by natural language, but also by the perceptual environment and by gestures. Evidence for this requirement can be found by studying corpora like the "Ozkan" corpus (Ozkan, 1994). Even within a very limited task universe – two persons, A and B, have to reconstruct together simple pictures (like pyramids in the desert…) composed of geometrical figures (triangles, lines,…) – not only do we find all types of referring expressions, but also many examples of interaction between perception, language and gestures:

A first case frequently observed concerns perceptual antecedents for referring expressions. The objects referred to are visible on the screen or manipulated before, but not mentioned before in the discourse. Exemple (1) shows one of theses cases: *en (them)* in B2 refers to the three triangles on the screen (Fig. 1), but only the first two were introduced in the discourse (A1, A3), whereas the third one was introduced by the gesture in A5.

(1) A1 pyramides dans le désert … alors une grosse pyramide
pyramids in the desert… so, a big pyramid
A2 *[gesture : A takes a first big triangle and put it on the screen]*
A3 alors tu mets un grand triangle sur la droite de l'autre et un petit peu au dessus
now you put a big triangle on the right of the other one and a little bit higher
B1 *[gesture : B takes a second big triangle and put it on the screen]*
A4 ouais
okay
A5 [gesture : A takes a third small triangle and put it on the screen]
B2 tu veux pas que j'<u>en</u> enlève un
do you want to delete one of <u>them</u>

A second observation leads to the hypothesis that perceptual groups function not only as antecedents for pronouns, but also as domains of quantifiying or interpretation for definite noun phrases. In example (2), *la*

pyramide de droite (*the pyramide on the right*) in A3 has to be interpreted within the perceptual context of the two big triangles, rather than within the group of the three visible triangles (Fig.1). Otherwise, it would designate the small triangle on the right, contrary to what the speaker intended, and indeed the hearer understood.

(2) A1 il faut prendre une grande horizontale et la placer à la pointe des deux grands triangles
take a big horizontal line and put it on the top of the two big triangles
A2 … et tu en prends une deuxième petite … et tu la places à gauche de la pyramide de gauche
… and take a second, small one… and put it on the left of the pyramid on the left
A3 voilà comme ça… et t'en prends une autre petite et tu la places à droite de <u>la pyramide de droite</u>
yes, like this … and take another small one and put it on the right <u>of the pyramid on the right</u>
A4 … et tu prends une autre petite horizontale et puis tu la places à droite de la petite pyramide
… and take another small horizontal line and put it on the right of the small pyramid

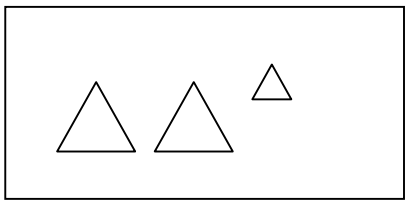

Fig. 1 – Visual context for (1) and (2)

A third group of examples shows that certain referring expressions cannot be interpreted without taking into account gestures. This is often the case for demonstrative pronouns, but also for still more interesting configurations like in example 3: here, the use of the context-dependent expression *l'autre* (*the other one*) in A2 has to be understood within a domain of two objects introduced in A1 and structured only by a gesture in B1.

(3) A1 alors il va falloir que tu prennes deux grandes barres
now, you have to take two big lines
B1 *[gesture : B takes a line and put it on the screen]*
A2 bon voilà…et <u>l'autre</u> tu vas la mettre parallèle à la première
okay… and put <u>the other one</u> parallel to the first one

All theses cases may be problematic for purely discursive approaches like Centering (Grosz et.al, 1995) or DRT (Kamp et Reyle, 1993). What we need here is a mechanism able to integrate discursive, perceptual and gestural information into a same context model.

2 Context modelling and the Incremental Algorithm (Dale & Reiter, 1995)

Dale & Reiter (1995) proposed their Incremental Algorithm with the aim to generate distinguishing descriptions. A distinguishing description has to characterize an entity *R* being referred to, but not any other entity – or distractor – in a current context set *C* (for a linguistic approach of the problem, see also Corblin, 1987). Given such a context set *C*, the algorithm goes through an ordered list of attributes characterizing the entities in *C* and retains the value of an attribute as a descriptor for *R* if it rules out at least one of the distractors.

A problem not adressed by the authors is the construction and updating mechanism of the context set. Even if they compare the context set to the focus spaces introduced by Grosz & Sidner (1986), it seems to be in practice "the global working set" (Dale, 1992). But our example (2), turn A3 – *la pyramide de droite* (*the pyramid on the right*) – shows that setting the context set to the global working set does not always

work. It rather suggests that the appropriateness of an attribute as a descriptor can change dynamically during the dialogue processing, depending on the currently activated context set. What we would like to show here is that we need a context dynamically structured into more local context sets (*domains*) with different activation weights. We assume that the dynamic context construction has to be multimodal, since the use of a referring expression depends not only on previous mentions in discourse, but also on the perceptual environment and on previous gestures, as shown in (1), (2) or (3).

3 A unified context model for generating referring expressions

3.1 Discourse, perception and gestures for dynamic context construction

Following Sanford & Garrod (1982) and Reboul (1998), the basic elements entering into our global context set are mental representations (MRs) for available entities of the application. They are introduced dynamically on the base of perceptual (e.g. visibility on the screen), gestural (introduction of a new entity) or discursive (mention of a new entity) criteria. An MR is modelled by an uniquely identified object (`@MR`), with at least a type (`@MR.type`). Type information is derived from a set of "generic" MRs, organized as a type hierarchy, which include general encyclopedic knowledge or knowledge specific to the application and is assumed to exist prior to discourse processing. Other information (`@MR.property`) provided via the discourse and via perception (size, color, position,…) may be added as necessary.

But as indicated before, human speakers do not seem to use always the entire context set when they are generating distinguishing descriptions. Thus, what we need is to construct and to represent subsets of the global context set, which can work as local *domains of reference* or *DRs.* Therefore, we consider that each MR may function as a DR. This is for example the case in (2), turn A1, where the MR introduced by *les deux grands triangles* (*the two big triangles*) becomes the DR for the interpretation of *la pyramide de gauche* (*the pyramid on the left*) in A2.

```
// discursive coordination
   If ("NP1 and NP2")
      @MR1(NP1)
      @MR2(NP2)
      @MR3 <= grouping (@MR1, @MR2)
   EndIf

// perceptual group
   If(percept_group(Fig1,Fig2))
      @MR1(Figure1)
      @MR2(Figure2)
      @MR3 <= grouping (@MR1, @MR2)
   EndIf
```

Fig. 2 – Algorithm for triggering groupings

It is also possible to create dynamically new DRs (Fig. 2) by a grouping operation. Depending on discursive criteria like coordination (*Maintenant il faut faire des maisons et une route*. / *Now, you have to make houses and a road*.) or on perceptual criteria like similarity or proximity (see the algorithms for perceptual grouping in Thorisson, 1994), two or more existing or newly created MRs will be grouped into a new MR, which can then function as a DR.

Within this newly created DR, the grouped entities have first to be characterized by a common type. This could be the type of the entities, if they are of the same type or the first common super type with respect to the type hierarchy, if they are different. Second, the entities have to be distinguished each from another with respect to a common

property like *type*, *size*, *position*,…. This property is the so-called the *differenciation criterion* (*DC*) – `@MR.dc` – and takes different values for different entities, like *big* or *small* for the DC *size*. The value of the DC for an item 1 in a DR MR_1 is referred to by `@MR1.dc.item1.dc_value`. Third, based on discursive, perceptual, gestural or task-related information, one of the elements of the DR may be focussed (see for example Hajicová, 1993; Grosz & Sidner, 1986; Grosz & al., 1995). This leads to the availability of the entity for pronominalization. The algorithm for theses operations is given in Fig. 3.

To sum up, our global context is composed of MRs, grouped into different DRs composing the global context are ordered according to their activation. An DR is activated when it is created or used for isolating a referent. In the following, we assume that this kind of context structure allows more precise predictions about the use of referring expressions in the "Ozkan" corpus.

3.2 Generating referring expressions : Type and distinguishing attributes

Given, on the one hand, the context including an active domain DR_A (a MR_A grouping elements of type *N* distinguished by a differenciation criterion DC_A) and, on the other hand, an entity *R* being referred to, we have to generate a distinguishing description for *R*. Our basic hypothesis is that the choice of the type of the expression (pronoun, definite,…) and of the distinguishing attributes for *R* depends on the relation between *R* and DR_A and, if $R \in DR_A$, on the existence or not of a focussed element in DR_A. Fig. 4 gives an outline of the algorithm.

```
@MR3 <= grouping (@MR1, @MR2)

      // common type calculus
      @MR3.type <= common_type(@MR1, @MR2)

      // calculus of the differenciation criterion and values
      If (@MR3.type = super_type(@MR1, @MR2))
         @MR3.dc  <= "type"
         @MR3.dc.item1.dc_value  <= @MR1.type
         @MR3.dc.item2.dc_value  <= @MR2.type
      Else
         If(@MR3.type = @MR1.type)
            @MR3.dc <= {"size"|"horizontal position"|…}
            @MR3.dc.item1.dc_value <= @MR1.{"size"|"horizontal position"|…}.value
            @MR3.dc.item2.dc_value  <= @MR2.{"size" | "horizontal position"|…}.value
         EndIf
         …// other cases…
      End If

      // focus calculus
      If(salient(@MR1))
         @MR3.dc.item1.focus = 1
      Else
         If(salient(@MR2))
            @MR3.dc.item2.focus = 1
         EndIf
      EndIf
```

Fig. 3 - Algorithm for the grouping of two MRs

```
// R is within the active DR (@MR_A)
If (R = (@MR_A.dc.item1 ∨ … ∨ @MR_A.dc.item_n))

   // Domain contains a focussed element
   If( (@MR_A.dc.item1.focus ∨ … ∨ @MR_A.dc.item_n.focus) = 1 )

      // the focussed element is R
      If(@MR_A.dc.item_R.focus = 1)
         Pronominalization

      // the focussed element is not R
      Else

         // R is the entire rest of the domain
         If(@MR_A.dc.item_R = (@MR_A.dc.item1 + … + @MR_A.dc.item_n) – focussed_element)
            Definite description :  the other one ∨
                                    the other @MR_A.type ∨
                                    the @MR_A.dc.item_R.dc_value @MR_A.type

         // R is not the entire rest of the domain
         Else
             Definite description : the @MR_A.dc.item_R.dc_value @MR_A.type
         EndIf

      EndIf

   // Domain does not contain a focussed element
   Else
      Definite description  : the @MR_A.dc.item_R.dc_value @MR_A.type
   EndIf

// R is not within the active DR : changing the active DR implies changing the DC…
Else…
```

Fig.4 - Algorithm for generating

3.3 Treatment of the examples

Example (1) shows the need of creating complex antecedents for pronominal expressions : In A1, a new MR $@T_1$ is introduced for *une grosse pyramide* (*a big pyramid*). Similarly, we create an MR $@T_2$ for *un grand triangle* (*a big triangle*) in A3. After B1, the two triangles form, based on the principles of similarity and proximity, a perceptual group on the screen. The grouping algorithm triggers the grouping of $@T_1$ and $@T_2$ into a new MR, let us say *@2BT*, with the structure in Fig. 5. $@T_2$ keeps the focus, because it is the last manipulated figure. In A4, a new small triangle $@T_3$ is introduced by a gesture. This small triangle forms, together with the two big triangles, a new perceptual group, represented by the MR *@3T* in Fig. 6, with a focussed element $@T_3$. In B2, the speaker uses a pronoun for referring to *@3T*. Following our algorithm for the generation of referring expressions, a pronoun can be used only if the referent is in the focus. This is not the case here, but the example shows that this constraint has to be to relaxed under certain conditions : in particular, the corpus shows that plural pronouns can be used to refer to the entire active domain, if there is an incompatibility between a plural pronoun and a single object in the focus. This is what happens in B2.

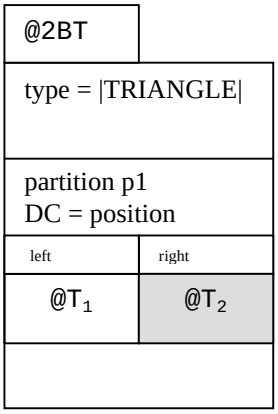

Fig. 5

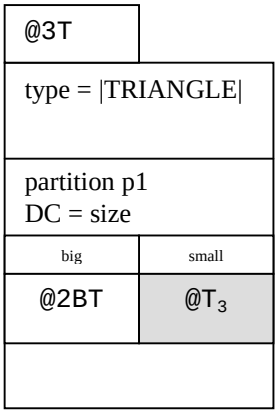

Fig. 6

Example (2) stresses the role of perception for establishing differenciation criteria, and thus, for finding distinguishing attributes of an entity. It illustrates also our hypothesis that running through the active DR goes preferentially with keeping the same differenciation criterion, whereas a change of the active DR will be expressed preferentially by using a new differenciation criterion :

A1 introduces in the context a MR for the group of *les deux grands triangles (the two big triangles)*, let us say, *@2BT*. Within this group, the two triangles, $@T_1$ and $@T_2$, are distinghuished each from the other by a perceptual differenciation criterion *horizontal_position*. The context structure before generating A2 is given in Fig. 7. Based on this context structure and following the algorithm, referring to $@T_1$ will be realized by *la pyramide de gauche* (*the pyramid on the left*) and leads, after the focalization of $@T_1$, to the context structure in Fig 8.

In A3, the algorithm is applied to this context structure in order to refer to $@T_2$. It generates *l'autre* (*the other one*) or *la pyramide de droite* (*the pyramid on the right*) – the second solution being realized in our corpus. In A4, things become yet more interesting: here, the speaker has to refer to the third – small – pyramid on the screen, but this pyramid, let us say $@T_3$, is not included in the active domain of reference, *@2BT*. Rather, it is a member of a domain *@3T* including *@2BT* and $@T_3$. This domain could be partitioned by two perceptual differenciation criteria: a first one distinguishes *@2BT* from *@T3* by a criterion *size* (*big* vs. *small*), a second one by *horizontal_position* (*left* vs. *right*). This context structure is showed in Fig. 9. What we suppose then is that the change of the

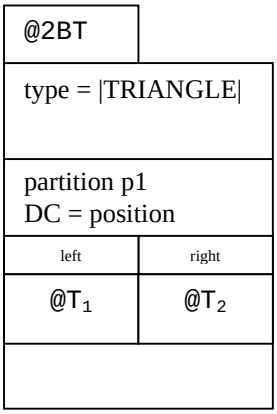

Fig. 7

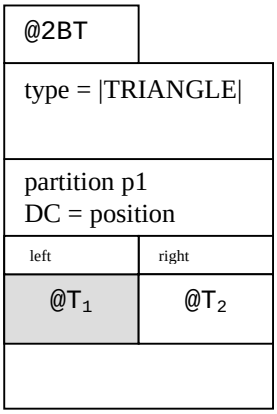

Fig. 8

active domain (from *@2BT* to $@T_3$) is expressed preferentially by a change of the active differenciation criterion. This leads to the choice of the differenciation criterion *size* and produces the expression *la petite pyramide* (*the small pyramid*), like in the corpus example.

Example (3), and more precisely the use of *l'autre* (*the other one*) in A2, shows the importance of taking into account gestures for the context construction during processing A1 and B1: In A1, *deux grandes barres* (*two big lines*) introduces a new referent and, consequently, a new MR *@2BL* in the context set. In the next turn B1, B manipulates a first line $@L_1$. This line is an element of *@2BL*, which is considered as the active DR, even if in this case the verbal reference act is substituted by a gestural one

(Siroux et al., 1995). The extraction of $@L_1$ partitions the DR *@2BL* into two lines, on the base of a differenciation criterion *manipulated* vs. ¬*manipulated*. Since $@L_1$ is the last element having been manipulated, it will be the focussed element of *@2BL*. The

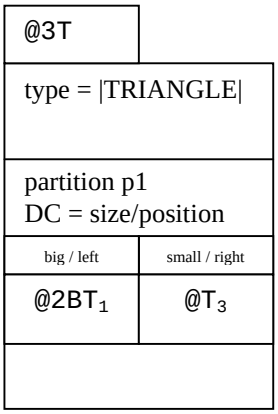

Fig. 9

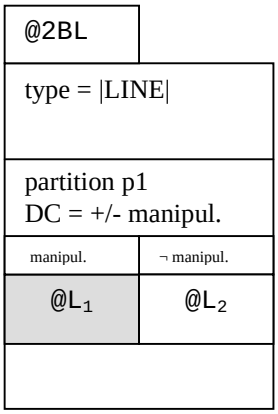

Fig. 10

context structure at this stage is schematized in Fig. 10. Now, in A2, A has to generate a referring expression for $@L_2$. Following the algorithm presented in the previous section, one solution will be *l'autre* (*the other one)*. This fits to what happens in our corpus example.

4 Discussion and further work

In this paper, we adressed the need of structuring the entire context set into subsets or *domains* in order to explain adequately the use of referring expressions in a multimodal corpus. We underlined in particular the importance of taking into account not only the discourse, but also perception and gestures for the construction of these domains. Therefore, we proposed a context model where the context is built up dynamically from different information sources. This manner of context modelling predicts correctly the use of referring expressions, which are "problematic" for other approaches. Further work consists of extending our theoretical framework to other kinds of referring expressions (demonstratives) and of exploring the hypotheses by testing our algorithms over the entire corpus. This includes, among other things, the adaptation of existing coreferential coding schemes (like Poesio, 2000) to multimedia corpora.

References

Corblin F. (1987) *Indéfini, Défini et Démonstratif*, Droz, Genève.

Dale R. (1992) *Generating Referring Expressions. Constructing Descriptions in a Domain of Objects and Processes.* The MIT Press, Cambridge, London.

Dale R. and Reiter E. (1995) Computational Interpretations of the Gricean Maxims in Generating Referring Expressions. *Cognitive Science 18, 233-263.*

Grosz B.J. and Sidner, C. (1986) Attention, Intention and the Structure of Discourse. *Computational Linguistics*, 12, 175-204.

Grosz B.J., Joshi A.K. and Weinstein S. (1995) Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 12(2), 203-225.

Hajicová E. (1993) Issues of Sentence Structure and Discourse Patterns. *Theoretical and Computational Linguistics*, 2, Charles University, Prague.

Kamp H. and Reyle U. (1993), *From Discourse to Logic.* Kluwer Academic Publishers. Dordrecht, Boston, London.

Poesio M. (2000) Coreference. *MATE Dialogue Annotation Guidelines-Deliverable D2.1*, January 2000, 126-182. (http://www.ims.uni-stuttgart.de/projekte/mate/mdag/)

Ozkan N. (1994) *Vers un modèle dynamique du dialogue : analyse de dialogues finalisés dans une perspective communicationnelle*. Ph.D.Thesis, Cognitive Science, INP Grenoble.

Reboul A. (1998) A relevance theoretic approach to reference. *Relevance Theory Workshop*, Luton (UK), Sept 8-10.

Sanford S.C. and Garrod A.J. (1982) The mental representation of discourse in a focussed memory system : implication for the interpretation of anaphoric noun phrases. *Journal of Semantics*, 1(1), 21-41.

Siroux J., Guyomard M., Multon F. and Remondeau C. (1995) Oral and gestural activities of the users

in the GEORAL system. *First International Workshop on Intelligence and Multimedia Interfaces*, Edinburgh.

Thorisson K.R. (1994) Simulated Perceptual Grouping : An Application to Human Computer Interaction. *16th Annual Conference of Cognitive Science Society*, Atlanta.