

# Reference reversibility with Reference Domain Theory

Alexandre Denis

TALARIS team / UMR 7503 LORIA/INRIA

Lorraine. Campus scientifique, BP 239

F-54506 Vandoeuvre-lès-Nancy cedex

alexandre.denis@loria.fr

## Abstract

In this paper we present a reference model based on Reference Domain Theory that can work both in interpretation and generation. We introduce a formalization of key concepts of RDT, the interpretation and generation algorithms and show an example of behavior in the dynamic, asymmetric and multimodal GIVE environment.

## 1 Introduction

The reference task in a dialogue system is two-fold. On the one hand the system has to *interpret* the referring expressions (RE) produced by the user in his utterances. On the other hand the system has to *generate* the REs for the objects it aims to refer to. We present in this paper a framework that considers that reference interpretation and generation are two sides of the same coin, hence avoiding any potential misunderstanding arising from the two modules discrepancies. Reference Domain Theory (RDT) (Salmon-Alt and Romary, 2000; Salmon-Alt and Romary, 2001) proposes to represent the diversity of referring acts by the diversity of constraints they impose on their context of use. The reversibility then lies in the possibility to express these constraints *independently of the considered task*.

In (Denis, 2010) we described the generation side of RDT in the context of the GIVE-2 challenge (Koller et al., 2010) which is an evaluation of instruction generation systems in a 3D maze. In this paper we propose the interpretation counterpart and show the required modeling to consider the dynamic, asymmetric and multimodal context of GIVE. We first present the reference model in section 2 and 3, discuss the interpretation problems in GIVE in section 4, detail an example in section 5 and present evaluation results in section 6.

## 2 Reference Domains

A rich contextual structure is required to give an account for the different kinds of discrimination we observe in REs such as semantic discrimination (e.g. “the blue button”), focus discrimination

(e.g. “this button”) and salience discrimination (e.g. “this one”). We introduce here the structure of *reference domain* which is a local context supporting these different discriminations.

We assume that *Props* is the set of unary predicate names e.g.  $\{blue, left, \dots\}$ , *Types* is the set of types of predicates e.g.  $\{color, position, \dots\}$ , and *val* is the function  $val : Types \rightarrow 2^{Props}$  which maps a type on the predicates names. Finally, *E* is the set of all objects and *V* the set of ground predicates e.g.  $\{blue(b1), \dots\}$ .

A *reference domain* *D* is then a tuple

$$\langle G_D, S_D, \sigma_D, (c, P, F) \rangle$$

where  $G_D \subseteq E$  is the set of objects of the domain, called the *ground of the domain*;  $S_D \subseteq Props$  is the *semantic description* of the domain, satisfied by all elements of the ground;  $\sigma_D \in \mathbb{N}$  is the *salience* of the domain. And  $(c, P, F)$  is a *partition structure* where  $c \in Types$  is a *differentiation criterion*;  $P$  is the *partition* generated by  $c$ ; and  $F \subseteq P$  is the *focus* of  $P$ .

For instance, a domain composed of a blue button  $b_1$  and a red button  $b_2$ , with a salience equal to 3, where  $b_1$  and  $b_2$  are differentiated using the color, and where  $b_1$  is in focus, would be noted as:

$$D = (\{b_1, b_2\}, \{button\}, 3, (color, \{\{b_1\}, \{b_2\}\}, \{\{b_1\}\}))$$

Finally we define a *referential space* (RS) as a set of reference domains (RD) ordered by salience.

## 3 Referring

A RE impose some constraints on the context in which it can be uttered, that is in which RD the interpretation has to be made. The constraints are represented as *underspecified domains* (UD), specifying the structure of the suitable RD in terms of ground, salience or partition. The explicit definitions of the UD makes possible to share these definitions between the interpretation and the generation modules, hence allowing the implementation of a *type B reversible reference module* (Klarner, 2005), that is a module in which both directions share the same resources.

| Expression    | $U(N, t)$ matches $D$ iff $\exists(c, P, F) \in D$ ;   |
|---------------|--|
| this one      | $F = \{\{t\}\} \wedge \text{msd}(D)$   |
| this N        | $F = \{\{t\}\} \wedge t \in N^{\mathcal{I}}$   |
| the N         | $t \in N^{\mathcal{I}} \wedge \{t\} \in P \wedge \forall X \in P, X \neq \{t\} \Rightarrow X \cap N^{\mathcal{I}} = \emptyset$ |
| the other one | $F \neq \emptyset \wedge P \setminus F = \{\{t\}\} \wedge \text{msd}(D)$   |
| the other N   | $F \neq \emptyset \wedge P \setminus F = \{\{t\}\} \wedge G_D \subseteq N^{\mathcal{I}}$                                       |
| another one   | $F \neq \emptyset \wedge \{t\} \in P \setminus F \wedge \text{msd}(D)$   |
| another N     | $F \neq \emptyset \wedge \{t\} \in P \setminus F \wedge G_D \subseteq N^{\mathcal{I}}$   |
| a N           | $t \in N^{\mathcal{I}} \wedge t \in G_D$   |

Table 1: Underspecified domains for each type of referring expression

### 3.1 Underspecified domains

The different types of UD are presented in table 1. Each UD is a parametric conjunction of constraints on a RD, noted  $U(N, t)$ , where  $t$  is the intended referent and  $N \subseteq \text{Props}$  is a semantic description.  $N^{\mathcal{I}}$  stands for the *extension* of  $N$ , and  $\text{msd}(D)$  stands for *most salient description*, that is, there is no more or equally salient domain than  $D$  in the current RS with a different description. Each UD is associated to a *wording* combining a determiner and a wording of the semantic description, for instance “the N” is a shortcut for a definite expression whose head noun and modifiers are provided by the wording of  $N$ . Finally we say that an UD *matches* a RD if all the constraints of the UD are satisfied by the RD.

### 3.2 Referring processes

Interpretation and generation can now be defined in terms of UD. The two processes are illustrated in figure 1 and the algorithms are presented in figure 2.

The *interpretation* algorithm consists in finding or creating a RD from the input UD,  $U(N, .)$  created from the input RE type and description  $N$ . The algorithm then iterates through the RS in salience order, and through all the individuals  $t$  of the tested domain to retrieve the first one matching  $U(N, t)$ . If a matching domain  $D$  is found, a restructuring operation is applied and the referent  $t$  is focused in the partition of  $D$ . On the other hand, if no domain is found, the UD is *accommodated*, that is a new domain and a new referent satisfying the constraints of  $U(N, t)$  are created. According to the task, this accommodation may not be possible for all REs, but for sake of simplicity we assume here this operation is always possible.

The *generation* side is the opposite, that is it finds an UD from an input RD. It first selects a RD containing the target referent to generate  $t$ , assuming here that the most salient domain has to be preferred. The description  $N$  used to instantiate the UD is composed of the description of the domain and the description of the referent in the partition (line 2). It then iterates through the

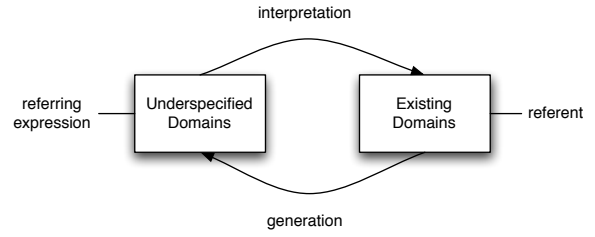


Figure 1: Reference processes

different UD by Givenness order (Gundel et al., 1993) and selects the first one that matches. A restructuring operation is applied and the found UD is returned, eventually providing the RE.

The restructuring operation, detailed in (Denis, 2010), aims to restrict the current context by creating a new domain around the referent in the referential space or by increasing the salience of the domain containing the referent. This operation helps to perform focalization in restricted domains.

## 4 The complex context of GIVE

The dynamic, asymmetric and multimodal context of GIVE requires additional mechanisms for interpretation. Asymmetry causes the *late visual context integration*, when the direction giver produces a RE to objects not yet known by the direction follower, that are only visually discovered later on. Space prevents us to describe in details the late integration algorithm, but the idea is, given a new physical object  $t$ , to scan existing domains of the actual RS to check if  $t$  can be merged semantically with any previous object  $t'$ . If this could be the case, the integration leads to create two parallel RS, one in which  $t = t'$  (the *fusion* hypothesis) and one in which  $t \neq t'$  (the *separation* hypothesis). If this cannot be the case,  $t$  is added as a new object. Following (DeVault and Stone, 2007), these alternative contexts can persist across time and further referring expressions may reject one or the other hypothesis as illustrated in section 5.

The second required mechanism is the proper handling of the *multimodal dynamic focus*, that is the combination of the linguistic focus resulting from RE, and the visual focus. It is possible to have two referential spaces for the linguistic or visual context as in (Kelleher et al., 2005; Byron et al., 2005), or to have two foci in a partition. We can also model *interleaved focus*, that is, only one focus per domain but that dynamically corresponds to the linguistic focus or the visual focus. The idea is that after each RE, the referent receives the focus as described in algorithm 1, but whenever the visual context changes, the focus is updated to the visible objects. Although interleaved focus prevents anaphora while the visual context changes, its complexity is enough for our setup.

**Algorithm 1** interpret( $U(N, \cdot), RS$ )

```

1: for all domain  $D$  in  $RS$  by salience order do
2:   for all  $t \in G_D$  do
3:     if  $U(N, t)$  matches  $D$  then
4:       restructure( $D, N, RS$ )
5:       focus  $t$  in  $D$ 
6:       return  $t$ 
7:     end if
8:   end for
9: end for
10: return accommodate( $U(N, \cdot), RS$ )

```

**Algorithm 2** generate( $t, RS$ )

```

1:  $D \leftarrow$  most salient domain containing  $t$ 
2:  $N \leftarrow S_D \cup \{p | p \in val(c), p(t) \in V\}$ 
3: for all  $U(N, t)$  sorted by Givenness do
4:   if  $U(N, t)$  matches  $D$  then
5:     restructure( $D, N, RS$ )
6:     return  $U(N, t)$ 
7:   end if
8: end for
9: return failure

```

Figure 2: Reference algorithms, relying on the same underspecified domains

## 5 Example

In this section we present the interpretation side of some expressions we generated in the GIVE setting (table 2). The detailed generation side of this example can be found in (Denis, 2010).  $S$  is the system that interprets the RE of  $U$  the user. The situation is:  $S$  enters a room with two blue buttons  $b_1$  and  $b_2$ , none of them being visible when he enters and  $U$  wants to refer to  $b_1$ .

| state of $S$ | utterance of $U$                      |
|--------------|---------------------------------------|
|              | Push a blue button ( $b_1$ )          |
| see( $b_2$ ) | Not this one! Look for the other one! |
| see( $b_1$ ) | Yeah! This one!                       |

Table 2: Utterances produced by  $U$ 

When  $S$  enters the room,  $U$  generates an indefinite RE “Push a blue button”.  $S$  first constructs an indefinite UD “a  $N$ ” with  $N = \{blue, button\}$ . However, because there exists no RD at first, he has to accommodate the UD, hence creating a new domain  $D_1$  containing a new linguistically focused individual  $t$ :

$$D_1 = \langle \{t\}, \{button, blue\}, 1, (\text{id}, \{\{t\}\}, \{\{t\}\}) \rangle$$

We assume that  $S$  moves and now sees the blue button  $b_2$  without knowing yet if this is the intended one. The integration of this new physical object then leads to two hypothesis. In the *fusion* hypothesis,  $b_2 = t$ , and in the *separation* hypothesis,  $b_2 \neq t$ . In both cases, the visible button is focused in the two versions of  $D_1$ ,  $D_{1FUS}$  and  $D_{1SEP}$ :

$$D_{1FUS} = \langle \{t\}, \{button, blue\}, 2, (\text{id}, \{\{t\}\}, \{\{t\}\}) \rangle$$

$$D_{1SEP} = \langle \{t, b_2\}, \{button, blue\}, 2, (\text{id}, \{\{t\}, \{b_2\}\}, \{\{b_2\}\}) \rangle$$

However,  $U$  utters “Not this one!” rejecting then the fusion hypothesis. To be able to consider the effects of this utterance, we have to take into account the ellipsis. This can be done by assuming that  $U$  is asserting properties of the target of his first RE, that is, he is actually stating that “[ $t$  is] not this one!”. The RE “this one” leads to the construction of a demonstrative one-anaphora UD that matches  $t$  in  $D_{1FUS}$  but  $b_2$  in  $D_{1SEP}$ . The following schema shows the contradiction in the fusion hypothesis:

|            |     |        |          |
|------------|-----|--------|----------|
|            | $t$ | is not | this one |
| fusion     | $t$ | $\neq$ | $t$      |
| separation | $t$ | $\neq$ | $b_2$    |

Being contradictory, the fusion hypothesis is rejected and only  $D_{1SEP}$  is maintained. For the readability of the presentation,  $D_{1SEP}$  is rewritten as  $D_1$ .

The interpretation of “Look for the other one!” is straightforward. A definite alternative one-anaphora UD is built, and both  $t$  and  $b_2$  are tested in  $D_1$  but only  $t$  is matched because it is unfocused (see the definition of the alternative one-anaphora in table 1).

Now  $S$  moves again and sees  $b_1$ . As for  $b_2$ , the integration of  $b_1$  in the referential space leads to two alternative RS. The buttons  $b_2$  and  $b_1$  cannot be merged (we assume here that  $S$  can clearly see they are two different buttons), thus the two alternative RS are whether  $b_1 = t$  or  $b_1 \neq t$ :

$$D_{1FUS} = \langle \{t, b_2\}, \{button, blue\}, 3, (\text{id}, \{\{t\}, \{b_2\}\}, \{\{t\}\}) \rangle$$

$$D_{1SEP} = \langle \{t, b_1, b_2\}, \{button, blue\}, 3, (\text{id}, \{\{t\}, \{b_1\}, \{b_2\}\}, \{\{b_1\}\}) \rangle$$

Eventually  $S$  has to interpret “this one”. Like previously, in order to take into account the effects of this utterance,  $S$  has to resolve the ellipsis and must consider “[ $t$  is] this one”. The RE “this one” is resolved on  $t$  in  $D_{1FUS}$  but on  $b_1$  in  $D_{1SEP}$ .

|            |     |    |          |
|------------|-----|----|----------|
|            | $t$ | is | this one |
| fusion     | $t$ | =  | $t$      |
| separation | $t$ | =  | $b_1$    |

This is now the separation hypothesis which is inconsistent because we assumed that  $b_1 \neq t$ . This RS is then ruled out, and only the fusion RS remains.

## 6 Evaluation

Only the generation direction has been evaluated in the GIVE challenge. The results (Koller et al., 2010) show that the system embedding Reference Domain Theory proves to rely on less instructions than other systems (224) and proves to be the most successful (47% of task success) while being the fastest (344 seconds). We conjecture that the good results of RDT can be explained by the low cognitive load resulting from the use of demonstrative NPs and one-anaphoras, but the role of the overall generation strategy has also to be taken into account in these good results (Denis et al., 2010).

Although it would be very interesting, the interpretation side has not yet been evaluated in the GIVE setting, but only in the MEDIA campaign (Bonneau Maynard et al., 2009) which is an unimodal setting. The results show that the interpretation side of RDT achieves a fair precision in identification (75.2%) but a low recall (44.7%). We assume that the low recall of the module is caused by the cascade of errors, one error at the start of a reference chain leading to several other errors. Nonetheless, we estimate that error cascading would be less problematic in the GIVE setting because of its dynamicity.

## 7 Conclusions

We presented a reference framework extending (Salmon-Alt and Romary, 2001) in which interpretation and generation can be defined in terms of the constraints imposed by the referring expressions on their context of use. The two modules sharing the same library of constraints, the model is then said *reversible*. However, because of the asymmetry and dynamicity of our setup, the GIVE challenge, additional mechanisms such as uncertainty have to be modeled. In particular, we have to maintain different interpretation contexts like (DeVault and Stone, 2007) to take into account the ambiguity arising from the late integration of the visual context. It would be interesting now to explore deeper our reversibility claim by evaluating the interaction between the two reference algorithms in the GIVE setting.

## References

- Hélène Bonneau Maynard, Matthieu Quignard, and Alexandre Denis. 2009. MEDIA: a semantically annotated corpus of task oriented dialogs in French. *Language Resources and Evaluation*, 43(4):329–354.
- Donna K. Byron, Thomas Mampilly, Vinay Sharma, and Tianfang Xu. 2005. Utilizing visual attention for cross-modal coreference interpretation. In *Proceedings of Context-05*, pages 83–96.
- Alexandre Denis, Marilisa Amoia, Luciana Benotti, Laura Perez-Beltrachini, Claire Gardent, and Tarik Osswald. 2010. The GIVE-2 Nancy Generation Systems NA and NM. Technical report.
- Alexandre Denis. 2010. Generating referring expressions with Reference Domain Theory. In *Proceedings of the 6th International Natural Language Generation Conference - INLG 2010*, Dublin, Ireland.
- David DeVault and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue. In *Proceedings of the 2007 Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Trento, Italy.
- Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.
- John Kelleher, Fintan Costello, and Josef van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167(1-2):62–102.
- Martin Klärner. 2005. Reversibility and reusability of resources in NLG and natural language dialog systems. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland.
- Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. Report on the second NLG challenge on generating instructions in virtual environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference - INLG 2010*, Dublin, Ireland.
- Susanne Salmon-Alt and Laurent Romary. 2000. Generating referring expressions in multimodal contexts. In *Workshop on Coherence in Generated Multimedia - INLG 2000*, Israel.
- Susanne Salmon-Alt and Laurent Romary. 2001. Reference resolution within the framework of cognitive grammar. In *Proceeding of the International Colloquium on Cognitive Science*, San Sebastian, Spain.