Ancrage référentiel en situation de dialogue

Frédéric Landragin, Susanne Salmon-Alt, Laurent Romary

Ancrage référentiel en situation de dialogue

F. Landragin* — S. Salmon-Alt — L. Romary***

* *LORIA – UMR 7503*
Laboratoire Lorrain de Recherche en Informatique et ses Applications
Campus scientifique, BP 239, F-54506 Vandœuvre-lès-Nancy cedex
{Frederic.Landragin, Laurent.Romary}@loria.fr

** *ATILF – UMR 7118*
Analyse et Traitement Informatique de la Langue Française
44, avenue de la Libération, BP 30687, F-54063 Nancy cedex
Susanne.Salmon-Alt@inalf.fr

RÉSUMÉ. *A partir d'extraits de corpus, nous montrons qu'un modèle d'interprétation de la référence basé exclusivement sur des contraintes imposées par l'expression linguistique (éventuellement accompagnée d'un geste de désignation) est insuffisant dans le cadre du dialogue avec support visuel. Pour exploiter finement ces contraintes, un tel modèle doit nécessairement s'appuyer sur une représentation des contextes intégrant les principes de fonctionnement de la perception visuelle, de la tâche et de la mémoire de l'utilisateur. Nous nous focalisons ici sur les mécanismes de construction et d'exploitation de ces contextes hétérogènes dans un cadre unifié. Nous proposons ainsi un modèle d'interprétation de la référence suffisamment générique pour pouvoir être appliqué à différents types d'interactions et à différents types d'applications.*

ABSTRACT. *From corpus extracts, we show that modeling the interpretation of referring expressions (eventually with designation gestures) only from linguistic constraints is not sufficient for the dialogue with visual aids. A fine exploitation of these constraints must go through a representation of the contexts implying the visual perception, task and user memory working principles. We focus here on the mechanisms of the construction and the exploitation of these heterogeneous contexts in an unified framework. We therefore propose a model for reference interpretation which is generic enough to be applicable to different types of interactions and different types of tasks.*

MOTS-CLÉS : *interprétation de la référence, dialogue multimodal, modélisation du contexte, perception visuelle, saillance.*

KEYWORDS: *interpreting referring expressions, multimodal dialogue, context modelling, visual perception, salience.*

1. Introduction

Notre objectif est de mener une réflexion pluridisciplinaire sur la nature des facteurs intervenant lors de l'interprétation des actions de référence. Nous ne présentons pas ici un système de dialogue homme-machine implémenté, mais nous souhaitons montrer comment une étape de spécification peut faire tendre vers un tel système[1]. Le calcul de la référence aux objets consiste à identifier la « relation qui unit une expression de la langue (dite en général *expression référentielle*) en emploi dans un énoncé et l'objet dans le monde que cette expression désigne » (Moeschler & Reboul, 1994 : p. 534). Contrairement à la conception structuraliste du signe linguistique en termes de *signifiant* et *signifié*, l'étude de la référence ne peut donc s'arrêter à l'analyse de la relation entre une chaîne sonore (le signifiant) et un concept (le signifié), relation qui est traditionnellement l'objet de la sémantique, et plus particulièrement de la question du sens. Si le sens peut être considéré comme la potentialité d'un signe à évoquer les entités concrètes ou abstraites de l'univers, la référence est l'instanciation de cette faculté dans un contexte particulier. Cela a une conséquence majeure pour le traitement automatique de la référence : l'établissement du lien entre une expression de la langue et son référent passe nécessairement par la prise en compte du contexte – linguistique et extra-linguistique – de l'énonciation.

Dans le cadre du dialogue homme-machine spontané permettant à l'utilisateur de s'exprimer en langue naturelle et d'accompagner ses énoncés oraux de gestes de désignation, les expressions référentielles sont matérialisées par une grande variété de syntagmes : groupes nominaux à différentes déterminations, avec ou sans tête nominale, et groupes pronominaux[2]. De plus, quel que soit le type de référent cherché – objet concret ou abstrait, objet fictif, groupe d'objets, espèce – le calcul de la référence dépend à la fois d'informations linguistiques (détermination et sémantique de l'expression référentielle, historique langagier) et d'informations non linguistiques (contexte perceptif, gestes de désignation, connaissances sur l'état de la tâche, habitudes des interlocuteurs). Notre réflexion cherche à définir un modèle formel de résolution de la référence tenant compte de ces paramètres, des particularités structurelles des modalités d'interaction et des facteurs cognitifs nécessaires à la compréhension de celles-ci : intention, perception, mémoire, attention. Notre première ambition dans l'élaboration de ce modèle est son application à toute forme de référence, et par conséquent à tout type d'application. Pour cette raison, une implantation informatique pour une application particulière nous semble à ce stade moins adaptée pour la validation de notre modèle que l'explication et la prédiction de phénomènes référentiels observables. Notre deuxième ambition est de faire intervenir à la fois les paramètres du langage et du geste spontanés, ainsi que de la perception visuelle. Or, que ce soit pour le recueil ou

1. Notre réflexion s'inscrit en cela dans le projet MIAMM (Multimedia Information Access using Multiple Modalities : http://www.loria.fr/projets/miamm).

2. Ces caractéristiques ne suffisent pas à en faire des expressions référentielles : certaines expressions comme par exemple « je vais acheter *un chat* » ne réfèrent pas, l'identification d'un référent spécifique n'étant pas indispensable à l'interprétation.

pour la validation de ces paramètres, il s'avère difficile de constituer et d'exploiter des corpus regroupant ces trois modalités. La cause en est la pluralité des paramètres et en particulier la complexité des paramètres visuels, qui se révèlent variés, approximatifs, parfois subjectifs. L'exploitation quantitative de toutes ces données s'avère donc difficile, comme le montrent les corpus Ozkan (1994) et Magnét'Oz (Wolff, 1999) dont nous tirons nos exemples. Ces corpus ont été constitués à partir de dialogues spontanés et, bien que les situations soient orientées par la tâche, ils comportent une grande variété de phénomènes référentiels. C'est l'analyse détaillée de ceux-ci qui nous semble devoir diriger l'élaboration d'un modèle. Ils ont en effet l'avantage de mettre l'accent sur des mécanismes fondamentaux de compréhension, qui pourront à plus long terme être validés par des expérimentations ciblées, paramètre par paramètre, en suivant les protocoles de la psycholinguistique.

Après avoir montré dans un premier temps la diversité des informations intervenant lors de l'ancrage référentiel, nous détaillons la nature structurée de ces informations et nous proposons une modélisation du calcul de la référence basée sur ces structures et sur la notion de saillance.

2. Diversité de l'ancrage référentiel et critique des approches classiques

La figure 1 reproduit l'extrait d'un corpus de dialogues finalisés homme-homme (Ozkan, 1994). Cet extrait illustre la complexité des phénomènes référentiels dans un dialogue homme-homme multimodal combinant langage, perception visuelle et gestes, même dans un univers restreint par une tâche simple (la conception de dessins à partir de primitives géométriques). Un manipulateur M suit les instructions d'un instructeur I qui est le seul à connaître le dessin final. I et M sont placés dans des salles séparées. Ils partagent le même écran avec les scènes en construction et la palette des figures disponibles, non représentée dans la figure 1.

Parmi les expressions référentielles à interpréter au cours de cet extrait consistant à construire une ligne d'horizon, nous nous concentrerons sur celles (soulignées dans l'extrait) qui correspondent aux références dans la zone de manipulation, d'une part pour montrer certaines failles des approches courantes, d'autre part pour faire apparaître tous les facteurs qui devront, selon nous, être pris en compte dans un modèle de la référence : l'intérêt de la suite *les deux grands triangles* (I2) – *la pyramide de gauche* (I5) – *la pyramide de droite* (I7) – *la petite pyramide* (I10) réside dans l'interprétation de *la pyramide de droite* (I7) qui ne désigne pas, comme on pourrait s'y attendre hors contexte, le petit triangle le plus à droite de la scène, mais celui le plus à droite parmi les deux grands triangles, et ceci sans ambiguïté pour les interlocuteurs.

Cette interprétation échappe en effet aux modèles référentiels fondés prioritairement sur l'appariement des propriétés exprimées linguistiquement (*pyramide*, *de droite*) avec les propriétés des objets décrites dans une base de

données (type, couleur, coordonnées de positionnement, etc.). La majorité de ces systèmes – dont SHRDLU (Winograd, 1972) est le précurseur et DenK (Kievit & Piwek, 2000) ou l'Atelier de la Référence (Popescu-Belis, 1999) sont des réalisations récentes – sont d'ailleurs dotés d'heuristiques particulières pour choisir un référent parmi plusieurs, au cas où l'expression à interpréter serait ambiguë. En revanche, l'absence d'une gestion dynamique des espaces de recherche sur des critères autres que la récence d'une éventuelle mention précédente ne leur permet pas de résoudre correctement la référence pour l'expression *la pyramide de droite*.

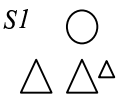
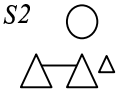
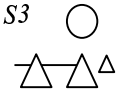
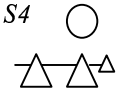
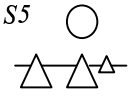
(*S1*)	I1	il faut prendre une grande horizontale	*S1* 
	I2	et la placer à la pointe des deux grands triangles […]	
	M1	*(geste de manipulation directe)*	
(*S2*)	I3	voilà comme ça … et tu en prends une deuxième	*S2* 
	I4	une petite […]	
	I5	tu la places à gauche de la pyramide de gauche […]	
	M3	*(geste de manipulation directe)*	
(*S3*)	I6	voilà comme ça […] et t'en prends une autre petite […]	*S3* 
	I7	et tu la places à droite de la pyramide de droite […]	
	M4	*(geste de manipulation directe)*	
(*S4*)	I8	voilà comme ça… et tu prends une autre petite verticale	*S4* 
	M5	une autre petite verticale	
	I9	euh horizontale pardon	
	I10	et puis tu la places dans la même lignée à droite de la petite pyramide	*S5* 
(*S5*)			

Figure 1. *Extrait du corpus Ozkan (transcription et scènes visuelles)*

Les mécanismes à l'œuvre ici se rapprochent plutôt des notions d'« espace focal » (Grosz & Sidner, 1986 ; Beun & Cremers, 1998) ou de « quantification contextuelle » (Westerstahl, 1984 ; Dekker, 1998). L'idée commune de ces travaux est de vouloir restreindre la résolution référentielle à l'appariement de propriétés à un sous-ensemble contextuel saillant. Cette stratégie permettrait en effet d'interpréter *la pyramide de droite* dans le sous-ensemble des deux grands triangles visibles à l'écran et d'aboutir ainsi à identifier le « bon » référent.

Le problème principal devient alors la définition des conditions de création de ces sous-ensembles. Beaucoup de travaux citent le problème sans proposer de solution effective. Proposer des solutions à ce problème demande en effet de s'interroger sur la prise en compte, dans l'historique du dialogue, d'autres facteurs

que ceux purement linguistiques. En plus d'un historique langagier – chose courante dans les systèmes actuels – il doit y avoir d'autres informations, en particulier d'ordre perceptif ou venant de la tâche, permettant de restreindre ou d'élargir dynamiquement l'espace d'interprétation d'une expression référentielle dans un dialogue de commande avec support visuel. Du côté de la perception visuelle, Thórisson (1994) montre par exemple comment des critères issus de la théorie de la Gestalt (Wertheimer, 1923) permettent de structurer l'espace d'interprétation en sous-ensembles, dont certains sont, à un moment donné, visuellement plus saillants que d'autres. Du côté de la tâche, Grosz & Sidner (1986) ont montré comment la structuration de la tâche permet de créer des sous-ensembles contextuels servant à lever l'ambiguïté des descriptions définies (cf. aussi Wright, 1990).

L'extrait de dialogue de la figure 1 illustre bien la collaboration de tous ces critères lors de l'interprétation des expressions référentielles dans un dialogue en langue naturelle. Si l'historique langagier seul n'aide pas à interpréter correctement l'expression *la pyramide de droite*, il est néanmoins indispensable : d'une part, il permet la réutilisation des algorithmes s'appuyant sur le contexte linguistique pour la résolution des pronoms personnels (Grosz *et al.*, 1995) ; d'autre part, il peut introduire, avec des groupes nominaux pluriels ou coordonnés, des sous-ensembles contextuels contraignant éventuellement la résolution référentielle dans la suite du dialogue, comme ici la mention *les deux grands triangles* en I2. Mais les seules indications linguistiques sur de possibles contextes d'interprétation sont souvent insuffisantes.

Dans l'exemple qui nous intéresse ici, au moment de l'interprétation de *la pyramide de droite*, la perception ainsi que le modèle de la tâche viennent renforcer l'hypothèse de la saillance du sous-ensemble des deux grands triangles. La perception fournit un ensemble visuellement homogène, celui des deux grands triangles reliés par une barre horizontale. La tâche – la construction d'un horizon, formé de lignes horizontales, autour des pyramides représentées par des triangles – a été décomposée par I en différentes sous-étapes (pose d'une barre, puis extension à gauche) qui permettent de prédire que l'expression *la pyramide de droite* désigne la grande pyramide de droite plutôt que la petite pyramide tout à droite de la scène.

A partir des faits mis en relief lors du commentaire de l'exemple ci-dessus, notre objectif est de proposer une synthèse reposant sur deux constats :

– une expression référentielle impose, par sa détermination et la sémantique de ses composantes linguistiques, un certain nombre de contraintes qui doivent être projetées dans un contexte approprié afin d'identifier le référent ;

– ce contexte dépasse les informations de nature linguistique et inclut également des informations de nature perceptive, de la tâche et du modèle de la mémoire de l'utilisateur.

La problématique qui nous intéresse ici est alors de modéliser de façon uniforme les contraintes linguistiques ainsi que les différentes dimensions contextuelles. Le

projet CERVICAL (Reboul *et al.*, 1997) a abordé cette problématique en proposant une modélisation du contexte sous forme de « domaines de référence », représentant des ensembles de référents, éventuellement partitionnés par une propriété distinctive ou un « critère de différenciation ». Nous avons prolongé ces travaux dans deux directions : d'une part, en considérant que l'expression référentielle impose elle-même, par sa détermination et sa sémantique, des contraintes à la fois sur les propriétés de son référent et sur celles de son contexte d'interprétation ou domaine de référence (Salmon-Alt, 2001) ; d'autre part, en affinant le rôle des critères langagiers, visuels et liés à la tâche permettant de construire les domaines contextuels qui forment des espaces de projection potentiels pour ces contraintes, avec une attention particulière pour les domaines perceptifs, très importants dans le dialogue homme-machine avec support visuel (Landragin *et al.*, 2001).

Dans la suite, nous partons d'une analyse fine des contraintes linguistiques pour montrer qu'une exploitation pertinente de ces contraintes passe par leur projection sur un contexte structuré, comportant des sources d'informations hétérogènes. La section 3, inspirée essentiellement de travaux de linguistique descriptive, sera donc consacrée à une étude des contraintes imposées par une expression à interpréter. La section 4 présente les différentes facettes de l'espace de projection : historique langagier, perception, tâche. Les mécanismes de sélection, et plus particulièrement les critères de choix en cas d'ambiguïté entre différents contextes disponibles, seront étudiés dans la section 5.

3. Contraintes venant de l'expression référentielle

3.1. *L'idée directrice : les domaines de référence*

La différence essentielle de notre modélisation par rapport aux modèles existants est l'hypothèse d'un cadre référentiel impliqué dans tout acte de référence. Alors que la majorité des travaux précédents (cf. par exemple Kamp & Reyle (1993), avec l'extension qu'en proposent Bos *et al.* (1995) pour traiter les descriptions définies) ramènent le problème de la référence à une relation dont le prototype serait le « liage » d'une expression référentielle à un référent accessible, nous pensons que l'accès au « bon » référent se fait systématiquement via l'activation d'un sous-ensemble contextuel, le « domaine de référence ». Cette idée est déjà présente dans Olson (1970), qui défend une théorie cognitive de la sémantique, affirmant que « Words do not mean referents or stand for referents, they have a use – they specify perceived events relative to a set of alternatives[3] » (p. 263). La conséquence de ce point de vue est que l'interprétation d'une expression est plus que l'identification d'un référent : c'est aussi l'identification d'un ensemble de possibilités exclues. Ce point de vue a effectivement été intégré dans les travaux de génération automatique

3. « Des mots ne signifient ni ne remplacent des référents, ils ont un usage – ils spécifient des événements perçus par rapport à un ensemble de possibilités. »

d'expressions référentielles, en particulier par Dale & Reiter (1995). Dans les systèmes de compréhension, en revanche, l'approche la plus courante pour identifier le référent d'une expression référentielle définie consiste à filtrer successivement les entités de l'application jusqu'à n'en retenir qu'une seule, compatible avec la description (Kievit & Piwek, 2000). En principe, ce processus est réitéré pour chaque expression, avec l'aide d'heuristiques pour traiter les expressions d'altérité comme *l'autre triangle* (Compérobot, Vivier & Nicolle, 1997), les ellipses et les *one*-anaphora[4] (SHRDLU, Winograd, 1972), pour lesquelles il est nécessaire de revenir sur des hypothèses précédentes.

Nous pensons alors que l'identification systématique des possibilités exclues peut être mise au profit de processus de compréhension à la fois plus efficaces d'un point de vue informatique et plus proches du fonctionnement cognitif. Le fait qu'elles fournissent le domaine d'interprétation pour toutes les expressions elliptiques, d'altérité et les expressions ambiguës traduit, selon nous, un principe d'interprétation plus fondamental : ces domaines forment l'espace d'ancrage contextuel préférentiel pour toutes les expressions à interpréter.

Notre modélisation repose donc sur l'hypothèse fondamentale que tout acte de référence consiste à isoler un référent dans un ensemble de référents comprenant l'objet à identifier et les « alternatives » – objets desquels le référent se distingue par la valeur d'une propriété distinctive. Nous appellerons cet ensemble « domaine de référence » et la propriété distinctive « critère de différenciation ». L'expression référentielle impose d'elle-même, par sa détermination et sa sémantique, des contraintes non seulement sur des propriétés du référent, mais également sur des propriétés de son domaine de référence. Ces contraintes seront modélisées ci-dessous sous la forme de domaines de référence sous-spécifiés. En confrontant ces domaines sous-spécifiés aux sous-ensembles contextuels fournis par l'historique langagier, la perception visuelle et la tâche, un ou plusieurs domaines de référence seront ensuite identifiés, et les référents en seront extraits.

3.2. *La construction de domaines de référence sous-spécifiés*

La nécessité d'interprétation dans un domaine contextuel est commune à tout type d'expression (indéfini, défini, démonstratif, pronom). Mais il apparaît également, à partir des études linguistiques fines consacrées au fonctionnement propre des marqueurs référentiels du français (Kleiber, 1994 ; Corblin, 1987), que ces différentes formes sont à considérer comme autant d'instructions de traitement différentes pour l'ancrage référentiel. Ces instructions seront modélisées ici par des domaines de référence sous-spécifiés.

---

4. Terme qui désigne les structures elliptiques de l'anglais comportant une trace linguistique (« the big *one* ») et correspondant en français à des groupes nominaux sans nom (« le grand »).

Les domaines de référence sous-spécifiés formulent des contraintes structurelles et des contraintes de typage sur le ou les domaines contextuels appropriés. Une première contrainte structurelle porte sur l'existence ou non d'une partition préalable du domaine. Lorsqu'une partition préalable du domaine est présupposée, alors une deuxième contrainte structurelle peut porter sur l'existence ou non d'un élément particulièrement saillant ou « focalisé » à l'intérieur du domaine. Les contraintes de typage portent d'une part sur le type des objets du domaine et d'autre part sur la propriété justifiant, le cas échéant, la partition du domaine[5].

Alors que la détermination d'une expression permet de dériver des contraintes sur la partition et la focalisation de son domaine de référence, la sémantique de ses composantes permet de préciser des contraintes sur le typage des éléments et le critère de différenciation associé à une partition éventuelle.

3.2.1. *La détermination : contraintes sur la partition et la focalisation*

Indéfinis. Selon Corblin (1987), un indéfini de forme *n N* extrait *n* éléments d'un domaine comprenant des éléments de type *'N'*. Un indéfini tel que *une horizontale* cherchera donc à s'interpréter dans un domaine contextuel formé par des objets de type '*horizontale'*. L'extraction référentielle est possible sans qu'il soit nécessaire d'avoir, au préalable, partitionné ce domaine sur un critère quelconque. De même, l'interprétation de l'indéfini ne fait pas appel à une propriété de focalisation préalable.

Définis. Corblin (1987) montre qu'un défini s'interprète dans un domaine à l'intérieur duquel son contenu constitue un signalement singularisant. Un défini tel que *la pyramide de droite* s'interprète alors dans un domaine hétérogène, c'est-à-dire partitionné selon au moins un critère de différenciation : ici, on cherchera par exemple un domaine comprenant des objets de type '*pyramide'* pouvant être opposés selon leur position horizontale ('*droite'* versus '*non droite'*). En revanche, aucune focalisation préalable d'un des éléments du domaine n'est nécessaire.

Démonstratifs. Un démonstratif tel que *cette pyramide* cherchera à s'interpréter dans un domaine comportant obligatoirement un élément identifiable autrement que par la désignation elle-même, par exemple par une focalisation discursive (thème du discours, …) ou un geste de désignation associé (Corblin, 1987). Par ailleurs, il doit être possible de (re-)classifier cet élément comme *pyramide*. L'extension du domaine d'interprétation est ici déterminée a posteriori, puisqu'elle dépend du nombre d'objets pouvant être recouverts par la désignation employée. Référer à la petite pyramide de la figure 1 par *cette pyramide* (+ geste de désignation) ou par *cette figure* (+ geste de désignation) aurait en effet des conséquences différentes sur

5. Nous utilisons les notions « type » et « propriété » en tant que caractéristiques des objets par rapport à une ontologie extralinguistique. Le passage des lexèmes *N* (*triangle*) et *P* (*grand*) aux types et propriétés de l'ontologie '*N'* ('*triangle'*) et '*P'* ('*grand'*) se fait selon un lexique spécifique associée à l'ontologie (cf. section 4.3).

l'interprétation d'une expression telle que *les autres* dans la suite du dialogue : dans le premier cas, celle-ci désignerait tous les triangles, et dans le deuxième cas tous les objets de la scène.

Pronoms personnels. Un pronom demande toujours un domaine comportant un élément focalisé préalablement. Cette contrainte est compatible avec les observations linguistiques sur le fonctionnement des pronoms en français de Kleiber (1994) ainsi qu'avec les principes de la théorie du Centrage (Grosz *et al.*, 1995).

3.2.2. *La sémantique : contraintes sur le typage et le critère de différenciation*

Indéfinis. Le domaine dont un indéfini *n N P* (*un grand triangle*) extrait des éléments doit être un domaine comprenant des éléments de type '*N*' ('*triangle*') ayant, le cas échéant, la propriété *'P'* ('*grand*').

Définis. Pour un défini nu *le N* (*la pyramide*), la singularisation du référent dans son domaine passe par la catégorisation de l'objet en tant que *'N'*. Cela signifie qu'un défini nu s'interprétera dans un domaine opposant un élément de type *'N'* à des éléments de type *'non N'*. Ce domaine doit donc être partitionné selon un critère de différenciation qui est le type de ses éléments. Par conséquent, le type du domaine peut être soit un sur-type de *'N'* (le référent de *la pyramide* s'extrait d'un domaine de '*corps géométriques*'), soit le type correspondant à un objet dont le référent est une partie (le référent de *la pyramide* s'extrait d'un domaine '*désert*'). Le calcul des relations entre *type* et *sous-type* et des relations entre *partie* et *tout* se fait également sur la base des informations fournies par l'ontologie (section 4.3). Pour un défini déterminé par un adjectif (*le N P*), la singularisation passe par la valeur *'P'* pour un attribut particulier. Le critère de différenciation est donc cet attribut, le type commun des éléments du domaine est *'N'*.

Démonstratifs. Corblin (1987) fait l'hypothèse qu'un démonstratif *ce N* (*cette pyramide*) recrute son référent sur des critères externes à la désignation elle-même. La désignation *N* serait alors disponible pour une reclassification éventuelle du référent en tant que '*N*'. Cela signifie pour notre modélisation qu'un démonstratif n'impose pas de contraintes particulières sur le typage des éléments de son domaine, à partir du moment où sa désignation est jugée compatible avec le type du référent. Ce jugement de compatibilité peut reposer sur des informations fournies par l'ontologie associée à la tâche (cf. section 4.3) : il s'agit soit de relations hiérarchiques de typage (un type est compatible avec ses sous-types : '*chat*' – *'animal'*), soit de connaissances spécifiques à l'application (dans le corpus Ozkan, on observe des reclassifications entre éléments géométriques et figuratifs telles que : '*rond*' – *'soleil'*).

Pronoms personnels. Un pronom personnel n'impose pas de contraintes sur le type des éléments de son domaine. Les seules informations disponibles sont le genre grammatical qui doit être compatible avec une des désignations possibles pour le référent.

3.2.3. *Les instructions de parcours particulières*

Un premier aspect intéressant de notre modélisation est l'intégration aisée des contraintes liées à des expressions souvent considérées comme « particulières », telles que *les autres triangles*, *la première pyramide* ou *la pyramide la plus grande*. Or le point commun de ces expressions est de présupposer explicitement un domaine de référence partitionné selon des critères de différenciation spécifique : *les autres triangles* présuppose un domaine de type '*triangle'* ayant préalablement été partitionné selon un critère quelconque ; *la première pyramide* présuppose un domaine de type '*triangle'* dont les éléments peuvent être distingués selon un ordre ; *la pyramide la plus grande* présuppose un domaine de type '*triangle'* partitionné selon la taille de ses éléments.

Si on applique tous les principes présentés dans cette section à l'interprétation de la pyramide de droite (figure 1), on obtient comme domaine sous-spécifié un domaine de type 'triangle' présupposant une partition sur un critère de différenciation qui est l'alignement horizontal et permettant d'opposer un élément de droite à des éléments qui ne soient pas de droite (cf. figure 6).

4. Support contextuel

Dans la section précédente, nous avons exposé les principes qui nous permettent de dériver, à partir de la détermination et de la sémantique d'une expression référentielle, des domaines de référence sous-spécifiés, exprimant des contraintes à projeter sur des contextes disponibles. Nous examinons ici plus en détail les critères linguistiques, visuels et liés à la tâche qui sous-tendent la construction de tels espaces de projection.

4.1. *L'historique langagier*

4.1.1. *La création de domaines*

Au cours du dialogue, de nouveaux espaces de projection – des domaines de référence – sont créés dynamiquement à partir d'indices linguistiques : ainsi, en [1], *un triangle rouge* et *un triangle vert* forment un domaine dont la prise en compte est nécessaire pour opérer des extractions ultérieures lors de l'interprétation de *le triangle rouge* et de *l'autre* [1a]. Par ailleurs, ce type de groupement crée des domaines dont les entités sont, sous certaines conditions, inaccessibles individuellement pour des reprises pronominales atoniques [1b].

Prends *un triangle rouge* et *un triangle vert.* [1]

(a) Mets *le triangle rouge* sur la droite. Supprime *l'autre.*

(b) * Mets-*le* sur la droite.

La question porte, par conséquent, sur les facteurs susceptibles de déclencher ces opérations de groupement. Les principaux critères sont des facteurs linguistiques (essentiellement syntaxiques) et discursifs (structures rhétoriques).

Dans les facteurs linguistiques, l'énumération et la coordination par la conjonction *et* sont les critères les plus importants. Kleiber (1986 : p. 77) souligne que « ce ne sont pas deux nouveaux référents qui sont en fait introduits, mais bien un seul référent, en l'occurrence l'ensemble des référents constitué par la coordination ». Par conséquent, certaines modélisations en sémantique discursive proposent des mécanismes de groupement adéquats : la DRT (Kamp & Reyle, 1993) par exemple, prévoit une opération de sommation permettant de regrouper les référents introduits par un groupe nominal coordonné. Dans l'exemple [1a], cela conduirait à la création d'un nouveau référent discursif pour *un triangle rouge et un triangle vert*. L'avantage de cette opération est de mettre à disposition une entité complexe sur laquelle pourra ensuite être résolu un pronom pluriel *ils*. En revanche, l'accès individuel à ses composantes n'est pas pour autant bloqué. Cela signifie qu'un enchaînement par reprise pronominale tel que [1b] est autorisé, alors qu'il semble impossible ou très difficile, comme le constate Kleiber (1986).

En plus de la coordination, d'autres facteurs linguistiques contribuent à la création d'ensembles référentiels. Parmi ceux-ci, on trouve la structure argumentale d'un prédicat. En effet, beaucoup de modélisations – dont l'algorithme de Sidner (1979), la DRT ou la théorie du centrage (Grosz *et al.*, 1995) – prévoient, d'une façon ou d'une autre, la possibilité de regrouper les participants d'une même éventualité. La théorie du centrage en fait même son principe de structuration contextuelle principal. Parmi les travaux moins formalisés, on pourra mentionner la grammaire cognitive (Langacker, 1991) qui modélise les processus de compréhension par une élaboration de positions sous-spécifiées (une sorte de « remplissage ») dans des schémas abstraits représentant des éventualités (« relations » ou « processus ») par des schémas plus concrets (« choses »).

A un niveau supérieur au segment phrastique élémentaire, on trouve des propositions de regroupement sur des critères syntaxiques et discursifs. Les critères syntaxiques sont formulés de façon élégante par la DRT : en fonction de certaines constructions (conditionnelles, temporelles et aspectuelles), des segments forment des complexes qui restreignent l'accessibilité de leurs référents discursifs. L'avantage de ces conditions de groupement est leur calculabilité. L'inconvénient par rapport à notre objectif de dialogue homme-machine est leur taux d'occurrence très faible dans nos corpus de dialogues, ainsi que leur dépendance d'une grammaticalité normative, pas toujours assurée dans la langue orale.

Parmi les travaux les plus représentatifs des groupements discursifs, il y a ceux de Webber (1991) et de Asher (1993) : leur objectif est de construire une représentation hiérarchique du discours, reflétant des relations de coordination ou de subordination entre des segments. Idéalement, cette représentation serait prédictive quant à la reprise des entités discursives, mais la calculabilité de cette structure reste

un problème essentiel de ces approches : Webber (1991) suppose, pour résoudre ce problème, rien de moins qu'un « oracle ».

Nous concevons que tous ces critères sont complémentaires pour la création de domaines de référence d'origine linguistique et que les meilleurs systèmes de résolution référentielle seront ceux qui tenteront de les combiner. En revanche, dans l'état actuel des connaissances, les seuls groupements pouvant être calculés de façon réaliste sont ceux fondés sur les critères de coordination et d'énumération, et dans une moindre mesure, sur la structure argumentale.

4.1.2. *La focalisation d'un élément*

Le calcul d'un domaine résultant d'un groupement implique le calcul d'un type commun des entités regroupées ainsi que le calcul d'une partition. Parmi les informations liées à la partition, la structure focale – le marquage éventuel d'un des éléments d'une partition en tant qu'élément le plus saillant et donc le plus accessible – mérite une attention particulière, dans la mesure où elle est un des éléments-clés de notre modélisation.

Groupement par coordination ou énumération. Les travaux de Gernsbacher & Hargeaves (1988) suggèrent qu'il y a une dissymétrie entre les éléments groupés par coordination, dans la mesure où le référent de la première mention garde dans tous les cas le bénéfice de la saillance. Transposé à notre modélisation, cela signifierait qu'un domaine issu d'un groupement par coordination comporte toujours un élément focalisé, correspondant au référent de la première mention. Or, cette hypothèse est en contradiction avec les observations de Kleiber (1986). Elle ne parvient effectivement pas à prédire la difficulté de la reprise pronominale atonique d'un des éléments du groupement, fût-ce le premier mentionné (cf. l'exemple [1b]). Il semble donc bien que les groupes coordonnés bloquent l'accès individuel à leurs constituants pour des reprises pronominales. Nous proposons alors la modélisation suivante : aucun des éléments du domaine issu d'un groupement déclenché par coordination ne reçoit une focalisation particulière.

Groupement des arguments d'un même prédicat. En ce qui concerne le groupement déclenché par le prédicat regroupant ses arguments, nous pouvons rappeler ici l'algorithme de Sidner (1979) qui distingue déjà le focus des agents des autres focus, eux-mêmes ordonnés selon leur rôle thématique, ou encore la modélisation de la théorie du centrage (Grosz *et al.*, 1995) qui ordonne les centres en avant selon la hiérarchie des fonctions syntaxiques : *sujet* > *objet direct* > *objet indirect*. Par ailleurs, l'idée d'une hiérarchisation des rôles thématiques a aussi influencé des travaux dans le cadre de la grammaire cognitive (Langacker, 1991) : l'élaboration de schémas complexes s'accompagne d'une distinction entre le « profil » et la « base » à l'intérieur des schémas composés. Le profil correspond à des entités plus proéminentes, définies en fonction du schéma élaboré. Si cette hiérarchisation correspond, *grosso modo*, à la hiérarchie des rôles grammaticaux, le cadre théorique de la grammaire cognitive permet d'aller plus loin, en proposant des

explications unifiées de différents problèmes liés à l'interprétation anaphorique de pronoms : Van Hoek (1995) a ainsi montré que des phénomènes relevant à la fois de la syntaxe (C-commande ; Reinhart, 1983) et de la pragmatique pouvaient s'expliquer de façon uniforme sur la base de cette hiérarchisation, car elle est capable de dépasser le cadre phrastique. Pour résumer les travaux sur la focalisation lors d'un regroupement des arguments du même prédicat, on peut constater un consensus large sur l'importance accordée à une hiérarchisation en fonction du rôle thématique. Cela signifie pour notre modélisation que ce type de regroupement, contrairement au regroupement déclenché par la coordination, implique la création d'un domaine focalisé. Le choix de l'élément focalisé dépend de la structure argumentale du prédicat et suit une hiérarchisation des fonctions syntaxiques (*sujet* > *objet direct* > ...) ou des rôles thématiques (*agent* > *patient* > ...).

4.2. *La perception visuelle*

4.2.1. *La création de domaines*

La psychologie de la forme, ou Gestalt, postule que notre perception d'une scène visuelle a pour résultat immédiat, non pas la perception distincte de chacun des objets qu'elle comporte, mais la perception globale de groupements d'objets. A partir de (Wertheimer, 1923), un certain nombre de critères de groupement ont été proposés, les plus importants étant la proximité (des objets proches forment facilement un groupe perceptif), la ressemblance ou similarité (des objets similaires se regroupent), et la bonne continuité (des objets présentant une continuité dans leur disposition se regroupent).

De nombreux travaux s'attachent à formaliser ces notions psychologiques pour en produire des algorithmes. La plupart se limitent cependant à un seul critère, souvent la proximité. Citons par exemple les travaux très fouillés du *Kubovy Perception Lab* (Kubovy & Wagemans, 1995) ou de Feldman (1997). La thèse de Briffault (1992) présente également une méthode de groupement suivant la proximité des objets, consistant à construire un graphe à partir de la relation « est le plus proche de » et à identifier les sous-graphes connexes qui constituent alors les groupements perceptifs. Une structuration hiérarchique de groupes est obtenue en extrayant dans chaque sous-graphe connexe les deux objets pour lesquels la relation s'applique dans les deux sens, et – à un niveau supérieur – en regroupant deux à deux les groupes les plus proches. Les inconvénients de cette méthode sont liés à son aspect artificiel et mathématique : la relation « est le plus proche de » fonctionne en tout ou rien, sans degrés de proximité, et peut relier de la même façon des situations très différentes en terme de perception, aboutissant alors à des groupements peu plausibles. Un avantage de cette méthode est par contre le codage des relations entre objets. Il n'est pas systématique mais fonction d'une loi de probabilité : plus la distance entre deux objets est élevée, moins il y a de chances pour que la relation soit codée. La représentation partiellement hiérarchique obtenue permet d'optimiser les

performances de certaines opérations spatiales, ce qui contribue à une meilleure plausibilité cognitive. La caractéristique principale de l'approche de Briffault (1992) réside surtout dans la gestion de plusieurs structurations des objets. Les critères à l'origine de ces structurations sont, outre la proximité : les relations de contact, d'inclusion ou de contenance entre objets, la détection des configurations linéaires, triangulaires, rectangulaires ou circulaires (ce qui correspond en partie au critère de bonne continuité de la Gestalt). Si cette approche permet de traiter efficacement les expressions spatiales, elle ne nous semble pas adaptée à notre approche qui, d'une part n'aborde pas encore le traitement des expressions comportant une relation spatiale entre deux objets, et d'autre part a pour ambition l'intégration de plusieurs critères perceptifs dans une même structuration. Une autre formalisation souvent citée dans le domaine du dialogue homme-machine est celle de (Thórisson, 1994) qui présente les avantages de combiner deux critères importants, la proximité et la similarité, et de donner pour résultat une liste *ordonnée* de groupes. Deux inconvénients nous paraissent cependant importants compte tenu de notre approche : l'absence de structuration arborescente des groupes (nos partitions), et la perte de l'identité du facteur de groupement (notre critère de différenciation).

Nous avons proposé dans (Landragin *et al.*, 2001) une formalisation à base de dendrogrammes, focalisée sur la proximité. Nous voulons ici étendre ce principe aux trois critères cités. A partir de la liste des objets visibles, de leurs coordonnées et de leurs caractéristiques physiques, nous construisons un dendrogramme pour la proximité, un second pour la similarité, et un troisième pour la bonne continuité. Le premier est obtenu, après calcul des distances entre toutes les paires possibles d'objets, par un algorithme classique de classification automatique (Belaïd & Belaïd, 1992), et correspond à une hiérarchie de partitions possibles de la scène, chaque partition étant caractérisée par son indice d'agrégation. La figure 2 donne un exemple sur une scène comportant cinq objets. Le second dendrogramme est obtenu par le même algorithme, non pas à l'aide de la distance cartésienne, mais à l'aide d'une échelle de propriétés, en commençant par la plus susceptible de rendre deux objets similaires (par exemple : *forme* > *couleur* > *taille* >...). Le troisième dendrogramme est obtenu à l'aide d'un algorithme récursif étendant un groupe à l'objet le plus proche, et détectant les continuités à l'aide de régressions linéaires. Les groupes ainsi obtenus sont caractérisés par la qualité (ou pertinence) de l'alignement de leurs éléments. Cette pertinence est maximale lorsque par exemple les objets du groupe sont parfaitement alignés, ou répartis de manière circulaire et régulière. A ce stade, nous obtenons des domaines de référence, éventuellement étiquetés par un indice de pertinence, et structurés dans une forêt de dendrogrammes.

Dans l'exemple de la figure 1, les deux objets les plus proches sont les deux triangles de droite sur lesquels porte l'ambiguïté. Un premier niveau dans le dendrogramme de proximité consiste donc en une partition de la scène en trois groupes, un premier avec ces deux triangles, un deuxième avec le rond, et un troisième avec le triangle de gauche. Un deuxième niveau voit le regroupement des trois triangles face au rond. Le dendrogramme de similarité est plus intéressant car il

regroupe dans un premier temps les deux grands triangles, comme le fait l'utilisateur en I2, avant de regrouper les trois triangles. Le dendrogramme de bonne continuité regroupe très rapidement les trois triangles, ce qui renforce, pour cette tâche consistant à dessiner un horizon, l'ancrage référentiel sur les triangles.

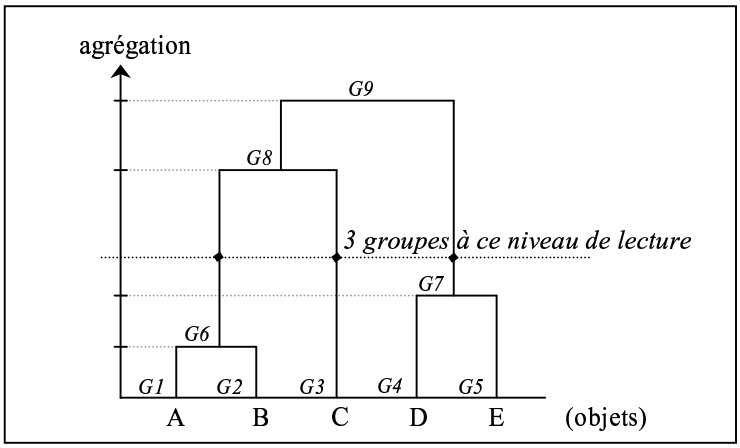


Figure 2. *Dendrogramme pour la proximité*

4.2.2. *La focalisation d'un élément*

L'exploitation des dendrogrammes a pour point de départ un objet (ou un ensemble d'objets) mis en avant, soit par un geste de désignation de la part de l'utilisateur (cf. section 5.1), soit par une particularité inhérente à la représentation visuelle de l'objet, qui le rend saillant dans le contexte courant. Cette saillance visuelle est exploitée par exemple lorsqu'il s'agit de trouver une interprétation privilégiée à l'expression *le N* dans une scène comportant plusieurs objets du type N. Par opposition au geste, la focalisation est ici implicite et doit être retrouvée par le système à partir des données qu'il possède sur les objets visibles et sur les particularités de leur affichage dans la scène. La méthode que nous proposons consiste à tester, dans un ordre précis, les caractéristiques visuelles des objets. Dès que l'on identifie un objet qui prend une valeur singulière pour la caractéristique correspondant à l'étape courante, cet objet est étiqueté comme étant le plus saillant. L'ordre des caractéristiques est le suivant (Landragin *et al.*, 2001) :

– au niveau des propriétés de l'objet : catégorie ; luminosité ; caractéristiques physiques (taille, géométrie, matériau, couleur, texture) ; orientation ; incongruité ; aspect énigmatique ; dynamique ;

– au niveau de la disposition spatiale de l'objet dans la scène : proche ; isolé ;

– au niveau de la lecture visuelle de la scène : à la place d'un point fort (c'est-à-dire aux intersections des lignes horizontales et verticales qui partagent le cadre en tiers) ; dans le prolongement ou à l'intersection de lignes directrices qui contraignent le parcours du regard ; à la place d'un point de fuite (lorsque la perspective est très présente dans la scène) ; à la place d'un point d'équilibre (lorsque la scène présente une symétrie qui repose sur cet équilibre).

Dans l'exemple de la figure 1, aucune des quatre expressions référentielles retenues ne fait appel à la saillance. Si on applique néanmoins notre algorithme aux quatre objets que sont les trois triangles et le rond, c'est ce dernier qui est tout d'abord identifié comme saillant, étant le seul de sa catégorie. En imaginant que l'utilisateur ait employé l'expression *le triangle*, c'est le petit triangle qui serait alors identifié comme saillant, étant le seul de sa taille.

4.2.3. *Les différentes façons d'exploiter les dendrogrammes*

Une première façon de faire appel au module relatif à la perception visuelle consiste à lui demander de construire les dendrogrammes sans focaliser d'élément. Dans certains cas d'expressions référentielles, seul un dendrogramme a besoin d'être construit. C'est le cas des expressions comportant une mention explicite du critère de groupement des référents, telles que : *les objets agglutinés* ou *les objets qui forment une ligne* (exemples tirés du corpus Magnét'Oz). Les domaines obtenus par le critère de bonne continuité ont un intérêt particulier : l'ordonnancement marqué de leurs éléments peut être exploité lors d'expressions telles que *le premier en partant de la gauche* ; *la troisième boîte* ou *la boîte suivante* (exemples de Magnét'Oz).

Une deuxième façon de faire appel à notre module consiste à lui demander de construire les dendrogrammes à partir de la focalisation d'un élément (sur la base d'un geste de désignation effectué sur cet élément, ou de sa saillance inhérente). Nous présentons ici un algorithme permettant de combiner les résultats des trois dendrogrammes en partant de cet élément. Le but est d'obtenir une liste de domaines de référence possibles, chacun de ces domaines étant caractérisé par l'élément focalisé et par un ou plusieurs facteurs de groupement perceptif. Deux méthodes sont possibles selon que l'on dispose ou non d'une correspondance entre les mesures des trois dendrogrammes.

Sans correspondance des mesures. Étendre la liste des objets focalisés à un ensemble cohérent de référents potentiels se fait par l'intersection des trois groupements de plus bas niveau (un pour chaque critère de la Gestalt), ce qui fait jouer les trois critères simultanément. Ceci correspond au premier niveau d'extension, le second consistant à intersecter deux groupements, et le troisième à étendre à l'un ou l'autre des groupements (ce qui revient à ne plus faire jouer qu'un seul critère). Cette méthode propose autant de résultats qu'il y a de combinaisons de critères, à savoir sept résultats pour trois critères. Les problèmes qu'elle présente sont liés à la pertinence des critères : d'une part un critère peut être plus pertinent que celui identifié comme prépondérant lors d'une intersection (en particulier

lorsque l'intersection revient à une inclusion) ; d'autre part un critère peut rester pertinent en remontant un niveau dans son dendrogramme, processus que l'on n'a fait pas ici, car il augmente considérablement le nombre de résultats.

Avec correspondance des mesures. Une solution aux problèmes de la méthode précédente consiste à pondérer les critères selon leur pertinence compte tenu de l'application et des types d'objets. Si l'application consiste par exemple en une succession de scènes basées sur une même logique à base d'ordonnancements des objets, le critère de bonne continuité sera privilégié. Si les objets sont tous très similaires, la proximité jouera sans doute un rôle plus important que la similarité et devra donc être pondérée de façon à être privilégiée. Lorsque les spécifications de l'application permettent de définir les poids de chaque critère, c'est-à-dire lorsque le système dispose d'une correspondance entre les unités de chaque critère et peut comparer les résultats des trois dendrogrammes, l'extension de l'ensemble des objets focalisés à un ensemble de référents potentiels se fait de la manière suivante : on classe sur un même axe les partitions fournies selon les trois critères ; on part de l'ensemble des objets focalisés et on réunit les groupes qui les contiennent ; on remonte selon l'axe et on prend en compte – de manière additive – le critère correspondant à la partition suivante (qui peut être le même critère que précédemment, mais avec une agrégation supérieure) ; on obtient ainsi un résultat étiqueté par une combinaison de critères ; on remonte selon l'axe jusqu'à ce que l'on atteigne la dernière partition dans le classement. Les avantages de cette méthode sont multiples : elle est ouverte à la prise en compte d'un critère supplémentaire de groupement ; plusieurs niveaux de granularité pour un même critère sont correctement gérés ; et, surtout, l'ordre d'obtention des résultats est celui de leur pertinence. Comme inconvénient, citons qu'une exécution en temps réel suppose un nombre raisonnable d'objets (pour une application gérant un très grand nombre d'objets, c'est tout le module perceptif avec le choix de l'algorithme de classification automatique qui est à revoir).

Dans l'exemple de la figure 1, en considérant un énoncé tel que *ces triangles* avec un geste de désignation sur le triangle le plus à gauche, le premier résultat sera l'ensemble des deux grands triangles ou l'ensemble des trois triangles, selon que l'on privilégie le critère de bonne continuité ou celui de similarité : la bonne continuité donne en premier lieu l'ensemble des trois triangles ; la similarité (combinée avec la proximité) donne en premier lieu l'ensemble des deux grands triangles.

4.3. *La tâche*

Dans une application de commande telle que celle que nous considérons ici, à savoir le positionnement d'objets sur une scène, la tâche joue un rôle important pour contraindre les objets disponibles et les actions exprimables. En partant des travaux de Grisvard (2000) qui ont montré comment différents niveaux dialogiques, y compris celui de la tâche, peuvent s'exprimer dans un formalisme proche de nos

domaines de référence, il convient de déterminer ce que peut être précisément un modèle de la tâche et comment celui-ci peut intervenir dans la construction des domaines et dans la focalisation d'éléments dans ces domaines. Trois notions se distinguent dans un modèle de la tâche : la gestion des fonctionnalités, l'ontologie des objets, et la structuration en buts et sous-buts. Elles interviennent toutes dans la construction de domaines de référence. Ainsi, des domaines vont rassembler les objets auxquels on peut appliquer les mêmes opérations. Les fonctionnalités jouent alors le même rôle de filtrage que la catégorie au niveau langagier.

L'ontologie des objets repose essentiellement sur une hiérarchie de types, assortie de mécanismes d'inférence[6] permettant par exemple de comprendre des reclassifications telles que *un chat – l'animal*. Les décompositions partie-tout apportent une dimension supplémentaire en autorisant la création d'une partition permettant de relier deux référents faisant l'objet d'une désignation associative (*le triangle – les segments* ; *la maison – la porte*). L'ontologie va contraindre les possibilités de création de nouvelles partitions à l'intérieur d'un objet donné aux seules structures décrites pour l'un des types dont relève l'objet. Enfin, l'ontologie devrait, idéalement, permettre de comprendre des reclassifications plus complexes : le passage de *triangle* à *pyramide* dans l'exemple de la figure 1 montre les limites d'une ontologie statique et générique (ces reclassifications ne peuvent pas être généralisées). Même à l'intérieur d'une tâche simple comme celle du corpus Ozkan, la reprise de *triangle* par *pyramide* n'est pas valide dans les mêmes conditions que celle de *triangle* par *toit*. Se pose donc le problème de l'acquisition d'ontologies spécifiques, rendu encore plus difficile par la dynamique du processus : on observe en effet que ces reclassifications dépendent largement de l'état d'avancement des connaissances du manipulateur et de l'évolution du contexte visuel.

La structuration en buts et sous-buts est sans doute la notion la plus importante, car elle va permettre de contraindre ou de privilégier une interprétation pour l'énoncé à venir. Les travaux de Grosz ont fait apparaître que la structure de la tâche, c'est-à-dire une hiérarchie intentionnelle, permet de créer des espaces attentionnels, limitant l'ensemble de recherche pour les antécédents d'une expression donnée. Or, comme l'a fait remarquer Gaiffe (1992), l'application de cette proposition suppose une tâche fortement hiérarchique, bien définie, connue et exécutée dans un ordre strict par les sujets. Cela est loin d'être le cas pour des tâches tout-venant, mais ce fait ne doit pas empêcher de faire jouer ce facteur, en combinaison avec d'autres, lorsque cela est possible. En pensant au corpus Ozkan, il mériterait en effet d'être exploité, car l'ordre de construction des éléments des dessins a été prédéfini (bien que pas toujours respecté), et de plus, la clôture des sous-tâches (pose d'un élément au bon endroit, fin de la construction d'un élément figuratif) est très souvent linguistiquement marquée. La prise en compte de tels facteurs permettrait par exemple de résoudre l'ambiguïté sur l'interprétation du pronom *le* dans

6. Le format de représentation des ontologies (réseaux sémantiques, graphes conceptuels, etc.) dépend du contexte concret de l'implémentation. Le projet MIAMM utilise par exemple des logiques de descriptions (RACER : http://kogs-www.informatik.uni-hamburg.de/~race/).

l'exemple [2] : étant donné la marque de clôture pour la pose d'un premier élément (*voilà*), le pronom peut être considéré comme ayant moins de probabilité de référer à l'instance qui vient d'être manipulé qu'au type dont relève cette instance.

Ensuite on a les pyramides dans le désert [...] [2]

Il faut que tu prennes *le gros triangle*.

Voilà et tu *le* reprends une deuxième fois un peu décalé par rapport au premier.

D'autre part, la structuration en une succession d'actions permet une structuration en une succession d'ancrages référentiels. Si l'on considère à nouveau l'exemple de la figure 1, on remarque que l'instructeur construit la ligne d'horizon en commençant par le segment ancré sur les deux grands triangles, en continuant avec le segment ancré sur celui de gauche, puis avec celui ancré sur celui de droite et sur le troisième triangle, pour terminer par le segment ancré sur ce troisième triangle. Présentée ainsi, la succession d'ancrages est presque prévisible : il semble logique qu'après s'être appuyé sur deux objets puis sur l'un des deux, l'instructeur continue par un acte de référence ancré sur le deuxième. Ceci va dans le sens de l'interprétation de *la pyramide de droite* dans le domaine comportant les deux grands triangles.

5. Mécanisme d'ancrage

Nous nous focalisons maintenant sur la réalisation de l'ancrage référentiel, par confrontation du domaine de référence sous-spécifié venant de l'expression référentielle et des domaines obtenus par les trois sources étudiées précédemment. Suivant le modèle du producteur-consommateur et selon une architecture modulaire avec noyau central, nous nous plaçons du point de vue de ce noyau – ou contrôleur de dialogue – qui va consommer les informations produites par le module lié à la sous-spécification, envoyer des requêtes aux modules liés aux sources contextuelles, confronter les résultats de ceux-ci, et, en cas d'ambiguïté, leur demander de refaire leur calcul sur la base de nouvelles directives.

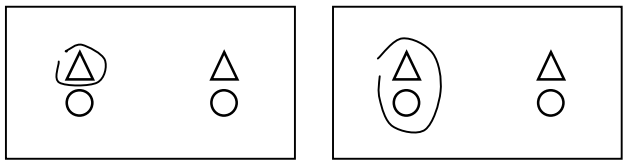
5.1. *L'exploitation du geste de désignation*

L'ancrage d'une expression référentielle dans un contexte a pour point de départ la focalisation d'une partie de ce contexte. En considérant le contexte visuel, il existe deux critères importants de focalisation : le geste de désignation (explicite) et la saillance visuelle (implicite). Nous présentons dans cette section le geste de désignation comme un point de départ dans notre modélisation. Nous nous appuyons pour cela sur l'étude du corpus Magnét'Oz (Wolff, 1999) qui regroupe énoncés oraux, trajectoires gestuelles et caractéristiques successives de la scène visuelle. Ces

données ont été enregistrées au cours d'une simulation de type magicien d'Oz (un humain simule la machine dans le but de tester le comportement d'un utilisateur avant le développement d'un système de dialogue). L'interaction – spontanée – se basait sur un écran tactile, ce qui a permis d'enregistrer les gestes effectués. Les scènes visuelles comportaient des objets manipulables et des « distracteurs » ni manipulables ni même catégorisables (formes abstraites). Cette expérimentation a permis le recueil de situations de références multimodales, et l'étude de Wolff (1999) a abouti à une caractérisation des trajectoires gestuelles et à un modèle d'interprétation de ces trajectoires en contexte visuel.

A partir de la catégorisation des trajectoires gestuelles (essentiellement en pointages et en entourages) ; de la catégorisation des expressions référentielles selon leur détermination (essentiellement en démonstratifs et en définis dans Magnét'Oz) et selon la présence ou non d'une catégorie et de modifieurs ; à partir aussi des critères de la Gestalt (présence ou non d'un groupe perceptif), nous avons extrait des situations prototypiques, par exemple : pointage + démonstratif + catégorie + groupe perceptif. Plus que représenter le corpus, ces situations nous semblent décrire des configurations correspondant à des modes de fonctionnement cognitif. Pour chacune d'elles, nous avons reconstruit un exemple prototypique, généralement en reprenant et simplifiant un extrait du corpus. Nous faisons l'hypothèse qu'un modèle d'interprétation de la référence fonctionnant correctement sur ces exemples prototypiques fonctionnera sur l'ensemble des situations qui en sont à l'origine.

La figure 3 présente deux de ces exemples. Une attention particulière dans leur construction a été donnée à la présence d'au moins un deuxième objet de la catégorie du référent et d'au moins un objet d'une autre catégorie, ceci pour vérifier le pouvoir prédictif de notre modèle face aux différentes formes de reprise et de continuité que peut prendre l'énoncé ultérieur : dans l'exemple [3a], l'expression référentielle *l'autre* se vérifiera dans le domaine de référence correspondant au contexte visuel complet et désignera le deuxième triangle, tandis que l'expression *le rond* pourra se vérifier de manière privilégiée dans le domaine correspondant au groupe perceptif de gauche, groupe focalisé par le geste de la première expression référentielle.

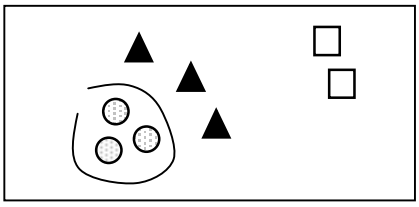


(a) *ce triangle* (+ geste) (b) *le triangle* (+ geste)

Figure 3. *Vérification de la détermination avec un geste d'entourage*

Les deux exemples de la figure 3 illustrent également le fonctionnement du défini et du démonstratif : dans l'exemple [3a] le geste isole un des deux triangles et prend ainsi avec le démonstratif le rôle de critère de différenciation dans le domaine correspondant au contexte visuel complet ; dans l'exemple [3b] le geste désigne cette fois un domaine de référence dans lequel le défini s'applique pour extraire l'unique objet de type '*triangle*'.

Le geste peut donc intervenir à deux niveaux : celui de la délimitation d'un domaine, ou celui de la focalisation d'un élément dans un domaine implicite. La vérification des contraintes de l'expression référentielle permet dans le premier cas d'extraire les référents du domaine délimité, et dans le second cas d'identifier le domaine. Cette dernière opération est cependant difficile : si dans l'exemple [3a] et les exemples similaires on identifie par défaut le contexte visuel complet, l'exemple de la figure 4 (tiré également de Magnét'Oz) montre un cas extrême d'exploitation des contraintes de l'expression référentielle. En effet, quand on l'applique à l'ensemble des objets délimités par l'entourage, l'expression *les formes les plus claires* ne se vérifie pas car les trois ronds ont la même couleur ; et quand on l'applique au contexte visuel complet, elle désigne plutôt les deux carrés blancs que les trois ronds (qui sont légèrement grisés). On se trouve dans une situation intermédiaire, où le geste identifie un ensemble de référents, et où le domaine de référence se situe quelque part entre cet ensemble et le contexte visuel complet. Son identification se fait alors en étendant l'ensemble désigné au groupe perceptif dans lequel il est inclus (et ce en remontant dans les dendrogrammes jusqu'à ce que le superlatif soit vérifié). On obtient un domaine de référence comprenant les trois ronds et les trois triangles, dans lequel les formes les plus claires sont bien les ronds désignés.



Les formes les plus claires (+ geste)

Figure 4. *Geste désignant les référents et initiant la construction du domaine*

L'identification du niveau auquel le geste intervient peut cependant être difficile, ceci à cause de l'ambiguïté de portée inhérente au geste. D'une part il est souvent imprécis, surtout en communication homme-machine lorsqu'il est effectué sur un écran tactile ou via tout dispositif nécessitant un calibrage délicat (gant numérique). D'autre part, contrairement au geste qui délimite, le geste qui pointe ne donne

aucune indication sur sa portée : il peut indiquer aussi bien un point précis ou l'objet le plus proche de ce point, qu'une zone spatiale très étendue (cf. l'exemple de (Romary, 1993) : *mets de la moquette ici* + pointage). C'est son association avec l'expression référentielle et les particularités du contexte visuel qui permettent de déterminer si le geste désigne des référents ou un domaine. En particulier, une expression au pluriel peut s'interpréter comme la référence à un groupe perceptif complet, faisant ainsi intervenir efficacement notre structuration du contexte visuel en domaines.

Cette caractérisation des rôles que peut prendre le geste donne un début de stratégie d'interprétation des références multimodales, que nous développons maintenant en tenant compte de toutes les sources contextuelles.

5.2. *Le contrôleur de dialogue et ses requêtes*

Le point de départ est la réception par le système d'un énoncé, langagier ou multimodal, de la part de l'utilisateur. Avant cela, aucun calcul n'est effectué par le module s'occupant de la perception visuelle, et les différents historiques sont prêts, ayant été mis à jour suite au traitement de l'énoncé précédent. Comme nous l'avons fait depuis le début, nous nous focalisons ici sur l'interprétation des expressions référentielles et nous laissons de côté le reste de l'énoncé oral. Notons seulement qu'une contrainte sur la fonctionnalité des objets cherchés va être construite à partir du prédicat, cette contrainte étant ensuite exploitée lors des différents filtrages appliqués dans les domaines pour en extraire les référents.

Pour élaborer le domaine sous-spécifié à la base des requêtes du contrôleur de dialogue, un module, noté *module sous-spécification* dans la figure 5, s'appuie sur notre modèle décrit section 3.2. Comme l'indiquent les liens en pointillé, ce module collabore avec le *module langagier* et le *module visuel*, le premier pour l'indication des principes linguistiques de construction de domaine, le second pour une première interprétation du geste en contexte visuel : selon la forme du geste et la disposition des objets par rapport à cette forme, c'est la délimitation du domaine, la focalisation des référents, ou les deux hypothèses qui vont être attribuées au geste, comme nous l'avons vu section 5.1.

A partir du domaine sous-spécifié obtenu, le contrôleur de dialogue émet une requête en parallèle aux trois modules : *langagier*, *visuel* et *tâche*. Il reçoit en retour trois listes ordonnées de domaines, certains d'entre eux ayant éventuellement un référent focalisé. Il confronte ces résultats, éventuellement par le moyen de scores numériques : selon la place d'un résultat dans les différentes listes, un nombre entre 0 et 1 est attribué à ce résultat, indiquant sa pertinence et permettant de le classer dans une seule liste par rapport aux autres résultats possibles. En étudiant dans chacun d'eux l'identité des référents et l'étendue du domaine de référence, les résultats sont exploités de la manière suivante :

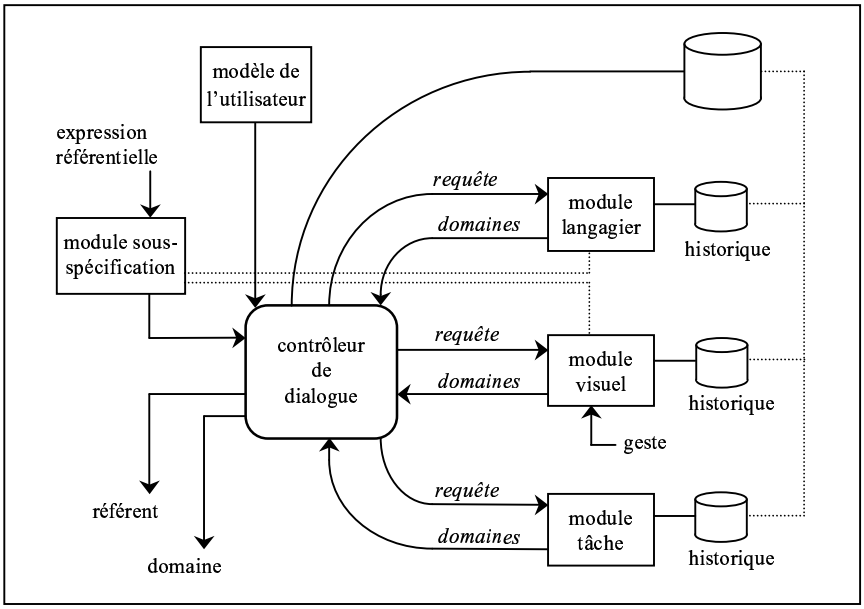


Figure 5. *Architecture logicielle pour la résolution de la référence*

– si des référents différents sont obtenus, il y a ambiguïté et le système doit choisir entre poser une question à l'utilisateur ou privilégier une interprétation qui sera prise en compte dans le traitement complet de l'énoncé (dans ce dernier cas, l'interprétation correspond au premier résultat dans la liste, en particulier s'il se détache nettement des autres en apparaissant plusieurs fois dans les premiers rangs) ;

– s'il n'y a pas ambiguïté sur l'identité des référents, il peut y avoir ambiguïté sur l'étendue du domaine de référence (c'est le cas typique : l'ancrage sur un domaine étant implicite, la portée de ce domaine est vague et peut conduire à plusieurs possibilités pertinentes) ; dans ce cas, les possibilités sont gardées et conduiront éventuellement à détecter une ambiguïté lors de l'analyse de l'énoncé ultérieur ;

– si aucun résultat n'est trouvé, le contrôleur de dialogue doit relâcher une contrainte dans sa requête et demander aux modules de refaire leurs calculs (les seules contraintes qui puissent alors être relâchées sont celles concernant le complément d'une partition : pour un défini par exemple, on autorisera son instanciation par l'ensemble vide) ;

– si, après ce relâchement de contrainte, aucun résultat n'est encore trouvé, cela privilégie l'hypothèse de l'erreur, soit de la part de l'utilisateur (comme le lapsus transformant *horizontale* en *verticale*), soit de la part du système (comme une erreur

de reconnaissance, par exemple un défini pris pour un démonstratif, ce qui conduit à l'élaboration de contraintes trop fortes dans le contexte).

Une fois la résolution des références de l'énoncé effectuée, le système doit encore résoudre les références aux actions, générer la réponse ou effectuer l'action adéquates, et mettre à jour les différents historiques. Un historique est affecté à chacun des trois modules (cf. figure 5). L'historique langagier a pour rôle de conserver les différentes expressions référentielles utilisées, ainsi que les domaines de références sous-spécifiés générés par son module. Un exemple d'utilisation des anciennes expressions référentielles a été donné dans la section 4.1.1. L'historique visuel conserve les états et structurations successifs de la scène. Un exemple d'utilisation de cet historique est la référence à un objet qui a disparu de la scène, phénomène qui peut nécessiter un retour à l'état antérieur de la scène pour retrouver le domaine visuel à la source de la référence. L'historique de tâche conserve les actions effectuées et les référents sur lesquels elles ont porté. Un historique global en lien direct avec le contrôleur de dialogue conserve les caractéristiques des différentes phases du dialogue. Sa description n'entre pas dans la problématique qui nous intéresse ici, car elle fait intervenir les stratégies de communication des interlocuteurs. Notons seulement, comme le montre la figure 5, qu'il comprend des pointeurs vers les historiques locaux que nous avons décrits.

5.3. *Une modélisation par structures de traits*

La construction de domaines sous-spécifiés par combinaison d'informations issues des modalités d'interaction, aussi bien que l'ancrage d'informations linguistiques sur des informations contextuelles, peuvent se faire sur la base d'un même mécanisme d'unification de structures de traits. Nous proposons ici une modélisation de la notion de domaine de référence sous la forme de deux structures de traits, une pour le domaine lui-même, une autre pour la notion de partition. Cette approche permet d'envisager une architecture ouverte de système de dialogue où les contraintes elles-mêmes, représentées sous la forme de structures de traits, peuvent transiter d'un module d'analyse à un autre, et cumuler de manière incrémentale les informations disponibles pour chacun de ces modules.

La structure de traits correspondant à un domaine de référence est la suivante :

– un *identifiant* garantissant l'unicité de cette structure pour l'ensemble des espaces de représentation du système de dialogue considéré ;

– un *facteur de groupement* indiquant la nature du module qui justifie l'existence de ce domaine de référence ;

– un *type* qui, par référence à une ontologie du domaine concerné, subsume l'ensemble des types des entités composant le domaine ;

– un *modifieur* ou ensemble de propriétés pouvant préciser le type ;

– une *cardinalité* exprimée soit sous la forme d'un entier, soit par un code de numération (*simple*, *pluriel*, *massif*, *inconnu*, etc.) ;

– une séquence éventuellement vide de *partitions*.

Une partition est une sous-structure d'un domaine de référence, et est pour sa part représentée à l'aide de quatre caractéristiques principales :

– un *critère de différenciation*, représentant la ou les caractéristiques justifiant la discrimination des différentes entités composant le domaine pour la partition considérée, et pris au sein d'une ontologie spécifique de tels critères (regroupant les critères linguistiques, les différentes combinaisons des critères de la Gestalt, les critères liés à la structuration en sous-buts, et ceux spécifiques à l'application) ;

– une *marque d'ordonnancement* à valeur binaire (*oui/non*), indiquant si les composantes de la partition peuvent être vues comme un ensemble ou une séquence ordonnée d'éléments (cette information dépend du critère de différenciation et peut dans certains cas résulter directement de la connaissance de celui-ci : un critère représentant par exemple la répartition horizontale de gauche à droite d'éléments dans le contexte visuel sera forcément associé à une séquence ordonnée de composantes) ;

– un *contenu*, formé d'une suite de références à des domaines de référence ou à des entités individuelles ;

– une *marque de focalisation*, éventuellement vide, correspondant à un index sur le contenu de la partition.

La combinaison des structures de traits s'effectue à l'aide d'un mécanisme classique d'unification, adapté pour d'une part gérer les différents appariements possibles entre séquences de partitions, et, d'autre part, traiter la comparaison ordonnée ou non des composantes d'une partition donnée. Enfin, on peut noter que la représentation d'entités individuelles peut très bien s'intégrer dans ce formalisme en considérant celles-ci comme des domaines à un seul élément, marqués par une cardinalité égale à *simple*. Bien que la description fine des mécanismes correspondants sorte du cadre de cet article, on notera qu'une telle représentation permet d'intégrer aisément le traitement des anaphores associatives par décomposition (par exemple méronimique) à l'aide de partitions particulières des entités élémentaires.

5.4. *Illustration du mécanisme*

La figure 6 montre les différents domaines sous-spécifiés disponibles et créés lors de l'interprétation de *la pyramide de droite* dans l'exemple de la figure 1. Dans certains d'entre eux, un point d'interrogation montre ce que l'on cherche à instancier. Le module langagier renvoie trois domaines de référence en mettant l'accent sur celui pour lequel une partition reste à instancier (il s'agit de @2 sur le

schéma). L'unification de ce domaine avec celui correspondant à la requête se fait grâce à l'ontologie assimilant d'une part *pyramide* à *triangle* tout en acceptant le modifieur *grand*, et assimilant d'autre part *gauche* à ¬*droite* (et ¬*gauche* à *droite*). Le module visuel renvoie quant à lui deux domaines @4 et @5 qui constituent deux résultats possibles et montrent l'ambiguité entre l'ancrage référentiel sur le domaine des deux grands triangles ou sur le domaine des trois triangles. Enfin, le module tâche, qui a décomposé l'action de contruction d'un horizon en une succession d'actions de pose de segment de droite, met l'accent sur un domaine distinguant la cible de l'action qu'est le segment, à son site constitué par les objets sur lesquels repose ce segment. Les sites correspondant aux actions précédentes étant le domaine des deux grands triangles puis un de ces triangles, l'accent est mis sur le second qu'est l'objet o2.

L'unification de ces domaines conduit bien à l'identification du référent o2 dans le domaine constitué par o1 et o2 : le module linguistique met en avant ce résultat ; le module visuel classe également ce résultat (correspondant à @4) avant l'autre possibilité (@5), car la similarité au niveau le plus immédiat intervient avant la combinaison de la bonne continuité et de la similarité à un niveau plus faible ; le module tâche conduit également à ce résultat avec une focalisation directe sur o2. Si l'ambiguïté est détectée, la prise en compte de multiples critères permet d'interpréter comme l'a fait intuitivement le manipulateur.

6. Conclusion et perspectives

Le modèle à base de domaines de référence que nous proposons n'est dirigé que par des préoccupations liées à la spontanéité de la communication. La diversité de celle-ci est prise en compte dès la première étape de l'interprétation (à savoir la traduction d'une expression référentielle en contraintes structurées), puis exploitée dans toute sa complexité lors des étapes ultérieures (la structuration des différents contextes et la confrontation de ces structures). La mise en œuvre de notre modélisation sur des extraits de dialogues montre qu'une grande partie des phénomènes référentiels est traitable, et ceci dans un cadre unifiée pour toutes les expressions référentielles et pour toutes les sources contextuelles. Parmi ces phénomènes que nous considérons comme relevant à part entière de la communication spontanée, un certain nombre ont été exclus ou considérés comme problématiques par les modèles existants (emplois génériques, emplois associatifs, groupes nominaux définis accompagnés d'un geste de désignation, groupes nominaux sans noms, expressions d'altérité ou d'ordre).

Notre modèle n'est ainsi dirigé ni par le type d'interaction, ni par le type d'application. Il est utilisable aussi bien pour une vision en 2D que pour une vision en 3D, aussi bien avec un écran tactile qu'avec un dispositif à retour de force, aussi bien pour une application d'interrogation de base de donnée que pour tout type de dialogue de commande. De plus, la formalisation proposée à base de structures de

traits et de mécanisme d'unification peut être facilement implantée, une fois que sont spécifiées l'application et les ontologies qu'elle met en jeu. Comme perspective à court terme, cette formalisation doit permettre d'aborder une réflexion de fond sur la standardisation des formats d'échange de données, notamment dans la perspective de la mise en place du sous-comité TC37/SC4 de l'ISO sur les ressources linguistiques.

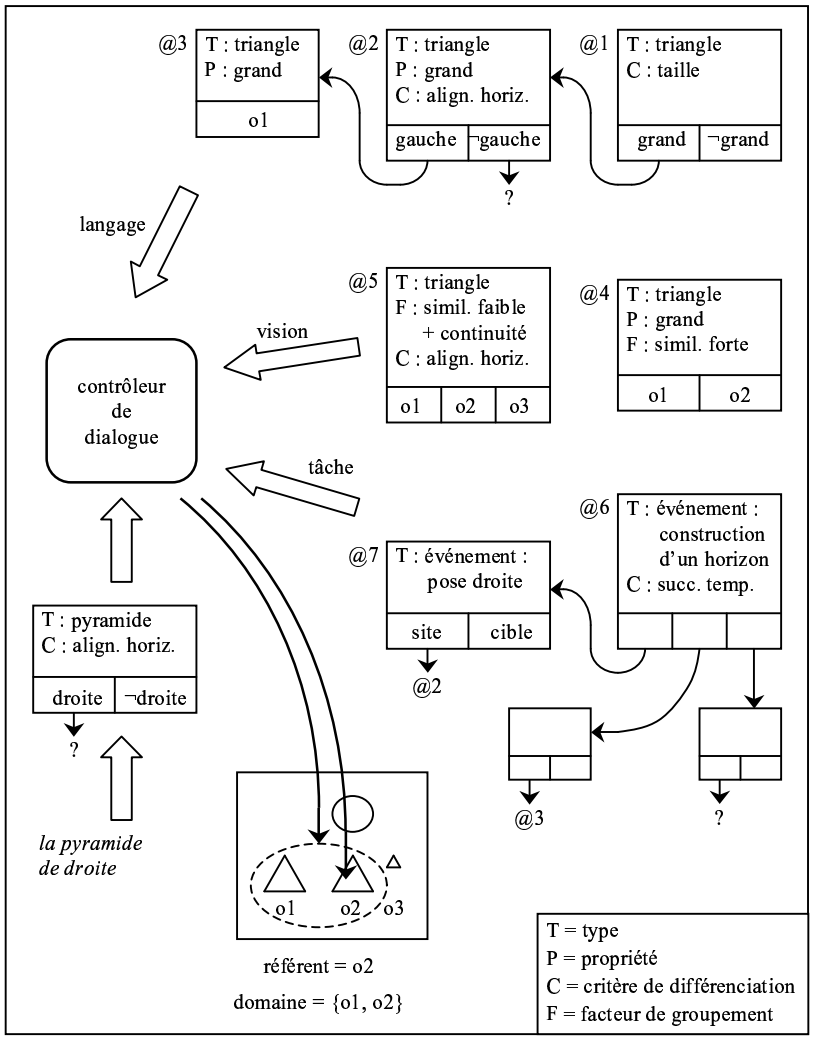


Figure 6. *Quelques domaines de référence pour l'exemple de la figure 1*

Cependant, certains aspects théoriques restent à affiner. D'abord, les domaines sont construits essentiellement à partir des propriétés des objets, alors que des informations supplémentaires devraient être prises en compte, en particulier celles concernant les événements auxquels les objets participent. Ce type d'information permettrait par exemple de traiter les coréférences entre indéfinis (Danlos & Gaiffe, 2000) et de formuler des contraintes supplémentaires sur l'interprétation des pronoms. Par ailleurs, ces connaissances sont nécessaires afin de résoudre des anaphores associatives à des participants d'une éventualité (*un meurtre – la victime*). Le traitement de ces phénomènes pourrait être intégré dans notre modélisation, à condition de disposer d'une représentation des événements et d'une définition des opérations possibles sur ces entités. Les travaux de Grisvard (2000) ont d'ores et déjà montré que notre cadre de modélisation est adapté à une telle extension.

Un autre aspect à affiner est l'ordonnancement des résultats renvoyés par les différents modules au contrôleur de dialogue. Nous identifions trois critères qui lui permettraient de remettre en cause ces ordonnancements : un critère de simplicité cognitive (le fait qu'un domaine contienne peu d'éléments et que ces éléments soient structurés de manière claire devrait le privilégier par rapport aux autres) ; un critère de pertinence (une estimation – éventuellement numérique – de l'effort nécessaire au traitement d'un domaine et des effets qu'il produit sur le contexte (Sperber & Wilson, 1995) permettrait de privilégier le domaine aboutissant au maximum d'effets pour le minimum d'effort) ; et un modèle de l'utilisateur : d'une part, un système capable de détecter les régularités dans la manière de référer de l'utilisateur serait capable de diriger l'interprétation par ces préférences, et d'autre part, le modèle de l'utilisateur interviendrait lors du calcul de la saillance visuelle, en privilégiant les objets qui lui sont familiers, les couleurs auxquelles il s'avère sensible, etc. Enfin, notre modèle axé sur la seule résolution de la référence devrait, à plus long terme, intégrer les actes de langage et une véritable gestion du dialogue (incluant une étude sur les stratégies de réponse et de réparation).

7. Bibliographie

Asher N., *Reference to Abstract Objects in Discourse*, Kluwer Academic Publishers, 1993.

Belaïd A., Belaïd Y., *Reconnaissance de formes, méthodes et applications*, Paris, Inter-Editions, 1992.

Beun R.-J., Cremers A. H. M., « Object Reference in a shared Domain of Conversation », *Pragmatics and Cognition*, vol. 6, n° 1/2, 1998.

Bos J., Mineur, A-M., Buitelaar, P., « Bridging as Coercive Accommodation », *Proceedings of the Workshop on Computational Logic for Natural Language Processing CLNLP'95*, Edinburgh, 1995.

Briffault X., Modélisation informatique de l'expression de la localisation en langage naturel, Thèse de doctorat, Université de Paris VI, 1992.

Corblin F., *Indéfini, défini et démonstratif*, Genève, Droz, 1987.

Dale R., Reiter E., « Computational Interpretations of the Gricean Maxims in Generating Referring Expressions », *Cognitive Science,* vol. 18, 1996.

Danlos L., Gaiffe B., « Coréférence événementielle et relations de discours », *TALN 2000*, Lausanne, Switzerland, 2000.

Dekker P., « Speaker's Reference, Descriptions, and Information Structure », *Journal of Semantics*, vol. 15, n° 4, 1998.

Feldman J., « Regularity-based Perceptual Grouping », *Computational Intelligence*, vol. 13, n° 4, 1997.

Gaiffe B., Référence et dialogue homme-machine : vers un modèle adapté au multi-modal, Thèse de doctorat, Université de Nancy I, 1992.

Gernsbacher M. A., Hargeaves D., « Accessing Sentence Participants: The Advantage of First Mention », *Journal of Memory and Language*, vol. 27, 1988.

Grisvard O., Modélisation et gestion du dialogue oral homme-machine de commande, Thèse de doctorat, Université de Nancy I, 2000.

Grosz B. J., Sidner C., « Attention, Intention and the Structure of Discourse », *Computational Linguistics*, n° 12, 1986.

Grosz B. J., Joshi A. K., Weinstein S., « Centering: A Framework for Modeling the Local Coherence of Discourse », *Computational Linguistics,* vol. 12, n° 2, 1995.

Kamp H., Reyle U., *From Discourse to Logic*, Kluwer Academic Publishers, 1993.

Kievit L., Piwek P., « Multimodal Cooperative Resolution of Referential Expressions in the DenK System », *In Bunt H., Beun R.-J. (Eds.), Proceedings of the Second International Conference on Cooperative Multimodal Communication*, Berlin, Springer Verlag, 2000.

Kleiber G., « Pour une explication du paradoxe de la reprise immédiate *Un N – Le N / Un N – Ce N* », *Langue française,* vol. 72, 1986.

Kleiber G., *Anaphores et pronoms*, Louvain-la-Neuve, Duculot, 1994.

Kubovy M., Wagemans J., « Grouping by Proximity and Multistability in Dot Lattices: A Quantitative Gestalt Theory », *Psychological Science*, vol. 6, n° 4, 1995.

Landragin F., Bellalem N., Romary L., « Visual Salience and Perceptual Grouping in Multimodal Interactivity », *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, 2001.

Landragin F., De Angeli A., Wolff F., Lopez P., Romary L., « Relevance and Perceptual Constraints in Multimodal Referring Actions », *In Van Deemter K., Kibble R. (Eds.), Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, to appear with CSLI Publications, Stanford, 2002.

Langacker R. W., *Concept, image, and symbol: the cognitive basis of grammar*, New York, Mouton de Gruyter, 1991.

Moeschler J., Reboul A., *Dictionnaire encyclopédique de pragmatique*, Paris, Seuil, 1994.

Olson D. R., « Language and Thought: Aspects of a Cognitive Theory of Semantics », *Psychological Review*, n° 77, 1970.

Ozkan N., Vers un modèle dynamique du dialogue : analyse de dialogues finalisés dans une perspective communicationnelle, Thèse de doctorat, INP Grenoble, 1994.

Popescu-Belis A., Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation, Thèse de doctorat, Université de Paris XI, Orsay, 1999.

Reboul A., Balkanski C., Briffault X., Gaiffe B., Popescu-Belis A., Robba I., Romary L., Sabah G., Le projet CERVICAL : représentations mentales, référence aux objets et aux événements, Rapport de Recherche, LORIA-LIMSI, 1997.

Reinhart T., « Coreference and Bound Anaphora: A Restatement of the Anaphora Questions », *Linguistics and Philosophy,* vol. 6, n° 1, 1983.

Romary L., « L'interprétation de *ici* dans les énoncés de positionnement », *Dans Vivier J. (Red.), Le dialogue homme-robot en langage naturel : problèmes psychologiques*, Presses Universitaires de Caen, 1993.

Salmon-Alt S., Référence et dialogue finalisé : de la linguistique à un modèle opérationnel, Thèse de doctorat, Université de Nancy I, 2001.

Sidner C., Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse, Thèse de doctorat, Massachusetts Institute of Technology, 1979.

Sperber D., Wilson D., *Relevance: Communication and Cognition*, Oxford, Blackwell, 1995.

Thórisson K. R., « Simulated Perceptual Grouping: An Application to Human-Computer Interaction », *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Atlanta, Georgia, 1994.

van Hoek K., « Conceptual Reference Points: A Cognitive Grammar Account of Pronominal Anaphora Constraints », *Language*, vol. 71, n° 2, 1995.

Vivier J., Nicolle A., « Questions de méthode en dialogue homme-machine : l'expérience Compèrobot », *Dans Sabah G., Vivier J., Vilnat A., Pierrel J.-M., Romary L., Nicolle A. (Reds.), Machine, Langage et Dialogue*, L'Harmattan, 1997.

Webber B. L., « Structure and ostension in the interpretation of discourse deixis », *Language and Cognitive Processes*, vol. 6, n° 2, 1991.

Wertheimer M., « Untersuchungen zur Lehre von der Gestalt II », *Psychologische Forschung*, n° 4, 1923.

Westerstahl D., « Determiners and Context Sets ». *In Van Benthem J., ter Meulen A. (Eds.), Generalized Quantifiers in Natural Language*, Dordrecht, Foris Publications, 1984.

Winograd T., *Understanding Natural Language*, Edinburgh University Press, 1972.

Wolff F., Analyse contextuelle des gestes de désignation en dialogue homme-machine, Thèse de doctorat, Université de Nancy I, 1999.

Wright P., « Using Constraints and Reference in Task-Oriented Dialogue », *Journal of Semantics*, n° 7, 1990.