Modélisation de la communication multimodale : vers une formalisation de la pertinence

Frédéric Landragin

Département de formation doctorale en informatique **École doctorale IAEM**
UFR STMIA

Modélisation de la communication multimodale

Vers une formalisation de la pertinence

THÈSE

présentée et soutenue publiquement le 2 avril 2003

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Frédéric Landragin

Composition du jury

Président :	J.-M. Pierrel	Professeur, Université Henri Poincaré – Nancy 1
Rapporteurs :	F. Alexandre	Directeur de Recherche, INRIA, Nancy
	J. Siroux	Professeur, IUT Lannion
	H. Zeevat	Professeur, Université d'Amsterdam
Examinateurs :	N. Bellalem	Maître de Conférence, Université de Nancy 2
	L. Romary	Directeur de Recherche, INRIA, Nancy

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Sommaire

Avant-propos

Ce manuscrit est le résultat d'un peu plus de quatre ans de travail répartis en trois étapes : un stage de fin d'étude d'école d'ingénieur[1]*, un stage de DEA et une thèse de doctorat*[2]*. Si le thème de recherche (le traitement pragmatique de la référence multimodale en dialogue homme-machine) n'a pas changé, l'approche et la nature du travail effectué ont pris différents aspects. Paradoxalement, c'est au cours de la première étape que s'est déroulé le principal travail d'implantation informatique*[3]*. La deuxième étape a ensuite essentiellement consisté en une analyse expérimentale*[4]*. Les résultats obtenus lors de ces deux phases, ainsi que la prise de conscience de la complexité du problème qui leur est liée, ont permis d'aborder la troisième étape, le travail de thèse proprement dit, avec les bases, la méthodologie et le recul nécessaires à une approche théorique.*

Le thème de recherche qu'est le traitement de la référence dans la communication homme-machine spontanée, du fait de la grande variété des phénomènes et des mécanismes cognitifs mis en jeu, nécessite une approche pluridisciplinaire faisant intervenir l'informatique, la linguistique, la psychologie, les sciences cognitives en général. L'enrichissement se fait non seulement au niveau des connaissances, mais également dans les collaborations qu'un tel sujet favorise[5]*. L'aspect théorique de mon travail a ainsi évolué et a abouti à une modélisation. Ce manuscrit présente cette modélisation ainsi que quelques aspects de sa validation*[6]*.*

---

1. Stage de recherche et développement de troisième année de l'Institut d'Informatique d'Entreprise, d'une durée de neuf mois, effectué au Laboratoire Central de Recherches de THOMSON-CSF, dans l'équipe « Communication Homme-Machine » dirigée par Célestin Sédogbo.

2. Réalisés tous les deux au LORIA dans l'équipe « Langue et Dialogue », sous la direction de Laurent Romary et l'encadrement de Nadia Bellalem.

3. Le stage a consisté en la spécification et l'implantation d'un composant du module de résolution de la référence, dans le cadre du projet ACTS-AC040 COVEN (*COllaborative Virtual ENvironments*). L'application consistait en l'aménagement d'un intérieur dans un environnement virtuel, avec des dispositifs permettant une visualisation en 3D et la capture de la parole et du geste. L'implantation réalisé dans ce cadre complexe a permis de mettre l'accent sur des aspects aussi bien pratiques (traitement du geste en 3D, traitement modulaire de la multimodalité) que théoriques (conséquences des contraintes induites par les dispositifs, étendue des phénomènes).

4. L'analyse a porté sur les phénomènes de référence apparaissant dans l'enregistrement d'une simulation de dialogue homme-machine multimodal (Magnét'Oz). Son apport a été aussi bien méthodologique (exploitation de corpus, avantages de la simulation comme méthode de recueil et de validation) que théorique (étendue des phénomènes possibles, consolidation de la bibliographie).

5. Plusieurs collaborations interdisciplinaires se sont révélées particulièrement fructueuses dans ce travail, d'une part celle avec Antonella de Angeli (psychologue), d'autre part celle avec Marion Fossard (doctorante en neurolinguistique/psycholinguistique), sans oublier les collaborations internes à l'équipe « Langue et Dialogue » et celles entre le LORIA et l'ATILF.

6. C'est dans le cadre du projet IST-2000-29487 MIAMM (*Multidimensional Information Access using Multiple Modalities*), coordonné par Laurent Romary, qu'une validation partielle a pu être menée au cours de ma troisième

Quelques courants théoriques et auteurs ont déterminé mon travail de recherche. Au départ, ce sont le livre de Jean-Marie Pierrel sur le dialogue homme-machine et la thèse de Bertrand Gaiffe qui m'ont ouvert les portes de ce domaine et confirmé ma volonté d'entrer dans le monde de la recherche scientifique. Ce sont ensuite deux importantes théories qui m'ont permis de construire mon travail, d'une part la théorie de la Gestalt avec l'article de Max Wertheimer, d'autre part la théorie de la Pertinence avec le livre de Dan Sperber et Deirdre Wilson. Quelques articles en linguistique computationnelle m'ont également été d'un grand apport, par exemple celui de Gérard Sabah, celui de Barbara J. Grosz et Candace L. Sidner, ou encore celui de Robbert-Jan Beun et Anita Cremers. Quelques essais et livres proches de la vulgarisation scientifique par leur facilité de lecture m'ont également aidé à prendre du recul, ceux de Boris Cyrulnik, ceux d'Anne Reboul et de Jacques Moeschler, ainsi que ceux traitant de l'intelligence artificielle, trop nombreux pour être cités ici. Je remercie tous ces auteurs pour avoir éveillé ma curiosité.

Je tiens à exprimer tous mes remerciements à Laurent Romary pour m'avoir accueilli dans l'équipe « Langue et Dialogue » du LORIA, m'ayant ainsi permis de travailler dans un environnement scientifique et humain très profitable. Grâce à lui, j'ai pu orienter librement mes recherches, dans des directions qui pouvaient parfois sembler aléatoires mais qu'il a toujours su contrôler et m'aider à exploiter. Je remercie vivement Frédéric Alexandre, Jacques Siroux et Henk Zeevat qui ont accepté d'être les rapporteurs de ce travail.

Je remercie également Nadia Bellalem qui a suivi le déroulement de ce travail et l'a agrémenté de ses remarques pertinentes. Je remercie tous ceux qui ont accepté de travailler avec moi, qui m'ont écouté et laissé le rôle principal dans les publications qui en ont résulté : Bertrand Gaiffe, Patrice Lopez, Susanne Salmon-Alt, ou encore Antonella de Angeli et Frédéric Wolff. Un grand merci, pour tous les moments d'échanges autour d'une table de travail, d'un pot ou d'un café, à tous mes collègues de l'équipe « Langue et Dialogue », en particulier Hélène Manuélian, Claire Gardent, Evelyne Jacquey, Benoît Crabbé, mais aussi Nadia, André, Jean-Luc, Etienne, Christine, Thi-Minh-Huyen, Djamé, Gérald, Azim, Patrick, Samuel, Olivier, Charles, Isabelle et j'en oublie. Merci également aux collègues du LORIA, en particulier à Armelle Brun et François Cuny qui ont partagé avec moi bien plus qu'un lieu de travail. Merci aussi aux personnes qui ont évalué mes publications, pour leurs remarques qui m'ont encouragé et pour leurs suggestions face aux difficultés que j'ai rencontrées. Merci à celles qui m'ont fait des remarques au cours de conférences, ainsi qu'à celles qui ont relu et corrigé mes brouillons. Merci ainsi à Eric Kow d'avoir bien voulu défranciser mon anglais.

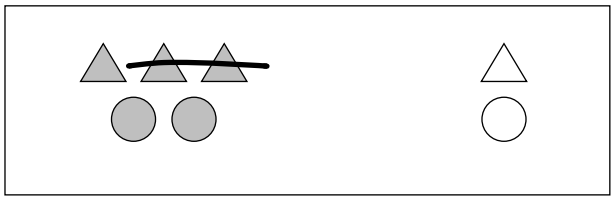
Enfin, merci à mes proches qui m'ont encouragé, soutenu, et aussi permis parfois de m'évader. Merci donc à ma famille, à Laurence Mullet, et à Pascal Ochs (guide de haute-montagne décédé en 2001).

année de thèse. Ce projet explore les possibilités référentielles d'une interaction incluant le retour tactile et le retour d'effort. J'ai ainsi pu confronter le modèle que je propose à un tel mode d'interaction.

Introduction

On utilise spontanément la parole et le geste dans nos conversations de tous les jours. Mais on n'a pas pour autant conscience de leur complexité, ni des années d'apprentissage qui ont été nécessaires à leur maîtrise. Les mécanismes mis en jeu à chaque fois qu'on exprime quelque chose sont ainsi nombreux, complexes, et même difficiles à discerner pour le chercheur qui s'y intéresse. Par exemple, les imprécisions de la parole se compensent avec celles du geste, cette compensation étant implicite. D'une manière générale, tout ce qui n'est pas exprimé, tout ce qui est sous-entendu est nécessaire à la communication. Cet implicite se devine, souvent grâce au choix subtil d'un mot plutôt qu'un autre, ou grâce à une petite particularité d'un geste. Cette efficacité se manifeste en particulier lorsque l'on désigne les objets du monde, objets à propos desquels on veut communiquer ou sur lesquels on veut agir.

Dans le cadre des travaux présentés ici, nous voulons exploiter cette efficacité dans la communication entre l'homme et la machine. Prenons un exemple mettant l'accent sur la désignation d'objets. Dans l'extrait de dialogue homme-machine présenté dans la figure 1, des formes géométriques de diverses couleurs sont affichées à l'écran et peuvent être déplacées, modifiées ou supprimées. Ce type de situation pourrait se trouver dans tout logiciel de dessin. Il s'agit en fait d'un exemple prototypique d'action sur des objets virtuels, exemple facilement extensible à d'autres types d'objets et à d'autres situations.



« *Colorie ces trois triangles en bleu...* » + geste → {△, △, △}

« *... et les deux cercles en rouge* » → {○, ○}

FIGURE 1 – *Exemple d'interprétation faisant intervenir perception visuelle, geste et langage.*

On imagine ici que l'interaction se fait non pas à l'aide d'une interface classique (menus et souris), mais par la voix et le geste, spontanément, comme dans la communication humaine, face à face. L'ordinateur est alors considéré comme un interlocuteur à qui l'on s'adresse en langage naturel, librement, en s'aidant quand on le désire de gestes pour désigner les objets de la scène. Pour traiter ces moyens d'expression, deux dispositifs sont exploités, un microphone et un écran tactile, tel qu'on peut en voir sur certaines bornes interactives publiques.

Dans l'extrait qui correspond à une telle interaction, l'utilisateur souhaite faire des modifications sur certains objets de la scène. La première partie de son énoncé est un ordre qui comprend

un énoncé oral (« *colorie ces trois triangles en bleu* ») et un geste de désignation (trajectoire plane représentée par un trait épais sur le schéma). Ce geste s'appuie sur le contexte visuel et s'associe à l'expression « *ces trois triangles* » pour désigner les trois triangles gris. Le système doit retrouver cette intention.

Or, compte tenu du fait que la trajectoire ne concerne que deux des trois triangles, il se trouve face au problème qui consiste à confronter les caractéristiques des trois modalités que sont perception visuelle (le fait que les trois triangles sont proches et alignés), parole et geste, pour identifier les triangles gris et pouvoir ensuite leur appliquer l'action demandée. Ce problème d'identification des objets (appelé aussi résolution de la référence aux objets) se retrouve dans la deuxième partie de l'énoncé, avec l'expression « *les deux cercles* » qui désigne, sans ambiguïté pour l'utilisateur et sans qu'il ait recours à un nouveau geste, les deux cercles gris situés dans la même zone spatiale que les triangles précédemment traités. Pour cette référence purement langagière, l'interprétation fait ainsi intervenir non seulement les caractéristiques de la perception visuelle et de la parole, mais également l'historique de l'interaction.

Les choix faits par l'utilisateur dans cet exemple ne sont pas les plus compréhensibles : la trajectoire gestuelle aurait pu recouvrir exactement les trois triangles visés, et des expressions telles que « *les trois triangles gris* », « *les trois triangles les plus à gauche* », ou « *les deux cercles gris* » pour la deuxième partie de l'énoncé auraient été plus claires. Pour un interlocuteur humain, l'énoncé de l'utilisateur ne pose cependant aucun problème de compréhension et un système de dialogue homme-machine doit pouvoir le traiter.

Si l'utilisateur n'a désigné par son geste que deux triangles, c'est moins par volonté que par précipitation : en employant l'adjectif numéral « *trois* », il s'appuie sur la précision de son expression langagière et s'autorise à négliger une certaine précision dans son geste. S'il voit que la trajectoire affichée ne correspond pas à son intention, si dans ce cas il ne la complète pas ni ne la recommence, c'est parce que cette forme d'expression lui semble pertinente compte tenu du contexte et de son expérience de locuteur, c'est parce qu'il sait que n'importe quel interlocuteur sera capable de retrouver son intention référentielle. La machine se doit d'être un tel interlocuteur. Il se peut également que l'écart entre le geste voulu et le geste réalisé soit dû à un biais imputable à l'écran tactile. La machine doit également tenir compte de ce biais.

Analyser finement les interactions entre les trois pôles que sont perception visuelle, parole et geste permettra de répondre à de tels défis, et ira dans le sens d'accroître les capacités d'interprétation de la machine. En prenant pour base un modèle de ces interactions tripolaires, celle-ci sera capable de comprendre l'exemple de la figure 1, pour lequel des systèmes basés sur des traitements séparés des modalités ou sur des interactions bipolaires s'avèrent insuffisants.

En effet, dans le cas de traitements séparés des modalités, rien ne permet d'exploiter l'appui du langage sur la situation, ainsi que, dans la première partie de l'exemple, de dépasser l'interprétation stricte de la trajectoire sur deux triangles. De plus, dans le cas de traitements basés sur les interactions bipolaires, l'association de la perception visuelle et du geste ne suffit pas à faire de la trajectoire une action de désignation imprécise et par conséquent extensible à un objet supplémentaire ; l'association de la parole et du geste pour cette même référence s'arrête sur une incompatibilité entre trois triangles d'un côté et deux de l'autre ; l'association de la perception visuelle et de la parole ne permet pas dans la deuxième partie de l'énoncé d'identifier les deux cercles visés parmi les trois affichés.

En revanche, une prise en compte des interactions tripolaires permet de résoudre, non seulement la première partie de l'énoncé en étendant l'ensemble des objets désignés au troisième qui appartient à la fois à la bonne catégorie et à la même zone visuelle, mais aussi la deuxième partie en restant focalisé dans cette zone. Cette connexion entre trois pôles qui interagissent s'approche

d'une compréhension globale. C'est ce que nous voulons explorer et formaliser. Disposer d'un même cadre formel pour les trois modalités nous semble constituer un point de départ important pour la modélisation. Pour gérer des groupes perceptifs d'objets, des ensembles correspondant à une focalisation de l'attention, et des listes d'hypothèses provenant de chacune des modalités, le système a besoin de structures d'objets qui puissent se confronter. Ces structures doivent donc appartenir à un même formalisme.

Nous avons ainsi l'objectif de proposer un modèle formel qui, tenant compte de différents points de vue (langagier, perceptif, gestuel) et se basant sur des résultats établis par la linguistique et par la psychologie cognitive, intègre de manière uniforme les trois modalités dans le but de résoudre toute forme de référence. A partir d'un tel modèle, il devient possible de commencer à réfléchir sur la communication humaine en mettant l'accent sur l'intégration d'informations et sur la modélisation.

Les principales questions auxquelles nous tenterons de répondre dans ce travail sont donc les suivantes : comment exploiter les informations transmises par les trois modalités pour spécifier des structures d'objets intervenant dans l'interprétation ? comment intégrer ces structures hétérogènes dans un même cadre formel ? quels sont les mécanismes de construction et de confrontation de ces structures ? comment exploiter les résultats obtenus ? comment mesurer leur validité ? comment passer de ces problèmes d'intégration et d'exploitation d'informations à un modèle de résolution de la référence aux objets ? quels sont les bénéfices d'une telle approche pour la modélisation de la communication humaine en général ?

Les points de réponse qui constituent les principaux apports de cette thèse sont répartis dans les trois parties qui la composent. La première partie présente notre objet d'étude et notre méthodologie. Elle contient nos propositions de points de départ pour la modélisation, points de départ obtenus suite à l'étude d'un phénomène selon plusieurs points de vue et souvent plusieurs disciplines. Ainsi, le premier chapitre décrit comment les actions de référence se traduisent dans le dialogue homme-machine. Il inclut une caractérisation des phénomènes de référence aux objets, en appliquant systématiquement les définitions et les distinctions classiques à des situations de dialogue homme-machine. Le chapitre 2 se consacre à l'interaction tripolaire des modalités. Nous y montrons que les modèles actuels sont principalement bipolaires, que cela ne permet pas d'en déduire des modèles tripolaires, et nous proposons des principes de base sur la nature des résultats bipolaires intermédiaires pour aboutir à un modèle tripolaire. La problématique ainsi posée, le chapitre 3 décrit la méthodologie adoptée. Un des apports de ce chapitre se trouve dans une réflexion sur les problèmes de validation dans le domaine de la communication homme-machine.

La deuxième partie présente notre modèle, en commençant par son principal concept, celui de « domaine de référence » (sous-ensemble structuré d'objets dans lequel a lieu l'interprétation), et en explorant successivement les critères qui permettent de l'exploiter. Après le chapitre 4 qui présente la notion de domaine de référence et les mécanismes de construction de domaines de référence propres à chacune des modalités, le chapitre 5 s'intéresse à la notion de « saillance » qui, en tant que point d'entrée dans un domaine, se trouve à la base de leur exploitation. A partir de travaux provenant de différentes disciplines, nous caractérisons cette notion qui reste à un niveau abstrait dans la majorité des travaux existants. Le chapitre 6 explore ensuite les critères de parcours de domaines de référence, en montrant comment un ordonnancement des éléments peut être détecté puis exploité pour la résolution de certaines expressions référentielles. Le chapitre 7 termine cette partie par la présentation synthétique de notre modèle, en l'abordant du point de vue de la pertinence. Nous montrons ici en quoi une présomption de pertinence dans la communication sous-tend notre modèle, et nous proposons des pistes de réflexion permettant d'aborder la formalisation de ce critère de pertinence, critère qui, comme celui de saillance,

est souvent exploité théoriquement mais rarement caractérisé formellement. Outre le modèle théorique proposé, les apports de cette partie se trouvent dans les caractérisations qui permettent d'aboutir à des algorithmes d'exploitation des critères ainsi identifiés.

La troisième partie décrit quelques applications du modèle. Le chapitre 8 s'intéresse aux architectures logicielles des systèmes de dialogue homme-machine et propose des directives d'implantation informatique pour notre modèle. Il identifie également les composants à ajouter pour pouvoir aboutir, à court ou moyen terme, à un système opérationnel. Si la nature même de notre modèle est directement en rapport avec une architecture logicielle, les applications illustrées dans la suite en sont des prolongements plus évolutifs. Le chapitre 9 montre ainsi comment l'adapter à un mode d'interaction particulier ou à une tâche particulière. Dans le chapitre 10 consacré aux exploitations connexes du modèle, exploitations pour lesquelles il n'a pas été conçu à la base, nous montrons comment il peut être « retourné » pour son utilisation dans le domaine de la génération automatique d'expressions référentielles. Pour cette problématique, nous montrons en quoi la gestion de domaines de référence, de critères de saillance et de pertinence constitue une stratégie efficace. Nous opérons enfin un retour à l'une des principales préoccupations de ce travail, celle de la plausibilité cognitive, en montrant en quoi nos propositions peuvent aboutir à l'élaboration d'un modèle cognitif de la communication. Nous posons les jalons d'une telle élaboration en spécifiant des protocoles expérimentaux pour valider du point de vue de la psychologie cognitive les plus importantes de nos propositions. Ces expérimentations restent à mettre en œuvre et constituent, avec quelques pistes d'amélioration théorique du modèle, nos perspectives de recherche par lesquelles nous terminons.

Première partie

Problématique et méthodologie

Chapitre 1

La référence aux objets dans le dialogue homme-machine

Comment se traduisent les actions de référence dans le dialogue homme-machine ? Quelles sont les principales caractéristiques des modalités qui interviennent dans la référence aux objets ? Les classifications de la littérature sont-elles pertinentes pour la référence dans le dialogue homme-machine ? A partir de quels concepts et de quelles constatations peut-on commencer à formaliser ?

Pour répondre à ces questions, ce premier chapitre présente quelques concepts fondamentaux, ainsi que quelques éléments de réponse obtenus suite à des observations personnelles. Les domaines concernés ne se limitent pas aux applications de l'informatique. Si la communication entre l'homme et la machine est un thème de recherche appartenant à l'informatique, c'est aussi un sujet pluridisciplinaire dans la mesure où ses préoccupations se rapprochent de celles de la linguistique et de la psychologie cognitive. Pour faire des systèmes informatiques qui comprennent l'humain, il est en effet nécessaire de se pencher sur ce qui caractérise la communication humaine. Ce chapitre présente, pour chacune des modalités intervenant dans les actions de référence, des définitions générales issues de plusieurs disciplines (§ 1.1), des caractérisations des phénomènes de référence applicables à la communication homme-machine (§ 1.2), et des formalisations (§ 1.3). Après cette première et indispensable étape, le chapitre suivant se focalisera sur la problématique liée aux interactions des modalités.

1.1 Définitions

Dans une première étape, nous nous focaliserons sur la *forme* de la communication : la parole, les gestes et leur classification selon leur intention communicative, ainsi que l'association de ces deux modalités, association appelée multimodalité. Nous explorerons ensuite le *contenu* de la communication, ce qui fera ressortir le problème de la référence aux objets dont nous aborderons les différentes facettes au cours d'une troisième étape. Les exemples qui illustreront chaque notion seront soit construits soit tirés de situations réelles de communication homme-machine.

1.1.1 Multimodalité et dialogue homme-machine

La parole. Deux distinctions restent fondamentales en linguistique : celle faite par Saussure (1916) entre *langue* et *parole*, et celle faite par Chomsky (1965) entre *compétence* et *performance*. La langue représente la connaissance que les membres d'une communauté ont du langage qui leur permet de communiquer entre eux. Ce concept, social chez Saussure, rejoint le concept inné de compétence de Chomsky, pour s'opposer à l'emploi effectif et individuel du langage. Cet emploi, qui relève de la performance selon le terme de Chomsky, a lieu aussi bien à l'oral qu'à l'écrit, ce que recouvre le terme « parole » de Saussure. Il inclut non seulement les variations individuelles, mais aussi les aléas de production, en particulier à l'oral (pour lequel nous gardons le terme « parole » que nous avons employé jusqu'ici). Ces aléas comprennent les hésitations, les répétitions ou encore les reformulations. Ils peuvent conduire à des incorrections grammaticales, la notion de grammaticalité provenant d'un usage normalisé de la langue se basant sur l'écrit.

Dans le but d'en faire ressortir les particularités, il est intéressant de comparer l'oral à l'écrit. A partir de la nature des mots, de la prosodie et de la syntaxe, Bellenger (1979) distingue plusieurs niveaux de langue, cités ici du plus au moins contrôlé : la langue *oratoire* (précieuse et pédante, elle se moule sur la composition écrite et se trouve dans certains discours emphatiques, dans des sermons) ; la langue *soutenue* (c'est celle des cours magistraux, des allocutions) ; la langue *courante* (c'est celle de la conversation, de la télévision en direct) ; la langue *familière* (vivante mais accumulant les incorrections grammaticales, c'est celle de la conversation non surveillée) ; la langue *populaire* (c'est celle de la conversation relâchée, caractérisée par une forte présence de l'argot). Ce qui apparaît dans cette hiérarchie, c'est l'apparition fréquente d'aléas dans la production à partir du niveau de la langue familière. Or, si l'écrit se situe généralement à des niveaux correspondant à la langue oratoire et à la langue soutenue, donc au-dessus, l'oral, particulièrement dans la communication homme-machine spontanée, se situe bien souvent entre la langue courante et la langue familière. La langue en entrée d'un système de dialogue homme-machine contient donc des aléas, et les modèles doivent en tenir compte, comme le fait par exemple celui de Lopez (1999). Nous garderons également cette préoccupation dans l'élaboration de notre modèle.

En communication homme-machine, le traitement automatique de la parole pose de nombreux autres problèmes. Les mots ainsi que leurs emplois sont par essence ambigus. La compréhension de certains mots, de l'association de mots ainsi que de la proposition énoncée font appel à de nombreux mécanismes qui sont devenus spontanés pour un humain mais qui doivent être programmés pour une machine. Parmi ces mécanismes, citons l'appel à des connaissances sur le monde et l'inférence de propositions à partir de celle de l'énoncé.

L'interprétation d'un énoncé fait en effet appel à des informations extra-linguistiques qui ne sont pas véhiculées par l'énoncé. Selon la formulation de Corblin (2002), le langage est par nature sous-spécifié, c'est-à-dire que « l'interprétation d'une structure n'est pas entièrement déterminée sur la base du matériau linguistique effectivement présent dans la structure » (p. 13). Pour interpréter correctement, une machine doit compléter dans la mesure du possible ce matériau linguistique par les informations extra-linguistiques. Cet objectif illustre la complexité du traitement automatique de la parole.

La parole est difficile à traiter et n'est pas adaptée à toutes les situations. Elle est bruyante et peut empêcher la confidentialité et gêner l'entourage ; elle est fatigante sur de longues durées ; elle est peu adaptée au spécialiste habitué à ses raccourcis au clavier. Elle est également moins efficace que des manipulations directes à la souris pour des processus continus comme le déplacement pas à pas d'un objet.

Malgré tous ces inconvénients, la parole a cependant de nombreux atouts. Elle est concise dans la mesure où un seul énoncé peut regrouper plusieurs commandes. Elle est rapide car elle se produit directement, sans exploration de menus et sous-menus. Elle est confortable car elle permet de prendre du recul par rapport aux actions : le locuteur ne *fait* pas lui-même mais il *fait faire* la machine en lui donnant des ordres (Pouteau 1994). Elle est efficace car elle permet de se concentrer sur la tâche tout en relâchant son attention sur les moyens de l'effectuer : l'application n'a pas à afficher de menus, l'utilisateur n'a pas à faire alterner son regard de l'écran au clavier ou à la souris (Mathieu 1997). Tout cela contribue à encourager l'exploitation de la parole en entrée des systèmes informatiques, que ce soit pour des traitements de texte ou pour des applications impliquant la manipulation d'objets comme l'aménagement de son intérieur, la spécification des caractéristiques de sa future voiture, et toute application de dessin ou de conception.

En fait, la parole est rarement produite seule. Bien au contraire, la conversation humaine est par essence multimodale. Kerbrat-Orecchioni (1996) distingue dans sa mise en œuvre trois types de matériau : le matériau *verbal* qui relève de la langue (informations phonologiques, lexicales, syntaxiques et sémantiques) ; le matériau *paraverbal* (prosodique et vocal : intonations, pauses, intensité articulatoire, débit, prononciation, caractéristiques de la voix) ; et le matériau *non verbal* transmis par le canal visuel (l'apparence physique des participants, les cinétiques lentes comme les postures et les attitudes, et les cinétiques rapides comme les regards et les gestes). Les énoncés, aussi bien dans la communication humaine que dans la communication homme-machine, sont le plus souvent une combinaison de ces trois types. Il est important pour un système de dialogue homme-machine de tous les exploiter. En effet, le sujet d'une communication orale est souvent relatif à l'endroit et au moment où sont produits les énoncés, et la langue orale emploie donc d'une façon très importante les informations contextuelles et les déictiques (termes qui renvoient directement à la situation d'énonciation, comme par exemple « *ça* », « *ici* », « *maintenant* »). Le matériau paraverbal, par exemple une accentuation lors de la prononciation de « *ça* », et le matériau non verbal, par exemple un geste de désignation lors de la prononciation de « *là* », sont alors aussi importants que le matériau verbal. Dans le matériau non verbal, nous donnerons une importance particulière aux gestes car ils sont utilisés de manière privilégiée pour la référence aux objets.

Le geste. A partir des nombreuses classifications de gestes que l'on peut trouver dans la littérature, Cosnier & Vaysse (1997) proposent un récapitulatif que nous présentons sous la forme synthétique suivante :

- Gestes communicatifs :
  1. Co-verbaux : ce sont des gestes non conventionnels qui dépendent systématiquement d'une production verbale simultanée. Ils se répartissent en :
     (a) référentiels (qui donnent de l'information sur les référents de l'énoncé) :
        - déictiques (gestes de pointage et de présentation),
        - illustratifs (ou iconiques), que l'on peut encore distinguer selon le concept qu'ils illustrent en spatiographiques (relatifs à une disposition spatiale) ; pictographiques (relatifs à une forme) ; kinémimiques (relatifs à une action) ; idéographiques ou métaphoriques (relatifs à un concept abstrait) ;

     (b) expressifs (ce sont en particulier les mimiques faciales, qui sont liées à l'énoncé et constituent la majeure partie de sa composante affective et émotionnelle) ;

     (c) paraverbaux (liés à l'énonciation, plus utiles au locuteur qu'à l'interlocuteur) :
        - battements (mouvements rythmant les paroles),

  - cohésifs (gestes de scansion renforçant la prosodie),
  - coordination (appuyant les connecteurs pragmatiques comme « *et* », « *puis* »).
2. Quasi-linguistiques : ils constituent un lexique d'une centaine de gestes par culture, sont donc conventionnels, et sont utilisés soit seuls, soit comme illustratifs. Ils ne dépendent donc pas nécessairement de la parole.
3. Synchronisateurs (ou régulateurs) : ils font partie de l'organisation conversationnelle, sont conventionnels et dépendent nécessairement de la parole. Ils se distinguent en :
   (a) gestes d'interaction (pour montrer ou vérifier que l'information passe) ;
   (b) gestes de passage de tour (pour prendre ou pour donner la parole) ;
   (c) gestes de maintenance de tour (pour garder la parole).

- Gestes extra-communicatifs : ils jouent le rôle d'indicateurs pour la communication émotive (ils montrent par exemple l'embarras). Ils se distinguent en :
  1. Automatiques : ce sont par exemple les gestes centrés sur le corps ou sur des objets.
  2. Planifiés (ou praxiques) : ce sont les gestes ludiques et utilitaires.

Dans cette classification, certains gestes sont spontanément inhibés par l'utilisateur d'une machine ; d'autres sont possibles mais peuvent être ignorés par le système de compréhension ; d'autres encore sont nécessaires à l'interprétation de la référence. Nous considérerons que les gestes synchronisateurs sont spontanément inhibés. Leur production implique en effet que l'interlocuteur soit capable de montrer qu'il écoute ou qu'il veut prendre la parole. Même si ces gestes sont théoriquement possibles lorsque l'interlocuteur est symbolisé par une représentation graphique telle qu'un avatar (comme dans le projet COVEN, cf. Normand *et al.* 1997), il est vraisemblable que le mode d'interaction induit par une application informatique diffère du débat politique et de ses passages de parole stratégiques. Nous considérerons que les gestes expressifs et extra-communicatifs sont possibles en dialogue homme-machine, mais peuvent être ignorés. En effet, ils n'apportent rien à l'information communiquée si ce n'est une légère connotation émotive pouvant être négligée. Parmi les gestes que l'on ne peut pas empêcher et qui sont nécessaires à la compréhension, restent donc les référentiels, les paraverbaux et les quasi-linguistiques. L'utilisation de ces derniers, qui apparaissent aussi dans les gestes co-verbaux en tant qu'illustratifs, dépend de l'application. Les gestes quasi-linguistiques appartiennent à un code qui reste très peu normalisé et dépend de communautés culturelles. Il est peu probable que l'utilisateur considère la machine comme un interlocuteur de sa communauté. L'application peut cependant inciter l'utilisateur à produire de tels gestes. Nous considérerons dans la suite que ce n'est pas le cas. Nous négligerons également les gestes paraverbaux, supposant en cela que l'utilisateur évitera de rythmer et de scander ses paroles par des gestes comme il le ferait en récitant un poème !

Nous mettons ainsi l'accent sur la seule catégorie restante, celle des gestes déictiques, qui sont parfois appelés gestes de démonstration, de monstration, d'indication, de désignation, d'ostension ou encore gestes ostensifs. La figure 1 de l'introduction présentait un tel geste sous la forme d'une trajectoire plane. Si ce type de trajectoire constitue notre contexte d'étude, nous retiendrons que les gestes ostensifs dans la communication humaine peuvent prendre aussi bien la forme d'un pointage de la main que d'un signe de la tête ou d'une indication par le regard.

La multimodalité telle que nous l'étudions dans le cadre de la référence aux objets se trouve au niveau de l'ostension. Elle fait intervenir non seulement les deux modalités d'expression que sont la parole et le geste, mais également la modalité de support qu'est la perception visuelle. Le geste ostensif s'appuie en effet sur le contexte visuel : sa précision, son étendue, sa forme dépendent des caractéristiques visuelles de l'objet visé et de sa disposition par rapport aux

objets non visés. D'une part, certains objets sont visuellement saillants, c'est-à-dire qu'ils sont perçus beaucoup plus rapidement et clairement que les autres objets. Un geste désignant un tel objet nécessite moins de précision qu'un geste désignant un objet non saillant. D'autre part, quelques objets peuvent former un groupe perceptif, dans le sens où, au premier regard, il est difficile de distinguer les éléments du groupe. Si l'objet visé appartient à un tel groupe, le geste a besoin d'être précis, voire de prendre une forme permettant de le séparer des autres éléments du groupe.

Ces deux notions de saillance et de groupe perceptif nous semblent fondamentales pour la compréhension de gestes et d'expressions référentielles multimodales en général. Elles n'interviennent pas seulement dans la perception visuelle mais également, comme nous allons y revenir tout au long de cette thèse, dans les phénomènes langagiers et attentionnels. Le geste lui-même a pour rôle principal de rendre un objet saillant. Selon Kleiber (1994), l'utilité du geste ostensif repose sur la conjonction de deux propriétés : le mode indexical de désignation (l'objet n'est pas appréhendé à travers le rôle ou les relations qu'il peut avoir dans la situation, mais est désigné directement) et l'apport de nouveau (le geste rend saillant pour l'interlocuteur un objet qui ne l'est pas encore). Pour qu'un geste ostensif soit utile, l'interlocuteur ne doit pas avoir l'objet visé à l'esprit, et cet objet ne doit pas être saillant dans la situation. Ce recours à l'interlocuteur se retrouve dans les travaux concernant l'évolution de la parole et du geste chez l'enfant. Cyrulnik (1995) voit ainsi la naissance du sens dans le geste d'ostension, et particulièrement dans le moment où l'enfant fait pour la première fois un tel geste tout en requérant l'attention de quelqu'un, par un regard insistant ou par une production verbale même incompréhensible. La multimodalité est ainsi vue comme le moyen originel d'expression de sens.

Le geste dans la communication homme-machine diffère du geste conversationnel à partir du moment où, en complément ou à la place de la commande vocale, est autorisée la manipulation directe. Un utilisateur ayant l'intention de déplacer un objet de l'application pourra vouloir agir directement sur l'objet à l'aide d'un geste. Celui-ci peut alors prendre plusieurs fonctions, au nombre de trois selon Cadoz (1994) :

- la fonction *épistémique* (prise de connaissance de l'environnement) ;
- la fonction *ergotique* (action matérielle : transformation de l'environnement) ;
- la fonction *sémiotique* (émission d'information à destination de l'environnement).

Si les gestes conversationnels ont tous par définition une fonction exclusivement sémiotique, le fait d'autoriser la manipulation directe introduit une ambiguïté : selon le dispositif d'acquisition choisi, même un geste intentionnellement ostensif pourra être interprété par la machine comme épistémique ou ergotique. Nous choisissons de considérer la communication homme-machine dans son aspect le plus spontané, le plus proche de la communication humaine, et nous ne tenons donc compte pour l'élaboration de notre modèle que de la fonction sémiotique. Dans le chapitre 9, nous étudierons l'adaptation de ce modèle à un type d'interaction particulier : celui induit par un dispositif à retour de force. Nous reparlerons alors des fonctions épistémique et ergotique et nous montrerons comment notre modèle s'y adapte.

La multimodalité. Nous avons montré comment la communication est par essence multimodale, et comment ce terme recouvre perception visuelle, parole et geste. Plus précisément, la communication est *multicanale*, les deux canaux principaux étant le canal visuo-gestuel et le canal audio-oral. Pour chaque canal, plusieurs *modes* ou *modalités* sont possibles. Pour le canal visuo-gestuel, on peut aussi bien considérer les gestes de la main que les mouvements de la tête ou la direction du regard. Ces trois modalités, que nous n'avons pas distinguées dans la classification des gestes communicatifs, se caractérisent premièrement par une dimension physiologique

correspondant à un sens humain et à une certaine intensité, et deuxièmement par un langage d'interaction. C'est ce langage d'interaction qu'il est nécessaire de définir dans un système de dialogue homme-machine. Si l'on veut que le système soit capable de traiter les mouvements de la main, de la tête et des yeux, autant de dispositifs d'acquisition et de langages d'interaction sont à prévoir. Un système exploitant plusieurs dispositifs ou *media* est appelé *multimedia*. Plusieurs dispositifs pouvant traiter une seule modalité (la reconnaissance de la parole peut s'effectuer grâce à un microphone couplé à une caméra qui lit sur les lèvres du locuteur), un seul dispositif comme une caméra pouvant traiter plusieurs modalités, il n'y a pas d'équivalence entre modalité et medium.

Dans notre modélisation, nous ne tenons pas compte des dispositifs mais seulement des modalités, qui reflètent les composantes fondamentales de la communication et non l'état actuel de la technologie. Parmi les modalités citées, nous nous concentrons sur la parole et le geste ostensif produit par la main, avec l'idée qu'elles véhiculent la quasi-totalité de la signification, du moins lors d'une action de référence. Cette focalisation sur le sens et non sur la forme se retrouve dans la connotation prise par les termes multimedia et multimodal : contrairement à un système multimedia, un système multimodal doit être capable d'interpréter les informations qu'il acquiert (Coutaz & Caelen 1991).

Le dialogue. Nous avons fréquemment employé le terme de système de dialogue homme-machine. Or ce terme recouvre plusieurs aspects, présentés ici du moins spécialisé au plus spécialisé en termes de possibilités d'interaction :

1. Dialogue homme-machine : il s'agit du dialogue au sens large, avec ou sans restriction d'expressivité. Le fait que l'interlocuteur soit une machine a des conséquences : l'absence de corps physique entraîne la réduction spontanée de la part de l'utilisateur de ses postures et de ses gestes ; l'aspect effectif de la communication entraîne la réduction des phénomènes caractérisant la conversation relâchée comme les répétitions ou les phrases inutiles. Kennedy *et al.* (1988) observent même une réduction spontanée des choix lexicaux. D'une manière générale, l'utilisateur sent qu'il doit aller droit au but, et cela renforce en lui l'application des principes de pertinence, comme ceux proposés par (Grice 1975) ou par (Sperber & Wilson 1995), principes que l'on suit spontanément dans notre façon de communiquer.
2. Dialogue finalisé : il s'agit du dialogue dirigé par un but à atteindre, dans le cadre d'une application particulière. Les objets et les actions appartiennent à un monde réduit, ce qui a pour conséquences la réduction du vocabulaire et de la syntaxe utilisée, la réduction des possibilités d'interaction, ainsi que la réduction des possibilités d'interprétation (cf. Pierrel 1987).
3. Dialogue à support visuel : contrairement au dialogue téléphonique (comme dans le cadre de l'automatisation des centres d'appel), le dialogue à support visuel entraîne un ancrage des actions dans un contexte visuel. La multimodalité s'avère alors incontournable : si l'on affiche des objets à l'écran, il est logique d'autoriser leur désignation par un geste. L'interaction prend certaines particularités dues au retour visuel des actions effectuées par la machine : l'utilisateur peut par exemple commenter verbalement une telle action.
4. Dialogue de commande : contrairement au dialogue de renseignement au cours duquel l'utilisateur pose des questions, il s'agit ici d'ordres que la machine exécute. Les énoncés prennent généralement la forme d'un prédicat suivi d'une ou de plusieurs expressions référentielles. Le retour visuel permet à l'utilisateur de constater le résultat et de continuer le dialogue. Nous nous focalisons ainsi sur les expressions langagières ou multimodales qui désignent

un objet ou un ensemble d'objets en exploitant les caractéristiques du contexte visuel.

Pour ces quatre types de dialogue, les étapes classiques de spécification du système de compréhension sont similaires. L'application permet de définir un modèle des actions et des possibilités d'interaction, un vocabulaire, une syntaxe. Une fois que les différents algorithmes de traitement des informations en entrée sont spécifiés et implantés, leur intégration se fait dans une architecture logicielle comprenant des modules pour la reconnaissance et la fusion des modalités, un module pour la gestion du dialogue, un historique de l'interaction. Nous y reviendrons dans le chapitre 8.

En ce qui concerne les phénomènes d'interaction, le dialogue de commande constitue notre contexte d'étude. Les exemples que nous avons construits ou que nous avons extraits d'enregistrements correspondent à un tel type de dialogue. Plus précisément, nous nous basons sur plusieurs types d'exemples :

1. Des exemples construits à partir de situations tirées d'un projet auquel nous avons participé : le projet COVEN. La tâche applicative impliquée consiste en l'aménagement de bureaux professionnels via une interaction multimodale dans un environnement virtuel. De manière simplifiée, il s'agit de manipuler des meubles : changer leurs propriétés, les déplacer, les remplacer par d'autres. Nous parlerons de ce projet dans le chapitre 9.
2. Des exemples imaginés lors de la phase de spécification d'un projet qui a commencé au cours de cette thèse : le projet MIAMM. La tâche consiste ici à accéder à des morceaux de musique (représentés par exemple par des icônes) via une interaction visuelle et *haptique* (c'est-à-dire incluant la perception tactile et le retour de force). Nous parlerons également de ce projet dans le chapitre 9.
3. Des exemples tirés de corpus multimodaux (extraits d'enregistrements) qui correspondent en fait à encore un autre type de dialogue, celui de la communication homme-homme multimodale *médiatisée* (ou *instrumentée*) : il s'agit de dialogues entre deux utilisateurs, l'un ayant une tâche à accomplir et l'autre jouant soit le rôle de la machine, soit le rôle d'un mentor dans le déroulement de l'interaction. Dans les deux cas (Wolff 1999 et Ozkan 1994), la tâche consiste à manipuler des objets abstraits tels que les formes géométriques de l'exemple de l'introduction. Ces extraits, ainsi que certains exemples modifiés que nous en avons déduit, nous serviront en particulier pour la caractérisation des phénomènes de référence en § 1.2.

Les exemples qui illustrent nos propos font ainsi intervenir aussi bien des chaises, que des icônes de fichiers musicaux ou des triangles. Nous montrons en cela que les phénomènes sur lesquels nous mettons l'accent apparaissent dans de telles tâches, et, de manière générale, dans toute tâche impliquant des références à des objets. Nous avons en effet l'ambition de modéliser la référence dans tout type de dialogue à support visuel.

1.1.2 Communication et interprétation automatique

La communication. Les deux points de départ régulièrement évoqués dans les recherches sur la communication sont, du côté des mathématiques, le modèle de Shannon & Weaver (1949), et du côté de la sociologie, la question de Lasswell (1948) : « [qui] [dit quoi] [dans quel canal] [à qui] [avec quel effet] ? ». Les deux mettent l'accent sur les différentes dimensions de la communication. Dans le domaine du dialogue de commande, c'est le *locuteur* (ou *utilisateur*) qui produit de la parole et des gestes (constituant un *message* et plus précisément un *énoncé*), dans les canaux

visuo-gestuel et audio-oral, et qui s'adresse ainsi à l'*interlocuteur* (ou *machine*) pour lui faire effectuer une action.

Le modèle de Shannon & Weaver fait en particulier ressortir l'aspect redondant de l'information communiquée. Ce point mérite d'être cité car la multimodalité est souvent vue comme redondante, particulièrement dans de nombreux travaux linguistiques qui considèrent le geste comme un moyen supplémentaire de désigner un objet déjà suffisamment décrit par le langage. Si un geste vers des objets de forme triangulaire n'a théoriquement pas besoin d'être associé au mot « *triangle* », il est cependant plus naturel que de l'associer à un mot abstrait comme « *chose* » ou « *objet* ». De plus, comme nous l'avons illustré dans l'introduction et comme nous le développerons tout au long de la suite, les informations véhiculées par la parole et le geste sont à la fois redondantes et complémentaires. Un système de dialogue homme-machine doit tenir compte de ce double aspect et ne pas considérer la redondance comme anormale.

Une étape importante dans les recherches sur la communication apparaît dans le schéma de Jakobson (1963), avec les six fonctions du langage qu'il identifie : la fonction *référentielle* servant à délivrer de façon neutre des informations brutes ; la fonction *phatique* regroupant les aspects de la communication qui prouvent une volonté d'adaptation du message aux destinataires et qui se traduisent par des phrases « pour ne rien dire » ou par des reformulations ; la fonction *émotive* mettant en jeu la sensibilité du locuteur qui juge et s'implique (permanence du « *je* ») ; la fonction *conative* permettant d'impliquer le destinataire par exhortation ou provocation (permanence du « *vous* ») ; la fonction *métalinguistique* permettant de parler du langage à l'aide du langage ; et la fonction *poétique* assurée par des effets de sonorités et de rythmes, ou par des métaphores. Bien que plusieurs des six fonctions puissent se retrouver simultanément dans un énoncé, certaines fonctions seront inhibées dans les situations de dialogue homme-machine à caractère finalisé. La fonction poétique s'avère rare à l'oral, et encore plus en situation de dialogue homme-machine. La fonction métalinguistique n'est de même pas très attendue dans un contexte finalisé. Elle apparaît néanmoins dans certaines situations, en particulier en réaction à une incompréhension ou à une erreur de la machine : « *je voulais parler du triangle et non du cercle* ». Dans le cadre du dialogue de commande, la fonction conative se traduit principalement par l'emploi de l'impératif. La fonction émotive n'a que peu d'effet sur une machine insensible aux émotions et programmée pour écouter. La fonction phatique s'oppose à l'efficacité qui caractérise la communication homme-machine et y est donc spontanément inhibée. Quant à la fonction référentielle, elle est fondamentale car elle correspond au mode d'interaction dans lequel se place l'utilisateur : chaque mot est porteur d'information et la référence est au cœur de la communication. Le dialogue homme-machine met donc l'accent sur la fonction référentielle.

Le référent. Tel qu'il ressort des travaux en sémiotique, le terme « référent » désigne une facette des signes. D'après (Klinkenberg 1996), un signe regroupe généralement quatre composantes : au départ, le *stimulus* est la manifestation concrète et sensible du signe (les ondes sonores ou lumineuses, par exemple). Il ne véhicule de signification que s'il correspond à un modèle abstrait pris dans un code. Ce modèle est le *signifiant* (par exemple les mots du langage), un signifiant n'ayant d'intérêt que s'il renvoie à quelque chose qui n'est pas lui-même. Le *signifié* est le concept ou l'image mentale suscitée par le signifiant, et renvoie à son tour à un *référent*, qui est ce à propos de quoi on communique : un objet du monde (« *une chaise* » ou « *le carré bleu* ») ; un groupe d'objet « *les triangles verts* » ; une espèce ou une classe d'objet (celle des carrés dans « *un carré a quatre côtés* ») ; un objet inexistant ou fictif (un cercle carré) ; une abstraction (la plausibilité mathématique) ; une qualité (la couleur) ; un événement (la suppression des triangles) ; un état ; un fait ; etc. Ces quatre composantes sont illustrées dans la figure 1.1 à partir de l'exemple de

l'introduction. Elles ne peuvent exister indépendamment les unes des autres : un stimulus n'est un stimulus que parce qu'il actualise le modèle qu'est le signifiant. Ce dernier n'a un statut que parce qu'il est associé à un signifié qui permet de le ranger dans une classe. Ce sont ces relations qui forment le signe.

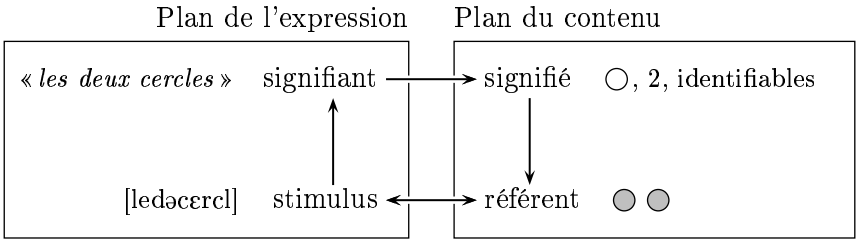


Figure 1.1 – *Modèle du signe adapté à l'exemple de l'introduction.*

Parmi les signes qui ont un référent, ceux pour lesquels la forme prise par le stimulus est indépendante de celle du référent sont dit *arbitraires*. Ils sont conventionnels et on doit en apprendre les règles (c'est le cas de la plupart des signes linguistiques). Parmi eux, les *index* ont pour fonction d'attirer l'attention sur un objet déterminé (Peirce 1960). C'est le cas des gestes ostensifs et, dans le langage, des déictiques (définis page 11). Les signes dont la forme entretient un rapport avec le référent sont appelés *motivés*. C'est le cas de certains gestes illustratifs.

Dans sa théorie générale des signes, Morris (1974) distingue trois niveaux d'analyse des signes : le niveau *syntaxique*, étudiant les relations des signes entre eux ; le niveau *sémantique*, portant sur la relation des signes à ce qu'ils désignent ; et le niveau *pragmatique*, prenant pour objet les relations entre les signes et leurs utilisateurs. Pour la sémantique, parler consiste à transmettre des significations. Pour la pragmatique, parler consiste à utiliser le langage d'une façon adaptée au contexte, à l'interlocuteur, et aux buts de la communication. Notre approche s'intéressant à l'interprétation de la référence en faisant large part à tous les aspects du contexte et au caractère finalisé de la communication s'incrit dans la pragmatique.

L'interprétation. Dans le cadre du dialogue homme-machine, l'interprétation commence par la réception d'un flux de parole, éventuellement couplé à une ou plusieurs trajectoires gestuelles. Une approche classique, qui suit la distinction de Morris et illustre bien les différentes étapes de l'interprétation de ces signaux, se caractérise par le traitement séquentiel suivant :

1. Analyse *lexicale* : à partir d'un lexique regroupant des termes appartenant au langage courant et des termes spécifiques à l'application, le but est de passer d'une suite de sons à une suite de mots de ce lexique.
2. Analyse *syntaxique* : elle consiste à structurer cette suite de mots, par exemple en une structure arborescente identifiant les groupes de mots selon leur fonction syntaxique : sujet, verbe, complément d'objet direct, etc.
3. Analyse *sémantique* : à partir de cette structure, elle a pour but de décrire le sens de l'énoncé, par exemple sous une forme logique traduisant la *compositionnalité* du sens : de manière simplifiée, l'expression « *les cercles gris* » conduit à une conjonction logique des sens des mots « *cercle* » et « *gris* » (pour cette étape comme pour la suivante, nous entrerons plus en détail dans les travaux de formalisation en § 1.3).
4. Analyse *pragmatique* : elle a pour but d'interpréter ce sens en tenant compte de la situation contextuelle, et consiste plus formellement à passer de la forme logique à un contenu

propositionnel qui décrit la proposition véhiculée par l'énoncé. C'est à cette étape que les références aux objets sont résolues. L'analyse pragmatique comprend également le calcul des *implicitations* (ce qui est calculé par inférence à partir du contenu propositionnel et du contexte, comme par exemple les présuppositions) et le traitement de l'*acte de langage* (assertion, interrogation, etc.). Dans le cas d'un ordre, l'application doit par exemple trouver les fonctions à appliquer pour rendre vrai le contenu propositionnel.

Dans la modélisation de ce processus, plusieurs phases de conception informatique sont nécessaires : la spécification du *modèle du domaine* (ou modèle de la tâche), c'est-à-dire des données de l'application (la base des objets et la base des fonctions) ; la spécification du *modèle du langage* ou modélisation du langage naturel (le signifiant) ; la modélisation des représentations correspondantes (le signifié) ; la spécification des correspondances entre ces représentations et les données de l'application.

Compte tenu de la richesse du langage naturel, cette dernière phase nécessite une modélisation complexe qui ne peut pas prendre la forme d'un *algorithme* (exploration systématique de toutes les situations possibles), mais fait intervenir diverses *heuristiques* (sélections dans les branches d'exploration). L'objectif est de spécifier un modèle suffisamment complet au départ pour limiter le nombre d'heuristiques nécessaires à la prise en compte de phénomènes particuliers. Sans encore entrer dans les détails de notre approche (cf. chapitre 3), nous faisons l'hypothèse que ce sera le cas si nous gardons des préoccupations cognitives dans l'élaboration de notre modèle. Les représentations que nous modélisons dans le système se rapprochent alors idéalement des représentations cognitives humaines.

Pour la résolution de la référence, notre objectif consiste à construire de telles représentations, en nous fondant sur l'exploitation des indices référentiels donnés par la situation et par l'énoncé langagier ou multimodal de l'utilisateur. La situation regroupe le contexte visuel, le caractère finalisé (ou intentionnel) de la communication, les contraintes dues à la tâche applicative, ainsi que la persistance d'une situation qui correspond au passé immédiat et qui peut se traduire par une certaine continuité dans les critères d'interprétation. L'énoncé comprend une expression référentielle langagière, qui prend généralement la forme d'un groupe nominal ou d'un pronom, par exemple : « *le fauteuil* » ; « *ce triangle* » ; « *le petit triangle bleu qui se trouve en haut à droite de la scène* » ; ou « *celui-ci* ». Une *expression référentielle multimodale* regroupe une expression référentielle langagière et un geste ostensif (« *cet objet* » associé à un geste vers un objet).

La génération. Dans le dialogue à support visuel, le résultat de l'interprétation faite par le système se traduit généralement par un retour visuel, une constatation visible de l'action effectuée. Le système peut de plus vouloir répondre à l'énoncé interprété. Quand cette réponse inclut la désignation d'un objet particulier à l'utilisateur, le système peut faire clignoter cet objet, l'afficher avec un rendu particulier, ou encore, si un avatar le représente en tant qu'interlocuteur, faire faire à cet avatar l'équivalent d'un geste ostensif (cf. par exemple Rist *et al.* 1997).

Un véritable dialogue implique de plus une réponse orale en langage naturel. La génération d'une telle réponse ou d'un énoncé en général est un processus que l'on peut décomposer en quatre étapes : l'étape pragmatique qui correspond à la sélection du contenu et fait intervenir le contexte situationnel ainsi que le but communicatif ; l'étape sémantique du « quoi dire » qui consiste à déterminer une représentation sémantique (comme une forme logique) de ce contenu ; l'étape du « comment le dire » qui consiste à sélectionner des unités lexicales et à leur appliquer des règles syntaxiques et morphologiques pour obtenir un énoncé sous la forme d'une suite grammaticalement correcte de mots ; et l'étape de synthèse qui consiste à « lire » cet énoncé (cf. en particulier Reiter & Dale 1997). Dans le chapitre 10, nous reviendrons sur ce domaine dans

le cadre de la référence aux objets, problème dont nous allons maintenant définir en détail les différentes facettes.

1.1.3 Référence et ostension

La référence. La distinction de Frege (1892) entre *sens* et *référence* constitue le fondement de nombreux débats philosophiques sur la référence[1]. Le sens d'une expression (ou *intension*) correspond, pour « *les deux cercles* », à quelque chose comme « *les objets, au nombre de deux, qui ont une forme circulaire* ». La référence (ou *extension*) renvoie dans le même exemple aux deux cercles gris de la scène en particulier. Certaines expressions ont un sens mais pas de référence, comme « *un triangle* » dans « *ajoute un triangle* », puisque le triangle en question n'existe pas encore. Une même expression peut prendre plusieurs références, dans des contextes faisant intervenir des objets différents. Deux expressions peuvent avoir la même référence avec des sens différents, par exemple « *le triangle vert* » et « *l'objet en haut à gauche* », qui réfèrent toutes les deux au même objet dans une scène contenant un seul triangle vert placé en haut à gauche.

Pour caractériser la référence, une méthodologie consiste à identifier les usages possibles d'une expression référentielle en fonction de ses composants. Un exemple célèbre est celui de Karttunen (1976), qui se pose des questions telles que : « sous quelles conditions un groupe nominal indéfini introduit un nouveau référent du discours ? ». Les exemples et classifications que nous citerons ci-après constituent la base pour une telle méthodologie. Une remarque préliminaire est que la forme du groupe nominal ne suffit pas à en faire une expression référentielle. En considérant l'expression « *le triangle* », on distingue :

- l'usage *référentiel* (ou *de re*), pour lequel l'expression est faite pour que l'interlocuteur retrouve le référent (« *le triange vert en haut à gauche* ») ou la classe complète (« *le triangle est une figure géométrique simple* ») ;
- l'usage *attributif* (ou *de dicto*), pour lequel l'expression ne réfère pas à un triangle en particulier mais à un triangle, quel qu'il soit (« *le triangle que je verrais bien ici* »).

Si le type d'usage peut souvent être retrouvé grâce au reste de l'énoncé et parfois même à partir du seul groupe nominal (comme dans le cas de l'article indéfini), il arrive que ce ne soit pas suffisant. C'est le cas de l'exemple « *l'objet en haut à gauche est vert* » adapté de Donnellan (1966). Selon la situation, l'énoncé peut se comprendre dans l'usage référentiel (« *l'objet qui se trouve actuellement en haut à gauche de la scène est vert* ») ou dans l'usage attributif (« *l'objet en haut à gauche, quel qu'il soit, triangle ou carré, doit toujours être vert* »). A l'intérieur de l'usage référentiel, une ambiguïté est parfois possible entre l'interprétation *spécifique* (qui aboutit au référent) et l'interprétation *générique* (qui aboutit à la classe complète). C'est le cas de « *un objet qui clignote est visible* » qui, selon la situation, peut se comprendre comme « *je vois un objet qui clignote* » ou comme « *d'une manière générale, un objet qui clignote est forcément visible* ».

Parmi les travaux français, ceux de Milner (1982) sont souvent présentés comme les premiers à proposer un modèle de traitement de la référence. Milner part d'une distinction proche de celle de Frege. Selon lui, une expression référentielle conduit à :

1. une référence *virtuelle* (la signification lexicale, donc indépendante de l'emploi) ;

1. Certains débats ont porté sur la nature directe ou indirecte des mécanismes d'attribution des référents. En considérant qu'il n'y a pas de sens intermédiaire mais une relation directe entre le mot et la chose, Russell incarne le point de vue opposé de Frege. Il l'illustre avec sa théorie des descriptions. Nous n'entrerons pas plus loin dans ces débats philosophiques, renvoyant à (Linsky 1967) et (Corazza 1995), pour nous concentrer sur ce qui est à la base de notre traitement des expressions référentielles.

2. une référence *actuelle* (le référent, donc en emploi).

Comme le sens chez Frege, la référence virtuelle est compositionnelle. Nous avons vu ci-dessus que certaines expressions ont par exemple une référence sans avoir de sens. De même ici, certaines expressions référentielles comme les pronoms n'ont pas de référence virtuelle. Elles sont incapables de déterminer par elles-mêmes leur référence actuelle, et sont dites privées d'*autonomie référentielle*. Milner précise cette notion dans (Milner 1989) : le processus qui permet d'attribuer un référent est un processus de *saturation sémantique*, et il y a faible saturation sémantique quand il y a manque d'autonomie référentielle. Le manque doit alors être suppléé par des informations que l'on va puiser ailleurs. Un démonstratif va par exemple puiser dans le contexte linguistique pour une reprise (comme dans : « *le cercle gris* » repris ensuite par « *ce cercle* ») ou dans la situation lors de son association avec un geste ostensif.

Selon Corblin (1987), la référence virtuelle se définit en fonction des propriétés qu'un référent doit posséder pour pouvoir être désigné. Si cela se conçoit bien pour une expression référentielle avec un article défini, il n'en est pas de même pour le démonstratif qui réfère par reprise et permet la *reclassification* (comme dans : « *le cercle gris* » repris ensuite par « *cette forme* »), ni pour l'indéfini qui ne désigne pas. Dans son livre dont le titre regroupe les trois grands types de déterminants (*Indéfini, défini et démonstratif*), Corblin se focalise sur les caractéristiques des références en fonction de leur détermination. Nous y trouvons ainsi la méthodologie descriptive indispensable pour aborder la formalisation, et nous y reviendrons.

La coréférence. Nous avons évoqué deux emplois du démonstratif : la reprise langagière et l'association avec un geste ostensif. Il s'agit de deux types de *coréférence*, ce terme désignant de manière plus générale toute référence double (deux expressions langagières, ou une expression langagière et un geste) sur le même référent. Plus précisément, la reprise constitue une coréférence *intra-mode* et l'association avec un geste une coréférence *inter-mode*. La première est nécessairement asynchrone (c'est-à-dire que les deux productions ne peuvent pas être simultanées mais successives) ; la seconde peut être synchrone ou asynchrone selon que le geste est produit simultanément, avant ou peu de temps après le démonstratif.

L'expression référentielle constituant la reprise et l'expression référentielle constituant l'antécédent sont les deux éléments d'une *anaphore*. Contrairement à l'antécédent, l'élément de reprise (ou élément anaphorique) est dépourvu d'autonomie référentielle. Les deux éléments ne sont pas forcément dans la même phrase. Si l'expression anaphorique apparaît avant son antécédent, on parle de *cataphore*. C'est le cas de : « *quand il était rouge, le triangle était plus visible* ». On parle d'anaphore nulle quand la reprise est élidée, comme dans « *devrait être là, le triangle rouge* » où « *il* » n'apparaît pas mais se déduit facilement.

Les deux éléments de l'anaphore sont souvent dans une double relation de coréférence et de reprise. C'est le cas de « *le cercle gris* » repris par « *ce cercle* », qui, du fait de l'identité des têtes nominales, est appelé anaphore *fidèle*. C'est aussi le cas de « *le cercle gris* » repris par « *cette forme* », qui, du fait de la relation conceptuelle ascendante de « *cercle* » vers « *forme* », est appelé anaphore hyperonymique. La coréférence est généralement virtuelle mais pas toujours actuelle : dans « *j'ai effacé le triangle vert mais il est revenu* », « *il* » réfère à nouveau triangle. Il existe des exemples d'anaphore sans coréférence. C'est le cas des référents évolutifs : dans « *peins le triangle rouge en bleu et mets-le ici* », « *le* » ne reprend plus « *le triangle rouge* » du début. C'est aussi le cas des anaphores associatives : dans « *mets la chaise ici* » suivi de « *recouvre le dossier* », l'élément anaphorique est une partie de son antécédent.

L'étendue des phénomènes anaphoriques est l'objet d'un débat que nous n'aborderons pas plus, renvoyant par exemple à (Krahmer & Piwek 2000). Retenons que le terme anaphore peut

même englober les groupes nominaux avec un antécédent non pas linguistique mais situationnel. On parle alors d'anaphore situationnelle, des exemples étant « *la scène* » ou « *ces trois triangles* » associé à un geste qui constitue alors l'antécédent. Toute référence multimodale peut ainsi être vue comme une anaphore.

Nous entrons alors dans un autre débat très présent dans les travaux linguistiques, concernant la compatibilité entre anaphore et deixis. Ce dernier terme désigne tout recours à la situation par un déictique ou un geste ostensif. Selon Corblin dans (Morel & Danon-Boileau 1990), la deixis est linguistiquement portée par le démonstratif : dans « *ce triangle* », la référence est créée par « *ce* » et explicitée ensuite par « *triangle* » : « *cela, qui, par ailleurs est triangle* ». La deixis ne se distingue alors de l'anaphore que par le geste ostensif associé à « *ce* ». Selon Danon-Boileau (*Ibid.*), il n'y a nullement opposition entre deixis et anaphore mais continuité graduée : ce qu'elles font, c'est présenter à l'autre un objet pour lui dire tout à la fois qu'il le connaît déjà, mais qu'il n'y a pas prêté d'attention suffisante.

Reboul dans (Moeschler & Reboul 1994) distingue dans la deixis la référence déictique (« *je* », « *ici* », « *maintenant* ») qui n'a pas besoin de geste, et la référence démonstrative ou association d'un groupe nominal démonstratif à un geste ostensif. Elle classe ainsi cinq modes de référence :

(1) référence directe : « *enlève la chaise qui est derrière le bureau* » ;

(2) référence indirecte : « *le support de l'écran* » (à propos d'une chaise qui a ce rôle de support) ;

(3) référence démonstrative : « *cet objet est de trop* » associé à un geste désignant une chaise ;

(4) référence déictique (ou deixis) : « ***je** parle de bureaux* » ;

(5) référence anaphorique (ou anaphore) : « *la chaise à gauche est de trop, supprime-**la*** ».

Dans cette classification, l'identification du référent se fait grâce à la référence virtuelle pour (1) et (2), grâce au geste ostensif qui permet d'attribuer une référence actuelle pour (3), grâce au recours à l'environnement physique lors du processus de saturation sémantique pour (4), et grâce au contexte linguistique lors du processus de saturation sémantique pour (5), le point commun entre (4) et (5) étant le manque d'autonomie référentielle. Reboul analyse les points communs et différences entre les cinq modes de référence et conclut qu'une définition de l'anaphore ne peut être en fin de compte que négative : ni (1), ni (2), ni (3) (d'où incompatibilité de l'anaphore avec le geste ostensif), ni (4). Nous retiendrons également que la simple présence d'un geste ostensif suffit à faire de la référence multimodale une référence démonstrative, même si l'expression verbale associée n'est pas un pronom ou un groupe nominal démonstratif.

Reboul (*Ibid.*) distingue enfin saturation sémantique et saturation référentielle : la première, dans la lignée de Milner, dépend de l'expression et de la référence virtuelle ; la seconde dépend de la capacité de l'expression en emploi à identifier un référent. Ainsi, un groupe nominal démonstratif indique par lui-même que la saturation sémantique ne suffit pas à identifier un référent (ce qui rejoint pour l'instant le point de vue de Milner). Or, si ce démonstratif est associé à un geste ostensif, il peut s'avérer de plus que cette association ne suffise pas à identifier le référent. Dans ce cas, l'expression n'est pas saturée référentiellement. C'était le cas, dans l'exemple de l'introduction, avec « *ces trois triangles* » associé à un geste ne désignant en première interprétation que deux triangles. Ce sont de telles expressions référentielles multimodales que nous allons étudier.

L'ostension non coréférente. De même qu'il existe des cas d'anaphore sans coréférence, il existe également des cas d'ostension sans coréférence. Ces cas, désignés par le terme *référence ostensive différée* dans (Quine 1971), correspondent à une référence multimodale indirecte, c'est-à-dire à une situation où l'objet désigné par le geste ne correspond pas au référent linguistique.

Un premier exemple, dans notre cadre de l'aménagement d'un intérieur professionnel, associe l'énoncé verbal « *cet employé a besoin de place* » avec un geste désignant un bureau. Le principe est que le geste désigne un objet qui n'est pas le référent réellement visé (l'« *employé* » qui n'est pas là mais à qui le bureau est destiné), mais conduit à ce référent (Kleiber 1994). Le geste joue ainsi le rôle d'une *métonymie*, terme qui désigne habituellement la substitution d'un mot par un autre dans le langage. Un deuxième exemple associe l'expression « *ce support pour écran* » à un geste désignant une chaise. La référence est ici directe par le geste et métonymique par le langage. Le référent réellement visé, celui sur lequel s'applique la proposition véhiculée par l'énoncé, est en effet celui de l'ostension.

D'autres exemples font intervenir la capacité de certains mots à connoter plusieurs concepts. Dans le cadre de l'accès à des morceaux de musique, le mot « *icône* » peut ainsi désigner la représentation visuelle ou le contenu du fichier. Dans l'association de « *cette icône* » avec un geste, deux cas se présentent : soit une référence démonstrative classique comme dans « *déplace cette icône ici* », soit une référence multimodale indirecte comme dans « *ouvre cette icône* ». Dans ce dernier cas, le geste désigne en effet la représentation alors que l'expression verbale désigne le contenu. Cette ambiguïté apparaît particulièrement dans le cadre du dialogue à support visuel faisant intervenir des représentations visuelles de concepts informatiques.

D'un point de vue conceptuel, la diversité des rapports référentiels entre langage et geste nous amène à employer un nouveau terme, celui de *demonstratum* (objet montré par le geste et ne correspondant pas forcément au référent). La figure 1.2 montre quel emploi nous ferons de ce terme par rapport au terme *référent*. Nous parlerons toujours d'expressions référentielles multimodales, directes ou indirectes, en appelant la partie langagière expression référentielle verbale ou démonstrative (même si elle ne contient ni déterminant ni pronom démonstratif), et la partie gestuelle ostension ou démonstration. Le terme référent sera utilisé pour référent intentionnel.

Dans les exemples du paragraphe précédent, les référents sont successivement un employé, une chaise, une icône-représentation et une icône-contenu. Ils diffèrent parfois des demonstrata : le bureau au lieu de l'employé, l'icône-représentation au lieu de l'icône-contenu. Il y a ici identité entre demonstrata intentionnels et demonstrata. Une différence apparaît par contre dans l'exemple de l'introduction, page 3 : les demonstrata sont les deux triangles concernés par la trajectoire gestuelle, et les demonstrata intentionnels, implicites, sont vraisemblablement les trois triangles qui correspondent clairement aux référents intentionnels. Cette différence n'est pas un cas exceptionnel mais apparaît au contraire facilement lors de situations de dialogue mettant en jeu des scènes visuelles complexes.

Terme	Utilisation
référents intentionnels	les référents de l'expression référentielle, multimodale ou non
référents langagiers	les référents de l'expression référentielle verbale seule
demonstrata intentionnels	les référents intentionnels du geste
demonstrata	les référents effectifs du geste

Figure 1.2 – *Distinctions entre demonstratum et référent, et entre intentionnel et effectif.*

Notre terminologie suit celle de Kaplan (1989), utilisée plus récemment, par exemple par Roberts (2002). Elle rejoint également la distinction de Corazza (1995) entre *référent intentionnel* (objet à propos duquel le locuteur a l'intention de parler) et *référent sémantique* d'un démons-

tratif (objet montré par le geste d'ostension qui accompagne le démonstratif). Le deuxième correspond au demonstratum.

La distinction entre demonstratum intentionnel et demonstratum illustre les conséquences de l'imprécision du geste, c'est-à-dire sa capacité à désigner autre chose que ce qui est intentionnel. Elle est quasiment inexistante dans les travaux linguistiques. On la trouve dans (Roberts 2002), qui fait une distinction entre *geste déictique* et *démonstration* : contrairement au geste déictique, la démonstration doit garantir que l'interlocuteur va retrouver le demonstratum intentionnel. Elle comprend donc une sorte de présomption de réussite, notion que l'on retrouvera plus loin avec la Théorie de la Pertinence.

1.2 Caractérisation des phénomènes de références aux objets

Toutes les définitions nécessaires ayant été établies, nous pouvons maintenant nous intéresser à une caractérisation systématique des phénomènes de référence, en nous restreignant désormais à la référence aux objets. Lors d'une première étape, nous nous intéresserons aux formes possibles d'expressions référentielles langagières, avec une attention particulière sur la détermination, c'est-à-dire sur les différentes possibilités d'articles et de déterminants qui viennent déterminer et quantifier l'accès aux référents. Lors d'une deuxième étape, nous nous focaliserons sur les phénomènes propres à la multimodalité, et en particulier sur la complétion des informations, c'est-à-dire sur les différentes manières selon lesquelles l'information référentielle se répartit entre le langage et le geste. Nous conclurons sur quelques conséquences qu'entraîne la prise en compte de ces phénomènes pour un système de dialogue homme-machine.

1.2.1 Phénomènes langagiers

Les composants d'une référence langagière. Il ne suffit pas *a priori* de prendre une grammaire du français pour, à partir des formes que peut prendre un groupe nominal, en déduire les formes possibles d'une expression référentielle verbale. Nous sommes donc parti d'un corpus de dialogue homme-machine pour en extraire les expressions référentielles et les classer ensuite en fonction de leurs composants. Ce corpus, dénommé Magnét'Oz, est celui élaboré par Wolff *et al.* (1998) et présenté en détails dans (Wolff 1999) : il consiste en une interaction multimodale avec microphone et écran tactile, telle que décrite dans l'introduction. Sans entrer pour l'instant dans les détails du protocole expérimental à l'origine de ce corpus, notons premièrement que la communication est spontanée et, en particulier, que la production de parole n'est pas contrainte ; et deuxièmement que la tâche applicative est très restreinte, se limitant au rangement de diverses formes géométriques dans des boîtes appropriées. Une copie d'écran est montrée figure 1.3.

L'expérimentation étant ciblée sur l'étude du geste, rien n'était attendu à propos des phénomènes langagiers. Tout portait à croire que le caractère fortement finalisé de la tâche allait entraîner des restrictions spontanées de la part des sujets. Loin de là, l'analyse du corpus que nous avons faite dans (Landragin 1999) a montré une grande richesse dans les expressions référentielles verbales. Cette richesse se rapproche de celle d'une grammaire des groupes nominaux du français, comme le montre la liste des composants observés (liste étendue parfois à quelques cas ponctuels pour compléter une classe grammaticale[1]) :

- Déterminants et articles (éventuellement associés à un marqueur déictique) :
  1. Déterminants du nom (ou articles) : articles indéfinis (« *un* », « *des* ») ; articles définis (« *le* ») ; articles partitifs* (« *du* », « *de la* »).

1. Ces cas non observés dans le corpus sont repérés dans la liste par un astérisque.

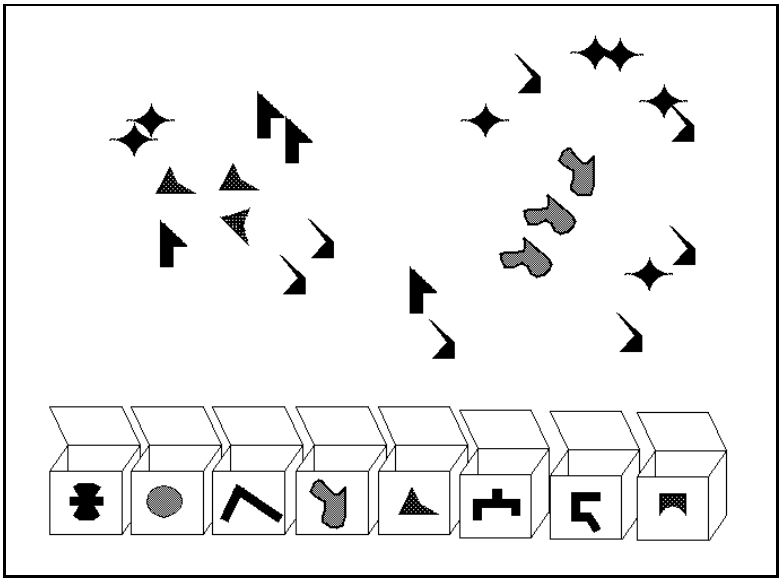


FIGURE 1.3 – *Exemple de scène visuelle dans l'enregistrement Magnét'Oz (Wolff 1999).*

2. Déterminants démonstratifs : formes simples (« *ce carré* », « *cet objet* ») et formes renforcées par un marqueur déictique (« *ce carré-ci* », « *cet objet-là* »).
3. Déterminants possessifs (« *mon* », « *notre* »*).
4. Déterminants indéfinis : quantifiants évoquant une quantité nulle* (« *aucun* », « *nul* ») ; quantifiants évoquant une quantité égale à un (« *chaque* », « *tout* » suivi d'un singulier, « *un* », « *une* ») ; quantifiants évoquant la pluralité (« *plusieurs* », « *quelques* », « *la plupart* »*, « *des* ») ; quantifiants évoquant la totalité (« *tous* ») ; quantifiants et, simultanément, caractérisants (« *certains* », « *divers* », « *différents* ») ; caractérisants (« *même* » [1], « *autre* » [2], « *tel* »*, « *quel* »*).
5. Déterminants numéraux cardinaux (« *un* », « *deux* », « *trois* »), qui peuvent être employés par exemple avec un article défini (« *les deux triangles* »), ou sans rien, c'est-à-dire avec un indéfini absent en surface mais existant dans la structure de l'expression (« *deux carrés* »).
6. Déterminants à base adverbiale (« *beaucoup de* », « *trop de* »*, « *peu de* », « *assez de* », « *suffisamment de* »*, « *plus de* »*, « *moins de* »*).
7. Déterminants formés sur des noms de quantité (« *une rangée de* »).

- Pronoms :
  1. Pronoms personnels représentants : formes non réfléchies (« *il* », « *le* », « *lui* ») et pronoms adverbiaux (« *y* », « *en* »).

1. Mot dont le fonctionnement est particulièrement intéressant : si « *les objets de même couleur* » ne pose pas de problème d'interprétation de « *le même* », il arrive souvent que « *les mêmes objets* » ne désigne justement pas les mêmes objets mais d'autres objets du même type.

2. Mot également intéressant, dont le fonctionnement consiste généralement à reprendre la catégorie d'un premier élément cité, et à en chercher les représentants (à l'exclusion de ce premier élément). Nous noterons que ce premier élément peut parfois être élidé sans gêner la compréhension. Dans le corpus Ozkan, Salmon-Alt (2001b) remarque ainsi l'exemple suivant : « *l'église est là, et à sa gauche, il y a une autre maison* ». Nous aurons l'occasion de revenir sur le fonctionnement de ce mot, par exemple dans le chapitre 10 pour la validation d'un aspect de notre modèle.

2. Pronoms démonstratifs (« *celui* », « *ce* ») : formes simples pouvant être complétées par un complément du pronom (« *celui au-dessous* »), par une proposition subordonnée relative (« *celle qui* ») ; formes renforcées (« *celui-ci* », « *ceci* », « *cela* », « *ça* »).
3. Pronoms possessifs (« *la sienne* », « *le nôtre* »*).

- Noms et substantifs : catégories (« *triangle* », ou tout simplement « *objet* »[1]).
  La suite de l'énumération contient les restricteurs de sens d'un substantif.
- Adjectifs qualificatifs :
  1. Adjectifs qui fonctionnent comme déterminants du nom : adjectifs numéraux ordinaux (« *premier* », « *deuxième* », « *dernier* ») ; adjectifs possessifs ; adjectifs indéfinis ; adjectifs démonstratifs.
  2. Adjectifs qui fonctionnent comme modifieurs du nom (« *grand* », « *rouge* », « *rayé* », « *brillant* », « *restant* »).
- Compléments du nom ou de l'adjectif, en particulier les subordonnées relatives (« *ces objets qui forment un triangle* ») et les adverbes : adverbe de degré d'intensité (« *les formes très claires* ») et adverbe de degré de comparaison (« *les carrés les plus petits* »).

Ces composants se combinent pour former des constructions telles que :

- Groupes nominaux (au moins un déterminant et un nom), avec ou sans adjectif, avec ou sans complément du nom ou de l'adjectif, avec ou sans coordination de noms ou d'adjectifs (« *cet objet et celui-ci* », « *cet objet plus celui-là* »).
- Groupes nominaux sans nom (cf. aussi Salmon-Alt 2001b) :
  1. Groupes nominaux elliptiques : ellipse de la tête nominale (« *le rouge* ») ; ellipse du complément du nom (« *cette ligne* » juste après « *cette ligne d'objets* »).
  2. Références mentionnelles (« *le premier* » puis « *le second* », « *l'un* » puis « *l'autre* »).
  3. Pronoms.
- Locutions prépositionnelles, c'est-à-dire expressions ayant une forme réductible à : « *groupe-nominal*$_1$ *préposition groupe-nominal*$_2$ ». Elles regroupent deux références, une première (« *groupe-nominal*$_1$ ») se basant sur le résultat de la seconde (« *groupe-nominal*$_2$ »). Les mécanismes que ces locutions mettent en jeu dépendent du type de la préposition (Briffault 1992) :
  1. Prépositions topologiques (« *à* », « *contre* », « *sur/sous* », « *dans/chez/en/hors de* »).
  2. Prépositions projectives (« *à gauche/à droite* », « *devant/derrière* », « *au dessus/au dessous* », « *à l'ouest/est/nord/sud* »*).
  3. Prépositions géométriques (« *entre* », « *parmi* », « *outre* »*, « *de l'autre côté de* », « *les deux formes de part et d'autre de celle-ci* »).

Dans un contexte visuel, ces différents mots et constructions syntaxiques peuvent apporter plusieurs types d'information sémantique pour la détermination du référent. En prenant l'appli-

1. Le corpus contient un cas extrême d'utilisation de ce mot particulièrement abstrait : « *cet objet et cet objet* », où la première occurence du mot est associé à un geste désignant un objet et la seconde à un geste désignant une entité correspondant à un groupe de trois objets.

cation Magnét'Oz comme source de phénomènes, ces informations peuvent être classées de la manière suivante :

- Aucune information sur les référents, à part le fait qu'ils soient directement identifiables : « *ça* » n'indique ni leur catégorie ni leur nombre (même pas le singulier ou le pluriel).
- Informations liées au mode de référence et au nombre de référents : le déterminant ou le pronom, éventuellement un marqueur déictique, avec éventuellement un adjectif numéral. Le mode de référence et le nombre de référents sont deux informations qui peuvent être portées par le même mot, le cas typique étant l'article défini au singulier. Le corpus présente* quelques cas d'ambiguïté sur le nombre, par exemple entre « *celle-là* » et « *celles-là* » qui ne se distinguent pas à l'oral. De nombreux cas d'association entre un article défini et un marqueur déictique apparaissent (« *la forme-ci* », « *l'objet-là* »).
- Catégorie des référents, souvent imprécise dans le corpus du fait du choix de formes géométriques difficile à dénommer (« *objet* », « *forme* », « *forme géométrique* », « *figure* », « *pièce* », « *cercle* », « *triangle* », « *pointe* », « *flèche* », « *bout de flèche* »).
- Informations liées au contexte visuel :
  1. Informations liées aux caractéristiques visuelles des référents, soit par la mention directe d'une propriété (« *gris* », « *gris clair* », « *clair* », « *en pointillés* », « *à petits pois* », « *qui comporte un angle droit* », « *en forme de L* », « *en forme de hache* »), soit par la mention d'une homogénéité par rapport à d'anciens référents (composés de « *celui* » qui reprend la catégorie, « *les objets de même couleur* » qui reprend une propriété), soit par la comparaison sur une propriété des référents à d'autres objets (« *les trois formes les plus claires* »).
  2. Informations liées à la disposition spatiale des référents, soit par la description de la disposition des référents entre eux (« *cette ligne d'objets* », « *ces objets formant un triangle* », « *les trois objets accolés* », « *ce groupement d'objets* »), soit par la description de leur disposition par rapport à d'autres objets (« *l'objet le plus à droite* »), soit par la description de leur disposition dans la scène (« *les deux objets au centre* »).
  3. Informations liées au parcours d'un ensemble délimité par le contexte visuel (« *le premier* », « *le deuxième* », « *le dernier* », « *l'autre* »).
- Informations liées au contexte linguistique, qui correspondent à des instructions de parcours d'un ensemble préalablement délimité par le langage (« *ce dernier* », « *le premier* », « *le second* », « *l'autre* »).
- Informations liées au contexte applicatif (« *le suivant* », « *le dernier* », « *ceux qui restent* », « *les autres* »).

Ce qui apparaît avant tout, c'est qu'un type de mot n'est pas lié à un type d'information. Nous avons évoqué cette idée plusieurs fois en § 1.1.3 avec les différentes interprétations que nous avons données de groupes nominaux similaires, ou encore avec les deux utilisations fondamentales du démonstratif que sont la reprise anaphorique et l'association avec un geste ostensif. Nous le voyons ici en particulier avec les références mentionnelles, qui peuvent être interprétées comme des instructions de parcours d'un ensemble préalablement délimité par la perception visuelle, par le langage ou encore par des contraintes applicatives. Nous le verrons également avec l'étude des déterminants et de leur possible association avec un geste ostensif (§ 1.2.2).

Les spécificités de l'oral. En se plaçant au niveau du vocabulaire utilisé à l'oral et en considérant les actions de référence, nous noterons la présence, relativement fréquente par rapport à un niveau de langue écrite, de termes vagues tels que « *truc* » ou « *machin* ». Ces mots ne fournissent que peu d'indications pour la résolution de la référence. Les expressions référentielles verbales peuvent également comporter quelques créations de mots, souvent à partir de mots d'autres catégories (« *le bureau est-il déplaçable ?* »).

Au niveau de la syntaxe, nous noterons deux phénomènes propres à la production de l'oral et pouvant apparaître à l'intérieur d'une expression référentielle : les périphrases, dues aux problèmes liées à la recherche du terme exact, et les phrases *emphatiques* (« *le triangle, le rouge, il faut le supprimer* » ; « *le bureau, sa chaise, elle doit se mettre ici* »). L'expression référentielle à prendre en compte combine alors les deux groupes nominaux utilisés.

L'utilisation spontanée du langage dans un niveau de langue non surveillée entraîne des ratés. Ces ratés se traduisent dans le matériau verbal par des bafouillements, des bégaiements, des lapsus, des marques d'hésitation, des répétitions ou encore des reformulations. Ils sont souvent accompagnés de marques particulières dans les matériaux paraverbal et non verbal, telles que des changements de rythme ou d'intonation au niveau de la prosodie, ou encore des gestes paraverbaux particulièrement appuyés. En nous inspirant de (Lopez 1999), nous distinguons :

1. Les phénomènes de bruit : les hésitations et les interruptions, qui se traduisent par des pauses ou par des interjections (« *euh* ») ; et les répétitions (« *le triangle rouge, le triangle rouge, lui je le mets ici* »). Le système doit ignorer ces phénomènes.
2. Les phénomènes d'omission : il s'agit de manques dans la structure syntaxique, par exemple la conjonction dans une coordination ou l'ellipse nominale. Le système doit compléter ces omissions, généralement à partir des structures employées dans les énoncés précédents.
3. Les phénomènes de distorsion : les précisions, qui enrichissent une première expression choisie rapidement, et les corrections, qui effacent une première expression choisie rapidement. Dans les deux cas, la première expression peut être interrompue, parfois même au milieu d'un mot. Les corrections se repèrent souvent par la présence d'une marque d'hésitation, d'une particule négative (« *non* ») ou d'une marque d'excuse (« *pardon* »). Le système doit interpréter ces phénomènes avant de résoudre les références.

Pour un système de dialogue homme-machine, le problème se situe au niveau de l'identification de la nature du raté. Ainsi, face à une expression telle que « *le triangle rouge, le triangle rouge* », le système peut faire plusieurs interprétations : celle de la répétition et celle de l'énumération de plusieurs références. De même, face à une expression référentielle telle que « *le cercle, le machin rouge* », le système doit déterminer s'il s'agit d'une précision, auquel cas il recherche un cercle rouge ; d'une correction, auquel cas il recherche une forme géométrique rouge qui n'est sans doute pas un cercle ; ou encore d'une énumération, auquel cas il recherche deux objets, un cercle et une forme rouge. Nous ferons l'hypothèse que les matériaux paraverbal et non verbal permettent dans la majorité des cas de lever l'ambiguïté et d'aborder la résolution de la référence sur une base correcte. Pour ce faire, le module de reconnaissance de la parole doit détecter les particularités prosodiques d'un énoncé pour les transmettre jusqu'aux modules sémantiques et pragmatiques qui pourront alors les exploiter. Si l'ambiguïté ne peut être levée, les différentes hypothèses seront testées et celle aboutissant au meilleur résultat sera privilégiée (cf. chapitre 8).

Les spécificités du dialogue. La référence dans un contexte interactionnel est plus complexe que la référence dans le texte ou dans le discours oral. Nous venons de le voir avec les phénomènes

liés à la production de parole, et nous le voyons ici avec les phénomènes liés à la situation de l'énonciation.

La référence en situation de dialogue dépend en effet de la complexité des organisations d'échange. Son interprétation repose sur la prise en compte du contexte d'énonciation, des intentions des interlocuteurs, de l'implicite dû à la situation. Pour illustrer cette complexité, Vivier (2001) parle de *référenciation* et non de référence : « en choisissant le terme de référenciation et non pas celui de référence, nous attirons l'attention sur tout ce qui fait la complexité psychologique d'un tel processus. Ce terme ne se réduit pas, en effet, au sens linguistique classique de *référence* : mettre en rapport des signes avec des objets (et surtout pas à des rapports de bijection entre les deux). La référenciation, ce n'est pas simplement *repérer de quoi on parle ?*, c'est aussi pour chacun des interlocuteurs *repérer qui parle à qui ?*, gérer l'interaction entre eux et co-construire la communication elle-même en ajustant leurs intentionnalités. [...] Un modèle de référenciation devrait proposer un véritable ancrage des référents introduits dans le discours sur un monde extérieur à celui-ci. » (p. 6–7).

Pierrel & Romary dans (Sabah *et al.* 1999) montrent qu'un traitement fin de la référence est indispensable pour un dialogue homme-machine naturel. Si les systèmes de dialogue homme-machine progressent, c'est bien parce qu'ils prennent en compte des propriétés de référenciation, et non parce qu'ils enrichissent leur vocabulaire ou leurs règles syntaxiques. Les progrès relèvent de la pragmatique, et la référenciation (ou référence comme nous continuerons à l'appeler) est au cœur du problème : elle fait intervenir toutes les composantes d'un système, particulièrement quand elle est multimodale.

1.2.2 Phénomènes multimodaux

Les composants d'une référence multimodale. Comme en § 1.2.1, nous partons du corpus Magnét'Oz et de l'analyse qui en a été faite dans (Wolff 1999) et (Landragin 1999). Une première phase a permis de dégager du corpus des catégories de formes prises par le geste (cf. Wolff 1999). Nous reviendrons plus tard sur ces formes qui dépendent très fortement de l'application et du dispositif d'acquisition qu'est l'écran tactile. Nous voulons rester ici au niveau du geste ostensif en général, et nous nous focalisons sur une deuxième phase d'étude du corpus (cf. Landragin 1999), ciblée sur les rapports entre parole et geste en termes de complétion d'information et de compensation d'imprécisions.

Considérons tout d'abord le cas le plus simple correspondant à l'assocation de « *ce triangle* » à un geste désignant un triangle. Au départ, une expression référentielle démonstrative peut s'interpréter comme une référence anaphorique ou démonstrative. La simple présence d'un geste, du moins s'il est bien marqué prosodiquement, suffit quasiment à rejeter l'anaphore. Même principe pour une expression telle que « *la forme bleue, celle-ci* », qui peut être considérée au départ comme une énumération ou une précision : selon la présence d'un ou de deux gestes, selon le fait que les deux gestes désignent la même forme bleue ou non, l'hypothèse d'énumération sera privilégiée ou rejetée en faveur de la précision.

Considérons ensuite que l'interprétation du geste permet de plus d'attribuer une ou plusieurs hypothèses de référents à l'expression, le geste étant par nature souvent imprécis. Dans le cas où les hypothèses du côté du geste diffèrent sur la nature du demonstratum ou des demonstrata, l'expression verbale permet de lever l'ambiguïté si elle comprend une mention de catégorie ou de propriété. C'est le cas de « *le triangle* » associé à un geste vers un amas de formes géométriques comprenant un seul triangle.

Dans le cas où les hypothèses gestuelles diffèrent sur le nombre des demonstrata, la présence dans l'expression verbale d'un adjectif numéral (« *ces trois triangles* » associé à un geste désignant

de manière imprécise deux ou trois triangles) ou d'une coordination (« *ce triangle, ce triangle et celui-ci* » associé à un geste similaire) permet d'en privilégier une.

Une véritable ambiguïté multimodale se présente quand l'expression verbale et le geste ostensif sont tous les deux imprécis, comme dans « *ces objets* » associé à un geste désignant de manière imprécise quelques triangles et quelques carrés.

L'utilisation du terme « *objet* », qui peut désigner aussi bien un triangle, une forme géométrique en général, ou même une entité telle qu'un amas de formes, peut également conduire à des ambiguïtés complexes. Dans le corpus Magnét'Oz, nous trouvons ainsi l'expression « *ce groupement d'objets et celui-ci* » associée à deux gestes, le premier vers un amas de formes et le second vers une forme assez proche d'un autre amas. Ce second geste peut alors s'interpréter comme désignant un seul objet ou comme désignant un groupement d'objets, et l'expression verbale n'apporte aucun indice fiable : on peut considérer que « *celui* » reprend la catégorie « *groupement* » et privilégier l'hypothèse de désignation d'un groupe, ou considérer que « *groupement* » est un quantificateur et que la tête nominale reprise par « *celui* » est « *objet* », auquel cas on privilégie plutôt l'hypothèse de désignation d'un seul objet. La configuration spatiale des objets en groupements plus ou moins perceptibles sera un indice plus pertinent et permettra de privilégier une des deux hypothèses.

Enfin, certaines ambiguïtés sont propres à la prononciation orale, en particulier lorsque l'interlocuteur ne peut pas distinguer à partir du matériau verbal le singulier ou le pluriel, comme dans « *celle(s)-ci* » associé à un geste imprécis, seule apparition d'une telle ambiguïté dans le corpus Magnét'Oz.

Les spécificités de la multimodalité spontanée. L'utilisation spontanée de la multimodalité a des conséquences sur la production de gestes. D'une manière générale, la continuité de la parole se retrouve dans la modalité gestuelle. Les gestes sont émis en continu et, en particulier dans le corpus Magnét'Oz, il est parfois difficile de segmenter le signal en trajectoires bien distinctes. Le corpus contient même un cas (unique) de passage en continu d'un geste ostensif à un geste illustrant un déplacement et inversement (« *cet objet doit aller ici* » associé à une seule trajectoire commençant par une désignation de l'objet en question, suivant le mouvement de déplacement et aboutissant à la désignation du lieu de destination). Nous ne tiendrons pas compte de ce cas dans notre modélisation, mais nous retiendrons que le geste spontané en communication homme-machine ne se réduit pas toujours à une série de désignations ponctuelles.

Le fait que la production orale présente des ratés a des répercussions sur l'utilisation de la multimodalité. Nous retrouvons les phénomènes d'hésitation ou de répétition évoqués précédemment : en même temps que la parole, le geste peut être interrompu, puis repris, en lien avec le faux départ ou non. Chen *et al.* (2002) montrent ainsi qu'il existe une corrélation entre le type de geste effectué pendant une réparation verbale et le type de réparation. Ils montrent également qu'un geste n'est pas toujours produit, et en particulier qu'une répétition n'est quasiment jamais accompagnée de geste. Des gestes apparaissent lors des ratés dans les actions de référence du corpus Magnét'Oz (cf. Landragin 1999). Les phénomènes observés sont les suivants :

1. Les phénomènes de bruit : une interruption dans la production d'un geste, parfois liée à une interruption dans la production verbale ; la répétition d'un geste, très souvent couplée à la répétition de l'expression verbale correspondante. Ces phénomènes doivent être ignorés par le système.
2. Les phénomènes de distorsion, c'est-à-dire la précision et la correction. La précision d'un geste n'est pas forcément liée à une précision langagière. Le système peut cependant la

détecter car le geste devient plus lent, plus réfléchi. Quant à la correction d'un geste, elle correspond généralement à un changement d'intention référentielle, qui se traduit à la fois par une correction verbale et une correction gestuelle, ce qui permet de l'identifier facilement.

Les ratés dans la production d'expressions référentielles multimodales peuvent se traduire également par une erreur de l'utilisateur dans l'une des deux modalités. Lorsque l'utilisateur porte son attention majoritairement sur une modalité, les chances de lapsus ou de décalage sur l'autre modalité en sont effectivement augmentées. Si l'erreur ne peut pas être compensée ou corrigée lors de l'appariement des modalités, elle est néanmoins détectée et une stratégie de réparation telle qu'une question peut être initiée par le système.

Un autre type de raté consiste en une mauvaise synchronisation temporelle (cf. Oviatt *et al.* 1997, Oviatt 1999). Le geste est en effet plus facile à produire et peut apparaître plusieurs secondes avant le segment verbal correspondant. Ce décalage pose problème au système dans la mesure où il rend plus difficile l'appariement d'un geste à l'expression référentielle verbale qui lui est associée. La prise en compte de critères pragmatiques nous semble ici aussi constituer une réponse pertinente à cet aspect de bas niveau dans le traitement.

Nous noterons également que la multimodalité n'est pas utilisée de la même façon par tous les utilisateurs. C'est aussi l'une des remarques d'Oviatt (1999). Certains utilisateurs privilégient clairement la parole ; d'autres se placent presque systématiquement dans une utilisation redondante (« *ce triangle rouge* » associé à un geste précis vers le seul triangle rouge de la scène) ; d'autres encore vont exploiter les caractéristiques du contexte visuel et la complétion d'information pour produire les gestes les plus rapides (et par conséquent imprécis) et les expressions les plus courtes possibles. Fondé sur la spontanéité de la communication, le système doit tout accepter. Il peut gérer un profil de l'utilisateur pour se familiariser avec les habitudes de celui-ci et pouvoir mieux les comprendre ultérieurement.

Les appariements multiples. Lorsque plusieurs références multimodales se suivent dans un même énoncé, on peut assister à un phénomène que nous avons appelé *référence multimodale combinée* (Landragin 1999) et qui consiste en l'association d'une désignation dans une modalité à plusieurs désignations dans l'autre modalité. Un exemple simple et fréquent dans le corpus Magnét'Oz est l'association de « *ces objets* » à plusieurs gestes ostensifs, un par objet. D'un point de vue combinatoire, plusieurs formes sont possibles :

1. L'association d'une expression référentielle *unitaire* (c'est-à-dire comportant un seul groupe nominal) à plusieurs gestes avec distribution parfaite (« *ces trois objets* » associé à trois gestes, chacun d'eux désignant un seul demonstratum).
2. L'association d'une expression référentielle unitaire à plusieurs gestes avec distribution complexe (« *ces trois objets* » associé à deux gestes, le premier englobant deux demonstrata, le second désignant un seul demonstratum).
3. L'association d'une expression référentielle *plurielle* (c'est-à-dire comportant plusieurs groupes nominaux coordonnés ou simplement énumérés) à un seul geste (« *cet objet, cet objet et celui-ci* » avec un geste englobant trois demonstrata ; « *ces deux objets et celui-ci* » avec un geste englobant trois demonstrata) ;
4. L'association d'une expression référentielle plurielle à plusieurs gestes, avec distribution parfaite (« *cet objet, cet objet et celui-ci* » associé à trois gestes désignant chacun un seul demonstratum).

5. L'association d'une expression référentielle plurielle à plusieurs gestes, avec distribution complexe (« *cet objet, cet objet et celui-ci* » associé à deux gestes, le premier désignant deux demonstrata et le second un seul demonstratum ; « *ces deux objets et celui-ci* » associé à trois gestes désignant chacun un demonstratum ; « *ces deux objets et cet objet* » associé à deux gestes, le premier désignant un seul demonstratum et le second deux demonstrata).

Parmi ces exemples, seul le tout dernier n'apparaît pas dans le corpus Magnét'Oz. Le traitement de tels phénomènes se situe au niveau du comptage des référents et des demonstrata. Le problème est surtout dû à l'imprécision du geste et à son incapacité à déterminer de lui-même le nombre de demonstrata. Du côté de l'expression référentielle verbale, seule l'utilisation du pluriel sans adjectif numéral cardinal (« *ces objets* ») pose problème à ce niveau.

Quelques conséquences pour la modélisation. Nous avons énuméré au cours des pages précédentes un grand nombre de phénomènes apparaissant lors d'actions de référence dans la communication spontanée. Un système de dialogue homme-machine véritablement intelligent doit traiter tous ces phénomènes, qui, comme l'ont montré nos exemples de triangles et de chaises, ne sont pas marginaux mais rendent compte au contraire de situations simples à la base de la compréhension. L'étendue particulière des phénomènes dans la communication multimodale illustre la complexité du problème.

De même que nous avons vu qu'aucun type de mot n'est réservé à un type d'information, qu'aucun type de détermination n'est réservé à un mode de référence (un groupe nominal défini tel que « *le triangle* » peut référer aussi bien directement qu'anaphoriquement), nous constatons ici qu'aucun type d'expression référentielle n'est réservé à l'association avec un geste ostensif. En particulier, le geste peut être associé aussi bien à un démonstratif qu'à un défini, et un groupe nominal démonstratif tel que « *ce triangle* » peut référer aussi bien anaphoriquement qu'avec un geste coréférent. Autrement dit, les phénomènes ne sont pas marqués linguistiquement.

Pour la conception d'un système de dialogue, l'enjeu est important car les indices linguistiques deviennent difficilement exploitables. Aucun indice linguistique ne permet de manière sûre de prévoir la présence d'un geste. Aucun indice linguistique ne permet de manière sûre de retrouver l'intention référentielle de l'utilisateur. Pour déterminer le mode de référence d'une expression référentielle dans une situation donnée, il est ainsi nécessaire d'analyser plusieurs types d'indices : des indices compris dans l'expression même, des indices compris dans le reste de l'énoncé, des indices compris dans l'historique du dialogue, dans le contexte visuel, bref, des indices compris dans la situation, en tenant compte de tous les aspects de celle-ci. Comme nous l'avons vu en fin de § 1.2.1, cette analyse constitue la facette principale du problème de l'interprétation en dialogue homme-machine. Pour l'aborder, il est nécessaire avant tout de traduire les informations constituant la situation dans des structures de données et dans des langages propres à la machine. C'est l'objet de la section suivante.

1.3 Formalisation des modalités pour la résolution des références

Avant de confronter les informations portées par les différentes modalités, il s'avère nécessaire de formaliser ces informations, à l'intérieur de chaque modalité. Nous commencerons par faire une revue des travaux portant sur la formalisation de la perception visuelle (modalité de support) avant de nous intéresser à celle de la parole et du geste (modalités d'expression). Nous décrirons à chaque fois les points de départ du système et les étapes permettant d'aboutir à une représentation formelle des contenus, en nous focalisant sur les contenus qui présentent un intérêt pour la résolution des références aux objets.

1.3.1 La perception visuelle, modalité de support

La reconnaissance d'objets. Sans entrer dans les détails du processus de reconnaissance des objets affichés dans la scène visuelle, nous noterons que ce processus met en rapport une structuration de la forme perçue, avec des représentations d'objets propres à l'utilisateur. Le traitement des contours, des jonctions, de la fermeture et des surfaces caractérise la structuration de la forme (Boucart 1996). La théorie de la Gestalt propose pour ce problème une liste de principes devenus incontournables dans les travaux sur la perception visuelle. Il s'agit des critères de :

- proximité (des unités proches se rassemblent) ;
- ressemblance ou similarité (des unités similaires se rassemblent) ;
- continuité (des unités disposées de manière continue et régulière se rassemblent) ;
- fermeture (par exemple, trois traits même non contigus forment un triangle).

Ces principes décrits par Wertheimer (1923) s'appliquent également au problème de la reconnaissance de groupes d'objets (ou *groupage*) comme nous le verrons très bientôt.

Dans le cadre du dialogue finalisé, nous considérons que l'utilisateur aussi bien que le système connaissent les objets de l'application, du fait de ce caractère finalisé de l'interaction. L'utilisateur en possède des représentations claires et les reconnaît immédiatement, quelle que soit la scène visuelle. Ses énoncés oraux comportent par conséquent des termes prévus par l'application, termes correspondant aux catégories des objets et à leurs propriétés. Le système n'est donc aucunement concerné par le problème de la structuration de la forme : il connaît les objets et, *a priori*, n'a pas besoin de les retrouver à partir d'énoncés imprécis traduisant une mauvaise reconnaissance.

Nous noterons également que certains aspects liés à la tâche sont compris dans les représentations : la classe d'appartenance de l'objet ; des informations concernant ses caractéristiques fonctionnelles ; ou encore le type de contexte dans lequel on le trouve et les objets avec lesquels il est fréquemment associé. Ces informations, quand elles apparaissent dans la base des objets de l'application, servent efficacement à la résolution des références.

La détection de groupes perceptifs. Ce problème nous intéresse particulièrement puisqu'il s'agit de structurer en groupes les objets de la scène visuelle. Comme nous l'avons illustré dans l'introduction avec l'expression « *les deux cercles* » qui réfère à deux cercles présents non pas dans la scène complète mais dans un groupe perceptif focalisé, disposer d'une telle structuration permet au système de tenir compte des focalisations spatiales de l'utilisateur et de comprendre les expressions reposant implicitement sur ces focalisations.

Suite à l'article fondamental de Wertheimer (1923) et aux présentations de la Gestalt par Köhler (1947) et Guillaume (1979), nous retenons que les trois principaux critères de groupage sont la proximité, la ressemblance, et la continuité. Ainsi, dans l'exemple de l'introduction, le groupe incluant les trois triangles gris et les deux cercles gris placés à gauche se justifie par le critère de proximité et par la ressemblance (au niveau de la couleur). A un niveau plus fin, on peut extraire de ce groupe celui des triangles et celui des cercles, sur le critère de ressemblance (au niveau de la forme) et sur le critère de continuité : les trois triangles forment un alignement parfait, ce qui concourt d'ailleurs à interpréter le geste comme désignant les trois.

Ces trois critères ont été l'objet de nombreux travaux, aussi bien théoriques que formels. La formalisation s'avère en effet réalisable : contrairement au domaine de la vision artificielle, le domaine de l'interaction homme-machine met en jeu une base de données comprenant la liste des objets avec leurs caractéristiques et les coordonnées de leur position à chaque instant. Or ces données peuvent être exploitées pour des comparaisons entre les objets. Les critères de proximité

et de continuité se formalisent à l'aide de simples calculs géométriques sur les coordonnées, et le critère de ressemblance se formalise grâce aux caractéristiques enregistrées des objets. Des données sur la configuration des objets peuvent émerger de ces calculs. Le système peut ainsi disposer de renseignements supplémentaires, informations dont il ne dispose pas au départ et qui se rapprochent des représentations cognitives élaborées par l'utilisateur suite à la perception visuelle de la scène.

Le laboratoire *Kubovy Perception Lab* se focalise ainsi sur la modélisation mathématique des critères de la Gestalt (Kubovy & Wagemans 1995, Kubovy *et al.* 1998). De son côté, Feldman (1999) définit des formes logiques pour un traitement informatique. Le défaut principal commun à ces travaux est de restreindre la formalisation à un seul critère de la Gestalt, sans véritable intégration des trois critères. Une ébauche d'intégration des critères de proximité et de ressemblance a été effectuée par Thórisson (1994) dans un article souvent cité dans le domaine de l'interaction homme-machine et à la base du modèle d'interprétation contextuelle de Wolff (1999). Nous y reviendrons au cours du chapitre 4.

La détection d'objets saillants. Un autre aspect important de la perception visuelle est la notion de *saillance* visuelle, le fait qu'un objet en particulier se distingue des autres objets et attire l'attention de l'utilisateur. Si le système est capable d'identifier à tout moment l'objet saillant dans la scène, il pourra d'une part modéliser une facette de l'attention de l'utilisateur, lui permettant ainsi de prévoir dans une certaine mesure à quel objet celui-ci va s'intéresser, et d'autre part interpréter correctement les actions de référence fondées sur cette saillance. Un cas typique d'expression référentielle reposant implicitement sur la saillance visuelle d'un objet est le suivant : « *enlève le triangle* », sans geste ni antécédent linguistique, dans le contexte de la figure 1.4-A. L'expression référentielle « *le triangle* », bien qu'ambiguë du fait de la présence de plusieurs triangles, s'interprète facilement comme référant au triangle gris à gauche.

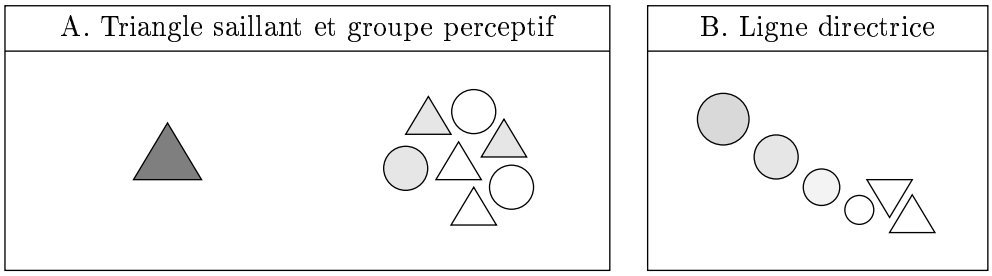


Figure 1.4 – *Groupement, saillance et ligne directrice dans la perception visuelle.*

Le problème pour le système de dialogue est l'identification de l'objet saillant. Il s'agit donc de caractériser la saillance visuelle. Or très peu de travaux ont pour but d'identifier des critères de saillance. Si de nombreux travaux s'intéressent à ce sujet, c'est dans le domaine des arts picturaux (Itten 1951, Kandinsky 1979) ou dans celui de la sémiotique de l'image fixe (Barthes 1964, Eco 1972). Or ces domaines s'intéressent à des catégories particulières d'images, essentiellement les peintures pour les premiers, essentiellement les images publicitaires pour les seconds. Aucun ne met véritablement en avant de classification générique de critères de saillance.

C'est du côté de la psychologie que de tels critères sont avancés. La théorie de la Gestalt, encore elle, propose ainsi des critères pour caractériser une forme qui vient en premier à l'esprit, ce qu'elle appelle une *bonne forme* et que nous pouvons considérer comme caractérisant un objet

visuellement saillant. Ces critères sont les lois suivantes :

- petitesse (une petite forme se démarque mieux qu'une forme de grande taille) ;
- contour (une forme à contour fermé se démarque mieux qu'une forme ouverte) ;
- simplicité (une forme simple se démarque mieux qu'une forme complexe) ;
- régularité et symétrie (une forme ayant une répartition régulière ou symétrique se démarque mieux qu'une forme n'en ayant pas) ;
- différenciation (une forme ayant une structuration originale se démarque mieux qu'une forme n'en ayant pas).

Nous constatons néanmoins l'absence de travaux étendant ces lois à des formalismes logiques ou mathématiques, ou proposant un modèle formel de saillance. Quand elle est intégrée dans une modélisation informatique, la saillance visuelle est extrêmement simplifiée et dépend fortement de l'application et des objets mis en jeu. C'est le cas par exemple des travaux de Dale (1992) ou d'Edmonds (1993) sur lesquels nous reviendrons dans le chapitre 5 consacré à la saillance.

La détection de lignes directrices. Un dernier aspect de la perception visuelle auquel nous nous intéresserons dans cette thèse est celui de ligne directrice. Il s'agit de lignes de force qui incitent le regard à suivre certains parcours lors de la perception. Ces lignes peuvent être calculées (comme dans la peinture et l'image publicitaire), ou au contraire involontaires (comme dans les scènes visuelles en dialogue homme-machine). Si tant est qu'elles existent, elles découlent alors du sens de lecture culturellement de gauche à droite, de la présence d'une perspective forte, d'une disposition particulièrement symétrique des objets dans l'espace, ou encore d'une hiérarchie marquée entre les objets saillants (le regard partant du plus saillant). La figure 1.4 B montre un exemple de parcours privilégié des objets, allant du plus grand cercle au plus petit pour finir sur le groupe des deux triangles.

L'intérêt en communication homme-machine réside dans le fait qu'un système capable de détecter à tout moment les lignes directrices pourra d'une part modéliser une deuxième facette de l'attention de l'utilisateur, lui permettant comme nous l'avons vu avec la saillance de prévoir les objets sur lesquels il va interagir, et d'autre part d'interpréter des références fondées sur cet ordonnancement des objets. Dans la situation de la figure 1.4-B, une expression référentielle telle que « *enlève le premier cercle* », sans geste ni mention langagière préalable, s'interprète facilement comme référant au cercle en haut à gauche de la scène.

Encore une fois, si le problème de la caractérisation des lignes directrices a été l'objet de quelques travaux dans le domaine des arts picturaux ou de celui de la sémiotique, nous n'en avons pas trouvé de répondant dans le domaine de l'informatique, que ce soit du côté des recherches sur les interfaces homme-machine ou du côté de la linguistique computationnelle. Nous proposerons dans le chapitre 6 quelques jalons dont nous tiendrons compte dans notre modélisation.

1.3.2 La parole et le geste, modalités d'expression

La reconnaissance de la parole et du geste. En dialogue homme-machine, la reconnaissance de la parole, c'est-à-dire la reconstitution du message transmis oralement, et la reconnaissance du geste, c'est-à-dire l'identification et la reconstitution de sa partie significative épurée de la phase d'approche et de la phase de retrait (Kendon 1980), constituent deux problèmes comparables. Matérialisés tous les deux par un flux continu de données dans lequel quelques informations significatives et structurées doivent être extraites, ces signaux sont soumis aux mêmes difficultés de traitement. Ces difficultés peuvent être distiguées en deux groupes, selon qu'elles sont

intrinsèques aux signaux ou spécifiques à leur traitement informatique.

Parmi les caractéristiques intrinsèques à la parole et au geste, nous noterons que l'un comme l'autre varient en nuance, en vitesse et en puissance. Ces variations prosodiques, intentionnelles ou non, ne sont qu'un cas particulier de ce qui constitue l'une des grandes caractéristiques à la fois de la parole et du geste, à savoir leur grande variabilité. Cette variabilité apparaît dans des contextes identiques, non seulement à un niveau inter-locuteur (chaque utilisateur a un timbre de voix qui le caractérise ainsi qu'une manière particulière de produire des gestes), mais également à un niveau intra-locuteur : en fonction des conditions de communication, de sa fatigue, ou même d'une manière générale, un utilisateur s'avère incapable de produire deux fois exactement le même son ou la même trajectoire gestuelle.

Nous noterons également que, pour les deux types de signal, la succession de deux éléments significatifs a une incidence sur ces éléments mêmes. Se retrouvent ainsi dans le geste les phénomènes de co-articulation de la parole, phénomènes qui résultent du fait que la réalisation acoustique d'un phonème varie selon les sons adjacents et se traduit par exemple par des assimilations (comme le [b] prononcé [p] dans « *la forme abstraite* » et le [d] prononcé [n] dans « *le carré de moquette* ») ou par des altérations (comme la dénasalisation du [õ] de « *bon* » en [ɔ] dans « *mets le carré de moquette au bon endroit* »). De même, la réalisation d'un geste varie selon les mouvements précédents et les mouvements suivants, la fin d'un geste pouvant par exemple altérer le début du geste suivant.

Parmi les caractéristiques spécifiques au traitement informatique de la parole et du geste, se trouvent tout d'abord les difficultés rencontrées lors de l'acquisition du signal. Les dispositifs sont peu efficaces compte tenu du milieu. Il s'avère ainsi nécessaire de distinguer la parole du bruit ambiant qui est capté avec elle, par exemple le bruit de fond comprenant celui de la machine. Pour l'acquisition du geste, tout dépend du dispositif utilisé, chacun ayant ses inconvénients : un système de caméras demande un traitement très complexe de détection de la main dans l'image ; un gant de désignation, en raison des perturbations magnétiques en particulier, entraîne des problèmes de calibrage et de discontinuité dans le signal. Le comportement même de l'utilisateur provoque une difficulté supplémentaire : pour compenser ses difficultés de perception, l'utilisateur peut être amené à modifier son articulation verbale et à corriger ses gestes au cours de leur production. C'est le cas de la compensation du bruit ambiant dans le cadre de la parole, et de la compensation du décalage entre le geste effectué et le geste perçu, décalage fréquent avec un écran tactile lorsqu'il fait apparaître la trajectoire, et qui incite l'utilisateur à corriger sa trajectoire.

Le système fait ensuite face au problème de la reconnaissance des unités (mot ou trajectoire gestuelle) dans les signaux. Deux approches classiques se confrontent : l'approche globale qui consiste à comparer chaque mot ou trajectoire identifiés à des formes de référence stockées, et l'approche analytique, qui consiste à segmenter en constituants élémentaires, au niveau le plus fin, pour recomposer ensuite l'unité. Que ce soit pour la parole ou pour le geste, la méthode globale se caractérise par un compromis entre la couverture et la précision : un équilibre reste à trouver entre un modèle exhaustif (impliquant peu de contraintes et par conséquent plus spontané) et donc une reconnaissance plus difficile, et un modèle réduit (impliquant de fortes contraintes et par conséquent artificiel et nécessitant un certain apprentissage) et une reconnaissance plus facile. D'une manière générale, la méthode globale s'avère insuffisante face aux grands vocabulaires que constituent le langage naturel spontané et les multiples trajectoires gestuelles possibles sur un écran tactile. Appliquer une méthode analytique pose également pour les deux modalités des difficultés comparables : l'identification des constituants élémentaires et la reconstitution descriptive du message. Nous avons vu qu'identifier un phonème s'avérait délicat compte tenu de leur grande variabilité en contexte (dépendance par rapport au phonème précédent et au

phonème suivant). Identifier des constituants élémentaires dans une trajectoire gestuelle s'avère également difficile compte tenu de l'absence de tout code à la base de la production de telles trajectoires. De même, si reconstruire le message oral sous la forme d'une suite de mots s'avère possible, reconstruire un message gestuel est autrement plus difficile. Cela nécessite de définir un lexique et une syntaxe du geste et de contraindre la reconnaissance, ce qui va à l'encontre de notre approche fondée sur la spontanéité de la communication. Si la reconnaissance automatique de la parole s'effectue actuellement le mieux à l'aide de méthodes statistiques, aucune méthode privilégiée ne s'est encore imposée pour la reconnaissance du geste.

L'étape suivante dans la reconnaissance consiste en l'identification des expressions référentielles. Les nombreuses recherches en syntaxe proposent diverses méthodes conduisant à diverses représentations. A la suite de (Lopez 1999), nous retiendrons les grammaires lexicalisées à base d'arbres parce qu'elles intègrent quelques aspects sémantiques qui nous seront utiles lors de l'évaluation des efforts cognitifs impliqués dans le processus de compréhension (cf. chapitre 7). Du côté du geste, cette étape consiste à segmenter correctement le flux gestuel en trajectoires indépendantes. Comme nous nous focaliserons sur les trajectoires effectuées sur un écran tactile, nous ne reviendrons pas sur ce problème[1].

La formalisation de la parole pour la résolution des références. En reprenant la séquence lexique, syntaxe, sémantique puis pragmatique, nous arrivons maintenant au niveau de la sémantique, c'est-à-dire de la représentation du sens de l'énoncé. Cette représentation se fait généralement sous une forme logique, ce qui permet de l'utiliser comme point de départ pour faire des inférences (un aspect de l'interprétation que nous n'aborderons pas ici).

Après les travaux de Montague focalisés sur l'élaboration d'algorithmes pour passer du langage naturel à une représentation formelle, la principale théorie proposant une représentation du sens d'un énoncé est la *Disourse Representation Theory* (DRT : Kamp & Reyle (1993), cf. aussi Corblin (2002) pour une adaptation aux déterminants du français). Elle consiste à construire des structures dans lesquelles sont repérées puis instanciées des variables correspondant aux référents. Ces structures peuvent se mettre à plat sous une forme logique. Sans entrer dans les détails, nous noterons que l'exemple de l'introduction « *colorie ces trois triangles en bleu* » conduirait à obtenir une forme logique telle que : $\exists$`e` $\exists$`x e-colorier_en_bleu(Interlocuteur,x)` $\wedge$ `démonstratif(x)` $\wedge$ `cardinal(x,3)` $\wedge$ `triangle(x)`. Dans cette formule, `e` représente l'événement et `x` l'ensemble de référents. Il s'agit en fait d'une adaptation de la DRT, telle que décrite dans (Landragin *et al.* 2000).

Le système de dialogue doit alors réagir face à l'acte de langage véhiculé par l'énoncé. Dans le cas de l'ordre précédent, l'application doit trouver les fonctions à appliquer pour rendre vrai le contenu propositionnel, celui-ci étant obtenu après identification des objets, comme on l'a vu page 17. Le résultat de cette analyse pragmatique peut être représenté par la formule suivante : $\exists$`e e-colorier_en_bleu(Machine,{`t_1`,`t_2`,`t_3`})`. L'application doit donc appliquer une affectation de couleur aux trois objets identifiés. Le résultat est alors montré à l'utilisateur par un retour visuel et un nouvel énoncé est attendu. Nous aurons l'occasion de revenir sur les représentations de l'énoncé au cours de cette thèse.

La formalisation du geste pour la résolution des références. En nous limitant au geste ostensif, nous sommes confrontés au problème de la description sémantique d'un tel geste, *a priori* dénué de tout code. L'hypothèse que nous faisons est qu'un tel code peut être construit en fonction des possibilités qu'offre le dispositif d'acquisition du geste. Dans le cas d'un écran tactile, un geste ostensif peut en effet prendre différentes formes. La plus simple est le point et correspond

1. Dans le chapitre 3, page 65, nous ferons le point sur les avantages de l'écran tactile.

à l'indication de direction telle que nous la connaissons dans la communication humaine. Les autres formes se caractérisent par une ligne continue, se traduisant au niveau du signal par un échantillon de points. L'objectif de la trajectoire est de désigner des objets, et les moyens utilisés pour y arriver varient, conduisant à d'innombrables possibilités de formes. Un exemple de trajectoire est l'entourage des demonstrata dans une courbe fermée. Or il s'avère possible de coder ce type d'information.

Le modèle sur lequel nous nous appuyons pour cela est celui de Bellalem (1995). Selon une approche analytique, il consiste à analyser le signal afin d'en extraire les éléments significatifs, par exemple un point d'arrêt ou une courbure particulière dans un entourage. Ces éléments se caractérisent par une *singularité*, c'est-à-dire par une rupture d'homogénéité pour une des propriétés de la trajectoire, les propriétés fondamentales étant la courbure et la vitesse. Un point d'arrêt se caractérise ainsi par une vitesse nulle entre deux segments de trajectoire à vitesse non nulle. Il constitue donc une singularité pour la propriété de vitesse. Le modèle de Bellalem comporte les étapes suivantes :

1. Détecter et identifier les singularités.
2. Modéliser la trajectoire par une succession minimale de courbes simples (B-splines).
3. A partir de cette modélisation, détecter des particularités supplémentaires telles que les points d'intersection (conduisant par exemple à une fermeture).
4. Décrire à partir de tous ces éléments la trajectoire en termes de courbes ouvertes ou fermées et de singularités. Cette description constitue la représentation sémantique du geste.
5. Interpréter le geste dans les contextes langagier et spatial. Cette étape pragmatique permet d'éliminer certaines hypothèses à propos des singularités. Par exemple, un point d'arrêt au milieu d'une trajectoire peut être interprété comme une hésitation de la part de l'utilisateur ou comme l'intention d'attirer l'attention de l'interlocuteur sur ce point particulier, par exemple pour désigner l'objet qui y est placé. Ce test de présence d'un objet en un point particulier constitue un recours au contexte visuel, et permet de privilégier l'une des deux hypothèses. Cette étape pragmatique aboutit enfin à l'identification des demonstrata.

S'il s'agit ici d'un modèle de traitement de trajectoires gestuelles planes, il existe également des travaux sur le geste en trois dimensions. C'est le cas de ceux traitant de la langue des signes, qui, bien que ne correspondant pas à un geste spontané tel que le geste ostensif, soulèvent des problèmes et proposent des solutions applicables au traitement de tout geste. Ainsi, beaucoup d'efforts sont faits actuellement sur la reconnaissance de la position et de la configuration des mains à l'aide d'un système de caméras. Ces travaux permettront peut-être à plus long terme de répandre ce dispositif dans la communauté du dialogue homme-machine. Nous pourrons ainsi enregistrer, analyser et proposer des modèles de reconnaissance du geste spontané de face à face, plutôt que de nous restreindre au geste spontané sur un dispositif limitatif. A plus court terme, cette restriction est malgré tout bénéfique puisqu'elle nous permet d'éviter de lourds problèmes de mise en œuvre et de nous concentrer sur les rapports entre perception visuelle, parole et geste, rapports que nous allons analyser maintenant en détails.

RÉCAPITULATIF

Dans le but de doter un système de dialogue de capacités interprétatives pour traiter les actes de référence de son utilisateur, nous avons détaillé dans ce chapitre l'étendue des phénomènes, à travers les différentes formes et interprétations possibles du côté du langage et du geste, et à travers les différentes façons d'exploiter les principales caractéristiques de la scène visuelle. Les classifications classiques portant sur le langage, le geste, la multimodalité, le dialogue et la référence nous ont permis d'établir un large éventail de phénomènes constituant la base de notre analyse. Nous avons ainsi montré à quel point la résolution de la référence multimodale est un problème complexe, notre conviction étant que seul un système capable de tenir compte de tous ces phénomènes aura une chance d'être un tant soit peu intelligent. Nous avons également posé les bases d'une formalisation, les principaux travaux sur lesquels nous nous appuyons à ce stade embryonnaire étant la théorie de la Gestalt pour l'analyse de la perception visuelle, la sémantique formelle pour la représentation du langage et le modèle de Bellalem pour celle du geste.

Chapitre 2

L'interaction des modalités

Comment le sens se répartit-il entre les trois modalités que sont la perception visuelle, la parole et le geste? Comment les caractéristiques de ces modalités interagissent-elles? Comment les modèles et systèmes existants tiennent-ils compte de ces interactions?

Maintenant que nous avons détaillé les caractéristiques référentielles des trois pôles que sont la perception visuelle, la parole et le geste, nous allons pouvoir explorer les conséquences de chaque caractéristique d'un pôle donné sur les deux autres pôles. Ce deuxième chapitre part ainsi d'une présentation des interactions tripolaires et de la problématique associée (§ 2.1), pour montrer ensuite que les modélisations et systèmes existants ne tiennent compte que d'une partie réductrice de ces interactions. Plus précisément, nous montrons qu'ils sont essentiellement bipolaires (§ 2.2) et ne peuvent s'étendre à des réalisations tripolaires, ce qui nous conduit à proposer des jalons pour aboutir à un tel objectif (§ 2.3). Nous exploiterons ces jalons lors de l'élaboration de notre modèle, dans la deuxième partie de cette thèse.

2.1 Le problème de la tripolarité

Récapitulatif des interactions entre modalités. La figure 2.1 tente de regrouper les principales interactions entre perception visuelle, parole et geste dans les actions de référence aux objets. Cette figure est l'occasion d'aborder un certain nombre de concepts nouveaux. Le terme *affordance* dénote le fait que certains objets, de par leurs propriétés visuelles, vont inciter l'utilisateur à leur appliquer une action privilégiée. Dans une interface à base d'icônes et de menus, un bouton incite ainsi fortement à cliquer dessus. Le terme *implicite* désigne tout ce qui n'est pas inclu dans l'énoncé mais qui participe à son interprétation : ce qui n'est pas dit, plus globalement ce qui n'est pas exprimé, ni par la voix, ni par le geste, ni même par un regard ou une posture particulière. D'autres termes ont déjà été employés au cours du chapitre précédent, et nous voulons ici les préciser, en commençant par détailler les caractéristiques de la perception visuelle qui influencent la production de parole et de geste.

La présence de *groupes perceptifs* dans le contexte visuel dirige la production de gestes au sens où tout pointage et toute trajectoire pourra désigner aussi bien les demonstrata que le groupe complet auquel ils appartiennent. Plus le groupe est compact et se distingue des autres objets de la scène, plus le geste pourra être effectué de manière rapide et imprécise sans

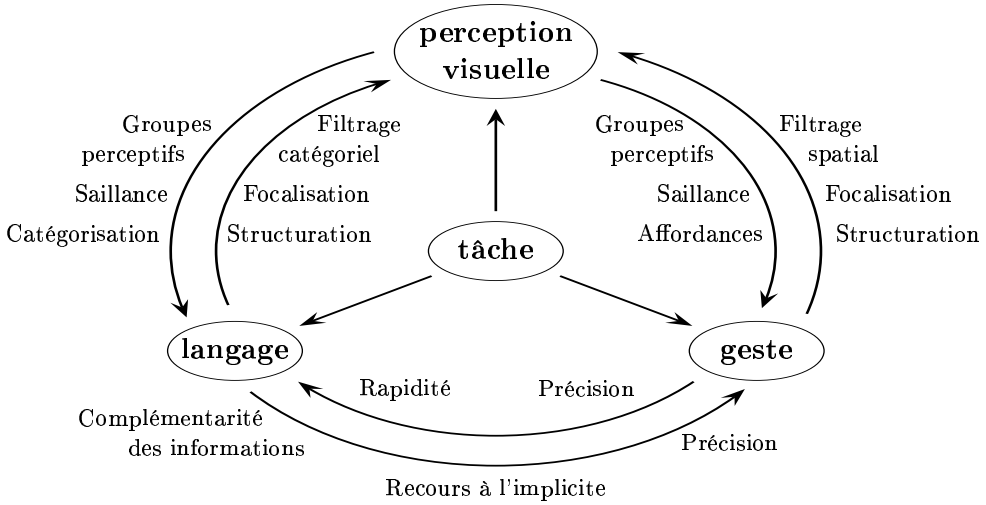


Figure 2.1 – *Les interactions entre modalités.*

que l'interprétation en soit gênée. Au contraire, si le geste a pour but de désigner un membre particulier du groupe, il devra être effectué avec suffisamment de précision et de rigueur pour que le demonstratum puisse être isolé perceptivement du groupe par l'interlocuteur. La production de parole est elle aussi affectée par la présence de groupes perceptifs, puisque la référence à un groupe pourra se faire par une expression référentielle appropriée (« *cet amas* » associé à un geste désignant le groupe), ou encore par un pluriel si les demonstrata appartiennent à une même classe d'objet (« *ces triangles* »). De même que le mot « *triangle* », le mot « *amas* » illustre le phénomène de catégorisation à la base du processus de perception. La notion de *saillance visuelle* intervient également dans l'expression multimodale, puisqu'elle autorise une certaine imprécision dans la description langagière et dans le geste, quand elle n'autorise pas tout simplement l'absence de geste. Elle incite d'autre part l'utilisateur à s'intéresser en premier lieu aux objets perçus en priorité, ce qui la rapproche de la notion de ligne directrice qui constitue également une caractéristique de la perception visuelle intervenant dans les autres modalités.

Le sens d'une action de référence multimodale se répartit entre l'expression verbale et le geste ostensif. Ainsi, une imprécision dans l'une de ces modalités peut se compenser par une précision particulière dans l'autre. Cependant, pour retrouver cette complémentarité, l'interlocuteur s'appuie sur des informations implicites. Parmi ces informations se trouve tout d'abord la présupposition qu'un geste est associé à l'expression verbale et inversement. Se trouve également la possibilité que le geste peut désigner un ensemble d'objets dans lequel le référent doit être extrait. C'est le cas d'un geste désignant un amas de triangles et de carrés lorsque l'expression référentielle associée est « *les triangles* ». L'implicite, qui pourrait être verbalisé comme « *les triangles parmi ces objets* » peut se retrouver grâce à l'emploi de l'article défini. La détermination et la classe d'objets souvent incluses dans l'expression référentielle constituent des critères de recherche d'objets. La focalisation spatiale est le critère de recherche d'objets fourni par le geste. L'implicite lié à la compensation du sens des modalités réside aussi dans la complétion de ces critères. Il en est ainsi de l'ostension non coréférente (cf. chapitre 1 avec l'exemple de « *cet employé* » associé à un geste désignant un bureau). La désignation d'un objet qui n'est pas le référent réellement visé mais conduit à ce référent constitue en effet un mécanisme implicite.

En quoi le fait de produire un geste et d'exprimer quelque chose influence notre perception ? La

focalisation spatiale induite par le geste ainsi que la focalisation catégorielle induite par les mots ont un rôle référentiel qui se traduit par un filtrage. Ce filtrage permet d'extraire dans l'ensemble des objets visibles les demonstrata ou les référents. Non seulement il focalise l'attention des deux interlocuteurs, mais il a également une action structurante sur le contexte visuel : suite au geste ou à la catégorie, les groupes perceptifs obtenus avec les critères de la Gestalt sont restructurés. Un geste qui extrait quelques objets d'un groupe entraîne ainsi la restructuration de ce groupe en deux parties désormais exploitables. De même, une catégorie met en avant le critère de la similarité et privilégie les groupes construits selon lui.

L'interaction tripolaire et le rôle de la tâche. Notre manière de voir influence nos actions de référence, ce qui se traduit par des choix dans la répartition du sens dans les modalités d'expression, par des choix de mots et de formes de geste, choix qui influent aussitôt sur notre manière de voir. Il s'agit bien d'une seule problématique qui doit être appréhendée globalement. Parmi les choix langagiers, le choix du déterminant illustre bien cette interaction tripolaire : un démonstratif associé à un geste déictique focalise sur un objet particulier de la scène, tout en structurant celle-ci en distinguant l'objet focalisé du ou des autres objets de la même catégorie. La présence de ces derniers justifie l'emploi du démonstratif : s'il n'y avait qu'un seul objet de la catégorie considérée, une description définie aurait suffit à permettre son identification.

Plusieurs facteurs cognitifs interviennent dans ce processus. L'attention et la mémorisation sont des facteurs particulièrement exploités lors des actions de référence. Une troisième facteur est celui d'intention (ou intentionnalité). Il est fortement lié à la tâche applicative : si celle-ci incite fortement l'utilisateur à exercer telle action particulière sur telle catégorie d'objet, la perception de la scène visuelle sera orientée vers ce type d'objet, et les filtrages spatial et catégoriel également.

Les difficultés dans l'approche. Cette problématique avec ses multiples interactions est difficile à appréhender. La conception d'un système de dialogue, pour être réalisable, nécessite une séparation des processus, séparation *a priori* incompatible avec notre objet d'étude. La plupart des approches s'oriente ainsi vers la formalisation du sens porté par une modalité, pour s'étendre ensuite à la prise en compte de phénomènes sémantiques ponctuels provenant d'une autre modalité. Comme nous allons le détailler dans la section suivante, certains travaux partent, à la base, d'une modélisation de deux modalités avec un même souci d'approfondissement pour les deux contenus et pour leurs interactions. En l'absence de travaux fondateurs tenant compte des trois modalités, nous utiliserons les modèles bipolaires comme base de travail, avec la volonté constante de garder une approche fondée sur la nature tripolaire des signes dans la communication, et avec le souci constant d'essayer d'appréhender et de comprendre les phénomènes sous cet angle.

2.2 Les interactions bipolaires : modèles existants

2.2.1 L'interaction entre la perception visuelle et la parole

Des modèles en psycholinguistique. Caron (1989) fait le point sur les rapports entre linguistique et psychologie, et présente un grand nombre de travaux ayant des préoccupations dans les deux disciplines. La psycholinguistique nous paraît *a priori* être d'un grand secours dans notre approche puisqu'elle étudie par définition les processus de codage et décodage mettant en relation les états des messages avec les états des interlocuteurs (ce qui est aussi appelé représentations mentales et qui inclut la notion de sous-contexte implicite d'interprétation). Son intérêt est de fournir des modèles psychologiques plausibles pouvant être pris comme exemples pour la

conception de modèles computationnels. Elle fournit également des procédures expérimentales sur lesquelles nous reviendrons. Parmi ces modèles, ceux qui offrent un intérêt pour le problème de la référence aux objets nous semblent être les suivants (ils sont ici présentés rapidement d'après l'analyse qu'en fait Caron) :

- Dans le domaine de la perception de la parole et de l'accès lexical :
  Le modèle des *logogènes* de Morton (1982) décrit l'activation de logogènes (ou représentations conscientes d'un mot) par des informations de nature auditives, visuelles ou liées au contexte (situation, connaissances générales). Bien que ce modèle soit dédié à l'accès aux mots, nous pouvons en imaginer une extension aux concepts ou aux objets de l'application dans le cadre d'un dialogue homme-machine. L'intérêt de l'approche réside dans l'interaction permanente de toutes les sources d'information pour le déclenchement d'un logogène. Ce déclenchement a lieu lorsqu'un certain seuil d'activation est atteint. Dans le modèle de Morton, le mot devient alors disponible. Dans notre contexte, l'objet de l'application deviendrait disponible, voire saillant et susceptible de constituer une interprétation privilégiée en cas d'ambiguïté. Ensuite, le logogène retrouve peu à peu son état initial. Il reste encore un certain temps partiellement activé, ce qui autorise un éventuel deuxième déclenchement plus rapide. De plus, et c'est encore un point intéressant pour notre approche, l'activation d'un logogène entraîne une activation partielle de ceux qui lui sont proches. Ainsi, dans notre cas, si un triangle rouge devient saillant, les autres triangles et les autres formes rouges deviennent quelque peu saillantes aussi. Nous exploiterons ce principe dans le chapitre 5 avec la notion de sailance indirecte, et dans le chapitre 8 pour la modélisation de l'attention.

- Dans le domaine de la sémantique lexicale :
  La sémantique *procédurale* considère le sens d'un mot comme un ensemble de procédures qui, selon le contexte où elles s'appliquent, peuvent donner lieu à des effets de sens différents. Comme le remarque Caron, cette approche est prévue pour s'adapter directement à l'informatique, mais reste encore trop vague. A notre sens, le contexte est tellement complexe (et particulièrement le contexte visuel) que les procédures seraient bien trop nombreuses pour pouvoir être spécifiées. La théorie *componentielle* de la signification, suivie entre autres par Fodor (1986), ramène quant à elle le sens des mots à un nombre fini de traits élémentaires. Pour être opérationnelle, cette approche suppose des concepts définis et délimités avec précision, ce qui est le cas du dialogue finalisé. D'une certaine manière, nous suivons cette approche en décomposant le contexte en une liste de traits.

- Dans le domaine de l'interprétation :
  Une critique que l'on peut adresser aux modèles précédents est de ne pas exploiter les mots grammaticaux tels que les déterminants, les pronoms ou les prépositions. Une étude exemplaire des déterminants est celle de Karmiloff-Smith (1979). Analysant les articles défini et indéfini, elle dégage les étapes dans l'acquisition de leurs différentes fonctions : discrimination, dénomination, deixis, passage au générique, etc.
  Parmi les grands résultats de la psycholinguistique en compréhension (tels que présentés par Caron), trois nous semblent encourager grandement notre approche :
  - l'importance de la référence et de la deixis dans l'interprétation ;
  - l'indépendance du sens par rapport à son support verbal (la représentation sémantique est conceptuelle et indépendante de la modalité, orale ou visuelle, sous laquelle sont transmises les informations) ;

- la nécessité d'un modèle mental lors de l'intégration sémantique (un modèle mental intègre les informations provenant des diverses modalités en un ensemble cohérent).

En conclusion, si la psycholinguistique est riche d'approches intéressantes et permettant une meilleure appréhension des phénomènes cognitifs intervenant dans la référence, nous constatons néanmoins, du point de vue de l'implantation informatique, que les propositions qu'elle avance sont encore trop vagues pour être opérationnelles.

Des systèmes en linguistique computationnelle. Pour la résolution de la référence aux objets, nous avons vu que les mots inclus dans l'énoncé servent de critères de filtrage perceptif. La catégorie, les propriétés évoquées avec les éventuels adjectifs qualificatifs, ou encore le nombre porté par le déterminant sont autant de critères. Compte tenu du fait qu'ils suffisent à traiter un grand nombre de situations, beaucoup de systèmes se limitent à ces critères. Peu de travaux traitent de l'interprétation dans un sous-ensemble implicite d'objets, et, parmi ceux-ci, on ne trouve que peu d'analyses des critères visuels à l'origine de tels sous-ensembles. Citons, dans le domaine de la génération automatique, ceux de Dale et de Reiter (Dale 1992, Dale & Reiter 1995, Reiter & Dale 1997), et surtout l'extension qu'en ont fait Krahmer & Theune (2002). En compréhension, nous nous baserons particulièrement sur l'approche de Salmon-Alt (2001b) qui partage avec eux quelques points communs que nous détaillerons dans le chapitre 4.

Nous nous baserons également sur quelques travaux qui nous semblent intégrer, avec une approche comparable à la nôtre, quelques critères tels que l'attention ou la saillance visuelle. Ainsi, parmi les travaux fondateurs en linguistique computationnelle sur la référence aux objets, ceux de Clark (Clark *et al.* 1983, Clark & Wilkes-Gibbs 1986) insistent sur la considération de la référence comme un processus coopératif soumis au principe de l'effort minimal, notion sur laquelle nous reviendrons dans le chapitre 7, et explorent l'importance de la saillance lors de références. Sur la base d'une expérimentation, ils montrent en particulier que la saillance influence la production d'expressions référentielles (qui, sans prise en compte de la saillance, seraient considérées comme ambiguës). De nombreux travaux ont abordé la notion de saillance, notion intervenant en particulier pour faciliter la compréhension de « *le N* » dans un environnement contenant plusieurs objets de type N. Wright (1990) montre ainsi que la tâche fournit des contraintes non-linguistiques qui œuvrent dans ce sens. D'une manière générale, la saillance dans ces travaux ne représente jamais la même étendue de phénomènes. Si elle tend à intégrer des informations de nature perceptive, elle reste néanmoins fortement liée à la forme linguistique. Nous nous inscrivons bien entendu dans cette voie, avec la volonté d'explorer les critères visuels et de les confronter aux critères linguistiques.

D'autres travaux sont beaucoup plus axés sur le contexte visuel tout en gardant une attention sur la forme linguistique. Les expériences de Kessler *et al.* (1996) montrent ainsi, avec un exemple proche de celui que nous avons présenté dans l'introduction, l'importance de sous-ensembles visuels dans l'interprétation des références. Beun & Cremers (1998) vont plus loin dans l'analyse en proposant un modèle qui gère à la fois la notion de saillance inhérente, celle de focus d'attention spatiale et celle de focus d'attention fonctionnelle. Les auteurs vérifient expérimentalement des hypothèses à propos de ces notions et aboutissent aux résultats suivants :

- un changement de focus se traduit par plus de redondance d'information dans les modalités d'expression (ce qui constitue un indice intéressant car détectable par un système) ;
- quand l'utilisateur reste dans un même focus, ses expressions référentielles sont courtes voire ambiguës ;
- le focus est une cause principale de cohérence dans le dialogue : les transitions d'une zone

spatiale à une autre et d'une sous-tâche à une autre suivent une certaine logique, théoriquement fournie par l'application.

Enfin, d'autres travaux s'intéressent essentiellement au contexte visuel, quitte à oublier les subtilités du langage. Ils se focalisent sur la notion de groupe perceptif ou sur celle de saillance visuelle, et nous en parlerons dans les sections appropriées des chapitre 4 et 5.

2.2.2 L'interaction entre la perception visuelle et le geste

Les trajectoires gestuelles. Dans notre contexte d'étude, à savoir le geste sur écran tactile, nous nous intéressons ici aux formes que peuvent prendre les trajectoires gestuelles planes et aux différentes étapes menant à leur interprétation. Les travaux dédiés au geste ostensif produit spontanément sur écran tactile en association avec la parole sont peu nombreux. En effet, lors de la phase de conception d'un système acceptant une entrée gestuelle, l'accent est souvent mis sur les possibilités d'action plutôt que d'ostension du geste. Des métaphores sont définies, par exemple pour la suppression d'un objet (trajectoire en forme de croix) ou pour un déplacement (trajectoire en forme de flèche), et remplacent ainsi une production verbale (cf. par exemple Johnston *et al.* 1997). Une telle méthode va à l'encontre de notre approche fondée sur la spontanéité de l'interaction, où le geste est quasiment réduit à un rôle référentiel et ne s'interprète qu'en association avec le langage.

Nous nous baserons ainsi sur les travaux de Bellalem et de Wolff (Bellalem 1995, Bellalem & Romary 1995, Wolff *et al.* 1998, Wolff 1999) qui sont à l'origine de notre approche. Suite à son étude du corpus Magnét'Oz, Wolff distingue quatre catégories de gestes ostensifs apparaissant spontanément sur écran tactile, répertoriées dans la figure 2.2. Lors de l'expérimentation à l'origine de ce corpus, les sujets pouvaient utiliser la parole et le geste spontanément, sachant que tout geste devait être effectué avec un stylet. Le choix de la désignation avec un stylet et non directement avec le doigt a été fait parce que le doigt est imprécis et parce qu'il accroche sur la surface de l'écran. Une autre raison pratique a conduit à ne pas afficher les trajectoires gestuelles en cours de production : c'est l'existence d'un décalage entre le geste effectué et la trajectoire perçue. Ce décalage, s'il est détecté par l'utilisateur, peut l'entraîner à modifier son geste en cours de production, comme on l'a vu page 35.

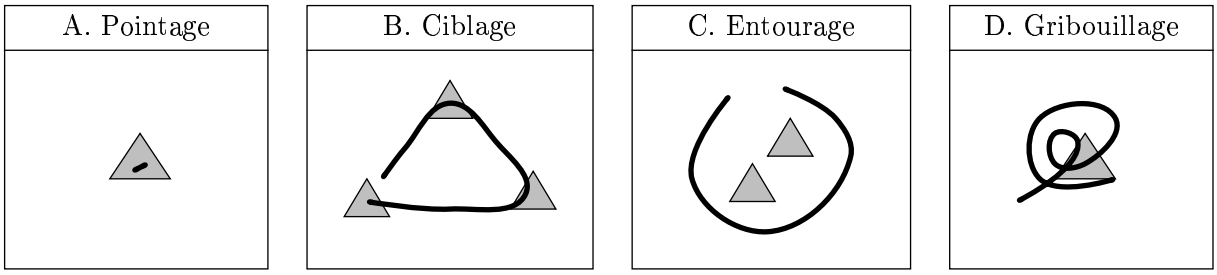


Figure 2.2 – *Catégories de trajectoires gestuelles en 2D.*

Caractéristiques lexicales et sémantiques des trajectoires. Bellalem propose un modèle partant des données captées par l'écran tactile et aboutissant à une description sémantique des trajectoires. Elle définit un lexique autour de la notion de singularité, c'est-à-dire de portion de trajectoire contrastant sur une ou plusieurs propriétés avec les portions précédente et suivante. Les propriétés d'une courbe continue étant la vitesse, la courbure et le recouvrement, les singularités sont les suivantes : point d'arrêt, point de rebroussement, point d'inflexion, point d'intersection, retour vers un point antérieur de la trajectoire.

Dénotant une rupture au milieu d'une régularité, ces singularités sont les parties saillantes de la trajectoire. Bellalem en déduit qu'elles sont intentionnelles et dénotent des significations potentielles. A partir de cette hypothèse, elle propose un modèle d'identification des propriétés sémantiques. Ainsi, un point d'arrêt peut être lié à la désignation d'un objet ; un changement de courbure peut être lié à un écart dans la trajectoire, écart causé par la volonté de l'utilisateur d'éviter un objet.

Notons que si les notions de lexique et de sémantique peuvent ainsi s'appliquer aux trajectoires gestuelles, la notion de syntaxe n'a par contre aucun sens. Il est en effet impossible de construire des règles syntaxiques permettant d'obtenir à partir du lexique défini toutes les trajectoires possibles, même dans un cadre applicatif restreint. Ainsi, contrairement à la parole pour laquelle un traitement structural est possible (il existe une structure derrière les mots, ce qui rend possible une analyse syntaxique éventuellement dirigée par la sémantique), on ne peut pas appliquer de traitement structural au geste mais seulement un traitement procédural. Autrement dit les procédures sont applicables directement à un niveau sémantique.

L'nterprétation pragmatique des trajectoires. Le recours au contexte visuel est l'étape suivante dans le traitement d'une trajectoire. En tenant compte de la disposition des objets par rapport à la trajectoire ainsi que la possibilité d'une imprécision quant à l'étendue de celle-ci, Wolff constate ainsi la possibilité d'ambiguïtés sur la catégorie de la trajectoire et sur les demonstrata, ambiguïtés décrites ci-dessous et illustrées dans la figure 2.3 :

- L'ambiguïté de forme :
  Une même forme de trajectoire peut correspondre à plusieurs intentions de désignation différentes. Ainsi, un geste ponctuel (pointage) peut désigner un seul objet ou un groupe d'objets. Si l'on ne tient pas compte de la disposition des objets présents dans le voisinage du pointage, on ne peut pas trancher. De même, un geste ciblant plusieurs objets disposés selon un cercle aura une forme circulaire. Si un autre objet se trouve au milieu du cercle, le geste pourra aussi correspondre à l'entourage de cet objet. Il en devient ambigu.

- L'ambiguïté de portée :
  Deux gestes ayant la même forme et correspondant à la même intention de catégorie (par exemple un ciblage) peuvent désigner un nombre d'objets différent. C'est le cas lorsque la trajectoire passe par deux objets et lorqu'un troisième objet appartenant au même groupe perceptif se trouve juste à côté : une première interprétation aboutit à la désignation de deux objets, une autre interprétation aboutit à la désignation du groupe perceptif complet, c'est-à-dire de trois objets.

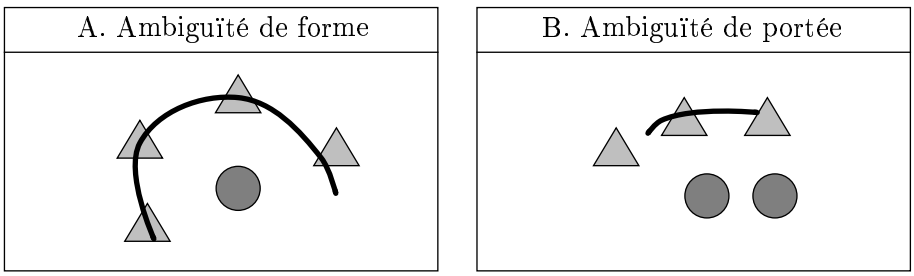


Figure 2.3 – *L'ambiguïté de forme et l'ambiguïté de portée.*

Pour formaliser ces constatations, Wolff explicite la notion de *groupement perceptif* : les objets affichés sur l'écran se partitionnent en groupements selon les deux critères principaux de

la Gestalt, la proximité et la similarité. L'utilisateur perçoit ces groupements, et les gestes qu'il va produire vont en dépendre. Wolff utilise l'algorithme de Thórisson (1994) pour obtenir une liste de groupes perceptifs à partir de la liste des objets visibles. Il partitionne alors la scène visuelle en plusieurs zones correspondant aux espaces de désignation de chaque objet et de chaque groupe. Ces zones, illustrées dans la figure 2.4, se répartissent selon deux catégories :

- Les zones élictives :
  A chaque objet et chaque groupement est liée une zone de sélection ou *zone élictive* qui le recouvre. Un geste dont la trajectoire reste dans la zone élictive d'un objet ou d'un groupement aura l'intention de désigner cet objet ou ce groupement. On l'appelera geste élictif. C'est typiquement le cas du pointage. La zone élictive s'étend un peu au-delà du gabarit de l'objet ou du groupement, afin de pallier l'imprécision des gestes.

- La zone séparatrice :
  Le reste de la scène, c'est-à-dire le fond qui ne peut correspondre à la sélection d'aucun objet, constitue la *zone séparatrice*. Un geste dont la trajectoire reste dans cette zone aura ainsi l'intention de séparer certains objets d'autres objets. On l'appelera geste séparateur. C'est le cas de l'entourage.

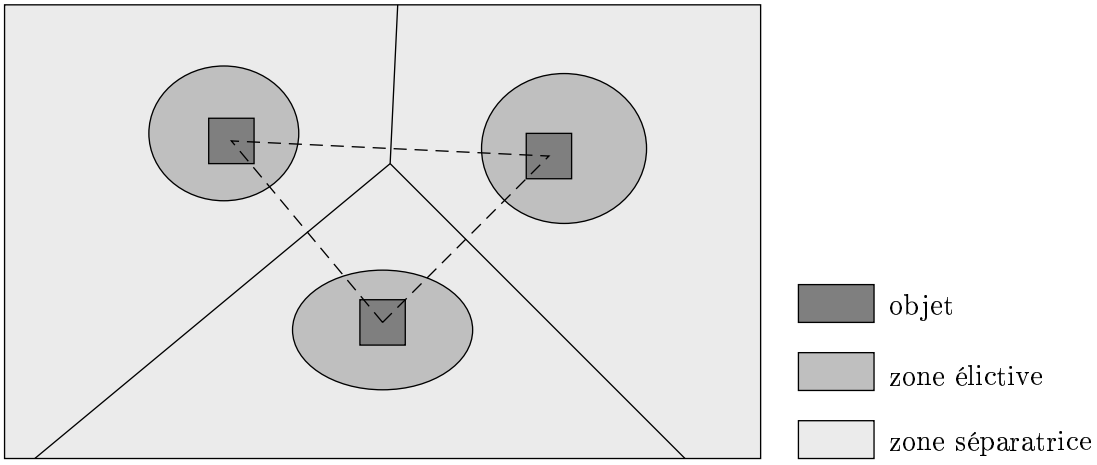


Figure 2.4 – *Les zones élictives et la zone séparatrice.*

L'interprétation contextuelle de Wolff consiste alors à déterminer les demonstrata, sans tenir compte de la production verbale, en envisageant les ambiguïtés de forme et de portée, et en envisageant la possibilité d'un geste élictif ou d'un geste séparateur. Plusieurs hypothèses peuvent être obtenues. Wolff n'en retient que deux : une correspondant à un objet isolé (ce qu'il appelle *accès individuel*), l'autre correspondant à la désignation d'un groupe perceptif complet (ce qu'il appelle *accès de groupe*). Ces hypothèses sont sensées être ensuite confrontées à l'énoncé langagier pour aboutir à l'identification des référents.

Limites du modèle présenté. Une première limite consiste dans les classifications avancées par Wolff. Deux points importants sont illustrés dans la figure 2.5 : d'une part certains gestes ne sont ni élictifs ni séparateurs ; d'autre part tout geste peut être une combinaison de parties de nature différente. Les distinctions de Wolff réduisent ainsi les phénomènes possibles.

C'est surtout dans sa dernière étape que le modèle de Wolff présente à notre sens un défaut majeur. Nous considérons d'une part que le langage peut diriger l'interprétation d'une trajectoire, au niveau même de l'étiquetage sémantique des singularités. D'autre part, nous rejetons la

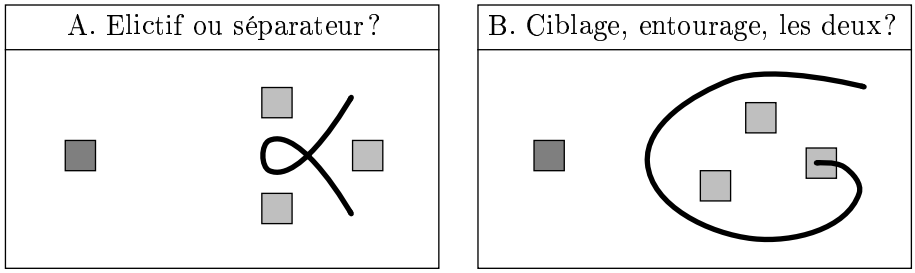


Figure 2.5 – *Limites des caractérisations de Wolff.*

nature des deux hypothèses émises : la présence d'un singulier ou d'un pluriel permet de rejeter respectivement l'accès de groupe et l'accès individuel, et ce rejet doit avoir lieu avant l'élaboration des hypothèses. Certaines hypothèses n'ont en effet aucun sens compte tenu de l'expression référentielle verbale. Elles sont émises avant même de savoir ce qui est cherché (un objet, plusieurs objets, une classe d'objet, un lieu, etc.), alors qu'elles devraient être dirigées en partie par les mots choisis par l'utilisateur. De plus, la référence à plusieurs objets n'implique pas forcément que les objets fassent partie d'un groupe perceptif, ce que le modèle de Wolff impose. Nous sommes encore loin d'une interaction tripolaire.

2.2.3 L'interaction entre la parole et le geste

L'appariement du geste et de la parole. Avant de résoudre la ou les références incluses dans l'énoncé courant, il est nécessaire de synchroniser les modalités, c'est-à-dire d'associer chaque geste à l'expression référentielle qui lui correspond sémantiquement. Ce problème peut s'avérer très complexe à partir du moment où l'on tient compte de la répartition des informations sémantiques entre la parole et le geste, répartition pouvant entraîner une certaine imprécision dans chacune des modalités et par conséquent l'incapacité du geste seul ou de l'expression verbale seule à identifier les référents. De plus, rien dans la parole ni dans le geste ne permet de les apparier : l'expression référentielle verbale peut être aussi bien démonstrative que définie, les deux pouvant aussi bien apparaître avec ou sans geste ; la forme du geste ne permet aucune hypothèse quant à l'identification de l'expression verbale qui lui correspond. Il semble néanmoins que l'utilisation du démonstratif, lorsque l'hypothèse de l'anaphore n'est pas envisageable, implique fréquemment l'utilisation d'un geste ostensif. L'appariement semble donc devoir être dirigé par le langage. Enfin, n'oublions pas la possibilité d'appariements multiples tels que nous les avons décrits page 30.

Dans les travaux existants, la mise en correspondance des gestes et des expressions référentielles verbales se fait essentiellement sur un critère de proximité temporelle. McNeill (1992) pose l'hypothèse que la phase préparatoire du geste précède le segment linguistique accentué, et que sa phase significative précède ou se termine au moment de l'accent maximum. Dans le domaine de l'ingénierie des interfaces, l'aspect prosodique est parfois négligé et cette hypothèse se retrouve souvent associée à des seuils déterminés expérimentalement. Oviatt *et al.* (1997) observent ainsi que le geste intervient dans un intervalle de trois à quatre secondes avant l'expression verbale associée, ces valeurs étant reprises dans (Johnston *et al.* 1997) pour un algorithme de synchronisation et d'intégration multimodale. Les autres critères généralement cités pour l'intégration multimodale sont la complémentarité logique et la compatibilité de type. Ils sont néanmoins souvent utilisés dans un but de gestion des manques d'informations. On retrouve ce principe

dans les interfaces pour lesquelles le geste peut venir prendre la place de la parole, avec des exemples tels que : « *colorie* » puis un geste désignant un objet, puis « *en rouge* », exemple tiré de (Brison 1997) et se situant à l'opposé de notre approche. Sans aller jusque-là, d'autres systèmes font une ségrégation entre actions réalisables à l'aide du geste et actions réalisables à l'aide du langage.

Selon Cohen (1992), les avantages des deux modalités se complètent et cette complémentarité doit diriger les technologies. Ainsi, le geste de manipulation est efficace pour une certaine classe de problèmes. Quant au langage naturel, il permet ce que le geste ne permet pas : la référence à des objets qui ne sont pas dans la scène, la description d'objets, la spécification de relations temporelles, la quantification, l'identification et la manipulation d'ensembles d'objets ou de parties d'objets, l'utilisation du contexte de l'interaction.

Dans notre cadre de l'interaction spontanée, le geste vient toujours en complément de la parole. Dans le corpus Magnét'Oz, nous observons ainsi une trace linguistique systématiquement associée au geste ostensif. Reste le critère de proximité temporelle : on observe dans le corpus que les hypothèses de McNeill et les différentes valeurs de seuil proposées ne sont pas toujours respectées. Ceci est peut-être dû à l'utilisation du dispositif de vis-à-vis qu'est l'écran tactile, ou au contexte applicatif très fort qui autorise une certaine imprécision. En tout cas, même lorsque les écarts temporels sont importants, les énoncés restent toujours compréhensibles. Un système doit donc comprendre ce type d'énoncés et nous partons de l'hypothèse que la proximité temporelle n'est pas un critère prépondérant pour la mise en correspondance des gestes et des expressions verbales. Par contre, suite à l'observation que l'ordre des gestes est identique à celui des expressions verbales dans tous les énoncés du corpus, nous partons de cette hypothèse qui nous semble en accord avec le caractère sponanté de l'interaction.

Deux types de stratégies de combinaison des informations portées par le geste et par le langage peuvent être envisagés :

1. La première stratégie consiste à : calculer pour chaque expression référentielle présente dans l'énoncé verbal un ensemble de candidats référents ; effectuer le même travail en ce qui concerne les candidats aux gestes ostensifs ; puis tenter l'appariement référence après référence. Chaque fois que l'intersection entre l'ensemble des hypothèses langagières et l'ensemble des hypothèses gestuelles se réduit à un singleton, on considère qu'on a obtenu le coréférent pour l'expression considérée.
2. La seconde stratégie consiste à ne pas travailler directement sur des ensembles d'objets, mais sur les contraintes visant à l'identification des référents.

La différence entre ces deux types de stratégies tient à la différence entre approche *extensionnelle* et approche *intensionnelle*. L'approche intensionnelle que nous suivrons tout au long de cette thèse s'intéresse aux caractères que partagent les membres d'un ensemble et non à la liste des membres de l'ensemble. Cette façon de voir les choses est particulièrement importante lorsque l'on veut gérer, comme nous allons le faire pour l'interprétation de la référence, des ensembles et des sous-ensembles d'objets. La différence entre les deux stratégies tient également à la différence entre forme logique et contenu propositionnel (cf. chapitre 1 page 17). C'est à ces formules que nous allons maintenant nous intéresser, en tentant d'y regrouper informations gestuelles et informations langagières.

La représentation sémantique couplée du geste et de la parole. Quel est l'avantage d'une forme logique multimodale par rapport à la seule forme logique langagière ? D'une part elle regroupe en une seule formule des informations émises dans un même but, ce qui est plus propre et plus

simple à gérer. D'autre part, l'utilité d'une forme logique réside dans ce que l'on peut en faire pour l'interprétation pragmatique, à savoir des tests de consistance logique, des déductions, des inférences. Avec une forme logique multimodale, toutes ces opérations peuvent se faire directement à un niveau multimodal. Bien que nous n'aborderons pas ces aspects du problème de la compréhension automatique, nous retiendrons les avantages qu'apporte une formule intégrant les deux modalités. Deux grandes approches se distinguent dans l'élaboration d'une telle formule :

1. Celle qui consiste à calculer une forme logique pour l'énoncé verbal et une forme logique pour l'énoncé gestuel, puis à fusionner les deux.
2. Celle de la sémantique dynamique qui consiste à calculer une forme logique pour l'énoncé oral et à l'augmenter et la modifier pour prendre en compte l'énoncé gestuel.

Si la première approche se conçoit bien pour des gestes communicatifs non réduits aux gestes ostensifs, la deuxième solution semble plus adaptée à notre contexte d'étude et à la référence aux objets en général. En effet, décrire sous forme logique les singularités et les portions de courbes d'une trajectoire ne présente que peu d'intérêt (face à l'ensemble des objets candidats) pour la résolution des références. Il n'en est pas de même pour un geste référentiel accompagnant une action. Pour une action de déplacement, par exemple « *mets ce triangle ici* » accompagné d'un seul geste partant du triangle en question et aboutissant au lieu de destination, la trajectoire du geste peut inclure non seulement la désignation de l'objet, la désignation du lieu de destination, mais également certaines caractéristiques du déplacement telles que la vitesse ou des lieux de passage. Dans ce cas, la forme logique du geste a une importance dans l'interprétation. Nous focalisant sur le geste ostensif, nous ne reviendrons pas sur cet exemple. Nous noterons cependant qu'il apparaît de manière ponctuelle dans le corpus Magnét'Oz.

Parmi les modèles et systèmes existants, nous citerons deux exemples. Le système Cubricon (Neal & Shapiro 1991) intègre les entrées des différents dispositifs (entrée vocale, entrée textuelle, gestes de désignation) dans un langage unifié. Tout est fusionné dans un seul flux qui contient les mots et les références gestuelles, et qui est interprété à l'aide de modèles adaptés pour le domaine, le langage et l'utilisateur. Le même langage unifié est utilisé pour la génération de réponses. Zeevat (1999) propose quant à lui une formalisation couplée du geste et du langage dans les structures de la DRT (présentée rapidement dans le chapitre 1 page 36). Avec la notion d'ancre intensionnelle à la place de l'ancre externe de (Kamp & Reyle 1993), il étend la DRT pour un meilleur traitement des démonstratifs. L'intérêt de son travail réside surtout dans la possibilité de confronter ensuite ces structures à des mécanismes faisant intervenir les présuppositions. Pour cette raison, l'intervention du geste reste assez simple. Comme dans Cubricon ou dans la majorité des systèmes existants, le geste est intégré sans tenir compte de toutes les caractéristiques que nous avons soulevées à son propos.

2.3 Vers une modélisation des interactions tripolaires

2.3.1 Possibilités d'adaptation des modèles bipolaires

Raisonner avec le chiffre trois. S'il s'avère relativement facile d'appréhender des interactions bipolaires, la tripolarité s'avère beaucoup plus délicate. Avec deux pôles, on arrive bien à imaginer que les interactions conduisent à un état d'équilibre que l'on peut concevoir. Avec trois pôles, l'état d'équilibre reste pour le moins mystérieux. C'est ainsi que nous constatons que les modèles et systèmes existants, lorsqu'ils modélisent les interactions entre modalités que nous avons présentées, se limitent à deux pôles. Nous l'avons vu au cours de la section précédente.

Nous le verrons également dans le chapitre 4 à propos de formalisation de la Gestalt : la grande majorité des travaux de formalisation se limitent à l'un ou aux deux critères principaux de la Gestalt, la proximité et la similarité. Personne ne s'aventure dans une modélisation tripolaire, en considérant par exemple la proximité, la similarité et la continuité. Nous nous y essayerons au cours du chapitre 4.

Le raisonnement avec trois pôles semble donc très différent du raisonnement avec deux pôles. C'est pourquoi, comme nous l'avons vu également dans la section précédente, les modèles et systèmes bipolaires ne peuvent pas s'étendre et devenir ainsi tripolaires. Il faut au contraire considérer globalement les trois pôles, en ayant conscience dès le départ de la complexité des interactions. Ainsi, en reprenant le modèle de Wolff et les critiques que nous avons formulées à son propos, nous considérons qu'étendre ce modèle bipolaire à la tripolarité (en ajoutant le langage) est impossible. Le langage ne peut pas se confronter avec le résultat obtenu suite au traitement du geste et de la perception visuelle, car ce résultat est déjà réducteur. Il est au contraire nécessaire de faire intervenir le langage en amont, ce qui enlève toute validité aux hypothèses bipolaires. Les hypothèses doivent être émises après considération des interactions tripolaires.

Autres polarités. Nous avons axé notre présentation des modèles et systèmes existants en prenant les trois paramètres que sont perception visuelle, parole et geste, car ils correspondent à des informations concrètes, tangibles. Nous aurions pu prendre d'autres paramètres, en particulier les trois qui nous semblent fondamentaux dans la conception d'un système de dialogue, à savoir l'intention, l'attention et la mémoire. Nous les avons relégués au second plan à cause de leur nature implicite et par conséquent du caractère très hypothétique des informations qu'ils regroupent. Une approche fondamentalement cognitive leur aurait sans doute donné plus d'importance.

Dans leur article désormais classique, Grosz & Sidner (1986) partent de ces paramètres, en particulier l'intention et l'attention. Des structures sont ainsi gérées pour ces deux critères, en plus d'une structure discursive. Le modèle obtenu comprend bien trois pôles. Il s'avère néanmoins difficilement exploitable pour le dialogue homme-machine à support visuel. Il reste en effet très orienté sur les indices langagiers et présente les inconvénients d'une méthode plus explicative qu'opérationnelle. Les auteurs avouent elles-mêmes que les structures intentionnelle et attentionnelle sont parasitaires. Nous voyons dans ce point la principale difficulté dans l'élaboration d'un modèle tripolaire. De plus, Salmon-Alt (2001b) avance une critique d'ordre méthodologique : les auteurs se servent d'une part des expressions référentielles pour calculer la structure discursive, et proposent d'autre part de considérer cette structure comme prédictive quant à l'interprétation de ces mêmes expressions. En conclusion, si les données du problème sont similaires aux nôtres (intention et attention), nous considérons que l'analyse doit aller plus loin et que les propositions doivent être plus formalisables.

2.3.2 Prolégomènes à un modèle tripolaire

Des prémices bipolaires pour la modélisation de la tripolarité. Comment appréhender dès le départ les multiples interactions entre les trois pôles que sont la perception visuelle, le langage et le geste ? Nous proposons ici de reprendre les trois bipolarités, en tentant de séparer pour chacune d'entre elles ce qui peut être appréhendé au niveau bipolaire et ne doit pas être remis en question ensuite, et ce qui ne peut être appréhendé qu'au niveau de la tripolarité. La première catégorie constitue en quelque sorte des prémices à la tripolarité. Nous tentons ici de les identifier à partir des considérations du chapitre précédent et du début de ce chapitre, pour les intégrer ensuite dans notre modélisation. Nous mettons l'accent sur leur aspect formel, pour que l'intégration

n'implique aucune réduction. Sans aller pour l'instant jusqu'à définir un formalisme, notons que le but est de disposer d'un tel formalisme, qu'il s'agisse de structures de traits incomplètes ou de structures de données ouvertes correspondant à des représentations mentales, pour faciliter l'intégration des informations.

Prémices relatives aux interactions entre perception visuelle et parole. Les catégories perçues par l'utilisateur ne sont pas forcément les mêmes que celles disponibles dans le langage, ni que celles internes au système de dialogue. Des termes tels que « *chaise* » et « *carré* » renvoient tout d'abord à des prototypes. La perception visuelle aboutit à l'identification de tels prototypes. On considère généralement que « *forme géométrique avec quatre côtés* » renvoit au même prototype que « *carré* ». Un point important pour la modélisation est que les prototypes de « *chaise* » et de « *carré* » ne sont pas définis de la même manière. En effet, le premier a un attribut « taille » que n'a pas le second. Ces prototypes permettent d'accéder aux objets de l'application, en passant par les catégories internes au systèmes, que l'on peut désigner par `chaise` et `carré` et qui correspondent à une organisation interne des concepts regroupant les objets de l'application.

De même, « *grand* » et « *rouge* » ne renvoient pas à des mécanismes identiques lors de la résolution de la référence : contrairement à la taille, la couleur est une propriété qui possède une valeur (la longueur d'onde) prototypique. Il est ainsi possible de calculer la distance entre les couleurs des référents et la valeur prototypique du mot utilisé. Ces mécanismes se combinent lorsqu'une expression référentielle comporte un nom et un adjectif qualificatif. Nous suivons en cela le principe de compositionnalité sémantique dont nous avons déjà parlé.

Un deuxième aspect dans les interactions entre la perception visuelle et la parole concerne les indices d'appel à un contexte visuel d'interprétation, et les contraintes que l'utilisation d'un mot particulier posent lors de l'identification d'un tel contexte. Ces indices nous semblent être portés en priorité par le déterminant. Nous suivons les grands principes établis par Corblin (1987) et repris en particulier par Salmon-Alt (2001b) pour une utilisation dans un cadre computationnel :

- Un indéfini de forme « *n N* » extrait n éléments d'un domaine comprenant des éléments de type N.
- Un défini s'interprète dans un domaine à l'intérieur duquel son contenu constitue un signalement singularisant.
- Un démonstratif s'interprète dans un domaine comportant obligatoirement un élément identifiable autrement que par l'expression référentielle (focalisation discursive ou par un geste d'ostension). Le démonstratif étant contrastif, il sous-entend l'existence dans le domaine d'un autre élément de même nature et non focalisé.
- Un pronom demande toujours un domaine comportant un élément focalisé préalablement.

Ces règles, bien qu'établies surtout dans un cadre purement linguistique, spécifient des contraintes sur des entités de nature extra-linguistique, et particulièrement sur des domaines d'interprétation. On retrouve de telles contraintes lors de l'interprétation de mots tels que « *autre* », « *premier* », « *suivant* » : ces mots s'interprètent dans un domaine comportant obligatoirement une focalisation préalable pour « *autre* », et un ordonnancement pour « *premier* » et « *suivant* ». Ces règles nous semblent parfaitement adaptées à notre approche et nous les exploiterons sans les remettre en question.

Prémices relatives aux interactions entre perception visuelle et geste. Nous allons ici reprendre quelques apports du modèle de Wolff pour élaborer les principes d'un premier traitement du geste

en contexte visuel, avant sa confrontation avec le langage. Il ne s'agit pas d'une interprétation mais d'une représentation unifiée des deux modalités pour faciliter une interprétation ultérieure à l'aide de l'énoncé verbal.

Pour déterminer les demonstrata d'une trajectoire gestuelle donnée, nous avons choisi d'utiliser des scores numériques : à chaque objet de la scène visuelle est attribué un score entre 0 et 1 dénotant la probabilité qu'a l'objet d'être un demonstratum. Des scores non égaux à 0 ou à 1 sont ainsi attribués pour pallier l'imprécision des trajectoires. Ces scores illustrent à notre sens le concept de traitement partiel : ils ne constituent pas une interprétation mais donnent une indication pour l'interprétation ultérieure. La détermination d'un score ne dépend pas du type de trajectoire (entre pointage, ciblage, entourage et gribouillage) mais du type d'accès aux objets (entre élictif et séparateur). Quelques calculs doivent précéder l'élaboration des deux hypothèses correspondantes :

1. La délimitation de la zone élictive de chaque objet et de chaque groupe perceptif (au sens de la Gestalt). Elle se fait sur la base de la disposition de l'objet ou du groupe perceptif par rapport aux autres objets (cf. figure 2.4), ainsi que de la compacité de l'objet : un objet parfaitement circulaire a une compacité maximale, tandis qu'un objet long et fin a une compacité minimale. Cet indice de compacité est aussi appelé agrégation.
2. Le taux de recouvrement d'une trajectoire, c'est-à-dire la proportion de pixels de la trajectoire recouvrant un objet et non le fond de l'image.
3. La pertinence pour une trajectoire de correspondre à un entourage. Pour que cette pertinence soit élevée, il faut que la trajectoire présente au moins une courbure significative, et que l'on puisse déterminer de quel côté sont les demonstrata.

Le taux de recouvrement permet, s'il est nul et si la trajectoire ne se réduit pas à un pointage, d'écarter immédiatement l'hypothèse du geste élictif. Dans le cas contraire, les deux hypothèses sont émises et le calcul des scores correspondants se fait de la manière suivante, illustrée dans la figure 2.6 :

- Dans l'hypothèse d'un geste élictif :
  Le score maximum de 1 est attribué à tout objet que la trajectoire superpose. Certaines nuances peuvent être traduites. Par exemple, un objet qui ne trouve pas directement sous la trajectoire mais en est très proche se voit attribuer un score quantifiant cette proximité. De même, un objet qui se trouve dans la continuité immédiate d'une trajectoire se voit attribuer un score proche de 1. D'autre part, le sens de production d'un geste élictif est un paramètre important. Il définit en effet un ordre dans les objets ciblés, et cet ordre est conservé pour l'interprétation tripolaire ultérieure.

- Dans l'hypothèse d'un geste séparateur :
  Le score maximum est attribué à tout objet se trouvant clairement à l'intérieur de la zone entourée par la trajectoire. Lorsque qu'un objet est très proche ou est recouvert par la trajectoire, le score traduit la proximité ou le taux de recouvrement.

L'algorithme de détection des demonstrata que nous avons ainsi spécifié émet des hypothèses qui ne pourront être interprétées qu'à l'aide de la parole. Si l'expression référentielle « *ces deux objets* » est associée à la trajectoire et au contexte visuel de la figure 2.6-A, les référents seront `c` et `e`, c'est-à-dire les deux objets à avoir le score maximal de 1. Si l'expression est « *ces quatre objets* », les quatre objets ayant un score non nul seront identifiés comme référents. En revanche, si l'expression est « *ces trois objets* », l'identification du troisième objet, plus complexe,

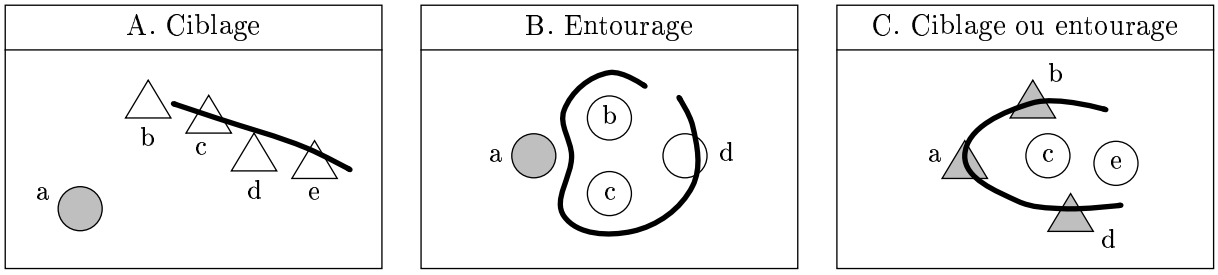


Figure 2.6 – *Scores numériques pour les hypothèses élictive et séparatrice.*

fait intervenir l'ordre de ciblage : l'objet `d` étant selon cet ordre entouré de deux référents, il sera identifié comme le troisième référent. Le principe est le même pour les trajectoires B et C de la même figure. La trajectoire B illustre en particulier l'intérêt de la sémantique de la trajectoire : la singularité constituée par l'évitement de l'objet `a` permet d'attribuer directement un score de 0 à cet objet, sans calcul de proximité par rapport à la trajectoire.

Quant aux deux cas limites présentés dans la figure 2.5 page 47, ils ne posent aucun problème avec ce système de scores. La trajectoire A ne peut pas être interprétée comme séparatrice, la seule zone entourée étant la boucle et ne contenant aucun objet. Dans l'hypothèse élictive, elle présente des scores certes faibles mais non nuls pour les trois carrés clairs, qui sont ainsi identifiés comme les demonstrata (ce qui correspond bien à ce que l'on comprend). La trajectoire B conduit à des scores médiocres dans l'hypothèse élictive et à des scores proches de 1 dans l'hypothèse sépatrice. Pour elle également, les demonstrata obtenus sont bien les trois carrés clairs. Nous ne reviendrons pas sur ces scores qui constituent en quelque sorte un pré-requis pour la suite de la modélisation.

Prémices relatives aux interactions entre parole et geste. Pour nous focaliser sur le problème de la référence aux objets dans la suite de ce travail, nous voulons proposer ici une stratégie algorithmique pour le problème de l'appariement des expressions référentielles et des gestes ostensifs. Notre algorithme montre que l'appariement ne se fait pas uniquement sur des informations temporelles mais a au contraire recours à des informations sémantiques. Pour tenir compte des phénomènes décrits dans le chapitre 1 et en particulier des appariements multiples, il tient compte en entrée des informations suivantes :

1. La synchronisation temporelle : dates de début et dates de fin de chaque geste et de chaque mot ; détection d'éventuelles régularités (indice important pour les énumérations).
2. Quelques informations prosodiques : détection des accents toniques, particulièrement sur les démonstratifs (indice permettant de renforcer certaines hypothèses de synchronisation).
3. Les catégories et les propriétés des objets, d'une part du côté des hypothèses de demonstrata, d'autre part du côté des expressions référentielles verbales. Elles constituent des indices de filtrage.
4. Les cardinalités, d'une part du côté des hypothèses de demonstrata (nombre d'objets concernés par l'ensemble des trajectoires gestuelles), d'autre part du côté des référents intentionnels (indication par un adjectif numéral, par un singulier ou un pluriel, ainsi que par les éventuelles énumérations et coordinations).

5. L'identité des objets dans les hypothèses de demonstrata. Cette information sert à vérifier les éventuelles hypothèses de répétition, hypothèses soulevées lors de l'analyse syntaxique. En effet, si deux gestes consécutifs désignent exactement les mêmes objets, et si de plus ces gestes sont produits parallèlement à une répétition langagière, l'hypothèse de répétition est confirmée.

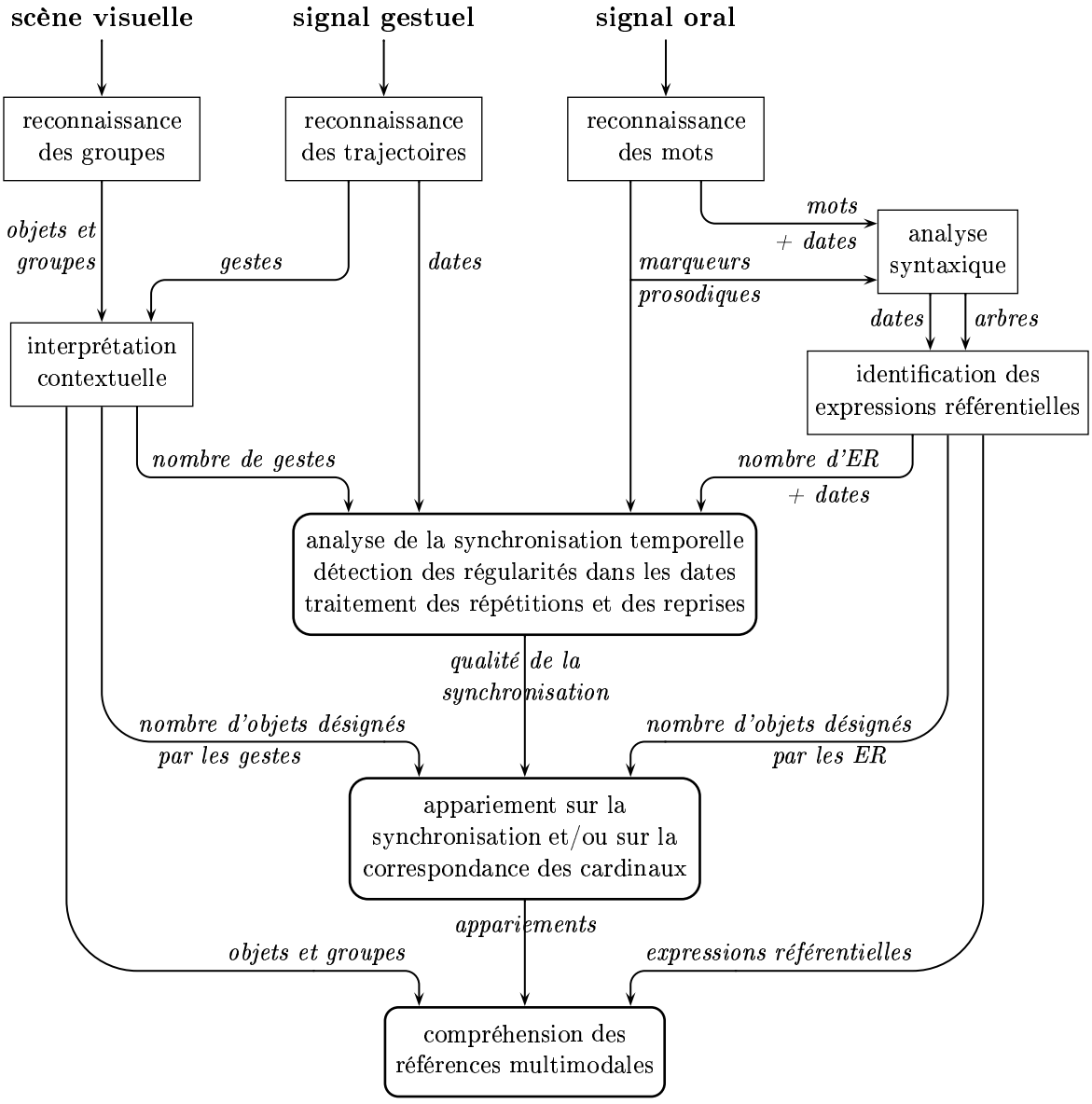


Figure 2.7 – *Appariement des expressions référentielles verbales et des gestes.*

La figure 2.7 illustre l'ordre de prise en compte de ces informations. Une fois que toutes les expressions référentielles verbales et tous les gestes ostensifs de l'énoncé courant sont identifiés, l'algorithme commence par traiter les désignations de lieu pour lesquelles les ambiguïtés sont moindres. Les gestes associés aux déictiques purs tels que « *ici* » sont ainsi traités et ignorés dans la suite du traitement. Pour les expressions référentielles et les gestes restants, les hypothèses d'appariement sont classées selon les indices temporels et prosodiques. Les possibles appariements

multiples sont testés en commençant par les hypothèses de gestes globaux plutôt que par les hypothèses de gestes individuels. Cette stratégie permet d'apparier en priorité ce qui concerne le maximum d'objets, et évite ainsi d'émettre un trop grand nombre d'hypothèses. Les cardinalités permettent alors de comparer le nombre de référents attendus avec le nombre de demonstrata dans les hypothèses émises. Ne sont pris en compte que les objets qui correspondent avec ce qui est attendu en termes de catégorie et de propriétés.

Cet algorithme faisant intervenir des hypothèses de demonstrata, on peut considérer qu'il tient compte du contexte visuel et correspond à une modélisation des interactions tripolaires pour le problème de l'appariement. Ce n'est pas encore le cas, pour la simple raison que les interactions entre la perception visuelle et les modalités d'expression ne sont pas toutes prises en compte. Notre algorithme s'appuie sur la nature des objets affichés et des groupes perçus, mais par encore sur la structuration du contexte visuel en sous-ensembles interprétatifs ou sur les phénomènes de saillance qui sont à la base de la résolution de la référence.

Avec cet algorithme pour le pré-traitement du couple parole–geste, avec les scores numériques que nous avons définis pour le pré-traitement du couple perception visuelle–geste, et avec les règles de correspondance entre mot et concept ainsi que les règles d'interprétation des déterminants pour le couple perception visuelle–parole, nous disposons du pré-requis nécessaire à l'élaboration de notre modèle tripolaire. Dans la deuxième partie de cette thèse, nous allons détailler les notions à la base de ce modèle. Nous commencerons par la notion de domaine de référence : la compréhension commence par la délimitation du domaine dans lequel appliquer les règles d'interprétation des déterminants et effectuer les pré-traitements que nous venons d'exposer. Nous nous focaliserons dans le chapitre 4 sur l'identification de tels domaines. Certains aspects méthodologiques sous-tendant notre modélisation méritent tout d'abord être éclaircis.

RÉCAPITULATIF
L'idée principale de ce chapitre est que la résolution de la référence aux objets passe par une compréhension fine des relations réciproques entre trois pôles : la perception visuelle, le langage et le geste. Une présentation des principaux modèles cognitifs et des principaux systèmes informatiques nous ont permis d'appréhender les facettes de ces relations. Nous avons montré que les travaux existants se limitent bien souvent à deux pôles, un raisonnement directement à trois pôles s'avérant beaucoup plus délicat. C'est néanmoins notre but, et nous nous sommes attaché dans ce chapitre à en poser les bases. Nous avons ainsi proposé une adaptation du modèle de Wolff pour qu'il devienne compatible avec notre approche, ainsi qu'un premier traitement des modalités. Avec ce chapitre, nous aboutissons à des interprétations partielles conciliables avec une future interprétation globale.

Chapitre 3

Positions théoriques et méthodologiques

Comment aborder le problème de l'intégration des sens portés par les trois modalités? Quelle méthodologie suivre dans le domaine du dialogue homme-machine? Quelles sont les autres disciplines concernées par cette problématique et quels ont été leurs principaux apports? Sur quels résultats pouvons-nous nous baser et dans quels domaines de recherche pouvons-nous remettre en cause certaines propositions et avancer nos propres propositions?

Ce chapitre présente tout d'abord (§ 3.1) les différentes facettes de la méthodologie de conception d'un système de dialogue homme-machine. Nous y verrons sur quels types de données peuvent se baser les spécifications des caractéristiques d'un système, et nous critiquerons les différentes méthodes de validation. Nous dépasserons ensuite (§ 3.2) les frontières de l'informatique-linguistique pour appréhender les principales caractéristiques de notre méthodologie et nous positionner face aux préoccupations et aux résultats de disciplines telles que l'intelligence artificielle, la linguistique, la pragmatique, la psychologie cognitive ou encore la sémiotique, qui sont elles aussi concernées par la problématique du dialogue homme-machine.

3.1 Méthodologie pour le dialogue homme-machine

Quelles questions se poserait un épistémologue analysant le domaine du dialogue homme-machine? Ce domaine peut-il constituer une science? Sa méthodologie est-elle scientifique?

Une science se caractérise par un objet et une démarche objective. Si l'objet est ici la conception de machines capables de dialoguer naturellement avec un être humain, qu'en est-il de la démarche? Plus précisément, une science se caractérise d'une part par des méthodes de recueil et d'analyse d'observations, et d'autre part par des méthodes de validation des interprétations (hypothèses, modélisations) qui sont faites de ces observations. Lors d'une première étape, nous étudierons le recueil d'observations. Compte tenu de la nature hétérogène de ces observations, nous présenterons dans une deuxième étape le travail d'intégration particulièrement délicat en dialogue homme-machine. Nous étudierons ensuite les méthodes de validation.

Enfin, une science aboutit à une théorie ou à un modèle de son objet, et, corrolairement,

produit des connaissances. L'intérêt des connaissances produites en dialogue homme-machine sera abordé dans la section suivante.

3.1.1 Le recueil de situations de dialogue

Restreindre les situations de dialogue. Quels types de dialogue peut-on étudier pour en tirer des hypothèses opérationnelles? Idéalement, tout type de dialogue entre deux humains devrait être utilisable. Les mots utilisés, les gestes produits, mais aussi toutes les composantes des matériaux paraverbal et non verbal devraient être observés, étant susceptibles d'être exploités en communication homme-machine. Bien entendu, les phénomènes correspondants sont infinis et certaines restrictions sont nécessaires, ces restrictions pouvant diriger soit le choix de dialogues à étudier s'ils sont disponibles, soit la mise en œuvre de situations de dialogue particulières, dans un but d'enregistrement et d'étude *a posteriori*.

Pour détailler ces restrictions, nous reprendrons les quatre types de dialogue que nous avons distingués dans le chapitre 1 (§ 1.1.1). Ainsi, dans le cas de la conception d'un système de dialogue finalisé, seule l'observation des situations correspondant à la tâche est vraiment nécessaire, ce qui réduit d'autant les possibilités d'interaction. Lors de la mise en œuvre d'une situation de dialogue finalisé, des instructions précises seront données aux interlocuteurs, pour les inciter à communiquer dans un but précis les dirigeant dans leurs productions verbales et gestuelles ainsi que dans leurs façons de gérer les interactions.

Pour la conception d'un système de dialogue à support visuel, une attention toute particulière au contexte visuel est nécessaire. Si ce sont des situations de communication humaine que l'on veut étudier, elles pourront faire intervenir des objets physiques qui seront filmés avec les participants. Si l'on veut se rapprocher de situations de communication homme-machine, une méthode consiste alors à médiatiser par une machine un dialogue entre deux humains. Tout se fait par l'intermédiaire d'un écran sur lequel sont affichés les objets virtuels de l'application. Avec cette méthode, c'est directement sur l'ordinateur que peuvent être enregistrées les caractéristiques du contexte visuel. De plus, si le logiciel permettant la communication médiatisée est capable de gérer des capteurs pour la voix et les gestes, ce sont également les modalités d'interaction qui peuvent être directement enregistrées, sans passer par des vidéos mais par un codage facilement exploitable des énoncés et de leur contexte d'interprétation.

Dans le cas particulier du dialogue de commande, l'un des participants se voit alors attribuer un rôle particulier consistant à exécuter les commandes de l'autre participant. Il joue en quelque sorte le rôle de la machine. On peut lui donner des directives pour qu'il n'accepte que certains types de commande. Cette méthode, connue sous le nom de magicien d'Oz en référence au roman de Frank Baum, consiste donc à simuler les capacités d'un système de dialogue pour étudier comment un utilisateur se comporterait face à lui. Le magicien est celui qui joue secrètement le rôle de la machine. Le corpus Magnét'Oz a été obtenu selon ce principe.

Simuler la machine. Plusieurs remarques critiques sur la méthode du magicien d'Oz s'imposent. Tout d'abord à propos du comportement du magicien (ou compère) : des raisons pratiques font qu'il s'agit souvent du concepteur du futur système. En effet, personne ne sait mieux que lui quels types d'acceptation et de réponse sont envisageables pour un système effectif sur la tâche considérée. Du coup, un biais par rapport à une communication vraiment spontanée peut apparaître. Si le magicien est une personne non concernée par la conception du système, on constate alors souvent qu'elle a du mal à respecter des contraintes supposées être celles d'une machine (cf. par exemple Salmon-Alt 2001b). Ce biais, auquel s'ajoute celui qui intervient inévi-

tablement lors de l'interprétation des données recueillies, soulève le problème de la validité de cette méthode.

Du côté de l'utilisateur de la simulation, plusieurs types de comportement apparaissent. L'utilisateur (ou sujet de l'expérimentation) peut communiquer naturellement ou au contraire être gêné par le fait que son interlocuteur soit une machine (Kennedy *et al.* 1988, Jönsson & Dählback 1988). Ainsi, l'absence de perception de l'attention de l'utilisateur de la part de la machine, ainsi que l'absence de retour visuel (par exemple du degré d'écoute de la machine), peuvent être troublantes et amener l'interaction à des échanges très impersonnels. Le sujet peut ainsi simplifier son langage jusqu'à le réduire à sa plus simple expression et à son plus simple contenu. On parlera alors de discours filiforme. Dans des cas extrêmes, certains mots grammaticaux, pourtant indispensables même dans un niveau de langue familier, peuvent être omis. On aboutit alors à un style *calepin* ou à un style *télégraphique*. D'une manière générale, la machine est souvent considérée comme inférieure : elle utilise le vouvoiement et on lui parle comme à un enfant.

Un autre type de comportement face à la machine est lié à la peur de provoquer un dysfonctionnement. Les ordinateurs font encore peur, dans le sens où on se sent démuni face à eux, et où on ne sait jamais quelles seront les conséquences de la plus petite erreur de manipulation. Certains utilisateurs sont tendus et communiquent comme en situation de stress. La qualité de l'enregistrement en est alors amoindrie. On retrouve très souvent, en particulier dans le corpus Magnét'Oz, un comportement lié à cette peur du dysfonctionnement : celui consistant à se limiter à ce qui a déjà fonctionné. L'utilisateur ose une commande orale ou multimodale, et constatant sa réussite, réitère exactement le même type d'énoncés pendant toute la durée de l'expérience. L'enregistrement obtenu est alors très pauvre en termes d'étendue des phénomènes observés. Qui plus est, l'importante fréquence d'un type d'interaction pourra biaiser les calculs statistiques à propos des fréquences d'apparition des différents phénomènes. Nous reparlerons des exploitations statistiques dans la section traitant de la validation (§ 3.1.3).

3.1.2 Le travail d'intégration

Des travaux parcellaires. Peut-on axer ses travaux de recherche sur le dialogue homme-machine dans sa globalité ? Les approches et travaux existants montrent que non. Les informaticiens s'occupant du traitement du signal sont confrontés à tellement de problèmes de reconnaissance qu'ils ne peuvent pas approfondir en parallèle des recherches en compréhension ou en génération. Plus encore, on trouve des spécialistes des lexiques informatiques, des spécialistes des formalismes syntaxiques, des spécialistes en sémantique formelle, des spécialistes en pragmatique focalisés sur l'exploitation du contexte, des spécialistes en pragmatique focalisés sur les actes de langage, etc. Dans le domaine des interfaces multimodales, c'est encore plus caricatural : on y trouve quasiment une spécialité pour chaque type de capteur, sans oublier le cloisonnement fréquent entre ceux qui s'intéressent à l'interprétation et ceux qui s'intéressent à la génération.

Les travaux sont d'une manière générale trop parcellaires. Chacun développe dans sa propre communauté scientifique sa partie de ce qui devrait être un jour un système de dialogue opérationnel. Beaucoup de problèmes se révèlent lorsqu'il s'agit de connecter tous les modules ainsi construits, de les intégrer pour en faire un système complet. Il arrive par exemple que les données spécifiées en entrée d'un module ne soient pas suffisamment exploitées. Il arrive que d'autres données, non prévues, s'avèrent utiles. La multitude des types de données, leur répartition large à des niveaux correspondant aux signaux bruts ou à des interprétations complexes, rendent leur confrontation problématique.

L'intégration des modules. L'intégration est ainsi une part entière du travail de conception. Plus que cela, c'est un axe de recherche comportant ses propres innovations : si l'on est capable de programmer des modules de reconnaissance de la parole, si l'on sait quelles informations placer dans un historique du dialogue ou dans le modèle de l'utilisateur, on a en revanche beaucoup de mal à spécifier comment les informations s'échangent et s'exploitent dans un système intégrant le module que l'on maîtrise. Ce problème est illustré par exemple avec l'expérience issue du projet VERBMOBIL (Wahlster 2000).

Compte tenu du fait que la spécification d'un module de compréhension langagière fait intervenir des connaissances linguistiques, il paraît évident que des échanges entre plusieurs disciplines sont nécessaires à l'élaboration d'un système de dialogue. Informaticiens et linguistes sont ainsi réunis par une même problématique, dans une approche de recherche *pluridisciplinaire*. Lorsque la problématique est purement informatique, la prise en compte de résultats issus d'une autre discipline peuvent aider. Cette approche *interdisciplinaire* se trouve par exemple dans l'exploitation informatique des critères de la Gestalt. Enfin, après les approches pluri et interdisciplinaires, se trouve la vision globale du système, toutes disciplines confondues, vision *transdisciplinaire* qui reste un idéal vers lequel tendre.

Deux principes méthodologiques : la transposition et la réduction. Dans le cadre de l'intégration du sens, les informations sémantiques provenant des différentes sources doivent pouvoir se comparer et se réduire. Ainsi, pour se confronter et constituer un nouveau sens plus global, les interprétations des signes visuels et des énoncés peuvent se comparer sur la base de leurs traits sémantiques. Si l'on veut garder une complexité non formalisable en traits sémantiques, il reste nécessaire de réduire les sens pour qu'ils puissent prendre la forme d'une structure formelle implantable dans un système informatique. Des structures telles que les domaines de référence à la base de notre modèle opèrent une réduction qui tente de conserver le maximum de propriétés sémantiques. Le problème posé par une telle réduction est commun à toutes les approches qui partent de modèles issues de la psychologie cognitive : ces modèles, construits souvent dans un but descriptif, sont peu formalisables pour un informaticien. Celui-ci est bien obligé d'en extraire les aspects formalisables et de négliger le reste. Cette réduction est à la fois nécessaire et destructive, détruisant la cohérence du modèle initial. L'important est de partir des modèles les plus formels pour minimiser la réduction. Une autre méthode consiste à spéculer sur la nature des aspects non formalisables pour les transformer en quelque chose d'implantable. Le risque est de faire perdre au modèle transformé toute validité psychologique, et nous éviterons cette démarche.

La transposition et la réduction interviennent également à un niveau méthodologique. Par exemple, les nombreux travaux faits sur le langage et en particulier la distinction de Morris (1974) entre syntaxe, sémantique et pragmatique, peuvent être utilisés pour l'analyse des trajectoires gestuelles sur écran tactile. La transposition permet de se poser des questions telles que : quelles peuvent être les composantes de la syntaxe d'une trajectoire gestuelle ? l'interprétation pragmatique d'une expression référentielle multimodale suit-elle les mêmes grands principes que celle d'une expression référentielle verbale ? La transposition peut ainsi permettre de prendre conscience des lacunes d'un sujet de recherche par rapport à un autre, et même d'identifier des problèmes. Quant à la réduction, elle sous-tend cette méthodologie puisque la comparaison ne se fait que dans le cadre de ce qui a déjà été identifié.

3.1.3 La validation

Valider un module. Pour la validation d'un composant lors d'une conception modulaire, les

méthodologies possibles sont classiquement les suivantes :

1. Valider de manière logique ou mathématique les fonctionnalités du composant.
2. Simuler les entrées et les sorties des composants connexes, en se fondant sur des données observées, par exemple issues de corpus.
3. Intégrer le composant dans une plate-forme existante, et faire fonctionner celle-ci dans des conditions réelles avec plusieurs types d'application.

La première méthode s'avère difficile compte tenu de la complexité des informations circulant entre les modules. Nous avons vu en effet que ces informations proviennent de sources diverses et que, pour certaines d'entre elles, leur formalisation était le résultat de la réduction de modèles issus de la psychologie. Il ne s'agit donc pas de variables mathématiques simples mais de structures complexes et hypothétiques sur lesquelles il est difficile d'appliquer une démonstration.

La deuxième méthode est quasiment inévitable lors de la conception. Si elle permet à chacun de tester son propre module, elle ne permet cependant pas d'appréhender le fonctionnement global du système. La qualité de la validation dépend des données fournies en entrée du module. Il ne s'agit pas seulement de fournir des données correspondant à tout ce que le module est capable de traiter, il s'agit également, pour les différents contenus possibles, de simuler leur fréquence d'apparition les uns par rapport aux autres, ainsi que leur fréquence d'entrée dans le module (par rapport au temps). Dans le cas contraire, les capacités du module seront validées mais pas le comportement de l'architecture. Lorsque les modules connexes sont implantés, cette méthode conduit naturellement à l'intégration logicielle.

Cette troisième méthode touche le domaine des architectures logicielles, faisant intervenir la répartition des capacités du système en modules et la spécification des interactions entre ces modules. D'une manière générale, une architecture doit être simple et ouverte, et l'interaction entre les modules doit avoir une base suffisamment normalisée pour que chaque concepteur s'y retrouve et puisse accéder à l'ensemble. Une fois le système opérationnel, il constitue un moyen de validation du modèle pour l'application instanciée. Deux inconvénients restent majeurs : d'une part la validation du modèle indépendamment de l'application est impossible, d'autre part l'étendue des phénomènes validés n'est pas complète puisqu'elle correspond aux situations testées. Les phénomènes validés correspondent ainsi à ceux qui apparaissent le plus fréquemment, autrement dit ceux que toute modélisation prend en compte en priorité, et non ceux qui sont à la limite d'un modèle ou qui sont à la base d'une distinction entre deux approches.

L'utilisation de corpus. Un corpus est un ensemble de données observées, servant par exemple pour la simulation des entrées d'un module. Il peut être constitué suite à une expérimentation, ou tout simplement à partir d'archives (de journaux, de bibliothèques). Dans notre cadre, la forme électronique est nécessaire pour rendre possible un traitement automatique. Elle rend également plus aisés l'échange de corpus dans la communauté de l'informatique-linguistique ainsi que les calculs statistiques permettant de déterminer la fréquence de tel ou tel phénomène.

Le recours aux corpus est un moyen efficace de validation. Cela nécessite de trouver un corpus correspondant au type d'application concernée par le système que l'on veut tester, éventuellement de construire soi-même un tel corpus à l'aide une expérimentation du type magicien d'Oz, et de fournir en entrée du système toutes les situations comprises dans le corpus. On calcule alors le pourcentage de situations correctement traitées et on obtient un indice de performance du système.

Seulement les résultats sont souvent insatisfaisants, en particulier lorsque le corpus n'a pas été élaboré pour le système à valider. La tâche applicative dans le corpus diffère généralement de celle du système, cette différence suffisant à rendre importante la fréquence de situations non

prévues et par conséquent mal traitées. Or ne considérer que la partie du corpus correspondant aux situations prévues revient à rendre la validation sans valeur.

Reste la solution consistant à construire un corpus spécialement pour le système. Le problème rencontré est alors d'ordre méthodologique : que vaut une validation faite sur des données élaborées pour cette validation ? Ce problème se retrouve au niveau du codage informatique du corpus : la façon dont on code un corpus ne doit théoriquement pas être dirigée par ce que l'on veut en tirer, mais doit au contraire garder une totale objectivité. Comme le montre (Habert *et al.* 1997), les biais sont difficiles à éviter et des précautions méthodologiques sont à prendre pour qu'ils ne soient pas trop nombreux.

Un autre problème qui reste ouvert dans le processus de validation sur corpus a trait à l'interprétation des phénomènes observés : pour que la validation recouvre un large panel de situations, il s'avère tentant de généraliser un phénomène observé à une classe de phénomènes. Si cela reste possible pour des phénomènes purement langagiers faisant intervenir peu de paramètres, il n'en est pas de même en multimodal. Toute scène visuelle est en effet unique : les paramètres visuels varient beaucoup plus que ceux du langage, et tous ces paramètres sont impossibles à maîtriser. Même la Gestalt est insuffisante : tout petit écart dans la disposition des objets peut par exemple avoir des conséquences sémantiques très fortes.

Enfin, une approche fondée sur la spontanéité de la communication homme-machine, et donc sur la nécessité pour le système de traiter tous les phénomènes susceptibles d'apparaître, se distingue d'une approche pour laquelle seuls les phénomènes les plus fréquents importent. Les corpus ne sont pas exploités de la même manière : si dans cette deuxième approche c'est essentiellement la *représentativité* des phénomènes qui compte, dans la première c'est aussi et surtout leur *significativité*, c'est-à-dire la nature des caractéristiques sémantiques qu'ils impliquent. La validation peut ainsi se faire non pas sur un échantillon représentatif du corpus, mais sur un échantillon significatif, où tous les doublons (situations sémantiquement identiques) ont été retirés.

La légitimité des corpus. Pour être exploitable informatiquement, un corpus ne correspond pas à un simple enregistrement, mais à un enregistrement transcrit et annoté. Le codage consiste à transcrire sous forme textuelle des données auditives, visuelles ou gestuelles. L'annotation consiste à interpréter ce codage pour le rendre lisible et exploitable. En se limitant au codage, on reste dans une approche *descriptive*. En annotant, on entre dans une approche *interprétative*, par projection d'une théorie. Pour un corpus langagier par exemple, le codage consiste à transcrire un enregistrement vocal en une suite de mots accompagnés éventuellement d'informations prosodiques. Si ce corpus est destiné à une étude des références, il est utile d'annoter les expressions référentielles qui y apparaissent. Seulement, c'est ici qu'apparaît un problème essentiel à l'annotation : pour annoter une expression référentielle, encore faut-il que cette notion soit clairement définie, pour éviter que l'annotateur ait à prendre des décisions face à des cas incertains, et ait donc à interpréter (cf. Poesio 2000).

On peut arriver à une situation où la même personne annote un corpus et valide son système à l'aide de ce corpus. On en revient alors au problème précédent : que vaut une telle validation ? Pour ce qui concerne l'annotation des références, Salmon-Alt (2001) en vient à une observation du même ordre : la validation d'une théorie sur des données présuppose un corpus annoté, et l'annotation d'un corpus doit reposer sur des principes issus d'une réflexion théorique. Face à ce paradoxe, la solution consiste à n'annoter que des phénomènes pour lesquels on dispose d'une base théorique consensuelle. Seulement un tel consensus s'avère très rare et le problème reste donc ouvert.

Dans le domaine de l'interaction multimodale, le codage et l'annotation de corpus sont des problèmes particulièrement délicats. Le codage d'une trajectoire gestuelle telle qu'elle peut être captée par un écran tactile, correspond en effet à une suite de points par lesquels passe la trajectoire au cours du temps. Cette transcription est inexploitable, d'une part car une suite de points ne dit rien sur la forme de la trajectoire (entourage ou ciblage), d'autre part car le contexte visuel sur lequel s'appuie la trajectoire n'intervient pas dans ce codage, mais, au mieux, dans un codage connexe. Pour pouvoir être exploitée, une trajectoire gestuelle doit donc être interprétée, cette étape nécessaire s'avérant difficile à la fois d'un point de vue méthodologique (sur quelles bases théoriques interpréter?) et d'un point de vue pratique (le travail requis est énorme et pas vraiment enrichissant). Nous voyons là l'origine du manque cruel de corpus multimodaux dans la communauté informatique, à la fois dans le domaine des interfaces homme-machine et dans celui de l'informatique-linguistique.

Dans le domaine du dialogue, l'interprétation est également une étape nécessaire dans l'annotation : une situation de dialogue fait appel à des informations implicites qui peuvent prendre une importance toute particulière dans une situation d'échange et dans un but applicatif précis. Pour que le dialogue ne soit pas mal interprété, le codage de certaines de ces informations implicites peut être utile. Mais qu'en est-il de leur statut ? Font-elles vraiment partie du corpus ? La solution consiste à ajouter au corpus des commentaires pour en faciliter la compréhension. Encore une fois, la qualité de la validation peut être remise en question par cette méthode. Nous pouvons conclure de tous ces problèmes non résolus d'une part que la validation sur corpus n'est pas encore une méthodologie acquise, d'autre part que l'extension progressive de l'interaction langagière vers l'interaction multimodale ne pose plus de problèmes méthodologiques qu'elle n'en résout. Au moins les problèmes posés permettent-ils de remettre en question quelques positions prises et donc de faire avancer le domaine.

3.2 Choix méthodologiques

Quelles sont les caractéristiques de notre approche ? En quoi les directives que nous nous sommes fixées sur la spontanéité de l'interaction et sur le recours aux particularités structurelles de la perception visuelle, du langage et du geste dirigent-elles notre approche ? Nous discuterons dans une première étape des caractéristiques d'une approche fondée sur la spontanéité dans l'interaction homme-machine, pour nous positionner ensuite face aux disciplines concernées, et aborder enfin un positionnement suite aux difficultés rencontrées lors de la phase de validation.

3.2.1 L'approche fondée sur la spontanéité de la communication

Effort pour la machine et non pour l'homme. « C'est à la machine de comprendre l'homme et non à l'homme de comprendre la machine ». Cette phrase, de plus en plus présente dans les publicités pour des ordinateurs, reflète la caractéristique principale de notre approche : ce n'est plus à l'utilisateur de faire l'effort de lire des manuels complexes et d'apprendre un langage artificiel de communication avec la machine (approche *technocentrée*), mais c'est au contraire à la machine de laisser l'utilisateur libre de s'exprimer comme il en a l'habitude, et de faire l'effort, par des algorithmes et des heuristiques complexes, de comprendre le sens de ce qu'il exprime et de le traduire dans un langage artificiel interne (approche *ethnocentrée*).

Que veut dire « spontané » ? Au départ, le terme utilisé est « naturel », comme pour le langage dans les expressions « langage naturel » ou « traitement automatique du langage naturel ». Or

naturalité peut sous-entendre innéisme [1]. Du fait que l'utilisation d'une machine n'est absolument pas innée mais nécessite au contraire un apprentissage, qu'il s'agisse d'utiliser une souris, un écran tactile ou même un microphone, nous éviterons le terme « naturel » pour parler de communication homme-machine. Nous lui préférerons le terme « spontané ».

Celui-ci sous-entend un grand nombre de choses, par exemple l'absence d'apprentissage, de contraintes ou encore d'efforts dans la production de gestes et de mots. Au contraire, l'utilisation d'une machine nécessite bien un apprentissage. Elle est donc bien liée à des contraintes et à des efforts, ne serait-ce que parce qu'il faut faire des efforts pour parler, et encore plus quand l'interlocuteur est une machine. L'important est de limiter ces trois aspects, et c'est ce caractère minimal qu'évoque la spontanéité. De plus, une communication homme-machine spontanée peut sous-entendre que l'utilisateur n'a plus conscience de la machine, et arrive à s'exprimer sans tenir compte de la nature de son interlocuteur. Ce n'est pas vraiment possible, surtout dans le cadre d'un dialogue à support visuel (par opposition au dialogue téléphonique). Qu'on le veuille ou non, on sera toujours conscient de parler à une machine. L'important est que cela ne nous empêche pas d'avoir un comportement communicatif normal. C'est là que nous voyons le sens principal de « spontané » : dans le fait que l'utilisateur arrive à considérer la machine comme un interlocuteur ayant les mêmes capacités de compréhension que l'homme, et arrive à parler avec elle comme il discute habituellement avec ses semblables, sans effort supplémentaire.

Quelles sont les caractéristiques d'une communication homme-machine spontanée ? La liberté dans les choix expressifs et dans les tours de parole, la possibilité non pénalisante de ratés dans la production de gestes ou de mots, la nécessité pour la machine d'avoir des réactions comparables à celle d'un homme (à la fois dans leur contenu que dans le temps de réponse), sont autant de caractéristiques. Le fait que l'utilisateur puisse couper la parole à la machine, et même le fait que la machine puisse couper la parole à l'utilisateur, jouent ainsi pour la spontanéité. Celle-ci est riche de détails simples tels que les hésitations ou les bafouillements,, petits détails qui peuvent sembler inutiles ou ridicules mais qui font partie de la communication spontanée.

Un autre point particulièrement important dans notre cadre du dialogue à support visuel est le recours encouragé à la multimodalité. En effet, le fait même d'afficher une scène visuelle incite inévitablement l'utilisateur à désigner les objets qui s'y trouvent. Le geste ostensif est ainsi encouragé, et nous pouvons dire que dialogue homme-machine à support visuel sous-entend multimodalité. Des expérimentations telles que celle de Hauptmann & McAvinney (1993) vont dans ce sens.

Mais est-il seulement possible de communiquer spontanément avec une machine ? Vivier (2001) trouve qu'actuellement, il est possible de vivre l'illusion, c'est-à-dire de prêter un rôle d'interlocuteur à une machine. Pour entretenir cette illusion, la machine doit avoir un visage, c'est-à-dire une représentation (virtuelle par un avatar, ou réelle par un corps construit à l'image de l'homme) de sa capacité à écouter. Selon les termes de Vivier, l'agent qui interagit avec un autre agent doit avoir une image de lui-même, une image de son interlocuteur et une image de la situation en jeu dans cette interaction. C'est à travers ces images qu'il peut adapter ses énoncés et construire avec lui une référence commune.

Beaucoup d'utilisateurs éprouvent cependant la peur du dysfonctionnement que nous avons déjà évoquée et qui nous semble liée à leur incompréhension des mécanismes internes de la machine. Avant de pouvoir communiquer spontanément avec une machine, il faut ne plus en avoir

1. Pour des considérations sur les liens nécessaires entre inné et acquis, ainsi que sur l'apparition du geste ostensif chez l'enfant, cf. en particulier (Cyrulnik 1995), qui montre comment le sens naît de l'association entre le geste ostensif et la parole.

peur. Le problème est peut-être insoluble pour nous qui ne sommes pas nés dans les ordinateurs. Par contre, pour les générations futures, pour les enfants qui naissent à côté d'eux, qui s'entrainent dès l'école primaire à les utiliser et qui jouent avec des jouets-robots, le problème ne se posera peut-être même pas. Ils n'auront sans doute aucune appréhension et ce n'est qu'avec eux qu'on pourra exploiter tout ce que l'on prépare actuellement pour la communication homme-machine spontanée.

Conséquences sur les recherches en dialogue homme-machine. Un système autorisant une communication spontanée ne doit imposer ni de restrictions ni d'apprentissages, et doit être capable de traiter tout type d'énoncé multimodal. Malheureusement, pour que le système le traite correctement, un type d'énoncé doit avoir été prévu par les concepteurs du système. Or il est impossible de tout prévoir. Comme un système fonctionne pour une tâche précise, on suppose que les mots et les gestes liés à la tâche (auquel on ajoute le langage d'interaction courant) suffiront à prévoir toutes les formes possibles d'énoncés. Mais, comme nous l'avons déjà vu, il reste impossible de prévoir tous les types de situations pragmatiques.

Nous avons montré dans le chapitre 1 que le dialogue homme-machine met l'accent sur les gestes ostensifs. Nous avons montré dans le chapitre 2 que ce type de geste peut prendre une multiplicité de formes lorsque l'interaction se fait à l'aide d'un écran tactile, mais que cette multiplicité peut être caractérisée avec un nombre limité de paramètres. Une première direction de recherche consiste ainsi à restreindre l'interaction par le choix d'un dispositif, pour ensuite analyser et exploiter la plus petite partie d'énoncé produit spontanément avec ce dispositif. Pour que les conditions d'interaction puissent reproduire les conditions propres à la communication humaine, il faudrait que la machine soit incarnée par un robot et, entre autres, que les gestes de l'utilisateur soient captés sans le gêner, par exemple à l'aide d'un système de caméras. Cette deuxième direction de recherche consiste à renforcer la technologie pour maximiser les chances de reconnaître le geste émis spontanément. Face à ces deux directions de recherche, nous avons choisi de ne pas nous préoccuper pour l'instant d'aspect techniques, mais de nous concentrer sur les significations du geste (même guidé par le dispositif) dans sa complexité et surtout dans sa complémentarité avec la parole. Notre choix se fonde sur les arguments suivants :

1. L'écran tactile est un dispositif qui ne requiert aucun apprentissage. La communication est spontanée car on est habitué au crayon et au papier.
2. Non seulement l'écran tactile met l'accent sur les gestes ostensifs, mais il permet de plus de les détecter facilement, et même d'acquérir directement leur partie signifiante, sans la phase de préparation qui a lieu avant ni la phase de retrait qui a lieu après, phases identifiées par Kendon (1980). Nous évitons ainsi un grand nombre de problèmes techniques.
3. L'acquisition et le traitement de trajectoires planes sont facilement réalisables et opérationnelles compte tenu de l'état de la technologie et des recherches sur le traitement du geste.
4. Les modèles de Bellalem et de Wolff sur lesquels nous nous appuyons ont été élaborés à partir d'expérimentations sur écran tactile.

Le problème de la compréhension à partir d'un microphone et d'un écran tactile est ainsi à la fois plus simple et plus compliqué que celui de la compréhension dans la communication humaine. C'est plus facile car le système ne reçoit que l'essentiel (les parties signifiantes des gestes). C'est plus complexe car le système perd beaucoup d'indices : la direction du regard, les attitudes et les postures, etc. Mettant l'accent sur les gestes ostensifs, c'est en tout cas un bon compromis pour l'analyse et la modélisation de la référence aux objets.

3.2.2 Position par rapport aux différents courants et disciplines

Nécessité d'une pluridisciplinarité. Face à la problématique de la compréhension automatique en dialogue homme-machine à support visuel, nous avons vu que l'informatique seule était impuissante, de même que la linguistique ou la psychologie. L'informatique ne se suffit pas, car pour modéliser des capacités de communication homomorphiques à une machine, il faut commencer par étudier l'homme et ses capacités à communiquer. La linguistique non plus, car elle ne prend pas assez en compte les caractéristiques structurelles des informations extra-linguistiques, surtout au niveau de la multimodalité. La linguistique considère souvent le geste comme déterminant directement les référents, et a tendance à négliger les subtilités de l'interaction entre langage et geste, ainsi que les particularités structurelles du contexte visuel. Il en est de même de la psychologie, du fait de l'absence d'un véritable cadre formel, permettant par exemple des confrontations autres qu'expérimentales. Face à des contraintes computationnelles, la psychologie apparaît trop floue, comprenant beaucoup de théories et de modèles, ainsi que beaucoup de contradictions difficiles à évaluer pour un non-psychologue.

Notre approche tente d'intégrer quelques apports de chacune de ces disciplines. Notre positionnement se caractérise par un travail de recherche dans certaines disciplines et par la formulation d'hypothèses ou tout simplement l'utilisation de résultats établis dans d'autres. Nous nous plaçons dans une optique à long terme, fondée sur l'aspect spontané de la communication homme-machine, sans tenir compte de l'état actuel de la technologie ou des problèmes ergonomiques actuels, mais en nous focalisant sur les mécanismes fondamentaux de la compréhension. Notre travail se situe donc dans la recherche fondamentale. Nous allons maintenant donner des précisions sur notre positionnement, discipline par discipline[1].

Positionnement par rapport à l'intelligence artificielle. Deux grands débats jalonnent l'histoire de l'intelligence artificielle : celui de l'existence d'états mentaux dans la cognition humaine et de l'intérêt d'en doter une machine, et celui de la possibilité pour une machine de manipuler du sens et pas seulement des formes.

Face au premier débat, notre position correspond à celle du cognitivisme faible. Contrairement au cognitivisme fort qui postule que les états mentaux existent, le cognitivisme faible voit les états mentaux comme n'existant pas forcément mais comme un moyen commode de modéliser une capacité cognitive. Plus précisément, nous adhérons au cognitivisme structural, qui aborde la représentation des états mentaux par des structures et des mécanismes de fonctionnement de ces structures, et qui illustré en particulier par la Gestalt. Nous nous opposons au cognitivisme computationnel, qui, entre autres, voit le cerveau comme un outil de traitement d'information et postule qu'une machine peut par conséquent être aussi intelligente qu'un humain. Loin de cette position, nous considérons que la modélisation d'états mentaux nous permet de simuler des capacités cognitives mais en aucun cas de rendre la machine intelligente.

Ceci nous conduit au deuxième débat : la machine peut-elle accéder au sens ? Nous considérons que non, que la capacité de manipuler des sens est liée à l'intuition humaine et n'est pas formalisable. Le sens peut-il se déduire de formes ? Peut-être. Dans notre approche, nous essayons de multiplier les sources possibles de formes dans le but d'approcher le sens. C'est en structurant le contexte visuel, c'est en tentant de formaliser des notions telles que la saillance que nous multiplions les sources de formes et que nous espérons nous rapprocher d'une intelligence artificielle.

1. L'informatique n'apparaît pas dans la liste des disciplines qui suivent, pour la simple raison que nous la considérons comme regroupant des disciplines s'intéressant à la modélisation d'une intelligence artificielle, à la conception d'interfaces homme-machine, ou encore à la modélisation de contraintes ergonomiques.

Positionnement par rapport à l'ingénierie des interfaces homme-machine. Les recherches dans ce domaine se caractérisent d'une part par des classifications des phénomènes liés à la multimodalité (surtout dans la communauté française), d'autre part par des algorithmes de fusion et par des architectures de systèmes multimodaux. Les classifications (cf. par exemple Coutaz & Caelen 1991) concernent en particulier les relations entre les informations langagières et gestuelles (complémentarité, redondance, équivalence), ainsi que les possibilités de synchronisation temporelle de ces informations (fonctionnement concurrent, alterné, synergique, exclusif). Elles n'ont aucun intérêt compte tenu de notre approche fondée sur la spontanéité : l'utilisateur étant libre dans sa production de gestes et de mots, tous les cas répertoriés par ces classifications sont possibles.

Les aspects algorithmiques et architecturaux ont par contre une grande utilité. Même s'ils sont souvent trop simples, les algorithmes mettent l'accent sur des structures de données (telles que les structures de traits) et sur des mécanismes (tels que l'unification ou la fusion d'information). Quant aux architectures logicielles, elles explorent les possibilités de gestion de modules, les précautions et les contraintes liées à la présence d'un noyau central dirigeant le traitement, la nécessité de spécifier un langage et des protocoles d'interaction entre les modules, etc. C'est donc dans leurs aspects techniques et non théoriques que les recherches sur les interfaces multimodales nous aident.

Positionnement par rapport à l'ergonomie. Contrairement à l'approche fondée sur la spontanéité dont le but est de faire évoluer les machines pour que les interactions soient les plus proches possibles des fonctionnements des dialogues humains, l'approche ergonomique considère que les utilisateurs auront toujours à s'adapter aux conditions de communication spécifiques à telle ou telle machine. Elle appuie ce point de vue sur les arguments suivants : les gens ne sont pas prêts à considérer une machine comme un interlocuteur ; des problèmes ergonomiques en découlent, et découlent également du caractère nécessairement applicatif de l'interaction ; il y a des limites technologiques, ne serait-ce que dans les dispositifs d'acquisition et dans leur mise en œuvre (Carbonell *et al.* 1997).

Ces arguments sont discutables. Nous avons déjà vu avec la remarque de Vivier (2001) et les nôtres que si nous ne sommes pas encore prêts à considérer une machine comme un interlocuteur, nos enfants le seront sans doute. Face au caractère orienté de la communication induit par la tâche applicative, nous rappelons que notre but est non pas d'étudier une interaction spécifique à une application, mais d'aboutir à un modèle de la communication qui soit plausible d'un point de vue cognitif et soit ainsi applicable à n'importe quelle application. C'est pourquoi nous nous centrons sur le problème de la référence aux objets, problème indépendant de l'application et situé au cœur de la conception d'un système de dialogue. Quant aux limites technologiques, elles nous paraissent avoir bien peu de poids : les progrès sont tels que les prendre en compte revient à travailler sur du court terme (et parfois à perte, par exemple lorsqu'un dispositif est remplacé par un autre), ce que nous nous refusons.

L'approche ergonomique a néanmoins un rôle important en amont, d'une part sur certains aspects méthodologiques qui servent à l'approche fondée sur la spontanéité, d'autre part sur l'identification de scénarios d'interaction cohérents. Elle s'avère également indispensable pour l'utilisation de certains dispositifs.

Positionnement par rapport à la psychologie. Les états mentaux dont nous avons parlé dans le paragraphe relatif à l'intelligence artificielle trouvent ici des fondements théoriques, avec la notion de représentation mentale. Le lien entre perception et langage d'un côté, et représentation mentale de l'autre côté est l'objet d'innombrables travaux qui peuvent être distingués en deux courants :

celui de la psychologie cognitive qui s'intéresse à la perception, la mémoire, la saillance ; et celui de la psychologie sociale qui traite du schéma de la communication, des processus interactionnels ou encore de la pertinence. Face à la diversité des approches, nous ne pouvons pas nous situer ici dans tel ou tel courant, d'autant plus que certaines caractéristiques de notre approche appartiennent à certains courants et d'autres à des courants parfois très différents. Nous le voyons ici avec la saillance et la pertinence, deux grands paramètres de notre travail, qui sont classées dans deux branches différentes.

Nous adhérons globalement au courant des grammaires cognitives, qui postule que le langage constitue une propriété émergente procédant des mécanismes généraux de la cognition et entretenant de nombreuses ressemblances avec d'autres activités cognitives, notamment la perception. Ce courant rejoint ainsi la Gestalt sur laquelle nous nous basons également. D'autre part, nous adhérons également, du moins dans une certaine mesure, au modularisme de Fodor (1986). Avec notre objectif computationnel et la presque nécessité d'une implantation informatique modulaire, il est en effet difficile de ne pas suivre les grands principes du modularisme tels que : la spécificité des modules à une opération précise et irrépressible (le fonctionnement d'un module ne peut pas être inhibé par un effort délibéré) ; l'encapsulation des informations (un module n'a accès qu'à une information limitée et ne prend pas en compte les informations internes aux autres modules) ; ou encore l'existence d'un système central chargé de coordonner et de centraliser les informations traitées par les modules. Cependant, nous ne suivons plus Fodor dans ses idées innéiste et surtout fonctionnaliste (à savoir qu'il est possible de dissocier les opérations effectuées des supports matériels qui les permettent), cette dernière entraînant une analogie du cerveau humain avec une machine, ce que nous avons déjà exclu.

Parmi les grands résultats de la psychologie cognitive sur lesquels nous nous appuyons dans nos hypothèses, citons les travaux sur la mémoire à court terme, en particulier ceux de Miller (1956) selon lesquels la capacité de cette mémoire tourne autour du chiffre sept, qu'il s'agisse d'objets ou de procédures amenant à l'identification d'objets. Citons également que le traitement des images est plus riche et donne lieu à plus d'interprétations que le traitement de la parole, et que des expérimentations incluant un protocole de rappel montrent que les images sont mieux mémorisées (Weil-Barais 1993). Parmi les résultats de travaux s'intéressant à la perception, nous noterons que les sens accaparent l'attention de la manière suivante : vue = 70 % (qui reste le principal sens pour la connaissance du monde), ouïe = 20 %, odorat = 5 % (Heilig 1992). Nous retiendrons également, pour la perception visuelle, la primauté de la couleur sur la taille, et, parmi les couleurs, la primauté du rouge sur les autres, ces constatations résultant d'expérimentations (Baticle 1985). Dans la suite de notre travail, nous utiliserons ces résultats sans les remettre en question.

Un point important dans notre approche est notre position face à la théorie de la Gestalt. Nous avons vu en 1.3.1 (page 32) l'usage que nous faisons des critères avancés par la Gestalt pour la détection de groupes perceptifs et d'objets saillants. Comme nous le verrons plus en détail dans le chapitre 4, nous structurons ainsi le contexte visuel en une partition de groupes perceptifs. Chaque groupe pouvant comporter des parties, la structure obtenue est hiérarchique. Or un des principes fondamentaux de la Gestalt est que la perception est globale, immédiate, inconsciente, et qu'elle se caractérise par un tout qui ne peut être assimilé à la simple addition de ses parties. Notre approche hiérarchique du groupage peut alors sembler contredire ce fondement. Il en est de même avec l'utilisation que nous ferons des lignes directrices : le fait que le regard parcourt la scène visuelle en suivant des lignes de force peut sembler contredire la perception globale de la Gestalt.

Notre réponse tient d'une part dans une distinction entre première perception inconsciente et

seconde perception semi-consciente, et d'autre part dans la multiplicité des perceptions globales possibles. Nous considérons en effet que les phénomènes de groupage et de saillance interviennent essentiellement lors d'une première étape perceptive correspondant à la perception globale de la Gestalt. Pour une même scène visuelle, les caractéristiques de cette perception globale peuvent varier d'un individu à l'autre. Plusieurs hypothèses sont par conséquent émises par la machine. Un autre argument pour justifier l'élaboration de plusieurs hypothèses réside dans la possibilité d'ambiguïtés visuelles. Pour pouvoir les conserver et les confronter, ce sont ces hypothèses qui sont regroupées dans nos structures hiérarchiques, et non les parties d'un tout correspondant à la perception globale. Quant au parcours du regard dans l'image, il requiert un certain intervalle de temps, qui s'avère incompatible avec l'aspect immédiat de la perception globale. Il s'inscrit donc dans une deuxième étape perceptive correspondant à une prise de connaissance consciente de l'environnement. Nous étudierons l'influence des groupes perceptifs dans ce phénomène, en montrant l'intérêt qu'apporte la structuration hiérarchique élaborée pour l'étape précédente. L'utilisation que nous ferons de cette structure ne contredit en rien les principes fondamentaux de la Gestalt qui ne sont pas concernés par cette deuxième étape perceptive.

Enfin, le principe fondamental de la Gestalt illustre bien notre approche de la tripolarité : les interactions tripolaires forment un tout qui est plus que la simple addition des interactions bipolaires. Nous ne suivons cependant plus ce principe lorsqu'il s'agit d'implantation informatique. Avec ce principe ,la Gestalt est en effet souvent associée au connexionnisme. Si cette approche informatique s'avère efficace pour la reconnaissance des formes et pour les problèmes liés à la vision artificielle, elle ne correspond pas à notre approche. D'une part parce que dans notre cas le système connaît déjà les objets affichés sur la scène et n'a pas à les reconnaître, d'autre part parce que nous avons besoin, pour l'intégration sémantique des trois modalités, d'états mentaux qui soient compatibles d'une modalité à l'autre, au moins dans leur structure. Notre structure hiérarchique de groupes perceptifs nous semble intéressante à ce point de vue.

Positionnement par rapport à la linguistique et à la pragmatique. Le principal inconvénient de beaucoup de travaux linguistiques pour notre problématique, c'est qu'ils relèguent l'extra-linguistique à ce qui sert de secours lorsque l'interprétation langagière n'aboutit pas. Trouver des travaux qui s'intéressent à la perception visuelle et au geste sans les reléguer à un rôle passif n'est pas immédiat. Ces travaux s'avèrent en fin de compte bien plus proches de la pragmatique que de la linguistique traitant du lexique, de la syntaxe, et même de la sémantique. Nous nous plaçons ainsi dans le courant (parfois appelé néo-fonctionnaliste) qui, se fondant sur les travaux du cercle de Prague, continue des recherches sur le contexte, le discours, la pragmatique, ou encore l'exploitation de corpus. Dans la pragmatique, deux approches peuvent être distinguées : l'approche cognitive et celle de l'analyse conversationnelle. Avec nos préoccupations de plausibilité cognitive et notre utilisation de la Théorie de la Pertinence, nous nous situons dans l'approche cognitive. L'analyse que nous faisons de corpus de dialogue n'est en rien liée à l'approche centrée sur l'analyse de la conversation, puisque nous n'y étudions que les phénomènes de référence et non les processus interactionnels.

Parmi les grands résultats sur lesquels nous nous appuyons, se trouvent les maximes de Grice (1975) et la Théorie de la Pertinence (Sperber & Wilson 1995), c'est-à-dire les principaux travaux concernant le fait que l'on communique spontanément de manière efficace, même quand c'est à une machine que l'on s'adresse. Les maximes de Grice décrivent au niveau du contenu d'un message les grandes règles à suivre pour éviter les incompréhensions et les malentendus : apporter suffisamment d'information, ne rien dire que l'on ne puisse démontrer, éviter d'utiliser des expressions ambiguës, etc. Quant à la Théorie de la Pertinence, elle s'appuie sur la présomption de pertinence d'un énoncé, c'est-à-dire le fait que tout locuteur donne à entendre que son

message est pertinent, même si l'interlocuteur doit recourir à des informations implicites pour le comprendre. Nous présenterons les grands principes de cette théorie dans le chapitre 7.

Un point important dans notre approche est notre position face à la sémantique formelle. Parmi les approches en sémantique, la sémantique formelle touche à la fois au couple linguistique–pragmatique et à l'informatique. Contrairement à la sémantique descriptive qui n'a recours qu'à la langue elle-même pour caractériser l'interprétation, la sémantique formelle exploite l'apport de formalismes logiques ou mathématiques aux considérations sémantiques, pour rendre possible leur modélisation rigoureuse et éventuellement leur implantation informatique.

A la suite de la théorie des quantificateurs généralisés, nous donnons une importance particulière aux déterminants, que nous considérons comme une fonction particulière s'appliquant à la dénotation du nom pour produire la dénotation de l'expression référentielle verbale. Nous nous situons surtout dans la lignée des approches dynamiques, à savoir la DRT (*Discourse Representation Theory*) dont nous avons déjà parlé, mais aussi de la théorie des changements de contexte (*File Change Semantics*), des logiques dynamiques (*Dynamic Predicate Logic*) ou encore de la sémantique des mises à jour (*Update Semantics*) (cf. Corblin 2002). Comme toutes ces approches, nous introduisons une représentation intermédiaire entre les expressions langagières et les concepts. Comme elles, nous considérons l'interprétation comme un processus incrémental : l'interprétation de chaque énoncé met à jour un état d'information sur le monde pour aboutir à un nouvel état d'information. Celui-ci se traduit dans notre approche par des domaines de référence. Nous suivons en cela l'approche de (Reboul *et al.* 1997) dont nous décrirons l'évolution au cours du chapitre 4.

Positionnement par rapport à la sémiotique. Etant la seule discipline à appréhender avec une importance égale la perception visuelle, le langage et le geste, la sémiotique nous fournit surtout un cadre théorique. C'est elle qui nous aide à prendre conscience des phénomènes liés à l'interprétation des signes visuels, des signes langagiers et des signes gestuels. C'est d'elle que viennent le modèle du signe et la notion de référent. C'est encore elle qui favorise les emprunts et les adaptations d'une discipline à l'autre. Le cadre qu'elle fournit permet de transposer d'un domaine à l'autre des mécanismes, des critères classificatoires, et d'explorer ainsi une branche peu abordée dans une discipline grâce aux résultats d'une branche similaire issue d'une autre discipline. Nous utilisons ce principe dans le chapitre 5 pour caractériser la notion de saillance langagière en nous aidant des résultats de travaux sur la saillance visuelle.

La sémiotique de l'image a une importance particulière dans notre approche. Joly (1993) montre que l'image peut être abordée par plusieurs théories : mathématiques, informatique, esthétique, psychologie, sociologie, etc. « Pour en sortir nous devrons donc faire appel à une théorie plus générale, plus globalisante, qui nous permette de dépasser les catégories fonctionnelles de l'image. Cette théorie est la théorie sémiotique » (p. 21). L'important dans l'approche sémiotique est d'aborder les phénomènes dans la façon dont ils provoquent des significations, dont ils provoquent une démarche interprétative. Dans notre cadre de la communication homme-machine, nous retrouvons cette même multiplicité d'approches et ce même intérêt pour les sources de significations. Les mots tels que les déterminants, les singularités des trajectoires gestuelles, les groupes perceptifs sont autant de sources de significations. La sémiotique de l'image nous permet de consolider notre approche de la saillance visuelle et des lignes directrices dans l'image fixe. Deux types d'approches se combinent dans l'analyse d'une image :

1. L'analyse selon l'axe syntagmatique, c'est-à-dire comment l'observateur passe d'un élément à l'autre dans la structure morphologique de l'image. Cette analyse est structurelle et fait intervenir les caractéristiques visuelles que sont point, ligne, forme, couleur, contraste,

masse, rythme, etc. Les arts picturaux et architecturaux en ont posé les principes, par exemple à travers les cours du Bauhaus (Itten 1961) ou les œuvres de Kandinsky (1979).

2. L'analyse selon l'axe paradigmatique, c'est-à-dire comment chaque élément de l'image fait penser à d'autres éléments, souvent symboliques, issus de la culture, du sacré, ou encore de l'expérience de l'observateur. Les travaux de Barthes (1964) ont ainsi donné des bases scientifiques à la composition d'images publicitaires.

L'analyse d'image (ou lecture d'image) regroupe ces deux approches, et la sémiotique considère généralement qu'une image ne peut pas être correctement interprétée sans la combinaison des deux. La bonne photographie est de même celle pour laquelle les deux arguments, syntagmatique et paradigmatique, se renforcent. Cette combinaison nous pose un premier problème dans notre analyse du contexte visuel : seule l'analyse selon l'axe syntagmatique semble pouvoir être modélisée pour être implantée dans un système. Les paramètres de l'analyse selon l'axe paradigmatique (iconicité, inconscient, culture) sont en effet pour le moins abstraits. Un deuxième problème est d'ordre méthodologique : comment peut-on généraliser des analyses faites à partir d'une image particulière, comme c'est souvent le cas des analyses sémiotiques ? Qui plus est, peut-on généraliser des principes d'analyse prévus pour des images travaillées et non pour les scènes visuelles d'un système de dialogue homme-machine ?

Aussi bien pour la saillance que pour les lignes directrices, nous faisons l'hypothèse que le caractère finalisé de la communication homme-machine inhibe chez l'utilisateur toute possibilité d'analyse sur l'axe paradigmatique. En effet, l'interaction finalisée se caractérise par des changements fréquents de contexte visuel, ainsi que par une rapidité et une efficacité incompatibles avec le temps nécessaire à l'analyse d'une image fixe. Nous nous focalisons donc sur l'analyse syntagmatique, c'est-à-dire sur les aspects morphologique du contexte visuel, comme nous le faisons également pour le groupage avec les critères de la Gestalt.

Pour répondre aux deuxième et troisième problèmes, nous considérons que les critères morphologiques identifiés à l'aide de quelques images structurées sont des critères inhérents à toute image. Dans toute scène visuelle d'une application informatique, se trouvent en effet un cadre avec les caractéristiques structurelles que cela implique, et des objets avec les conséquences morphologiques que leur disposition induit. De plus, nous ne voyons la modélisation de ces critères morphologiques que comme un argument supplémentaire venant appuyer la pertinence d'un domaine de référence ou d'une hypothèse de référent. Il ne s'agit en aucun cas de baser toute une interprétation sur de tels critères. Avec ces précautions, l'apport de la sémiotique nous paraît renforcer notre approche fondée sur l'exploitation d'indices hétérogènes pour l'interprétation.

La sémiotique nous apporte donc un cadre théorique unifié permettant la prise de conscience de phénomènes et les échanges interdisciplinaires. Les notions de signifié, de polysémie, de code ou encore de double articulation se retrouvent en effet dans le langage comme dans la perception visuelle, renforçant ainsi notre idée que la communication multimodale forme un tout devant être appréhendé globalement.

Conséquences sur notre travail de recherche. Ces divers positionnements ont induit une certaine démarche dans l'exploration des champs de recherche. La confrontation des diverses approches, s'est avérée complexe mais nécessaire pour étudier les éléments pertinents à notre but d'intégration. C'est dans le domaine de la linguistique computationnelle que l'exploration est la plus difficile, d'une part parce que les approches et les modèles proposés y sont nombreux, d'autre part parce que notre sujet s'intéresse particulièrement à la pragmatique du langage et requiert ainsi beaucoup de pré-requis en syntaxe et en sémantique. Plus que d'exploration, il s'agit aussi d'un travail de comparaison des phénomènes observés dans tel ou tel contexte étudié par telle

ou telle discipline, d'un travail de confrontation de théories, d'adaptation de modèles dans un contexte computationnel, ainsi que d'intégration de paramètres hétérogènes.

Au cours de l'élaboration de notre modèle, nous avons dû poser des hypothèses sans lesquelles nous n'aurions pas pu avancer dans l'intégration. Dans le dernier chapitre de cette thèse, nous proposons des protocoles expérimentaux pour valider ces hypothèses. La nécessaire complexité de ces protocoles prouve que leur mise en œuvre ne peut pas rentrer dans le cadre d'un travail de thèse. La validation constitue ainsi pour nous une perspective de recherche. Comme elle se trouve à l'origine de la principale difficulté rencontrée au cours de ce travail, nous allons maintenant montrer en quoi elle est problématique dans notre approche.

3.2.3 Choix face aux problèmes de validation

Les difficultés dans la validation. Nous avons évoqué en § 3.1.3 diverses méthodes de validation pour le domaine du dialogue homme-machine. Maintenant que notre approche est clarifiée, nous allons passer ces méthodes en revue pour préciser les difficultés qu'elles soulèvent et pour tester leur compatibilité avec nos impératifs sur la spontanéité de la communication et sur l'intégration de données hétérogènes pour l'interprétation. Ces méthodes sont les suivantes :

- La validation par élaboration d'un système de dialogue homme-machine :
  Cette méthode s'adapte mal à notre approche fondée sur la spontanéité et sur la multimodalité. Il faudrait en effet valider le système sur tous les cas de figure possibles. Or, du fait de la combinatoire des formes de référence, ces cas de figure sont innombrables et il ne semble possible d'en valider que les plus fréquents. Des tests suffisamment nombreux permettent en effet l'apparition de situations prototypiques, sans même forcer l'utilisateur à se mettre dans de telles situations (pour ne pas en compromettre la spontanéité). Or la subtilité d'un modèle repose justement sur sa capacité à traiter des phénomènes peu fréquents mais qui dénotent un fonctionnement significatif du langage naturel. A l'origine de notre critique se trouve également la notion d'acceptabilité. Le comportement d'un système de dialogue est validé lorsqu'il est jugé acceptable. Or cette notion n'est pas quantifiable mais est au contraire subjective. D'où notre position face à cette méthode de validation, avant même de considérer les difficultés pratiques liées à l'élaboration d'un système de dialogue, difficultés que nous avons présentées en § 3.1.3. Si nous ne proposons pas dans cette thèse de système complet ni même de morceaux d'implantation (qui n'auraient dans ce cas été produits que dans une optique à court terme), nous proposons néanmoins des directives d'implantation.

- La validation par élaboration d'un module d'interprétation :
  Cette méthode nécessite une simulation des données entrant et sortant de ce module, ce qui s'avère très complexe compte tenu de la fréquence et de la nature hétérogène de ces données. Il faudrait simuler non seulement les modalités d'entrée (ce qui s'avère complexe pour le geste), mais également le contexte visuel. Il faudrait surtout synchroniser toutes ces données. Ces contraintes requièrent un travail de programmation presque aussi lourd que la conception d'un système complet. D'autre part, l'interprétation multimodale met en œuvre bien plus qu'un seul module : elle nécessite la gestion d'historiques de l'interaction (un par modalité), d'un modèle de l'application, d'un modèle de l'utilisateur, etc. Implanter cet ensemble revient quasiment à implanter un système de dialogue complet. Ici aussi, notre principale contribution est la proposition de directives opérationnelles.

- La validation à l'aide de corpus :

En plus des problèmes méthodologiques que nous avons détaillé en § 3.1.3, et du fait de l'importance que nous donnons à la multimodalité et à l'implicite dans la communication, nous ne pouvons que rejeter cette méthode. Nous voulons continuer ici la distinction entre phénomènes représentatifs et phénomènes significatifs, pour en déduire une méthodologie qui nous semble pouvoir compléter l'exploitation classique de corpus. Le but du dialogue homme-machine fondé sur la spontanéité est de comprendre tout le langage naturel et tous les gestes possibles. Une méthodologie adaptée doit donc passer par l'analyse de la variabilité des phénomènes. A partir de la catégorisation des expressions référentielles selon leur détermination et selon la présence ou non d'une catégorie et de modifieurs ; à partir de la catégorisation des trajectoires gestuelles selon leur mode d'accès aux demonstrata ; à partir aussi des caractéristiques structurelles du contexte visuel (présence ou non de groupes perceptifs, d'objets saillants, de lignes directrices), il nous semble possible de regrouper la grande majorité des phénomènes en une série de caractéristiques. Nous pouvons ainsi extraire les types de situations qui apparaissent dans un corpus, par exemple : groupe nominal démonstratif + singulier + catégorie + trajectoire élictive + groupe perceptif. Pour chacune de ces situations, nous pouvons reconstruire un exemple prototypique, généralement en reprenant et simplifiant un extrait du corpus. Nous faisons l'hypothèse qu'un modèle d'interprétation de la référence multimodale qui fonctionne correctement sur ces exemples prototypiques fonctionnera sur l'ensemble des situations qui en sont à l'origine. Ce n'est que dans une deuxième phase, si certains exemples prototypiques s'avèrent mal traités par notre modèle, que nous pouvons nous intéresser à la fréquence d'apparition des situations correspondantes dans le corpus, soit pour nous autoriser à les négliger, soit pour corriger le modèle.

- La validation par simulation selon la méthode du magicien d'Oz :

  En plus des critiques adressées en § 3.1.1, nous reprenons ici l'argument donné pour la validation d'un système de dialogue, à savoir la nécessité de simuler tous les cas de figure possibles, ce qui s'avère impossible avec une approche fondée sur la spontanéité. Ces critiques face à la validation ne s'appliquent cependant pas au recueil de données, toujours indispensable à la modélisation. Les données obtenues à l'aide de la méthode du magicien d'Oz sont en effet très utiles pour valider les classifications et les mécanismes à la base d'un modèle. Le corpus Magnét'Oz (Wolff 1999) a ainsi été utilisé de la manière suivante : la simulation a été lancée par Wolff pour l'aider dans son travail de modélisation ; les données obtenues ont permis de prendre conscience de certains phénomènes ; elles ont également permis la validation de certains aspects du modèle de Wolff ; elles sont de plus à la base de nos caractérisations, qui nous permettent de ce fait d'appréhender les conditions expérimentales d'une future simulation. On alterne ainsi entre phase théorique de modélisation et phase expérimentale de simulation, l'une apportant à l'autre les bases pour l'améliorer. Dans cette boucle, nous considérons que nous avons suffisamment exploité le corpus Magnét'Oz et que nous avons les connaissances et les hypothèses pour aborder l'élaboration d'une nouvelle simulation. Ce sera l'une de nos perspectives.

- La validation à l'aide d'expérimentations psycholinguistiques :

  Cette méthode s'avère difficilement compatible avec notre approche ciblée sur la modélisation des interactions tripolaires. En effet, valider un paramètre se fait en maîtrisant les autres paramètres, c'est-à-dire en utilisant des situations qui les inhibent, ou pour lesquelles ils peuvent être parfaitement décrits. Ces contraintes sur les conditions d'expérimentations semblent aller à l'encontre de la spontanéité. Il s'avère de plus difficile de maîtriser des pa-

ramètres cognitifs tels que la mémoire : pour chaque situation focalisée sur des paramètres visuels ou gestuels, rien ne dit que la mémoire du sujet ne vient pas interagir, par exemple suite à la répétition d'un même type de situation. Malgré ces problèmes, l'expérimentation selon des protocoles issus de la psycholinguistique s'avère intéressante pour la validation de notions telle que la saillance visuelle ou les domaines de référence. Des protocoles fondés sur le rappel et décrivant les aspects expérimentaux nécessaires ont en effet déjà été utilisés et nous semblent réutilisables dans notre contexte. De plus, une validation selon ces protocoles permettrait de donner une dimension cognitive à notre modèle. Pour cette raison, nous explorerons cette voie en proposant dans le chapitre 10 des expérimentations adaptées à nos paramètres.

Face à ces difficultés, nous avons défini quelques perspectives de recherche et nous avons choisi prioritairement la voie de l'expérimentation pour la validation. Cette voie n'est cependant pas totalement explorée dans cette thèse, du fait des moyens qu'elle demande. Il s'avère en effet que monter une expérimentation pour un seul paramètre requiert beaucoup de temps, réparti entre les spécifications du protocole, les séances successives avec les sujets, les enregistrements et les calculs qui y sont associés, ainsi que le dépouillement des données recueillies et la publication des résultats. Compte tenu de la multiplicité de nos paramètres, nous sommes contraint de reléguer le déroulement de ces expérimentations dans nos perspectives.

C'est finalement la validation par des échanges dans la communauté scientifique qui semble la seule réalisable à court terme, c'est-à-dire dans le cadre des trois années de la thèse. Même si elles n'en sont pas représentatives, nos publications reflètent ces échanges et constituent une façon provisoire de valider nos propositions (provisoire car l'évaluation de publications reste subjective et très liée au contexte scientifique). A titre d'indice, à chaque fois qu'une de nos publications viendra valider un aspect de notre modèle, nous l'indiquerons par une référence bibliographique. Nous noterons également qu'un cadre d'échanges pluridisciplinaires présente ses inconvénients. Les problématiques du dialogue homme-machine sont en effet complexes, et sont souvent mal perçues dans les communautés monodisciplinaires. Les problèmes qui se sont posés compte tenu de notre travail axé sur l'intégration peuvent être exposés de la manière suivante :

- présenter en centrant sur l'aspect théorique conduit les lecteurs à instancier le modèle pour leur propre application particulière et à produire des critiques inadaptées ;
- présenter en centrant sur un exemple pratique concret tend à réduire la complexité ;
- présenter à l'aide d'un panel d'exemples représentatifs s'avère lourd ;
- mettre l'accent sur l'intégration dévalorise les recherches faites dans chaque pôle ;
- mettre l'accent sur les pôles entraîne une sous-estimation du travail d'intégration.

Face à ces problèmes, nous avons choisi d'alterner entre présentations théoriques et présentations basées sur des exemples, en tentant de garder à chaque fois des préoccupations quant à la complexité des interactions entre les paramètres de notre modèle.

Conséquences de nos choix méthodologiques. Après ces considérations méthodologiques plutôt pessimistes, il peut sembler aléatoire de se lancer dans un tel travail de recherche. Nous voulons maintenant montrer que la recherche dans le domaine de la communication multimodale spontanée est au contraire très enrichissante, d'une part pour celui qui l'entreprend, d'autre part pour l'évolution des systèmes de dialogue homme-machine. Pour celui qui l'entreprend, parce que s'intéresser à la communication dans ses nombreux aspects et à travers toutes ses modalités s'avère très formateur. Ce travail permet en particulier d'acquérir une vision globale sur les mécanismes de référence, bien plus que ne le permettent des approches exclusivement linguistiques ou informatiques. Cette vision globale inclut une prise de conscience de phénomènes complexes que nous

avons essayé de transcrire dans les deux premiers chapitres. Elle permet d'élaborer des classifications plus générales et donc mieux exploitables que ne le sont celles obtenues en se concentrant sur un point de vue particulier. Nous allons ainsi proposer des caractérisations formelles qui nous semblent être les principaux apports de notre travail. Ces caractérisations constituent un moyen irremplaçable pour décrire précisément des phénomènes et pour préparer leur exploitation computationnelle. Notre but est ainsi de donner le maximum d'appuis théoriques nécessaires à la modélisation de la communication multimodale pour de futurs systèmes de dialogue. Notre contribution appartient en cela à l'informatique : formalisant des concepts issus d'autres disciplines, elle fournit des principes pouvant conduire directement à une implantation. Suivant cet objectif, la formalisation de la Théorie de la Pertinence constitue une ambition attrayante.

RÉCAPITULATIF
Nous avons montré dans ce chapitre que la conception d'un système de dialogue homme-machine est avant tout un travail d'intégration pluridisciplinaire. Il faut tenir compte des caractéristiques de la perception visuelle avec une approche psychologique, il faut tenir compte du pouvoir d'expression du langage et du geste avec une approche linguistique ou psycholinguistique. Nos analyses des phénomènes de référence se trouvent ainsi justifiées. Il nous faut également garder une rigueur informatique et des préoccupations méthodologiques pour adapter des théories cognitives dans un but computationnel. Nous avons ainsi montré en quoi la validation d'un système de dialogue est problématique et nous en avons déduit des pistes dans notre travail.

Deuxième partie

Les concepts, le modèle

Chapitre 4

Les contextes et les domaines de référence

Comment la notion de domaine de référence permet-elle de confronter les sources contextuelles que sont la perception visuelle, le langage et le geste? En quoi cette notion répond au problème lié aux multiples interactions entre ces trois modalités? A quel moment des domaines de référence apparaissent-ils? A partir de quelles données et selon quels mécanismes le système de dialogue peut-il les construire?

Le but de ce chapitre est de montrer l'utilité du cadre formel apporté par le modèle des domaines de référence pour intégrer des sources de connaissance hétérogènes. Dans une première étape (§ 4.1) nous présentons ce modèle avec les principes de bases et les mécanismes de construction communs à toutes les sources contextuelles, pour nous intéresser ensuite (§ 4.2) aux particularités de construction des domaines visuels et des domaines langagiers. Les relations entre les éléments d'un domaine seront étudiées dans les chapitres 5 et 6, et l'exploitation des domaines pour la résolution de la référence sera détaillée dans le chapitre 7, qui contiendra ainsi le noyau de notre modèle.

4.1 La notion de domaine de référence

4.1.1 L'ancrage référentiel dans un contexte

Mécanisme de l'ancrage référentiel. En énonçant « *dans l'ensemble des objets visibles, sélectionne les deux triangles rouges* », l'utilisateur fait une référence à deux objets et précise un ensemble contextuel dans lequel le système doit extraire ces deux objets. Cet ensemble est nécessaire à l'interprétation, pour ne pas résoudre la référence dans la base de données complète des objets de l'application. Dans l'exemple précédent, il est explicite et a trait au contexte visuel. Dans l'exemple « *dans l'ensemble des objets que je désigne par mon geste, sélectionne les deux triangles rouges* », il est également explicite et a trait au contexte lié au geste ostensif. Dans « *crée un triangle et un carré, et peins le triangle en rouge* », il est cette fois lié au contexte linguistique : c'est en effet l'expression langagière « *un triangle et un carré* » qui définit cet ensemble ou domaine de référence. Enfin, dans « *parmi les objets à traiter, commence par le triangle rouge* »,

c'est la tâche applicative qui permet de délimiter le domaine de référence.

Tout acte référentiel s'ancre ainsi dans un domaine de référence issu d'une source contextuelle particulière. Le problème réside dans le fait que la délimitation de ce domaine n'est que rarement explicite. « *sélectionne le triangle rouge* » ou « *les deux cercles* » est souvent utilisé sans autre précision. Une expression référentielle repose ainsi dans la plupart des cas sur un domaine de référence implicite. Le système de dialogue doit retrouver ce domaine pour interpréter. Pour ce faire, une méthode consiste à exploiter tout indice dans l'expression verbale ou multimodale, de façon à déterminer de quelle source contextuelle provient le domaine. Des hypothèses de domaines sont ensuite émises, pour la ou les sources retenues. L'objet de ce chapitre est de détailler la construction de ces hypothèses.

Positionnement face aux travaux existants. Les mécanismes à l'œuvre dans ce processus d'interprétation se rapprochent des notions d'« espace focal » (Grosz & Sidner 1986 ; Beun & Cremers 1998) ou de « quantification contextuelle » (Westerståhl 1985). L'idée commune de ces travaux est de vouloir restreindre la résolution référentielle à l'appariement de propriétés à un sous-ensemble contextuel. 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 (auxquels on peut ajouter également la DRT dont nous avons déjà parlé) ramènent le problème de la référence à la mise en relation d'une expression référentielle à un référent accessible, nous pensons que l'accès au 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 et affirme que les mots ne signifient ni ne remplacent des référents, mais spécifient des événements perçus par rapport à un ensemble de possibilités.

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 (d'alternatives). Ce point de vue, mis en avant par exemple par Corblin (1987), a été intégré dans les travaux de génération automatique 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 *et al.* 2001). En principe, ce processus est réitéré pour chaque expression, avec l'aide d'heuristiques pour traiter les expressions linguistiques particulières comme par exemple les expressions d'altérité (« *l'autre triangle* »), les ellipses et les *one-anaphora*[1], pour lesquelles il est nécessaire de revenir sur des hypothèses précédentes.

Nous pensons que l'identification systématique des possibilités exclues à l'aide de domaines de référence 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.

4.1.2 Genèse du modèle des domaines de référence

Des structures regroupant les informations liées aux références. D'après (Reboul *et al.* 1997), la construction de structures pour conserver des informations relatives aux actes de référence

1. 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* »).

effectués trouve son origine principale dans (Karttunen 1976). Dans le but de traiter les coréférences, Karttunen donne le principe consistant à construire des fichiers qui regroupent les renseignements sur les individus mentionnés. Il met l'accent sur le problème consistant à détecter les nouveaux individus. Il conclut ainsi que ce rôle revient à l'article indéfini, celui du défini étant de réactualiser un fichier. Heim (cité par Reboul *et al.* 1997) se base sur ce travail et décrit un peu plus précisément les informations à mettre dans les fichiers. Il distingue ainsi les informations sur l'individu et les relations entre fichiers. Evans (1985) introduit des informations perceptuelles et pose ainsi les bases pour le traitement du démonstratif. Enfin, Récanati (1993) distingue les informations descriptives de nature encyclopédique et les informations non descriptives et perceptuelles, qu'il range dans des dossiers d'objet à trois étages : un niveau inférieur correspondant à la perception, un niveau intermédiaire correspondant aux concepts, et un niveau supérieur correspondant à la connaissance générale sur le monde.

En tenant compte des travaux de Gaiffe (1992) qui montrent comment exploiter formellement ces informations, Reboul (Reboul *et al.* 1997, Reboul & Moeschler 1998b) reprend ces principes pour structurer ce qu'elle appelle une *représentation mentale*, structure attachée à chaque objet de l'application et contenant, dans la mesure du possible, l'ensemble des informations attachées à cet objet et susceptibles d'être activées lors d'une référence. Une représentation mentale correspond ainsi à un point de vue (ou perception) du locuteur sur l'objet, ce point de vue pouvant évoluer au cours du dialogue. Une représentation mentale se présente comme une suite d'informations regroupées sous divers champs :

- une entrée d'identification (pour que l'application puisse la repérer) ;
- une entrée logique (relations que la représentation mentale entretient avec les autres) ;
- une entrée visuelle (informations tirées de la perception de l'objet) ;
- une entrée spatiale (orientation intrinsèque, trace des déplacements) ;
- une entrée lexicale (expressions référentielles potentielles et effectivement utilisées) ;
- une entrée encyclopédique (informations disponibles sur le référent et non incluses dans les champs précédents : informations sémantiques dont la catégorie, informations fonctionnelles indiquant à quoi sert l'objet).

Le positionnement de cette théorie face aux travaux cités précédemment repose sur la volonté de traiter tous les phénomènes de référence et pas seulement de coréférence (critique envers Karttunen et les approches qui lui ont suivi), ainsi que sur l'importance de tenir compte des informations extra-linguistiques (principale critique de Reboul envers la DRT de l'époque). Ce positionnement n'est plus tout à fait valable, compte tenu des évolutions importantes par lesquelles est passée la DRT. Un avantage de la théorie des représentations mentales reste cependant le souci de plausibilité cognitive, ainsi que la possibilité d'extension à d'autres processus cognitifs que l'interprétation des références. Grisvard (2000) propose ainsi une extension pour la formalisation des actes de langage.

Des représentations mentales aux domaines de référence. Dans la théorie des représentations mentales, la résolution des références est considérée comme un processus d'assignation d'une représentation mentale identifiante à une expression référentielle. Dans (Reboul *et al.* 1997), l'existence de sous-ensembles contextuels est abordée avec une modélisation du contexte sous forme de domaines de référence, représentant des ensembles de référents, éventuellement partitionnés selon une propriété distinctive ou « critère de différenciation ». Un domaine de référence est créé lorsque l'expression référentielle réunit plusieurs objets (« *les deux triangles* »), quand son interprétation met en jeu une extraction (« *les autres triangles* »), et pour rendre inaccessibles les reprises pronominales ou démonstratives individuelles après la référence à un ensemble

d'objets (impossibilité de « *mets-le sur la droite* » après « *prend un triangle et un carré* »).

A la suite de (Salmon-Alt 2001b), nous avons prolongé dans (Landragin *et al.* 2002c) 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 ; 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.

Evolution du modèle des domaines de référence. 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 discriminante). Nous gardons les appellations « domaine de référence » et « critère de différenciation ». Nous introduisons la notion de « critère d'ordonnancement » pour fournir un cadre au traitement des expressions référentielles telles que « *le premier* », « *le second* » et « *le dernier* ». Nous parlerons également du type d'un domaine de référence comme le type subsumant l'ensemble des types des objets groupés dans le domaine. Un domaine regroupant des cercles se verra ainsi attribuer le type « *cercle* », alors qu'un domaine regroupant des cercles et des carrés se verra attribuer le type « *forme géométrique* ». Il ne s'agit pas d'une propriété du domaine mais seulement d'une propriété commune à ses éléments. Nous noterons au passage qu'aucune propriété du groupe ne se déduit des propriétés de ses éléments. En effet, bien que l'on ait « tous les éléments du groupe sont noirs donc le groupe est noir », on n'a pas « tous les éléments du groupe sont petits donc le groupe est petit ».

Nous donnons une importance toute particulière à la notion de partition : de notre point de vue, un domaine de référence regroupe plusieurs objets selon un facteur visuel, gestuel, linguistique ou applicatif, ce groupement pouvant être partitionné en fonction d'un critère de différenciation ou d'ordonnancement. Pour un même groupement, plusieurs façons de partitionner sont possibles : un ensemble de formes géométriques peut être partitionné selon la taille ou la couleur des objets, donnant lieu à autant de partitions. Les caractéristiques d'un domaine de référence sont donc un facteur de groupement et une séquence de partitions possibles, les caractéristiques d'une partition étant le critère de différenciation ou d'ordonnancement, ainsi que la liste des contenus, à savoir la liste des objets (en passant éventuellement par des domaines de référence lorsque certains objets sont groupés). Nous détaillerons en § 9.3.1 page 181 des structures de traits pour les domaines et leurs partitions.

Une autre évolution du modèle réside dans la modélisation des contraintes qu'impose l'expression référentielle linguistique sous la forme d'un domaine de référence sous-spécifié. Nous considérons ainsi que les contraintes imposées par la détermination et la sémantique de l'expression référentielle conduisent à la construction d'un domaine de référence sous-spécifié. La confrontation de ce domaine sous-spécifié aux domaines complets issus de la perception visuelle, de l'historique du dialogue ou de la tâche permettra de simplifier le processus de résolution de la référence, et en particulier l'identification de domaines de référence adéquats. Nous reviendrons sur les détails de cette confrontation dans le chapitre 7, pour nous concentrer ici sur la construction de domaines issus de la perception visuelle et de l'historique du dialogue.

Notre apport dans la construction de domaines de référence. Notre principal apport réside dans la modélisation de critères perceptifs pour la construction de domaines visuels. Nous proposons ainsi divers facteurs de groupement (liés aux critères de groupage de la Gestalt) qui viennent s'ajouter à ceux proposés par Salmon-Alt (2001b). Les principes de base que nous avons détaillés

restent communs à tous les types de domaines.

La figure 4.1 illustre à la fois les domaines de référence et les grands principes de notre approche. Un domaine de référence est représenté comme une boîte grisée, comportant éventuellement une ou plusieurs partitions. Les partitions sont ici représentées par des cases blanches et sont chacune associées à un critère de différenciation. Parmi les autres attributs d'un domaine de référence, on retrouve le type et le facteur de groupement. Enfin, tout domaine de référence se caractérise par un index (ou étiquette) qui permet au système de le repérer. Les étiquettes représentées dans la figure sont par exemple DR_1 et DR_2. Au niveau du processus de résolution de la référence, la figure illustre la modélisation des contraintes qu'impose l'expression référentielle « *ce cercle* ». Le domaine sous-spécifié correspondant met en avant un type « *cercle* », ainsi qu'une partition comprenant un élément focalisé, ce qui traduit le fonctionnement du démonstratif pour cette expression. Ce domaine sous-spécifié sera confronté aux domaines fournis par les analyses des expressions référentielles antérieures (par exemple « *le cercle bleu et le carré rouge* », dont un résultat est illustré par DR_1) et de la scène visuelle (dont un résultat est illustré par DR_2).

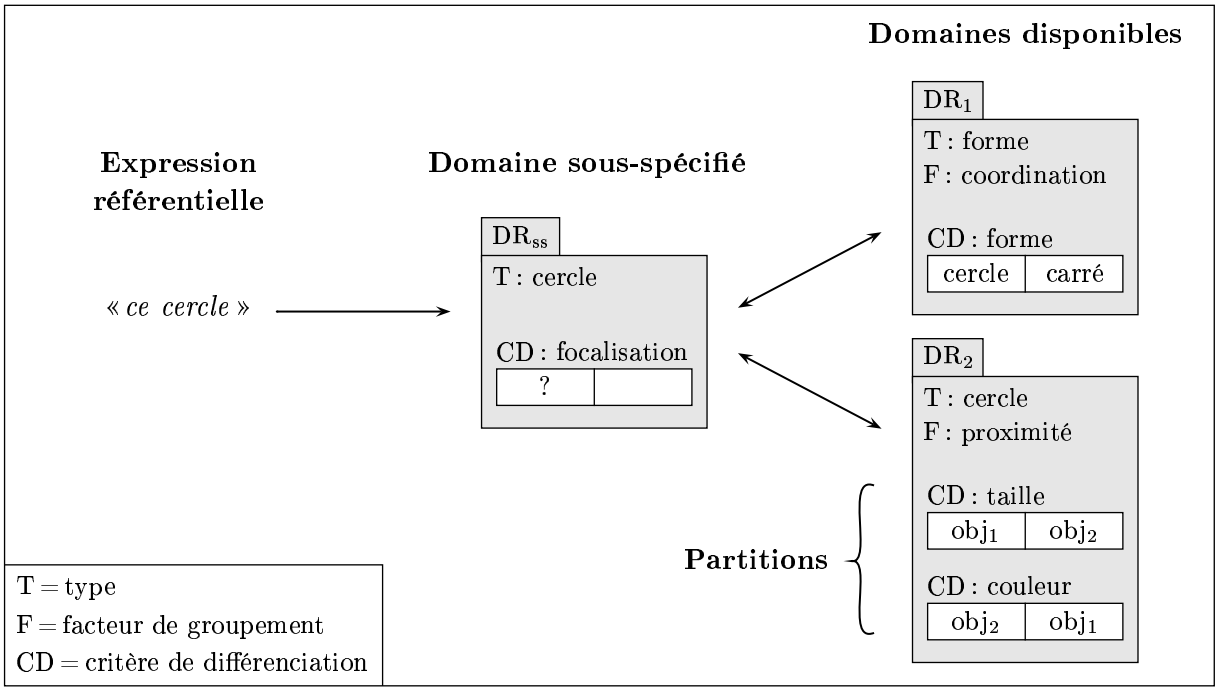


Figure 4.1 – *Domaines de référence.*

4.1.3 Plausibilité cognitive

Statut des domaines de référence. Compte tenu du terme « représentation mentale » qui leur est associé, compte tenu de leur rôle dans l'interprétation de la référence aux objets, des questions peuvent se poser à propos des domaines de référence. Correspondent-ils à des états mentaux représentatifs ? Le modèle qui décrit les mécanismes de leur construction et de leur exploitation a-t-il pour but la modélisation de la mémoire de travail ? celle de l'attention ? Sans aller jusque-là, nous considérons qu'un domaine de référence est un moyen de modéliser une partie du processus d'interprétation de la référence, moyen qui nous semble proche du fonctionnement cognitif sans en postuler une modélisation. Pour justifier ce point de vue, nous allons illustrer cette plausibilité

cognitive à l'aide de différents arguments.

Arguments psychologiques. La plausibilité des domaines de référence réside tout d'abord dans le fait qu'ils n'expriment qu'un minimum de contraintes dans le processus de résolution de la référence. Plutôt que des contraintes, ils apportent au contraire toutes les informations nécessaires compte tenu de la situation. Que ces informations soient conceptuelles ou liées à l'historique du dialogue, elles représentent un point de vue cognitif plausible sur les référents. La plausibilité des domaines réside également dans leur possibilité d'intégrer et de confronter dans une même représentation des informations hétérogènes. En effet, que les indices référentiels viennent d'un geste, du contexte visuel, de la mémoire ou encore de l'intention, ils sont utilisés de la même façon dans le processus de résolution de la référence. Cette intégration de multiples critères cognitifs nous semble caractériser les capacités humaines de compréhension.

Salmon-Alt (2001b) avance quant à elle l'argument de la stabilité des domaines de référence face à l'interprétation d'expressions référentielles ambiguës : la bonne interprétation correspond généralement au domaine activé. Elle fait également un parallèle entre la difficulté dans le modèle à passer d'un domaine clairement délimité à un domaine plus large, avec la difficulté du principe cognitif général consistant à dépasser un cadre instauré. Pour apporter un argument supplémentaire, elle se positionne face aux grammaires cognitives : suivant cette théorie, la compréhension est vue comme une opération cognitive de conceptualisation à partir de structures sémantiques. Ces structures sont fournies par des expériences perceptives, des connaissances conceptuelles ou encore des connaissances encyclopédiques, et sont comparables à nos domaines de référence. Salmon-Alt conclut ainsi que la prise en compte de domaines de référence au cours de l'interprétation constitue une application des principes de la grammaire cognitive à la résolution de la référence.

Arguments linguistiques. Le modèle des domaines de référence a l'ambition de traiter tous les phénomènes référentiels dans un même cadre. Il est ainsi prévu, à la base, pour traiter les références mentionnelles, les expressions ordinales, ou encore les expressions d'altérité, sans les considérer comme des expressions particulières avec leurs heuristiques particulières. Il est également prévu pour traiter les relations rhétoriques, temporelles, événementielles, même si ces aspects ne seront pas abordés dans ce travail. La compréhension est ainsi vue comme un tout, sans phénomènes marginaux. Un deuxième argument réside dans l'exploitation fine qui est faite des petits mots grammaticaux tels que les conjonctions et les déterminants. Notre modèle tient compte de l'importance des indices qu'ils constituent, de leurs différents usages référentiels, ainsi que des conséquences de leur emploi pour les interprétations ultérieures.

Arguments extra-linguistiques. Un autre intérêt des domaines de référence se trouve dans leur adéquation à la prise en compte de divers phénomènes extra-linguistiques, en particulier la perception visuelle et son influence sur la référence aux objets, ainsi que le geste ostensif et son influence sur les phénomènes d'indexicalité. La plausibilité du modèle réside selon nous dans son ouverture à tout type de modalité, modalité d'interaction ou de support dans la communication. Nous pouvons très bien imaginer des domaines de référence liés au regard et aux autres gestes communicatifs. D'une manière générale, tout phénomène attentionnel semble pouvoir être pris en compte. Le modèle des domaines de référence s'avère ainsi compatible avec les structures attentionnelles de (Grosz & Sidner 1986), avec les espaces focaux de (Beun & Cremers 1998) ou de (Kessler *et al.* 1996), avec le pavage de (Romary 1993) pour l'interprétation de « *ici* », avec les cadres de (Schang 1997) pour l'interprétation des références spatiales. Pour ces dernières, qui nécessitent un cadre de référence mettant en rapport la cible avec le site, les domaines de

référence s'avèrent particulièrement pertinents.

Arguments informatiques et mathématiques. Pour garder une certaine plausibilité, un modèle d'interprétation ne doit pas mettre en jeu des algorithmes extrêmement complexes, incluant d'innombrables comparaisons ou retours en arrière. De simple structures de traits nous semblent adéquates pour nos domaines, l'interprétation se faisant selon un mécanisme d'unification adapté.

En prenant le point de vue de la théorie des ensembles, le modèle d'interprétation lié aux domaines de référence consiste à passer d'un traitement en extension à un traitement en intension. Ceci ne se fait que si l'ensemble de départ est justifié sémantiquement. Dans un domaine de référence, ce rôle est joué par le facteur de groupement. Dans une partition, la procédure justifiant la constitution de groupes d'objets correspond à un ou à plusieurs critères de différenciation. De ce point de vue, le modèle des domaines de référence semble permettre une certaine optimisation au niveau des algorithmes, puisqu'il ne requiert pas la gestion de multiples ensembles d'objets, mais celle de procédures simples aboutissant à l'identification d'objets. Considérant la résolution des références aux objets comme intensionnelle et non extensionnelle, cette procédure semble de plus bien correspondre au fonctionnement cognitif humain.

4.2 Construction de domaines de référence

Dans la plupart des approches classiques, les sous-ensembles contextuels testés pour la résolution d'une référence entretiennent entre eux une relation d'inclusion : on commence par tester le sous-ensemble correspondant aux objets focalisés, on continue avec le sous-ensemble des objets visibles à l'instant de l'énonciation, pour terminer avec l'ensemble de tous les objets gérés par l'application. Chaque étape dans ce processus consiste à élargir un ensemble d'objets à un autre qui l'inclut. Nous nous opposons à ce principe en construisant des structures arborescentes de sous-ensembles contextuels, chacun d'eux correspondant à un domaine de référence. Nous donnons ici nos principes de construction de ces structures, en distinguant les principes spécifiques à la perception visuelle de ceux spécifiques au langage.

4.2.1 Domaines visuels

Les facteurs intervenant dans la perception visuelle. La perception visuelle est une activité cognitive faisant intervenir de nombreux facteurs, certains s'avérant plus faciles à modéliser que d'autres. Contrairement à la sensation, la perception est un processus actif. L'intention indissociable de la perception se traduit par un filtrage : percevoir, c'est filtrer (sinon on serait assailli de données). Avec une approche plus descriptive, nous pouvons distinguer plusieurs hypothèses de comportements dans la perception visuelle de l'utilisateur d'un système de dialogue finalisé :

1. L'utilisateur connaît l'objet sur lequel il veut agir. Soit il sait où se trouve cet objet, car son attention est déjà focalisée sur celui-ci, soit il en a représenté en mémoire un prototype visuel, image qu'il va essayer de trouver dans la scène. Dans ce cas, sa perception est dirigée par une catégorie et des propriétés précises telles qu'une taille et une couleur particulières.
2. L'utilisateur sait sur quel type d'objet il va agir. Il en a également représenté un prototype en mémoire, ce prototype étant cette fois moins contraint. Sa perception est alors dirigée de manière beaucoup plus souple.
3. L'utilisateur ne sait pas sur quel type d'objet il va agir. Sa perception n'en est pas totalement libre pour autant : les phénomènes de saillance visuelle et de ligne directrice vont en

effet intervenir, avec pour conséquence de privilégier la perception de tel ou tel objet, de tel ou tel groupe, de tel ou tel type d'objet.

Les trois cas se retrouvent dans le corpus Magnét'Oz. Si le troisième apparaît plutôt lors de la première action face à une nouvelle scène visuelle, les deux premiers se ressentent lors d'un grand nombre d'actes de référence, en particulier le second qui est lié à la tâche applicative de Magnét'Oz : l'utilisateur ayant à ranger des objets selon leur type, nous constatons qu'il traite tous les objets du même type avant de passer à un autre type. Sa perception est par conséquent dirigée par le type courant.

Les conséquences de ces comportements se trouvent dans la priorité donnée à un des critères de la Gestalt (par exemple la similarité des types) lors du groupage. Pour modéliser l'intention perceptive de l'utilisateur, il s'avérerait utile de privilégier le critère qui correspond à cette intention, intention qu'un système de dialogue devrait être capable d'identifier. Compte tenu que même une analyse de la direction du regard de l'utilisateur ne saurait privilégier l'une ou l'autre hypothèse, les seuls indices exploitables pour le système sont les caractéristiques de l'énoncé. Or ces indices apparaissent bien faibles. Peut-on vraiment affirmer que l'utilisation du terme « *triangle* » dans un énoncé du système va focaliser l'intention perceptive de l'utilisateur sur ce type d'objet ? Nous considérons que la part hypothétique est trop forte, et nous nous focalisons donc sur la conception d'un algorithme de groupage qui fonctionne quelque soit l'intention perceptive de l'utilisateur.

Critique des approches classiques du groupage. Sans revenir sur la présentation qui en a été faite dans le chapitre 1, rappelons que la théorie de la 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 ont été proposés pour le groupage d'objets, les plus importants étant la proximité, la similarité et la continuité. Chacun de ces critères a fait l'objet de nombreux travaux de formalisation. Nous en avons évoqué quelques-uns page 33. Nous voulons maintenant en étudier deux de manière plus approfondie, celui de Briffault (1992) et celui de Thórisson (1994) qui nous semblent s'accorder parfaitement au domaine du dialogue homme-machine et à nos préoccupations pluridisciplinaires.

Briffault (1992) présente une méthode de groupement selon 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 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. De plus, la hiérarchisation obtenue nous semble artificielle, en particulier l'extraction des deux objets les plus proches dans chaque groupe. Un avantage de cette méthode est par contre le codage non systématique des relations entre objets, mais en 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. Ce principe suit le résultat de l'expérimentation et de la modélisation de McNamara (1986). La représentation partiellement hiérarchique obtenue a pour avantage l'optimisation des performances de certaines opérations spatiales, ce qui contribue à une meilleure plausibilité cognitive. La caractéristique principale de l'approche de Briffault 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 locutions prépositionnelles (mais, dans le cas des expressions spatiales, seulement la référence directe ou démonstrative liée au site), et d'autre part a pour ambition l'intégration de plusieurs critères perceptifs dans une même structuration (plusieurs critères de différenciation pouvant être attribués à un même domaine de référence via la notion de partition).

De son côté, la formalisation de Thórisson (1994) présente les avantages de combiner deux critères importants, la proximité et la similarité, et de donner pour résultat une liste ordonnée de groupes, le critère d'ordonnancement étant lié à la notion de « bonne forme » de la Gestalt. Deux inconvénients nous paraissent cependant importants compte tenu de notre approche : la perte de l'identité du facteur de groupements et l'absence de structuration arborescente des groupes, plus précisément l'absence de liens entre les groupes proposés.

Notre approche du groupage. Comme nous l'avons vu dans le chapitre 1, la plupart des travaux de formalisation de la perception visuelle se basent sur la théorie de la Gestalt, mais ne prennent en compte qu'une partie des critères qu'elle propose. Pourtant, les trois principaux critères de groupement se complètent, dépendent les uns des autres et fonctionnent simultanément. Nous voulons explorer ici les problèmes posés par leur intégration formelle, l'objectif étant de partitionner les objets de la scène visuelle en groupes. Pour permettre d'aller plus loin dans l'intégration, nous tiendrons compte dans notre proposition de la future intégration avec le langage, c'est-à-dire la possible intervention dans le groupement de critères linguistiques tels qu'une catégorie des objets (lorsque l'expression référentielle en mentionne une) ou une zone attentionnelle de départ (lorsque l'expression référentielle contient un terme tel que « *à gauche* » ou lorsque l'interaction se situe dans un espace attentionnel clairement identifié). Les autres principes qui guideront notre modélisation sont la structuration hiérarchique des groupes et la possibilité de plusieurs partitions dans un même groupe et pour un même critère de la Gestalt. Nous considérerons également que certains critères savent mieux que d'autres exploiter différents niveaux de granularité, en particulier le plus important d'entre eux : la proximité.

Le rôle de la profondeur dans le groupage. Une remarque préliminaire au groupage a trait au débat entre 2D et 3D : dans le cas d'une scène en 3D telle qu'on peut en trouver dans un environnement virtuel, faut-il grouper les objets en considérant leurs coordonnées dans la scène, ou en considérant les coordonnées de leur projection sur l'écran ? L'écran est en effet le principal dispositif de visualisation et transmet une image plane qui peut être considérée comme la base de la perception visuelle. Cette image n'est cependant pas dépourvue de relief. N'importe quelle image peut très bien donner une sensation de profondeur, et pas seulement les images de type trompe-l'œil qui renforcent l'impression de 3D par leur perspective, les ombres portées ou le contraste volontaire entre le premier plan et l'arrière-plan. Le regard parcourt l'image et, par le truchement du cerveau, nous donne une vue d'ensemble de la scène, cette vue d'ensemble étant représentée cognitivement en 3D. Deux objets très proches sur l'image projectée mais en fait très éloignés dans la scène réelle (l'un au premier plan et l'autre dans l'arrière-plan) ne seront ainsi pas considérés comme proches dans l'esprit de l'observateur.

Pour cette raison, nous choisissons de procéder au groupage avec les coordonnées 3D des objets. Des considérations sur l'intervention des lois d'association de la Gestalt dans le processus de perception viennent appuyer notre choix. Si Wertheimer a énoncé son principe de proximité

en se référant probablement à la proximité rétinienne, Rock & Brosgole (1964) montrent que les lois d'association interviennent plus en aval dans le traitement de l'information visuelle, après l'effet créé par les conditions de profondeur et de luminosité. Ils s'appuient pour cela sur une expérimentation et concluent en remplaçant le critère de proximité de la Gestalt par la notion 3D de « proximité phénoménale ».

Le groupage en 3D a également des avantages algorithmiques, ces avantages ayant été analysés lors de l'implantation présentée dans (Landragin 1998). En effet, si l'on procède au groupage avec les coordonnées de projection 2D, on est obligé de faire les calculs de projection, ce qui représente un algorithme assez coûteux. Comme ces calculs sont cablés en *hardware* dans les ordinateurs graphiques et sont ainsi inaccessibles, leur reprogrammation s'avère nécessaire. Elle consiste en particulier en la gestion d'un *Z-buffer* (stock de la profondeur en chaque point de l'image), pour que les recouvrements d'objets soient pris en compte. Le processus prend du temps. Grouper les objets avec leurs coordonnées 3D est beaucoup plus direct.

Avec ce problème de recouvrement, de nouvelles questions sont posées par la 3D. Ainsi, doit-on prendre en compte les objets cachés ? Nous considérons que les objets cachés peuvent avoir été perçus précédemment par l'utilisateur, qu'ils sont par conséquent présents dans sa mémoire de travail, et qu'il faut donc les prendre en compte. L'interaction dans un environnement virtuel se caractérise en effet par des déplacements et par des changements d'angle de vue. Les objets cachés à un moment donné peuvent ainsi être parfaitement visibles quelques instants plus tôt ou plus tard. La question des limites de l'espace visuel se pose de la même façon : doit-on prendre en compte les objets en bordure de la scène visuelle ? Est-il même envisageable de prendre en compte les objets non visibles mais très proches de cette bordure ? Notre réponse à ces deux questions est affirmative : un objet qui apparaît seul en bordure de champ peut en fait être très proche voire accolé à des objets non visibles. Dans ce cas, l'algorithme de groupage ne doit pas le considérer comme isolé, les conséquences pouvant être son identification comme objet saillant si les autres objets visibles sont tous groupés.

Le débat entre groupage en 2D et groupage en 3D a ainsi des conséquences sur le groupage par proximité et sur le groupage par continuité, ces deux facteurs tenant compte des coordonnées des objets. Il a également des conséquences sur le groupage par similarité. En effet, tenir compte de l'image projetée revient à différencier deux objets identiques vus sous des angles différents. Nous considérons au contraire que l'utilisateur reconnaît un type d'objet, quel que soit l'angle sous lequel il le perçoit. L'orientation n'est donc pas un critère de similarité. Autrement dit, comme les groupages par proximité et par continuité, le groupage par similarité ne tient pas compte de la projection sur l'écran.

Le groupage selon la proximité. Nous nous plaçons dans le cadre d'une interaction dans un environnement virtuel, telle qu'étudiée dans le projet COVEN [1] pour une tâche d'aménagement d'un intérieur. Le groupage se fait avec les coordonnées 3D des objets de l'application, à savoir des chaises et des tables. Comme nous venons de le voir, nous tenons compte des objets cachés et des objets en limite du champ de vision. Nous tenons compte aussi du point de vue de l'utilisateur, avec deux critères : la distance entre les objets de la scène et la représentation de l'utilisateur (l'avatar), et la distance des objets par rapport à l'axe de visée de cet avatar : plus un objet s'éloigne du centre de visée, moins il importe dans le processus de perception. Pour l'adaptation

1. Les résultats donnés dans ce paragraphe et dans la suite de la section ont été obtenus suite à une implantation. Nous avons réalisé cette implantation pour un démonstrateur dans le cadre du projet COVEN évoqué dans l'avant-propos. Certains choix sur lesquels nous passons rapidement ici sont détaillés largement dans (Landragin 1998).

de l'algorithme à une interaction en 2D comme la manipulation de formes géométriques sur écran tactile, les seuls paramètres sont pour l'instant les coordonnées 2D des objets. Que les coordonnées soient 2D ou 3D, elles correspondent au centre de gravité de l'objet et s'avèrent insuffisantes. Nous les associons ainsi systématiquement à une indication quant au gabarit de l'objet. Ce gabarit est classiquement pris en compte par une boule (en 3D) ou un disque (en 2D) englobant l'objet. Leur rayon doit être enregistré dans la base de données des objets, avec les coordonnées du centre de gravité.

Belaïd & Belaïd (1992) font le point sur les algorithmes de groupage. La classification automatique non supervisée s'avère la seule adaptée à notre approche partant de coordonnées de points pour en déduire automatiquement des groupes. Classifier des points (ou vecteurs) consiste à regrouper ces vecteurs en classes vérifiant la propriété de compacité (les points représentant une classe donnée sont plus proches entre eux que des points de toutes les autres classes) et la propriété de séparabilité (les classes sont bornées et il n'y a pas de recouvrement entre elles). La constitution des classes se fonde sur les distances entre les points. Les calculs de distance pouvant se faire entre des points et des classes déjà constituées, il s'avère nécessaire de définir la distance entre un point et une classe. Plusieurs types de distance sont possibles.

Une distance vérifie les quatre propriétés mathématiques suivantes : la séparabilité (la distance entre deux points distincts est strictement positive), la réflexivité, la symétrie, et l'inégalité triangulaire. Les distances classiques sont les distances de Hamming, euclidienne et du maximum [1]. Nous considérons dans notre cadre que la distance entre deux objets est la distance euclidienne entre les deux points les plus proches de ces objets. La distance entre deux objets qui se touchent est ainsi quasiment nulle mais reste strictement positive, ce qui permet de vérifier que les quatre propriétés sont vérifiées. Pour calculer la distance entre un point et une classe, deux solutions sont possibles : celle du saut minimal qui consiste à prendre le minimum des distances, et celle du diamètre maximal qui consiste à prendre le maximum des distances. Afin de rester dans la logique du minimum, nous retiendrons le saut minimal. Nous pouvons maintenant nous intéresser au groupage, pour lequel deux grandes approches se distinguent :

- Le groupage hiérarchique :
  Cette méthode de classification automatique consiste à effectuer une suite de regroupements en agrégeant à chaque étape les groupes les plus proches. On part d'un état où chaque objet est un groupe, on réunit en un groupe les deux groupes les plus proches et ainsi de suite, jusqu'à n'obtenir qu'un seul groupe contenant tous les objets. Chaque étape donne lieu à une partition des objets en groupes qui est caractérisée par un indice d'agrégation (ou niveau). L'indice d'agrégation d'un groupe correspond ainsi à la distance entre ses deux sous-groupes. Il est à noter que le choix de la distance influe sur les regroupements. L'ensemble des partitions associées à leur niveau forme ce que l'on appelle une hiérarchie indicée et peut se représenter par un arbre ou un dendrogramme.

- Le groupage non hiérarchique :
  Là aussi, il s'agit de déterminer sur l'ensemble à classer une partition représentant au mieux les divers regroupements pouvant exister au sein de cet ensemble. Il existe pour cela deux méthodes : l'arbre de longueur minimale et les nuées dynamiques. La première consiste à extraire un graphe connexe de coût minimal de l'arbre valué dont une arête représente la distance entre deux objets (qui sont les nœuds). La seconde consiste à établir une partition à partir d'éléments suffisamment représentatifs de chaque classe par un procédé iteratif

1. Ici et dans la suite, nous renvoyons à (Belaïd & Belaïd 1992) pour les aspects mathématiques.

qui établit les meilleurs représentants et la meilleure partition à chaque étape. Ces deux méthodes sont surtout intéressantes lorsqu'il y a plusieurs milliers d'objets à classer et nous ne les utiliserons donc pas.

Une phase préparatoire pour le groupage hiérarchique consiste à construire un demi-tableau contenant les distances des objets deux à deux. Ces distances sont calculées de la manière suivante : à la distance euclidienne entre les centres de gravité des deux objets, nous retirons les rayons des sphères englobantes. La sphère englobante d'un objet ayant un volume plus grand que celui de l'objet, on corrige son rayon en le multipliant par un coefficient, afin de rendre plus réaliste la distance entre deux objets. Ce coefficient, de 0.5 dans notre implantation, est utile lorsque les objets sont de grande taille. C'est le cas par exemple lors du calcul de la distance entre deux tables : si celles-ci sont suffisamment proches, la distance corrigée par les rayons des sphères englobantes (donc avec un coefficient égal à 1) peut être négative et doit alors être corrigée. Pour cela, nous la rendons nulle, ce qui correspond bien à des objets proches. La distance obtenue est ensuite pondérée selon la position des objets par rapport au participant : pour chacun des deux objets, on effectue une pondération proportionnellement à l'inverse de la distance de l'objet au participant. C'est ici que se fait la prise en compte du point de vue du participant. Dans notre implantation, nous n'avons pas pris en compte l'écart par rapport à l'axe de visée (cf. Landragin 1998). Le demi-tableau une fois rempli, nous obtenons directement la première partition de la scène visuelle, c'est-à-dire celle où chaque objet forme un groupe.

L'algorithme proprement dit est très simple : à chaque étape, les deux groupes les plus proches sont regroupés, la distance entre les deux déterminant le niveau d'agrégation du groupe, et donc celui de la partition obtenue. Cette nouvelle partition est stockée avec celles obtenues précédemment. Le nouveau demi-tableau des distances prend la place de l'ancien, et on réitère jusqu'à n'obtenir plus qu'un seul groupe. La figure 4.2 illustre cet algorithme pour le groupage de cinq objets avec des distances euclidiennes prises de manière arbitraires.

En ce qui concerne la validation de cet algorithme, nous ne rentrerons pas ici dans les détails des tests réalisés, pour la simple raison qu'un test n'a d'intérêt que s'il permet de confronter les partitions obtenues à une expression référentielle. Nous noterons cependant que le temps de calcul est négligeable dans la mesure où il n'augmente que de très peu le temps pris par la résolution de la référence, et ce quel que soit le nombre d'objets présents dans la scène. Nous noterons également les effets du choix d'un coefficient de proportionnalité dans la pondération rendant compte de la distance entre le participant et les objets : plus on le diminue, plus on tend vers un groupage en 2D, et plus on l'augmente, plus des objets éparpillés dans le lointain auront tendance à être groupés avant les objets au premier plan. Ce coefficient reste à régler empiriquement, en fonction du type de scènes visuelles mises en jeu dans la tâche applicative.

Pour que cet algorithme soit ouvert au traitement du langage, il faut que les mots de l'expression référentielle n'interviennent pas dans le processus de groupage, mais après, lors de l'exploitation de son résultat. Pour cela, il faut que le groupage s'applique à tous les objets visibles (plus précisément à tous les objets dont la sphère englobante est incluse ou intersecte la pyramide de visée du participant), et non aux seuls objets vérifiant la catégorie et les propriétés présentes dans l'expression référentielle. Ce choix s'appuie sur le fait qu'un groupe peut être composé d'objets de natures différentes : si on enlève les objets dont la nature n'est pas celle recherchée, la cohésion du groupe peut être perdue. C'est le cas par exemple de quatre chaises entourant une table : ce n'est pas parce que la référence a trait à une chaise qu'il faut oublier la table (sans laquelle les chaises ne se justifient pas).

1. Première partition correspondant à autant de groupes que d'objets

objet	A	B	C	D	E
groupe	G_1	G_2	G_3	G_4	G_5
agrégation du groupe	–	–	–	–	–

Demi-tableau des distances entre objets (la distance minimale apparaît en gras)

d	G_1	G_2	G_3	G_4	G_5
G_1	–	10	8	10	13
G_2		–	34	2	41
G_3			–	26	**1**
G_4				–	29
G_5					–

2. Deuxième partition correspondant au groupage de G_3 et G_5 en G_6 :

objet	A	B	C	D	E
groupe	G_1	G_2	G_6	G_4	G_6
agrégation du groupe	–	–	1	–	1

d	G_1	G_2	G_4	G_6
G_1	–	10	10	8
G_2		–	**2**	34
G_4			–	26
G_6				–

3. Troisième partition correspondant au groupage de G_2 et G_4 en G_7 :

objet	A	B	C	D	E
groupe	G_1	G_7	G_6	G_7	G_6
agrégation du groupe	–	2	1	2	1

d	G_1	G_6	G_7
G_1	–	**8**	10
G_6		–	26
G_7			–

4. Quatrième partition correspondant au groupage de G_1 et G_6 en G_8 :

objet	A	B	C	D	E
groupe	G_8	G_7	G_8	G_7	G_8
agrégation du groupe	8	2	8	2	8

d	G_7	G_8
G_7	–	**10**
G_8		–

Figure 4.2 – *Algorithme de classification automatique pour le groupage.*

Le groupage selon la similarité. Si les calculs de proximité faisaient intervenir les coordonnées des objets, les calculs de similarité font, eux, intervenir leur catégorie et leurs propriétés physiques, telles qu'elles apparaissent dans la base de données gérée par l'application. Comme la proximité phénoménale, la similarité intervient assez tard dans le processus de perception visuelle et s'applique aux représentations mentales des objets, et non aux caractéristiques liées à leur projection en 2D. L'orientation ou la taille projetée des objets n'interviennent donc pas dans le groupage, et celui-ci se fait à partir des seules propriétés inhérentes aux objets.

Comme le groupage par proximité, le groupage par similarité permet d'obtenir une liste ordonnée de partitions de la scène. Ces partitions peuvent se représenter sous la forme d'un dendrogramme, et se caractérisent par des facteurs de groupement. Les principaux facteurs sont ici la forme, la couleur et la taille. D'autres facteurs tels que la texture ou la réflectance peuvent être négligés, pour la simple raison qu'il existe peu de mots pour les décrire et donc peu de situations où ils interviennent comme critères de différenciation.

Reste la question de l'importance relative des trois propriétés citées. Nous donnerons la primauté à la forme car elle correspond à la catégorisation et car elle intervient en premier dans le processus de perception. Il est en effet difficile d'appréhender les propriétés d'un objet tant qu'on n'a pas reconnu son type. Vient ensuite la couleur, puis la taille. En effet, comme nous

l'avons évoqué dans le chapitre 3 et comme des expérimentations avec des mesures de temps l'ont prouvé (Baticle 1985), la couleur est traitée plus rapidement que la taille et intervient avant elle dans le processus de perception. Nous retrouvons également cette priorité dans certains travaux en génération automatique. Reiter & Dale (1997) donnent ainsi un exemple de référence à un chien parmi deux, dont l'un est grand et blanc, et l'autre est petit et noir. Pour désigner ce dernier, ils génèrent plus facilement « *le chien noir* » que « *le petit chien* ».

Les regroupements d'objets ayant la même couleur se feront donc avant les regroupements d'objets ayant la même taille. Contrairement au groupage par proximité pour lesquels tous les groupes identifiés ont le même facteur de groupement (la proximité), les groupes obtenus ici peuvent se voir attribuer un facteur de groupement générique de similarité, ou un facteur plus précis de forme, de couleur ou de taille.

Le groupage selon la bonne continuité. Il s'agit ici de regrouper des objets présentant une certaine régularité dans leur disposition, que cette disposition soit linéaire, circulaire, ou suive n'importe quelle courbe, du moment que les écarts entre les objets restent réguliers. Deux méthodes nous semblent intéressantes : d'une part la régression linéaire pour détecter les dispositions linéaires, d'autre part une méthode particulière de groupage selon la proximité pour détecter les régularités dans les écarts entre objets. Elles aboutissent toutes les deux à un dendrogramme comparable à ceux des groupages précédents. Elles peuvent de plus être réunies en une méthode hybride.

La première méthode consiste à considérer toutes les droites possibles, c'est-à-dire, dans le pire des cas, tous les couples de points possibles, et à appliquer une méthode de classification automatique réunissant une droite et le point qui en est le plus proche pour obtenir une nouvelle droite. Le résultat de cette classification est confronté à celui du groupage selon la proximité, dans le but d'identifier les ensembles de points qui sont à la fois proches les uns des autres et qui forment une droite. Ne sont gardés que les ensembles d'au moins trois points, car grouper seulement deux objets selon la bonne continuité n'a aucun sens.

La deuxième méthode consiste à reprendre le groupage selon la proximité en modifiant une étape : lorsque l'on réunit deux objets (ou groupes), on ne spécifie pas les coordonnées du nouvel objet comme étant celle du barycentre, et on ne calcule plus la distance entre ce nouveau groupe et un autre objet de la scène à partir de ce barycentre. Au contraire, pour calculer cette distance, on retient la plus petite des distances entre chacun des éléments du groupe et l'autre objet de la scène. En suivant cette méthode, des ensembles de points équidistants sont groupés au même niveau. Lorsque beaucoup d'objets sont groupés à un même niveau, on en déduit qu'on a détecté un groupe présentant une régularité pertinente dans la disposition de ses éléments.

Le groupage selon les trois critères. 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. La figure 4.3 donne un exemple de ces constructions pour une scène comportant huit objets. Chaque groupe identifié pouvant jouer le rôle d'un domaine de référence, nous obtenons à ce stade des domaines de référence structurés dans une forêt de dendrogrammes.

Dans chacun de ces dendrogrammes, plusieurs niveaux de granularité apparaissent. Pour la proximité, on observe ainsi un premier niveau (`prox niv 1`) qui correspond à un carré isolé et à deux groupes hétérogènes (au niveau des catégories) ; un deuxième niveau (`prox niv 2`) qui correspond à deux groupes ; et un troisième niveau (`prox niv 3`) qui correspond à un seul groupe global. Dans le dendrogramme relatif à la similarité, le niveau `sim niv 1` regroupe les objets exactement identiques en termes de forme et de couleur (d'où deux groupes et trois objets

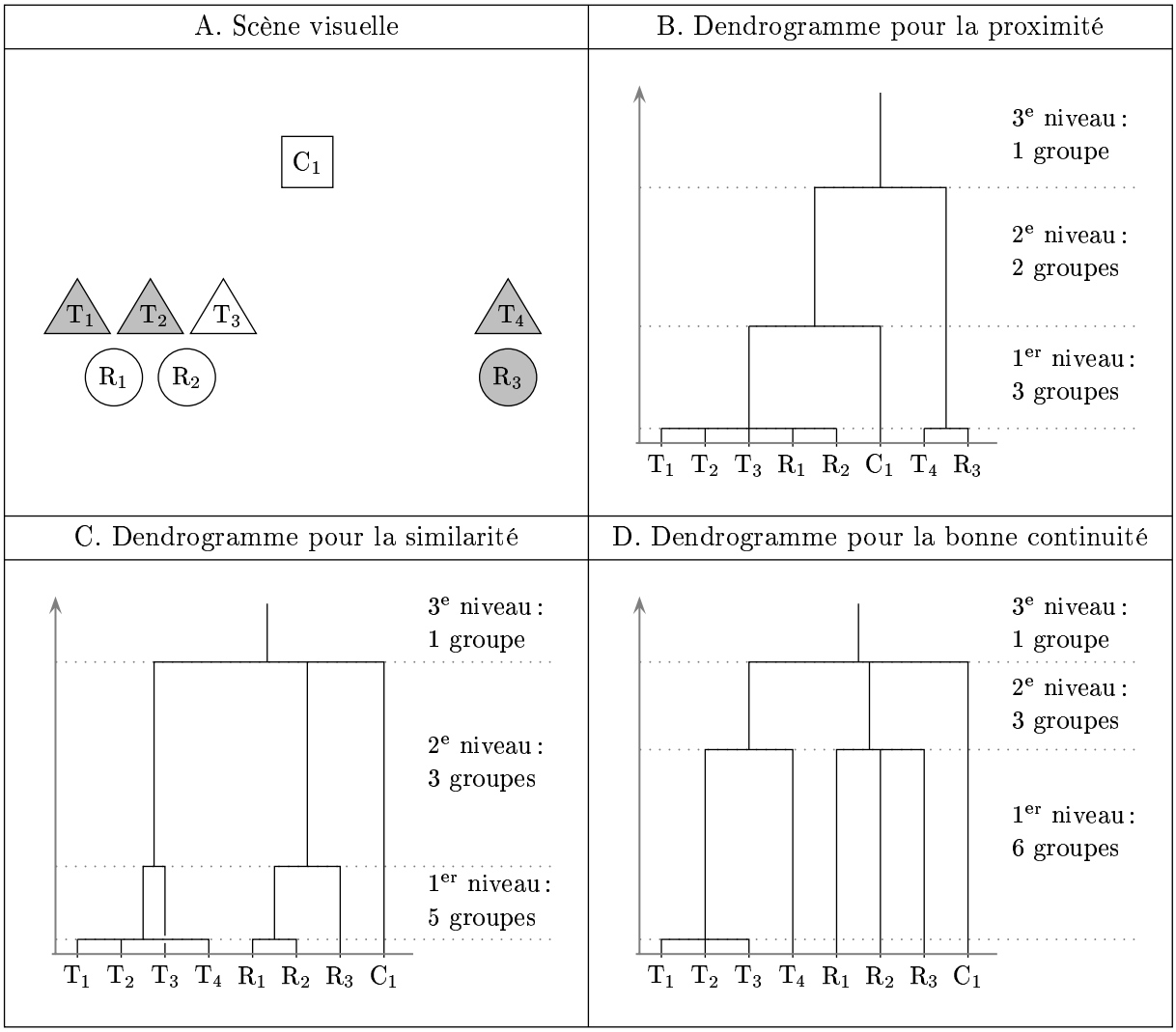


Figure 4.3 – *Exemple de scène avec les dendrogrammes de proximité et de similarité.*

isolés) ; `sim niv 2` regroupe les formes (d'où le groupe des triangles, le groupe des ronds et le carré isolé) ; et `sim niv 3` regroupe toutes les formes géométriques. Dans le dendrogramme relatif à la bonne continuité, seul le premier regroupement (un groupe de trois triangles parfaitement alignés) est pertinent, les autres intervenant beaucoup plus tard (`cont niv 2` correspond à un groupe de quatre triangles alignés mais non régulièrement répartis, et à un groupe de trois ronds ayant les mêmes caractéristiques). On constate dans ce dendrogramme que les groupes se font à partir de trois objets.

L'intégration des trois dendrogrammes obtenus n'est pas un problème simple : comment comparer des niveaux de similarité avec des niveaux de proximité ou de continuité ? Quel est le statut de groupes dont le facteur de groupement correspond en partie à la proximité et en partie à la similarité ? Pour répondre à ces questions, notre approche consiste à ne pas intégrer les dendrogrammes à cette étape, mais à utiliser le langage, le geste ou la saillance visuelle comme principe intégrateur. Pour cette raison, nous répondrons à nos deux questions après avoir étudié la notion de saillance. Nous renvoyons ainsi au chapitre 7 (plus précisément en § 7.2.1 à partir

de la page 135) qui présente une modélisation de la focalisation axée sur l'intégration des trois dendrogrammes.

4.2.2 Domaines linguistiques

Facteurs de construction de domaines. La question porte sur la nature des facteurs susceptibles de déclencher des opérations de groupement. Nous distinguerons des facteurs linguistiques, essentiellement syntaxiques, et des facteurs discursifs jouant au niveau de la structure rhétorique. Des indices peuvent être explicités dans l'expression référentielle ou dans l'énoncé verbal complet, comme dans « *déplace le groupe constitué par la table et ses trois chaises* ». Un domaine de référence linguistique groupant les quatre meubles est ainsi construit. Dans une telle expression, l'adjectif numéral « *trois* » suffit à grouper les chaises en question, et la conjonction de coordination « *la table* ***et*** *ses trois chaises* » suffit à ajouter la table à ce groupe. L'énoncé simplifié « *déplace la table et ses trois chaises* » conduira ainsi à la construction du même domaine. C'est à de tels indices que nous nous intéressons.

L'énumération et la coordination. Les facteurs lexicaux et syntaxiques semblent être l'utilisation d'un pluriel, d'un adjectif numéral, d'une coordination par une conjonction ou encore d'une énumération. Ces facteurs semblent également être sémantiques dans la mesure où ils se combinent pour spécifier ce que l'on peut appeler un référent unique. Comme le relève Salmon-Alt (2001b), Kleiber (1986) va dans ce sens : « 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 par exemple, prévoit une opération de sommation permettant de regrouper les référents introduits par un groupe nominal coordonné. Dans l'exemple « *Prends un triangle rouge et un triangle vert. Mets le triangle rouge sur la droite. Supprime l'autre* », cela conduit 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 interprété le 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 « *mets-le sur la droite* » est autorisé, alors que, comme le constate Kleiber (1986), il semble impossible ou au moins très difficile.

La participation à un même événement. 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 comme celle de (Sidner 1979), la DRT ou la Théorie du Centrage (Grosz *et al.* 1995), prévoient la possibilité de regrouper les participants d'une même éventualité. La Théorie du Centrage en fait même son principal critère de structuration contextuelle. Parmi les travaux moins formalisés, on pourra mentionner la grammaire cognitive de Langacker, qui modélise les processus de compréhension avec des schémas abstraits représentant des éventualités (cf. Salmon-Alt 2001a).

Ainsi, dans l'énoncé « *ajoute une table dans le bureau* », un domaine de référence linguistique regroupant une table et un bureau est créé, selon le facteur de groupement que constitue la participation à un même événement. Dans l'exemple « *un triangle recouvre un carré* », le domaine créé regroupe les deux formes géométriques citées, et, dans la partition qui correspond à ce domaine, le critère de différenciation qu'est le rôle thématique permet de distinguer le triangle du carré.

Choix pour la création et la gestion de domaines linguistiques. Nous n'avons pas considéré ici les phénomènes de saillance linguistique qui peuvent fournir des critères de différenciation pertinents. Nous n'avons pas non plus évoqué les phénomènes anaphoriques. Sur ce point, nous retiendrons pour l'instant qu'une anaphore revient généralement à la focalisation d'un élément dans un domaine linguistique déjà construit. Cette focalisation constitue un critère de différenciation qui, soit prend la place de l'ancien, soit justifie la création d'une nouvelle partition. Elle se rapproche en cela de la focalisation opérée par la saillance linguistique. Nous y reviendrons dans le chapitre suivant consacré à la saillance.

Nous concevons que les critères linguistiques et rhétoriques présentés sont complémentaires. A la suite de (Salmon-Alt 2001b), nous considérons que les meilleurs systèmes de résolution de la référence seront ceux qui tenteront de les combiner. En revanche, dans l'état actuel des connaissances, les seuls groupements pouvant être calculés et surtout exploités de façon réaliste sont ceux fondés sur les critères syntaxiques de nombre, de coordination et d'énumération. Nous nous restreindrons donc à ces facteurs de groupement. Il est à noter que, de même que la construction des dendrogrammes pour la perception visuelle ne contraint en rien leur intégration ultérieure avec le geste ou le langage, les domaines linguistiques ne sont qu'une base minimale pour une intégration ultérieure avec les autres modalités.

Nous avons détaillé dans ce chapitre les grands principes du modèle des domaines de référence, avec une attention toute particulière pour la construction de domaines visuels et linguistiques. Cette construction reste pour le moment cloisonnée : nous n'avons pas pris en compte les paramètres visuels dans la construction de domaines linguistiques et inversement. Nous avons par contre montré que l'intégration, et donc la modélisation des interactions entre les modalités, pouvait se faire à partir des structures construites ici. Cette intégration constitue le noyau de notre modèle et nous y reviendrons dans le chapitre 7. Comme elle fait intervenir des notions que nous n'avons pas encore caractérisées, il s'avère nécessaire de continuer notre présentation avec ces caractérisations, en commençant par la notion de saillance.

RÉCAPITULATIF
Nous avons montré dans ce chapitre que tout acte de référence fait intervenir l'activation d'un domaine de référence, ou sous-ensemble contextuel dans lequel se limite l'interprétation, c'est-à-dire dans lequel s'appliquent les principes de fonctionnement des déterminants et les filtres basés sur les composants de l'expression référentielle verbale. Si cette idée est déjà présente dans la thèse de Salmon-Alt, nous montrons ici que ces domaines de référence proviennent de diverses sources contextuelles, et que la richesse de l'interprétation réside dans l'exploitation de cette hétérogénéité. Nous proposons dans ce but la formalisation des caractéristiques référentielles des sources contextuelles dans un cadre unifié, fondé sur des structurations de domaines et sur la formulation de contraintes en termes de partitions dans un domaine. Dans notre formalisation de la perception visuelle, à laquelle nous nous intéressons particulièrement, nous intégrons trois critères de la Gestalt, la proximité, la similarité et la continuité. En utilisant un algorithme de classification automatique, nous structurons les domaines visuels dans des dendrogrammes.

Chapitre 5

La saillance, un point d'entrée dans un domaine

Quels sont les phénomènes recouverts par le terme « saillance » ? Quels sont les principaux paramètres qui caractérisent ces phénomènes ? Quels sont les critères calculables qui permettent de quantifier la saillance ? Comment la saillance peut-elle intervenir dans le modèle des domaines de référence ?

Si la saillance est souvent évoquée dans divers travaux relatifs à la perception visuelle ou au langage, les phénomènes désignés par ce terme sont souvent très différents. La saillance apparaît comme un concept fourre-tout regroupant tout ce qui est flou mais déterminant dans le processus de perception ou de compréhension. Nous essayons dans ce chapitre de regrouper les principaux critères évoqués dans la littérature, en faisant un parallèle entre la nature des critères visuels et celle des critères linguistiques. Nous aboutissons ainsi à une liste de critères calculables qui nous permet de proposer une méthode de quantification numérique de la saillance visuelle et une caractérisation formelle de la saillance linguistique. Dans ce chapitre, nous faisons tout d'abord le point sur la méthodologie de caractérisation de la saillance et sur le rôle de celle-ci dans l'interprétation de la référence (§ 5.1). Nous détaillons ensuite les rapports entre saillance, geste ostensif et domaine de référence (§ 5.2). Nous caractérisons enfin la saillance visuelle et la saillance linguistique (§ 5.3) pour permettre leur prise en compte dans notre modèle.

5.1 Saillance et référence

5.1.1 Définition générale de la saillance

La saillance comme mise en relief. Une entité de discours saillante se distingue des autres entités du discours. Un objet saillant est un objet qui se distingue des autres objets, qui se trouve mis en valeur. De la même façon, une zone visuelle saillante se distingue particulièrement de l'ensemble constitué par la scène visuelle. La notion de saillance dénote le degré de cette mise en valeur. Si le nom commun « saillance » n'existe pas dans le vocabulaire français, son utilisation et son sens se déduisent facilement de l'adjectif « saillant ». Un terme proche, et que nous utiliserons comme synonyme, est celui de « prégnance ».

Une première définition de ce qui est saillant est ce qui arrive en premier à l'esprit. Ce point de vue correspond souvent à considérer comme saillant ce qui est naturel, simple, clair. Selon Stevenson (2002), les premiers travaux dans les années 1970 se sont focalisés sur la saillance visuelle avec cette idée de simplicité naturelle. Elle cite Osgood & Bock (1977) qui identifient trois types de saillance linguistique applicables à la saillance en général :

1. La simplicité naturelle (*naturalness*) : l'ordre naturel des constituants dans une phrase reflète souvent l'ordre naturel des événements (agent–action–patient). Du côté de la vision, ce point de vue rejoint celui de la Gestalt à propos de bonne forme.
2. La clarté (*vividness*) : elle correspond à la saillance inhérente liée aux traits sémantiques de l'entité du discours, ou, du côté de la vision, à la saillance liée aux propriétés des objets.
3. La saillance liée à l'intérêt du locuteur (*motivation-of-speaker*) : il s'agit de la saillance due à l'objectif, à l'intention. Elle prend ici un caractère subjectif et s'applique aussi bien au langage qu'à la vision.

Une deuxième définition d'un objet saillant est un objet qui capte l'attention. Que ce soit par des propriétés inhérentes à cet objet, par une simplicité naturelle qui lui est propre ou par une mise en valeur qui lui est extérieure, un objet qui capte l'attention est saillant.

Une troisième définition pour la saillance est tout ce qui se trouve à l'origine d'un énoncé, d'un message, d'un acte de référence de la part du locuteur. La saillance ne fait pas partie du message mais tout le message se base sur elle, s'explique par elle, se structure en fonction d'elle. Dans un même ordre d'idée mais cette fois du point de vue de l'interlocuteur, la saillance peut être définie comme le point de départ pertinent pour l'interprétation d'un message, comme l'indice sur lequel repose tout le processus de compréhension.

D'une manière générale, la saillance ne caractérise pas l'attention ou la mémoire : elle entraîne une certaine attention, elle implique une certaine mémorisation, un rôle privilégié dans la mémoire de travail. La saillance peut prendre un rôle perturbateur : on peut se focaliser intentionnellement sur des objets précis tout en étant perturbé par un objet fortement saillant. La saillance ne caractérise pas la structure linguistique de l'énoncé ou de l'historique du dialogue : elle structure ceux-ci en mettant en avant certains de leurs éléments.

Différents points de vue sur l'émergence de saillance. Les propriétés structurelles de l'énoncé et les propriétés physiques de l'ensemble des objets de la scène permettent de distinguer dans cet ensemble certains éléments, et de les considérer comme saillants. La saillance n'est ainsi jamais propre à une seule unité mais émerge toujours dans un contexte regroupant des unités. Dans ce sens, nous considérons que la saillance s'applique aussi bien au locuteur qu'à l'interlocuteur, avec l'idée que la communication est coopérative et que ce qui est saillant pour l'un doit l'être pour l'autre. Nous ne ferons donc pas de distinction entre saillance pour le locuteur et saillance pour l'interlocuteur.

Parmi les distinctions possibles, nous tiendrons compte de la distinction entre saillance explicite et saillance implicite : typiquement, le geste ostensif rend un demonstratum saillant et constitue le critère explicite de saillance. Du côté du langage, un énoncé tel que « *considère le triangle rouge* » rend implicitement le référent saillant pour la suite du dialogue (« *peins-le en bleu* », par exemple). Le pronom est un indice fort de recours à la saillance pour l'interprétation. Ce recours à la saillance, qui a lieu également pour l'interprétation de « *le N* » dans le cas où la scène visuelle comprend plusieurs objets de type N, est la plupart du temps implicite.

La distinction entre saillance immédiate et saillance préalable rejoint les considérations précédentes, dans le sens où un geste ou une expression qui rend un objet saillant relève de la saillance

immédiate, alors que l'interprétation d'un pronom requiert une saillance préalable. A propos de saillance immédiate, nous noterons qu'un objet peut être rendu saillant non seulement par les modalités d'interaction, mais aussi par un comportement ou une intention manifeste, comme celle de l'enfant qui tend la main vers la cage d'un lion et qui se voit prévenir : « *attention,* ***il*** *risque de te mordre* » (exemple de Isard 1975).

Une autre distinction abordée dans la littérature est celle entre saillance globale et saillance locale : d'après Alquier (1998) qui traite de saillance dans la perception visuelle, « On peut définir deux sortes de saillance relatives à la perception. Un élément visuel peut présenter une saillance qui lui est propre, comme par exemple, un point blanc parmi un ensemble de points gris. Cette saillance locale est à distinguer de la saillance globale d'un groupe d'éléments visuels, qui traduit la structure d'ensemble de ce groupe. Cette saillance structurelle est voisine de l'idée de "bonne forme" de la Gestalt. » (p. 101). Nous n'exploiterons pas cette distinction car nous utilisons la saillance dans le processus d'interprétation de la référence. Notre approche consiste ainsi à proposer l'hypothèse d'un objet saillant face à une expression référentielle ambiguë. L'identification de cet objet se fait par comparaison de la saillance de chaque objet. Il s'agit donc nécessairement d'un calcul de saillance locale.

Nous considérerons enfin la distinction entre saillance directe et saillance indirecte : contrairement à la saillance directe, la saillance indirecte trouve son origine non pas dans des caractéristiques de l'objet même, mais dans des caractéristiques d'un objet qui lui est proche. La saillance indirecte correspond ainsi au transfert de saillance d'un objet à un autre. Visuellement, la chaise placée devant une table très saillante en devient saillante. Linguistiquement, une entité du discours en lien grammatical direct avec l'entité focalisée en devient saillante.

Différents facteurs pour l'émergence de saillance. Maintenant que nous avons exploré quelques phénomènes de saillance, nous pouvons nous interroger sur la nature des facteurs qui donnent naissance à de tels phénomènes. Sans entrer pour l'instant dans les facteurs propres à la saillance visuelle ou linguistique, nous nous focalisons sur des critères généraux, qui nous paraissent applicables à la saillance en dehors même de considérations liées au dialogue homme-machine ou à la communication multimodale.

Est ainsi saillant ce qui est original, ce qui est nouvellement introduit dans la situation. Ce facteur rejoint le concept de non-familiarité : selon Loftus & Mackworth (1978), des objets non familiers dans un environnement donné tendent à être fixés plus longtemps, et en deviennent saillants. Ce facteur rejoint également le concept d'inattendu : est saillant l'élément perturbateur, inattendu, curieux, intrigant, énigmatique. Il s'agit en effet de l'élément sur lequel on s'interroge, ou sur lequel le regard va s'attarder pour résoudre le problème qu'il pose. Par exemple, tout élément visuel ou langagier pour lequel l'activité perceptive de reconnaissance s'avère difficile en devient saillant.

Est également saillant ce qui est fort et stable. Ce facteur rejoint la conception de la Gestalt selon laquelle la prégnance dénote la force de résistance aux perturbations. Par exemple, le sujet grammatical d'un énoncé a une certaine force et une certaine stabilité, ce que prouvent les reprises pronominales. Au niveau de la perception visuelle, les éléments placés à des zones stratégiques de l'image (par exemple le centre du cadre) sont saillants. Autrement dit est saillant ce qui se trouve à des endroits clés, que ces endroits soient des positions grammaticales ou des zones spatiales.

Un autre facteur de saillance consiste en ce qui est simple, immédiat, percutant. Dans le langage, certains mots ont un tel caractère, par exemple le mot anglais « *time* », court et avec une diphtongue particulièrement sonore, que l'on retrouve fréquemment dans des chansons où cette saillance est très exploitée. Dans la perception visuelle, on trouve un tel facteur dans le concept de bonne forme de la Gestalt. Une bonne forme correspond ainsi à un minimum d'informations

sensorielles, qui, par conséquent, sont reconnues plus vite (meilleure distinction de cette forme par rapport au fond), sont mieux mémorisées, et sont décrites plus succinctement.

Notre principal facteur consiste en une rupture dans une continuité. Nous avons déjà vu cette caractérisation dans le chapitre 1, au moment de la présentation de la notion de singularité dans le modèle de Bellalem pour la formalisation des trajectoires gestuelles (cf. page 37). Cette notion de singularité pour au moins une propriété est un facteur de saillance. Au niveau du langage, elle correspond par exemple à un accent tonique au milieu de l'énoncé. Au niveau de la perception visuelle, elle correspond à une propriété de forme, de couleur ou de taille propre à un seul objet dans l'ensemble des objets visibles. Par exemple, un objet rouge dans un ensemble d'objets bleus est saillant (alors que l'unique objet rouge dans un ensemble d'objets multicolores ne l'est pas). Un objet dont la présence casse le rythme de l'image est également saillant. Dans un même ordre d'idée, un singleton dans une partition en groupes perceptifs correspond à un objet saillant par le fait qu'il est isolé de tout groupe.

Caractérisation générale de la saillance. Notre caractérisation fondée sur la singularité nous semble opérationnelle, dans le sens où repérer des irrégularités dans une liste de propriétés s'avère possible d'un point de vue computationnel, du moins à partir du moment où ces propriétés peuvent se représenter formellement. Notre objectif est d'identifier ces propriétés, en tenant compte des considérations précédentes et en comparant systématiquement les facteurs visuels et les facteurs linguistiques. Nous suivons ainsi une méthodologie consistant à élaborer une caractérisation parallèle. Les questions qui se posent sont les suivantes : quelle est la validité d'une telle caractérisation ? Peut-on comparer et unifier les critères de saillance visuelle et les critères de saillance langagière ?

S'il est encore trop tôt pour s'intéresser à cette dernière question, nous pouvons apporter une réponse à la première. Avant de les étudier en détail, notons déjà que les principaux critères visuels identifiés dans la littérature sont les points forts, les lignes de force, les répétitions et les symétries. Ils peuvent *a priori* apparaître dans toute scène visuelle, aussi bien avec des formes géométriques telles que des triangles ou des carrés qu'avec des meubles dans une tâche d'aménagement d'intérieur. La transposition de ces critères au langage permet d'identifier des facteurs de saillance linguistique liés à la structure syntaxique de l'énoncé, à la présence de répétitions ou de symétries. Or ces phénomènes n'apparaissent véritablement que dans le langage poétique. On les retrouve dans des figures de style telles que le chiasme ou le changement de rythme. Le langage poétique est structuré selon une volonté particulière et il est logique que de tels phénomènes y apparaissent. Ces phénomènes sont par contre peu probables dans un énoncé de dialogue homme-machine. Par conséquent, nous pouvons nous demander si les phénomènes visuels équivalents apparaissent de manière significative dans les scènes visuelles d'une application de dialogue homme-machine, ou, au contraire, sont spécifiques des images réfléchies et structurées comme le sont la plupart des images artistiques et publicitaires. Nous considérons que ces phénomènes visuels peuvent apparaître à tout moment dans le dialogue, pour la simple raison que ces scènes sont affichées sur un écran qui constitue un cadre, cadre duquel naissent des endroits stratégiques, donc des points forts. Pour une caractérisation générale de la saillance, nous tenterons donc de réunir ces phénomènes avec des phénomènes linguistiques plus généraux.

5.1.2 Importance de la saillance lors la référence aux objets

La saillance du point de vue cognitif. De même que nous l'avons fait en § 4.2.1 page 85 pour mieux appréhender les phénomènes, il est possible de distinguer plusieurs hypothèses d'exploi-

tation de la saillance :

1. L'utilisateur se prépare à agir sans savoir sur quel objet il va se focaliser. Pour choisir un objet, soit il se laisse guider par des contraintes liées à la tâche applicative, soit il se laisse guider par sa perception visuelle de la scène. Dans ce cas, il se laisse attirer par les objets visuellement saillants.
2. L'utilisateur se prépare à agir en réaction à un énoncé du système. Cet énoncé a construit une certaine saillance que l'utilisateur va réutiliser, soit en la confirmant, soit en la détruisant en rendant saillant un autre objet.
3. L'utilisateur est en train d'écouter un énoncé du sytème. Dans le processus de compréhension et plus précisément de résolution de la référence, il va recourir à la saillance pour comprendre de quels objets parle le système.
4. L'utilisateur est en train de réfléchir à l'énoncé qu'il va produire. Il exploite la saillance préalable, visuelle, linguistique ou liée à la tâche, pour simplifier la phrase qu'il est en train de préparer.

Ces quatre exemples mettent l'accent sur deux aspects de la saillance, celui d'aider à l'identification des objets susceptibles d'être traités, et celui de constituer un terrain commun entre les deux interlocuteurs pour la génération et la compréhension des expressions référentielles. La saillance apparaît ainsi comme une partie indissociable du reste de la cognition : elle est liée à l'intentionnalité, à l'attention, aux choix des mots et à l'interprétation contextuelle de leur sens.

D'un autre côté, la saillance peut être vue comme se confrontant avec le reste du processus cognitif. Selon Stevenson (2002), il y a confrontation entre d'une part la saillance, et d'autre part la cohérence du discours ainsi que l'accessibilité des informations. En effet, la simplification d'une expression référentielle, justifiée par la saillance du référent, a des conséquences sur l'énoncé. Si cette simplicité suffit à l'interlocuteur à retrouver le référent, elle peut néanmoins s'accompagner d'une imprécision quant à l'introduction du référent dans l'énoncé ou quant à son accessibilité pour d'éventuelles reprises ultérieures.

La saillance pour la résolution de la référence dans les travaux existants. Nous nous focalisons sur le rôle de la saillance dans la résolution de la référence aux objets. L'intérêt est de réduire l'interprétation aux seuls objets saillants, ou en tout cas de permettre cette possibilité pour résoudre une ambiguïté. Encore faut-il déterminer quels sont les objets saillants. D'une manière générale, il n'y a pas de consensus. Quelques critères sont avancés, mais soit ils restent flous, soit ils sont liés à la tâche applicative et ne sont pas généralisables.

Nous prendrons l'exemple de travaux s'intéressant à la description d'environnements inconnus dans lesquels l'un des interlocuteurs doit se repérer sur la base de ce que lui dit l'autre. C'est dans ce cadre que Lynch (cité par Edmonds 1993) identifie le point de vue, la familiarité et les buts de tâche courants comme des critères importants, et que Devlin (cité par Edmonds 1993) dit que la saillance peut être influencée par la manière d'identifier, la visibilité, la proéminence et l'importance fonctionnelle. Ces définitions mettent en valeur la subjectivité de la saillance (familiarité, manière d'identifier), l'importance du contexte de la tâche et la perception visuelle. Edmonds (1993) les utilise pour l'étude de la référence dans le dialogue lorsque les deux interlocuteurs n'ont pas de connaissance commune des référents. Il décrit une situation-type de dialogue : les deux interlocuteurs se parlent au téléphone et ne peuvent donc pas utiliser le geste de désignation ; l'un d'eux décrit un itinéraire routier à l'autre qui ne fait qu'écouter et demander des précisions. Pour que l'itinéraire soit clair, le premier (que nous désignerons comme étant le locuteur) utilise les caractéristiques qui lui semblent les plus saillantes. Si l'itinéraire passe par

exemple devant un immeuble particulièrement haut, la taille de cet objet sera saillante et donc il l'utilisera : « *tu verras un immeuble très haut* ». Le second (que nous désignerons comme étant l'interlocuteur) accepte la description ou demande des précisions lorsque la caractéristique ne lui semble pas pertinente : « *haut comment ?* ». Lorsque l'interlocuteur accepte la description de l'itinéraire, c'est qu'il a confiance dans sa validité.

A première vue, plus le locuteur emploie de descripteurs, plus il y a de chances pour que l'interlocuteur accepte la description. On doit cependant écarter la solution de messages trop longs dont l'effet néfaste est de gêner la compréhension. Le message peut être raccourci en enlevant les éléments qui n'apportent rien. L'adaptation computationnelle des maximes de Grice (1975) par Dale & Reiter (1995) va dans ce sens. L'idéal est de ne laisser dans le message que les informations pertinentes. On fait des descriptions courtes non pas en minimisant la longueur des expressions mais en utilisant des informations saillantes (Reiter & Dale 1992). Nous noterons que l'algorithme incrémental de Dale et Reiter a été complété par Krahmer & Theune (2002) pour tenir compte d'une échelle de saillance linguistique. L'algorithme obtenu est un exemple intéressant d'exploitation de la saillance pour la référence aux objets. Nous n'entrerons pas dans les détails de cet algorithme pour la simple raison qu'il est conçu pour la génération et non pour l'interprétation des expressions référentielles. Comme nous le verrons dans le chapitre 10, la saillance n'intervient pas de la même façon dans les deux processus.

A la suite de (Clark *et al.* 1983), nous pouvons aller plus loin en affirmant que les informations contenues dans l'expression référentielle peuvent tellement reposer sur la saillance qu'elles en paraissent à première vue ambiguës. Sur la base d'expérimentations, (Clark *et al.* 1983) montrent que des sujets comprennent des expressions référentielles ambiguës.

Un autre système ciblé sur la saillance est le système DenK présenté en particulier dans (Kievit *et al.* 2001). Il s'agit d'un modèle de résolution d'expressions référentielles multimodales dans le cadre de la manipulation d'un microscope avec la voix et le geste *via* la souris. Le système est capable de résoudre un pronom sur une entité disponible dans le contexte visuel. Il utilise la saillance comme critère de choix parmi les possibilités obtenues après les filtrages catégoriels classiques. Si la saillance visuelle est bien explorée, la saillance linguistique se limite à la récence, c'est-à-dire aux entités de discours mentionnées le plus récemment. Ce manque d'homogénéité dans le traitement de la saillance caractérise ainsi la plupart des systèmes.

Intervention de la saillance dans notre approche. Considérant qu'une bonne gestion de la saillance est un élément essentiel d'un système de résolution de la référence, nous donnons une importance particulière à la prise en compte de la saillance dans sa globalité, sans privilégier la saillance visuelle ou la saillance linguistique, et en apportant un cadre unifié pour leur exploitation. Le modèle des domaines de référence nous semble être un tel cadre. Plus que cela, du fait de la nature hétégorène des sources contextuelles à l'origine de domaines de référence, ce modèle nous semble idéal pour confronter les sources de saillance. Saillance visuelle, saillance linguistique, saillance applicative, saillance induite par le geste, saillance attentionnelle, saillance intentionnelle, saillance mémorielle : toutes semblent pouvoir être intégrées au modèle des domaines de référence.

5.1.3 La saillance dans le modèle des domaines de référence

Un paramètre peu considéré. Que ce soit dans le modèle de Reboul ou dans les améliorations qui en ont été faites, la saillance est un paramètre souvent oublié. Peut-être est-il considéré comme trop abstrait, ou comme découlant naturellement de faits concrets tels qu'un geste ostensif de la part de l'utilisateur ou une action ostensive de la part de la machine (mise en relief d'un objet à

l'aide d'un clignotement ou de n'importe quelle caractérisation perceptive exceptionnelle). Pour notre part, nous ne voyons pas la saillance comme un résultat mais comme un critère qui se déduit du contexte, et qui, soit amène à une certaine structuration de celui-ci, soit repose sur une telle structuration. Nos structurations se faisant en termes de domaines de référence, nous voulons ici préciser les rapports entre saillance et domaine de référence.

La saillance dans un domaine de référence. La saillance est-elle plutôt un facteur de groupement dans un domaine de référence ou un critère de différenciation dans une partition ? Dans le premier cas, cela veut dire que le système est capable de distinguer dans l'ensemble des objets de l'application ceux qui sont saillants et ceux qui ne le sont pas, quelle que soit l'origine de cette saillance. Des objets visuellement saillants se retrouvent ainsi dans un même domaine de référence que des objets linguistiquement saillants. Cette méthode a l'avantage de donner directement le domaine de référence dans lequel procéder à l'interprétation fondée sur la saillance, c'est-à-dire lorsqu'une première interprétation aboutit à une ambiguïté. Elle est en revanche réductrice et irréalisable, d'une part parce que la saillance ne fonctionne pas en tout ou rien mais met en jeu des degrés de saillance, d'autre part parce que la saillance d'un objet se calcule en fonction des autres objets. Ce dernier point nous amène à imaginer la saillance comme un critère de différenciation pour une partition dans un domaine déjà créé. Le calcul de la saillance pour tous les objets du domaine a alors un sens, ces objets étant liés par un facteur de groupement qui donne une cohésion à l'ensemble. Cette cohésion permet la comparaison des éléments. Si le facteur de groupement du domaine est lié à la perception visuelle, la saillance calculée sera la saillance visuelle. Le ou les éléments les plus saillants pourront être mis en avant dans une partition, le critère de différenciation retenu pour cette partition étant une combinaison du critère « saillance » et du ou des critères correspondant aux caractéristiques qui ont rendu l'élément saillant (la couleur et la taille par exemple). Ces derniers sont inévitables et la saillance apparaît ainsi, non pas comme un critère de différenciation à part entière, mais comme un critère supplémentaire dans le cadre d'une différenciation déjà justifiée. La saillance constitue en cela un point d'entrée dans un domaine : elle permet de mettre en avant une partition lorsque plusieurs ont été déterminées mais qu'une seule d'entre elles possède le critère de différenciation supplémentaire « saillance ».

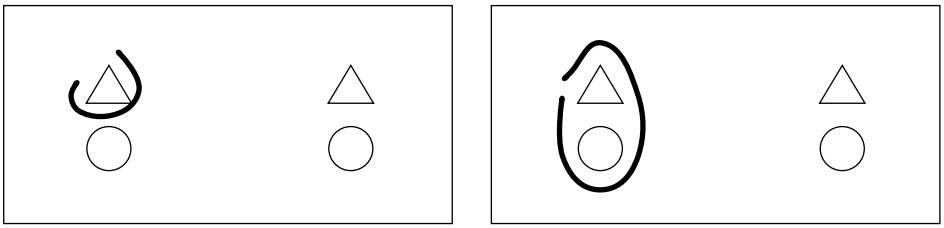
La saillance peut-elle être un critère d'ordonnancement ? La comparaison des saillances des éléments du domaine permet *a priori* de les ordonner dans une partition étiquetée par un critère d'ordonnancement. Nous rappelons cependant que l'utilité de ce critère est dans l'interprétation des références telles que « *le premier* », « *le second* » ou « *le suivant* ». Or rien ne vient confirmer que la saillance est un facteur suffisamment fort pour inciter à de telles expressions référentielles. Bien au contraire, nous verrons en § 6.1.2 qu'il existe d'autres facteurs, comme celui des lignes directrices dans la perception visuelle, qui déterminent d'autres ordonnancements plus pertinents. Une ligne directrice peut par exemple partir de l'objet le plus saillant visuellement et parcourir ensuite certains objets dans un ordre qui n'a rien à voir avec l'ordre décroissant de saillance.

La saillance d'un domaine de référence. Un domaine peut-il être globalement saillant ? Nous considérons que le cas est effectivement possible. Un domaine qui contient un élément particulièrement saillant en devient indirectement saillant dans sa globalité, du moins à partir du moment où la saillance initiale s'applique à un élément du domaine par rapport à des éléments extérieurs au domaine. Par exemple, les objets très proches d'un objet saillant attirent eux aussi l'attention, du moins en comparaison avec d'autres objets plus éloignés. De plus, la saillance peut s'appliquer directement à un domaine, sans passer par un élément particulier. Nous allons le voir tout de suite avec l'exemple d'un geste ostensif désignant un domaine de référence.

5.2 Saillance et geste ostensif

Liens entre demonstrata et domaine de référence. Le geste ostensif rend des demonstrata saillants. Il constitue ainsi le seul facteur véritablement explicite de saillance. Nous avons déjà montré que les référents pouvaient différer des demonstrata. Nous voulons montrer ici comment le geste ostensif et son association avec une expression référentielle verbale donne des indices sur les référents et sur les domaines de référence. Cette section ne présente pas de véritable algorithme mais plutôt un récapitulatif à la fois théorique et pratique sur les rapports entre demonstrata et domaines de référence. Ce récapitulatif a été validé dans (Landragin 2002).

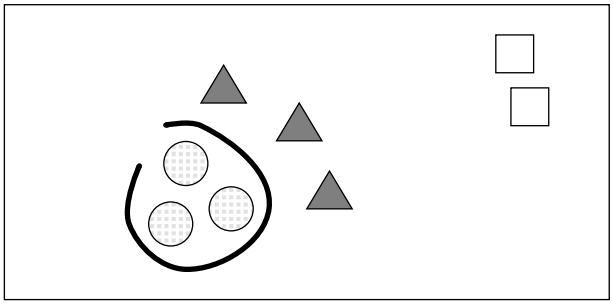
Dans le premier exemple de la figure 5.1, le geste isole l'un des deux triangles et prend ainsi avec le démonstratif le rôle de critère de différenciation dans le domaine de référence correspondant au contexte visuel complet. Dans le deuxième exemple, le geste désigne cette fois un domaine de référence dans lequel l'article défini s'applique pour extraire l'unique objet de type « *triangle* ».



A. « *Ce triangle* » + geste B. « *Le triangle* » + geste

Figure 5.1 – *Vérification de la détermination avec un geste d'entourage.*

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 le premier exemple de la figure 5.1 et les exemples similaires on identifie par défaut le contexte visuel complet, l'exemple de la figure 5.2 (tiré du corpus 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).



« *Les formes les plus claires* » + geste

Figure 5.2 – *Geste désignant les référents et initiant la construction du domaine.*

Nous nous trouvons donc 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. En considérant une structuration du contexte visuel en dendrogrammes tels que décrits dans le chapitre précédent, cette extension se fait 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, et dans lequel les formes les plus claires sont bien les ronds désignés. D'un point de vue cognitif, ce domaine de référence correspond à une zone attentionnelle prise en compte par l'utilisateur lors de la production de son geste, et justifie son choix d'utiliser un superlatif. La raison de ce choix peut se trouver dans la volonté d'exprimer un contraste saillant entre formes claires et formes sombres. L'expression référentielle ultérieure « *les autres* » exploite ce contraste en désignant les trois triangles sombres, c'est-à-dire les autres « *formes* » présentes dans le domaine de référence.

Une attention particulière dans la construction des exemples des figures 5.1 et 5.2 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. Le but est d'illustrer le pouvoir prédictif de notre modèle face aux différentes formes de reprise et de continuité que peut prendre l'énoncé ultérieur. Comme nous venons de le voir avec « *les autres* » pour la figure 5.2, nous pouvons montrer l'intérêt de nos domaines de référence pour la figure 5.1. Dans le premier exemple, 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. Dans le deuxième exemple, l'expression « *l'autre* » ne fonctionne pas dans le domaine délimité par le geste. Or une telle expression semble effectivement inacceptable après un défini au singulier. Par contre, l'expression « *le rond* » fonctionne parfaitement pour désigner le rond inclus dans le domaine délimité par la trajectoire gestuelle.

Identification du niveau auquel intervient le geste. Une étape importante dans l'interprétation d'une expression référentielle multimodale s'avère ainsi être l'interprétation d'une trajectoire gestuelle comme délimitant un domaine ou comme désignant un référent. Plusieurs paramètres rendent difficile cette interprétation. D'une part le geste est souvent imprécis, surtout en communication homme-machine lorsqu'il est effectué sur un écran tactile ou à l'aide de tout dispositif nécessitant un calibrage délicat (comme le 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. Ce problème a été illustré dans le cas d'une tâche d'aménagement d'intérieur dans (Romary 1993) avec l'énoncé de positionnement « *mets de la moquette ici* » associé à un pointage en un point du sol. Déterminer le rôle du geste se fait à l'aide de l'expression référentielle verbale et des particularités du contexte visuel. Nous avons montré le rôle particulièrement important de la détermination dans cette identification du rôle du geste. Nous y reviendrons lors de la présentation de notre modèle tripolaire dans le chapitre 7.

Récapitulatif sur les rôles du geste. Ce n'est pas parce qu'un geste délimite précisément une zone spatiale que le regard de l'utilisateur s'arrête aux limites de cette zone. Elle peut en effet être étendue jusqu'à inclure un contraste entre objets. Elle peut aussi être étendue jusqu'à contenir un objet facilement nominalisable, cet objet pouvant alors être celui nommé dans l'expression référentielle verbale. Or, si le regard de l'utilisateur ne s'arrête pas aux limites de la zone désignée par le geste, il est logique de dire que le regard de l'interlocuteur ne le fait pas non plus, et qu'un

système de dialogue homme-machine doit prendre en compte un tel facteur.

D'une manière générale, nous considérons que le geste est un point d'ancrage saillant et non un délimitateur précis. Cette position confirme notre point de vue à la fin du chapitre 2 que nous avons exposé avec des scores compris entre 0 et 1 pour la détermination des demonstrata. Ainsi, les rôles possibles d'un geste ostensif sont les suivants :

1. Identification directe d'un demonstratum ou de demonstrata (« *ce triangle* » avec un geste non ambigu vers un triangle) ;
2. Point d'ancrage (désignation d'un lieu ou d'un demonstratum) pour l'identification de demonstrata (« *ces objets* » avec un pointage vers un objet appartenant à un groupe perceptif saillant) ;
3. Point d'ancrage (désignation d'un demonstratum) pour une référence générique (« *ces fauteuils* » avec un geste vers un fauteuil) ;
4. Point d'ancrage (délimitation d'un domaine de référence) pour l'extraction de référents (« *le triangle* » avec un geste vers un triangle et un carré).

Dans les trois premiers cas, l'expression référentielle et éventuellement quelques particularités du contexte visuel peuvent contenir des indications pour la délimitation du domaine de référence. Le superlatif de la figure 5.2 constitue une telle indication. Un autre exemple plus fréquent d'indications propre au contexte visuel est la répartition des objets en groupes perceptifs. Dans le dernier cas, l'expression référentielle doit au contraire sous-entendre qu'un domaine de référence est délimité par ailleurs, en l'occurence par le geste. C'est l'article défini qui, dans son association avec un geste, joue un tel rôle.

Comme nous l'avons déjà fait à la fin du chapitre 2, nous avons spécifié ici un pré-requis pour la modélisation proprement dite, modélisation qui sera l'objet du chapitre 7. Maintenant que nous avons détaillé les rapports entre geste et domaine de référence, nous serons plus à même d'aborder l'intégration de ces notions avec le langage. Nous pourrons alors nous intéresser à un algorithme d'identification du rôle du geste.

5.3 Caractérisation des critères de saillance

Nous proposons dans cette section deux classifications de critères de saillance, la première pour la saillance visuelle et la seconde pour la saillance linguistique. Le but étant d'obtenir des critères concrets qui soient calculables par un système de dialogue homme-machine, nous donnerons à chaque fois les pistes pour de tels calculs.

5.3.1 Saillance visuelle

Etat des lieux. Les remarques faites à propos de saillance en général sont valables à propos de saillance visuelle. Les travaux existants ne se basent pas sur un consensus, et les critères avancés par les uns et les autres restent soit flous, soit liés au contexte visuel particulier d'une tâche applicative particulière. Nous nous intéressons ici aux caractérisations formelles de la saillance visuelle Certains travaux ont déjà été cités en § 5.1.2. Ils sont ici abordées d'un point de vue classificatoire et opérationnel.

Dans le cas d'une tâche de description d'un itinéraire routier dans une ville, Davis (cité par Edmonds 1993) classe selon leur saillance les repères visuels que l'on peut trouver dans une ville. Cette classification est sensée représenter ce que le locuteur pense que l'interlocuteur croit

saillant. Au niveau le plus saillant de la classification, on trouve par exemple les immeubles et les feux rouges. Cette approche est intéressante dans le sens où elle apporte une méthodologie pour aborder la saillance. Elle reste cependant tributaire de la tâche applicative. L'approche de Reiter & Dale (1992) consiste elle aussi à classer dans une hiérarchie les caractéristiques d'un objet devant être utilisées prioritairement pour décrire cet objet. Les caractéristiques étudiées sont le type, la taille et la couleur. L'intérêt de cette approche est d'être plus générique, son inconvénient est de ne considérer qu'un ensemble limité de caractéristiques.

Edmonds (1993) reprend le principe de hiérarchie en y ajoutant un coefficient de saillance : plus le coefficient est élevé, plus la caractéristique correspondante a de chances d'être utilisée pour décrire l'objet. Il affirme d'autre part que la saillance dépend du contexte entourant le référent. Il donne l'exemple d'un immeuble *a priori* saillant par sa taille importante, et qui perd toute saillance lorsqu'il est entouré d'immeubles encore plus grands. Il considère également que certaines caractéristiques sont saillantes pour certains objets et pas pour d'autres, de même que certaines caractéristiques sont saillantes dans un but précis du dialogue et non dans un autre but. Par exemple, la caractéristique « taille » est saillante lorsque le but du dialogue est la désignation d'un immeuble, mais n'est pas saillante lorsque le but du dialogue est la désignation d'une intersection de rues. Cette approche ne l'empêche pas de dresser une hiérarchie des caractéristiques candidates à la saillance, mais il le fait pour chaque catégorie possible d'objets : pour un immeuble, par exemple, il considère que les caractéristiques les plus saillantes sont le style d'architecture et la taille. Sa hiérarchie possède donc deux niveaux de saillances : le premier correspondant aux types (ou catégories) d'objets, le second correspondant aux caractéristiques autres que ce type et classées indépendamment selon chaque type. L'approche consistant à utiliser des coefficients de saillance est intéressante car elle pourrait permettre de traiter la liste des caractéristiques possibles de la manière suivante : si la caractéristique la plus saillante a un coefficient beaucoup plus élevé que ceux des caractéristiques suivantes, l'objet peut n'être décrit que par cette caractéristique ; dans le cas contraire et surtout si les coefficients sont faibles dans l'ensemble, il pourrait être nécessaire de décrire l'objet par plusieurs caractéristiques, grossièrement par les deux les plus saillantes. Nous retiendrons ces principes pour l'élaboration de notre classification.

Des critères physiologiques. L'étude de la saillance visuelle commence généralement par des considérations physiologiques, en particulier à propos de perception des couleurs. Le temps de latence, c'est-à-dire le décalage entre le début de l'excitation et celui de la sensation, varie selon les couleurs : très bref pour le rouge, un peu plus long pour le vert, maximal pour le jaune. Baticle (1985) donne ainsi les chiffres suivants : rouge (22,6 millièmes de seconde), vert (37,1 millièmes de seconde), gris (43,4 millièmes de seconde), bleu (59,8 millièmes de seconde), jaune (96,3 millièmes de seconde). Il note également que la sensibilité chromatique est fonction de l'éclairement : le jour, le maximum de sensibilité est dans le jaune et le rouge, le soir et la nuit dans le bleu. Cocula & Peyroutet (1989) notent que la visibilité de chaque couleur est moindre quand elle est associée à d'autres. Suite à une étude expérimentale, ils déterminent les couples de couleurs susceptibles du plus grand impact visuel. Ils obtiennent un classement dont les premiers éléments sont les suivants : 1. noir sur blanc, 2. noir sur jaune, 3. rouge sur blanc, 4. vert sur blanc, 5. blanc sur rouge, etc.

Ces données nous donnent les bases pour un calcul de saillance des couleurs. En fonction de la couleur du fond et de sa propre couleur, un objet d'une couleur unie pourra se voir attribuer un score de saillance chromatique. Le calcul se complique quand l'objet réunit plusieurs couleurs : il s'agit alors de déterminer quelle partie de l'objet a le plus grand impact visuel, pour favoriser la couleur de cette partie dans le calcul. Pour une chaise avec un dossier rouge et des pieds noirs,

il est clair que la couleur à prendre en compte (la couleur principale) est le rouge.

Ces calculs s'avèrent insuffisants dans un contexte où plusieurs objets ont la même couleur : dans une scène comportant huit chaises rouges et une bleue, les calculs aboutissent à considérer les huit chaises rouges comme saillantes. L'information, qui touche la majorité des chaises, n'a que peu d'intérêt. De plus, étant la seule à posséder cette caractéristique visuelle, la chaise bleue se distingue dans le groupe et présente pour cette raison une saillance importante. Notre approche consistera donc à détecter tout d'abord les objets ayant une couleur particulière, à les étiqueter comme chromatiquement saillants, puis, dans un deuxième temps, à classer tous les objets selon le temps de latence de leur couleur principale.

Des critères liés à la structure de la scène visuelle. Une image 2D se caractérise par des limites qui forment un cadre. Dans ce cadre, et avant même de considérer les objets physiques qui y sont représentés, certaines zones spatiales peuvent se distinguer et être considérées comme saillantes. D'après les études en arts picturaux, comme par exemple celles de (Itten 1961) ou de (Kandinsky 1979), les zones qui attirent le regard sont classiquement :

- les zones claires, les tâches colorées en contraste avec le reste de l'image ;
- les zones nettes qui se détachent des zones floues ;
- le quart inférieur gauche, qui correspond à la partie perçue le plus rapidement ;
- les zones au voisinage de points forts ;
- les zones qui tendent à équilibrer un autre point fort ;
- les zones situées dans le prolongement ou à l'intersection des lignes directrices.

Nous arrivons ainsi à la notion de point fort. D'une manière générale, le regard est attiré par les points forts. Toujours dans une première approche consistant à ne pas considérer les objets, la question est de déterminer quels sont les emplacements privilégiés des points forts. Ces emplacements, mis particulièrement en avant par les études sur la composition photographique (Sanmiguel 2000), sont classiquement :

- les quatre emplacements forts classiques, qui, selon le format, correspondent aux intersections des lignes horizontales et verticales situées aux tiers du cadre, ou aux intersections des lignes correspondant à la proportion du nombre d'or[1] ;
- le centre de l'image (mais dans ce cas le regard y reste) ;
- le ou les points de fuite, si la perspective est marquée ;
- le point d'équilibre ou le centre de symétrie, si l'image présente une symétrie ou un équilibre des masses qui repose sur ce point ;
- l'élément qui équilibre toute une composition, souvent le sujet.

Si certains éléments de ces deux listes sont facilement calculables, il n'en est pas de même d'un point d'équilibre ou d'une ligne directrice. Nous tenterons dans le chapitre 6 d'apporter des éléments pour la modélisation des lignes directrices. Avant cela, face aux difficultés posées par la détection de symétries et d'équilibre, nous faisons le choix de nous intéresser aux seuls objets de la scène. Comme le dernier élément de la liste précédente le mentionne, nous considérons que ce sont les objets présents dans la scène qui structurent celle-ci.

1. C'est de ces proportions que découlent tous les conseils de composition relatifs au placement du sujet, à la proportion relative de la terre et du ciel pour un paysage, etc. Les tiers de l'image s'appliquent parfaitement au cadre que constitue l'écran d'un ordinateur. Pour des formats particuliers comme l'image panoramique, les règles sont très différentes.

Des critères liés aux propriétés des objets. Le but est de déterminer les caractéristiques des objets qui contribuent à leur saillance visuelle. Ces caractéristiques constituent des critères calculables par le système, les calculs se faisant à partir des données présentes dans la base des objets de l'application, c'est-à-dire les coordonnées 2D ou 3D, les types et les propriétés physiques. A partir de ces caractéristiques, nous obtenons une hiérarchie de facteurs de saillance visuelle. Nous présentons ici une troisième version corrigée de cette hiérarchie, la première version se trouvant dans (Landragin 1998) et la deuxième dans (Landragin 1999).

- La saillance par une mise en évidence explicite :
  Lorsqu'un objet est mis explicitement en évidence, que ce soit grâce à une luminosité particulière, à une intention ostensive de la part du système (par un rendu visuel particulier[1]), ou à une intention ostensive de la part de l'utilisateur (production d'un geste ostensif), cet objet est clairement saillant. Si les autres objets du même type N ne sont pas saillants, une expression référentielle définie telle que « *le N* » suffit alors pour le désigner.

- La saillance par la catégorie :
  La catégorie est quasiment toujours explicite dans l'expression référentielle. En effet, même quand tous les objets visibles sont de la même catégorie, quand par exemple la scène ne comprend que des triangles, la mention de la catégorie n'est pas nécessaire : « *l'objet rouge* » suffit à désigner le seul triangle de cette couleur. Or nous avançons l'hypothèse qu'on dira plus spontanément « *le triangle rouge* ». Cette hypothèse se vérifie avec le corpus Magnét'Oz, où même les objets difficilement nominalisables se voient attribuer une catégorie par les sujets. Une des conséquences est que le recours à la saillance par la catégorie n'a que peu d'utilité : si un objet est saillant parce qu'il est le seul de sa catégorie N, une expression telle que « *le N* » s'interprètera sans ambiguïté et donc sans recours à la saillance. Celle-ci ne s'avère donc utile que lorsqu'il n'existe qu'un seul mot pour désigner deux catégories voisines. Par exemple, dans une scène ne comportant que des chaises dont une chaise à roulettes, cette dernière sera saillante et on pourra considérer que l'expression « *enlève la chaise* » suffit à la désigner.

- La saillance par les caractéristiques physiques :
  Parmi les caractéristiques physiques d'un objet, citons la forme, la couleur, la taille, le matériau et la texture. Comme nous l'avons vu dans le chapitre 4 lors du groupage par similarité, nous avons ordonné ces propriétés (en tout cas les trois premières). Ainsi, dans une scène comportant plusieurs triangles dont un rouge et les autres bleus, le triangle rouge est saillant. De même, dans une scène contenant des chaises dont une chaise pour enfant bien plus petite que les autres, cette dernière est saillante.

- La saillance spatiale, c'est-à-dire par la localisation dans la scène :
  La localisation dans la scène recouvre plusieurs caractéristiques : la position d'un objet par rapport aux autres objets (l'objet est-il isolé, appartient-il à un groupe perceptif, à une ligne directrice ?) ; la position de l'objet par rapport au participant (à quelle distance est-il situé du participant ? est-il proche ou latéralement distant de son axe de visée ? comment est-il orienté ?). Ainsi, un objet est saillant s'il est très proche du participant, du moins par rapport aux autres objets de la scène. Un objet isolé est saillant si tous les autres objets

1. Dans un environnement 3D, plusieurs rendus sont couramment utilisés pour dénoter des significations particulières liées à l'interaction ou à la tâche : rendu en fil de fer, avec des textures transparentes ou par affichage de la boîte englobante de l'objet. Dans le projet COVEN par exemple, les objets sélectionnés apparaissent avec leur boîte englobante, la couleur de celle-ci dépendant du participant à l'origine de la sélection.

visibles appartiennent à un groupe perceptif. Ce type de saillance est directement lié à notre structuration de l'espace visuel en groupes perceptifs.

- La saillance par l'incongruité ou l'aspect énigmatique :
  Un objet dans une situation incongrue est en infraction avec une règle implicite, culturelle ou fonctionnelle. Ainsi, une chaise renversée ou placée sur une table est incongrue et donc saillante si toutes les autres chaises sont sur le sol. De même, un objet qui se trouve dans une position inhabituelle est saillant : une chaise tournant le dos au participant est saillante si les autres lui font face.

- La saillance par les fonctionnalités :
  Nous considérons que les fonctionnalités d'un objet peuvent être perçues visuellement, et nous nous intéressons ici à l'effet de saillance dû aux propriétés fonctionnelles d'un objet au travers de la représentation visuelle de ces propriétés. Par exemple, dans une scène contenant plusieurs ordinateurs, un ordinateur allumé est saillant si tous les autres ordinateurs sont éteints. Dans une scène contenant un bureau et plusieurs chaises, une chaise faisant face au bureau et les autres étant détachées de tout meuble, la chaise face au bureau est saillante. Entre une table et un objet posé dessus, l'objet est probablement plus saillant que la table qui a un intérêt moindre mais juste une fonction de support.

- La saillance par la dynamique :
  Dans une scène contenant un objet animé et plusieurs objets inanimés, l'objet animé est saillant. De même, dans une scène contenant un objet en mouvement et plusieurs objets statiques, l'objet en mouvement est saillant.

- La saillance indirecte :
  Comme nous l'avons vu plus haut, la saillance indirecte correspond au transfert de saillance d'un objet à un autre. Ainsi, la chaise placée devant une table saillante en devient saillante. Le transfert de saillance intervient également avec les lignes directrices, comme nous le verrons dans le chapitre 6.

Des critères liés à la subjectivité. Nous avons vu que la subjectivité pouvait affecter ces caractéristiques. Toute perception visuelle est en effet liée à l'utilisateur et à ses buts communicatifs. Dans le cadre de sa description d'itinéraires routiers, Edmonds (1994) donne l'exemple un peu simple de panneaux indicateurs écrits en grec et donc saillants seulement pour des grecs. Cet exemple est critiquable (à notre avis le panneau indicateur reste saillant même pour un français mais n'est tout simplement pas compris). Son intérêt est d'insister sur le fait que même un objet fait par nature pour être saillant peut voir sa saillance remise en cause par la subjectivité. Nous distinguons deux critères pour la subjectivité : la familiarité ou dépendance à l'expérience de l'utilisateur, et l'attention de celui-ci au moment de la communication. La familiarité peut se décomposer en :

- La familiarité visuelle culturelle : l'exemple typique est qu'à partir du moment où une scène comporte un être humain, celui-ci est saillant. Dans une moindre mesure, un animal est aussi culturellement saillant. Nous noterons également qu'au cours de la perception visuelle d'un visage humain, le regard est d'abord porté sur les yeux, puis sur la bouche et le nez.

- La familiarité visuelle individuelle : l'exemple classique est celui du peintre et de l'informaticien qui entrent dans une salle et qui ne voient pas les mêmes objets (murs ou ordinateurs)

avec la même saillance. On acquiert tous ses propres sensibilités, sa propre vision des couleurs, etc. Dans la hiérarchie de facteurs de saillance visuelle, ce sont surtout la saillance par les catégories physiques et par la localisation dans la scène qui peuvent être affectées par cette subjectivité, les autres types de saillance faisant intervenir des caractéristiques trop précises pour qu'il y ait confusion. Ainsi, au niveau des couleurs par exemple, un daltonien percevra les objets différemment : un objet saillant parce qu'il est bicolore ne sera pas saillant pour un daltonien qui perçoit les deux couleurs de la même façon. Au niveau de la forme, de la taille ou du matériau, notre éducation et notre vie passée nous familiarisent avec certains types d'objets. La saillance par la localisation dans la scène est également subjective : la proximité peut être prépondérante pour un utilisateur alors que la distance à l'axe de visée peut l'être pour un autre.

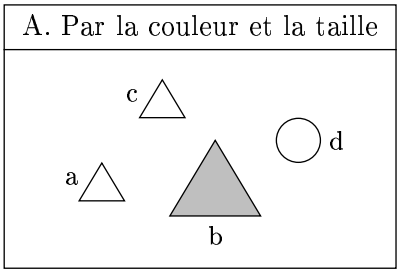
Un autre exemple de subjectivité est lié à l'attention de la personne, à son intérêt ou à son intention au moment de la communication. Est saillant l'élément qui a de l'intérêt compte tenu de l'objectif de la communication. Par exemple, quand on invite des collègues à entrer dans un bureau, les chaises sont saillantes car ils vont vouloir s'asseoir.

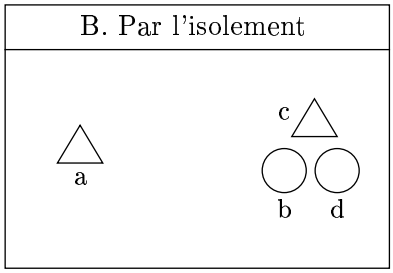
La plupart des phénomènes de saillance peuvent donc être affectés par la subjectivité. Face à ce problème, la seule solution est de trouver la méthode de résolution la plus générique possible et de ne pas conclure trop vite dans les cas particuliers. Dans l'état actuel de notre travail, nous choisissons d'ignorer ces critères.

Le calcul de la saillance visuelle dans notre approche. La saillance doit être calculée par le système à partir des données qu'il possède sur les objets visibles et sur les particularités de leur affichage dans la scène. Une première méthode 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 et le calcul est arrêté. Dans ce processus, l'ordre des caractéristiques, qui correspond à celui présenté dans (Landragin *et al.* 2001b), est le suivant :

1. Au niveau des propriétés de l'objet :
   - catégorie ;
   - caractéristiques physiques (forme, couleur, taille, matériau, texture).
2. Au niveau de la disposition spatiale de l'objet dans la scène :
   - proximité ;
   - isolement ;
   - orientation.
3. Au niveau de la lecture visuelle de la scène :
   - à la place d'un point fort calculable ;
   - à la place d'un point de fuite.

Cette méthode rapide met en avant les caractéristiques les plus pertinentes. Il arrive cependant qu'un objet soit saillant, non pas par une caractéristique qui apparaît en premier dans la liste ci-dessus, mais par une accumulation de caractéristiques moins bien placées dans la hiérarchie. Ainsi, dans le premier exemple de la figure 5.3 qui comporte trois triangles et un cercle, ce dernier est l'objet saillant car il est le seul de sa catégorie, la catégorie étant la première caractéristique considérée. Or un des triangles se distingue par sa taille et par sa couleur, et semble plus saillant que le cercle. Nous en concluons qu'il est nécessaire de prendre en compte toutes les caractéristiques dans le calcul de saillance.





	a	b	c	d
forme :	0	0	0	1
couleur :	0	1	0	0
taille :	0	1	0	0
isolement :	0	0	0	0
saillance :	**0**	**0.5**	**0**	**0.25**

	a	b	c	d
forme :	0	0	0	0
couleur :	0	0	0	0
taille :	0	0	0	0
isolement :	1	0	0	0
saillance :	**0.25**	**0**	**0**	**0**

Figure 5.3 – *Scores numériques pour un premier calcul opérationnel de la saillance visuelle.*

Nous proposons ainsi la méthode de calcul suivante : pour chacune des caractéristiques influentes dans la tâche applicative (par exemple la couleur), nous comparons les valeurs prises pour cette caractéristique par chacun des objets visibles (par exemple bleue, rouge et verte). Un objet qui est le seul à posséder une valeur (par exemple un objet qui est le seul à être rouge) se voit attribuer le chiffre 1 pour la caractéristique. Un objet qui partage sa valeur avec au moins un autre objet se voit attribuer le chiffre 0. Une fois que toutes les caractéristiques ont été traitées, à chaque objet correspond un vecteur composé de 0 et de 1. Il ne reste plus qu'à normer ces vecteurs et à comparer les valeurs obtenues. L'objet qui a la valeur la plus élevée est étiqueté comme étant l'objet saillant. Dans le premier exemple de la figure 5.3, c'est bien le grand triangle gris qui apparaît comme le plus saillant. Le deuxième exemple de la même figure illustre l'importance de la proximité et par conséquent des groupes perceptifs dans la saillance.

Pour conserver la hiérarchie entre les caractéristiques, une méthode simple consiste à pondérer chacune des caractéristiques par un coefficient inversement proportionnel à son rang dans la hiérarchie. Ainsi, si nous considérons comme pour le groupage que la couleur est plus importante que la taille, nous pondérons tous les chiffres relatifs à la couleur par un coefficient (entre 0 et 1) plus élevé que celui attribué à la taille. Nous ne l'avons pas fait dans la figure 5.3, car nous considérons que la détermination de ces coefficients revient à la tâche applicative. Nous y reviendrons donc dans le chapitre 9 (§ 9.1.1).

5.3.2 Saillance linguistique

Etat des lieux. Dans la phrase ou dans le discours, à l'écrit ou à l'oral, beaucoup de notions sont liées à la saillance : le focus, c'est-à-dire ce qui est rendu saillant par la structure de l'énoncé ; le thème, c'est-à-dire ce dont parle l'énoncé ; ou encore le sujet du dialogue (ou topic pour garder le terme anglais). Ces notions se recouvrent souvent et sont utilisées différemment par les différents auteurs. L'objet de cette section est de proposer une classification synthétique des facteurs de saillance linguistique, en mettant l'accent sur l'aspect formel et donc calculable par un système de dialogue homme-machine. Pour l'élaboration de cette classification, nous reprendrons les distinctions faites à propos de la saillance visuelle, ce qui nous permettra d'identifier de nouveaux facteurs.

A la suite de (Stevenson 2002), nous séparons les aspects formels, c'est-à-dire liés aux ca-

ractéristiques prosodiques et grammaticales de l'énoncé oral, des aspects sémantiques liés au contenu du message. Dans notre identification des critères de saillance, nous commencerons par les critères formels qui semblent les plus faciles à calculer. Nous étudierons ensuite les critères sémantiques, a priori moins formalisables, pour finir par les critères subjectifs, c'est-à-dire faisant intervenir des hypothèses sur les facteurs cognitifs mis en jeu, et pas seulement sur l'énoncé.

Des critères liés à la forme de l'énoncé. Dans la majorité des travaux en linguistique computationnelle, par exemple dans (Alshawi 1987), (Lappin & Leass 1994), (Grosz *et al.* 1995), ou encore dans les travaux de Hajičová cités dans (Krahmer & Theune 2002), ce sont essentiellement des critères formels qui définissent la saillance. Comme le soulignent Krahmer et Theune, la récence est souvent mise en avant, les entités les plus saillantes étant définies comme étant les plus récemment mentionnées. A partir de ces travaux et en nous appuyant sur les critères classificatoires que nous avons utilisés pour la caractérisation de la saillance visuelle, nous proposons la classification suivante pour les critères formels de saillance linguistique d'un mot ou d'un groupe de mots :

- La saillance intrinsèque au mot :
  Certains mots sont saillants de par leur nature même. Dans le cadre du dialogue oral, il s'agit par exemple de mots constitués de phonèmes particulièrement sonores ; de mots particuliers dans leur prononciation (une combinaison de phonèmes peu fréquente). Les mots qui contiennent beaucoup d'allitérations (consonnes), ceux qui contiennent beaucoup d'assonances (voyelles), sont ainsi des mots intrinsèquement saillants. En dehors de considérations phonétiques, certains mots comme « *ça* » et les déictiques purs sont saillants, du fait cette fois de leur manque d'autonomie référentielle, et de l'habitude qu'ils entraînent ainsi chez l'interlocuteur à faire particulièrement attention aux conditions de leur énonciation. Sont aussi intrinsèquement saillants les noms propres. Ce critère a souvent été évoqué dans la littérature, par exemple dans (Garrod & Sanford 1988).

- La saillance par une mise en avant explicite lors de l'énonciation :
  Nous distinguons ici quatre critères pour l'oral, certains d'entre eux ayant des équivalents à l'écrit. Le premier est l'association avec un geste, que ce geste soit ostensif ou expressif. Le deuxième est une prosodie particulière, par exemple un accent tonique bien marqué, une lenteur inhabituelle et intentionnelle, ou encore la transmission d'une émotion comme la colère ou l'ironie par une prononciation ou une intonation adéquate. Le mot ou groupe de mots ainsi mis en avant en devient saillant. L'équivalent à l'écrit est une typographie en caractères gras ou italiques. Un autre exemple de prosodie particulière est la présence d'une pause avant et après la prononciation d'un mot ou d'un groupe de mots : l'élément ainsi détaché du reste de l'énoncé en devient saillant. L'équivalent à l'écrit est la mise en apposition, entre deux virgules par exemple. Le troisième critère est une erreur de prononciation, un bégaiement par exemple. A l'écrit, les fautes d'orthographe jouent le même rôle : on les repère facilement, et ces fautes rendent saillants les mots concernés. Enfin, notre quatrième critère de saillance est la rupture dans une continuité de rythme : lorsque la phrase se caractérise par un rythme et qu'un mot vient casser ce rythme, ce mot en devient saillant. Ce phénomène s'avère néanmoins plus fréquent dans la poésie que dans l'énonciation en situation de dialogue homme-machine.

- La saillance par une construction syntaxique dédiée :
  Les constructions clivées en « *c'est ... qui ...* » constituent l'exemple typique de mise en relief explicite d'un élément. On dit aussi que l'élément est focalisé. « *Le triangle rouge* »

est ainsi le groupe nominal saillant dans « *c'est le triangle rouge que tu dois mettre à côté du bleu* ». Une autre construction dédiée est le détachement en tête de phrase, comme dans « *le triangle, le rouge, tu dois le mettre à côté du bleu* ».

- La saillance syntaxique liée à l'ordre d'apparition des mots :
  Un énoncé est une suite de mots caractérisée par des positions stratégiques. Le début, la fin, ou encore une position en rejet, sont prédisposés pour rendre saillant le mot ou le groupe de mot qui y prend place. L'intérêt des deux constructions précédentes par rapport à « *tu dois mettre le triangle rouge à côté du bleu* » est aussi de placer l'expression référentielle « *le triangle rouge* » à un endroit stratégique pour la rendre saillante. Cet exemple va à l'encontre du critère de récence souvent avancé. Il tend au contraire à montrer que le début de l'énoncé, c'est-à-dire la partie la moins récemment mentionnée, est la plus saillante. (Kessler *et al.* 1996) confirme ce point de vue en montrant que la première entité focalisée reste fortement saillante, parfois plus que celle focalisée en dernier. Dans les critères syntaxiques, nous noterons également les répétitions. Il est clair en effet qu'un mot répété en devient saillant, ce phénomène n'étant pas rare dans le dialogue homme-machine, comme nous avons pu le constater dans notre étude du corpus Magnét'Oz. Un troisième critère, cette fois plus rare, est la présence d'une symétrie. Ainsi, la figure de style qu'est le chiasme ou encore les constructions comportant le rappel symétrique d'un terme tendent à rendre saillant ce terme. Un exemple repris dans (Stevenson 2002) montre que l'information marquée par une symétrie comme « *deer* » dans « *deer in deer hunting* » est plus saillante que la même information dans une configuration moins marquée (« *deer in hunting deer* »).

- La saillance grammaticale, c'est-à-dire liée aux fonctions grammaticales des mots :
  La notion d'actant recouvre les éléments sous-catégorisés par le verbe de l'énoncé, donc par exemple le sujet, le complément d'objet direct et le complément d'objet indirect. Cette notion reste large et classer ces éléments s'avère nécessaire pour identifier l'élément saillant. Le sujet grammatical est souvent considéré comme l'élément le plus saillant. Il se trouve généralement au début et renforce ainsi la saillance liée à cette place. Les constructions passives, qui permettent d'inverser les fonctions grammaticales, se justifient ainsi. La Théorie du Centrage (Grosz *et al.* 1995) propose une hiérarchie des fonctions syntaxiques : sujet, puis complément d'objet direct, puis complément d'objet indirect. Une autre fonction grammaticale, moins fréquente mais constituant un facteur quasiment explicite de saillance linguistique, est la fonction vocative. Lambrecht (1996) montre par exemple que les propriétés des groupes nominaux vocatifs se rapprochent de celles des anaphores et des cataphores nulles (« *devrait être là, le triangle rouge* »), renforçant ainsi la saillance de leur antécédent.

L'approche de Lappin & Leass (1994) récapitule un peu tous ces critères formels. Les auteurs présentent un algorithme de résolution des anaphores faisant intervenir un calcul de saillance des groupes nominaux. Le principe de l'algorithme est de partir d'un seuil initial et de le faire varier en fonction de différents facteurs syntaxiques. La figure 5.4 présente les valeurs numériques utilisées, montrant l'importance à la fois du sujet grammatical et de la récence (les critères syntaxiques et grammaticaux sont mélangés).

Des critères liés au sens de l'énoncé. C'est quand nous abordons la sémantique que nous considérons les notions de thème, de focus ou encore de topic. Nous les présentons ici en distinguant celles relevant de la sémantique du mot ou du groupe de mots, celles relevant de la sémantique de la phrase ou de l'énoncé, et celles relevant de la sémantique de la conversation,

Type de facteur	Seuil initial de saillance
récence phrastique	100
emphase sur le sujet	80
emphase existentielle	70
objet direct	50
objet indirect	40
tête d'un groupe nominal	80
emphase non adverbiale	50

Figure 5.4 – *Saillance linguistique dans* (Lappin & Leass 1994).

c'est-à-dire se construisant au cours du dialogue :

- La saillance liée à la sémantique des mots :
  Les composants de l'énoncé se caractérisent par un rôle thématique : l'entité qui fait l'action est l'agent, l'entité qui la subit le patient. Dans « *le triangle rouge cache le bleu* », l'agent est « *le triangle rouge* ». Dans « *le triangle bleu a disparu* », « *le triangle bleu* » ne fait pas l'action : on dit qu'il s'agit du thème. Dans la suite, nous assimilerons thème et agent. Classiquement, par exemple dans (Sidner 1979) ou dans la Théorie du Centrage (Grosz *et al.* 1995), l'agent est considéré comme plus saillant que le patient, lui-même considéré comme plus saillant que les autres rôles thématiques. Ces derniers, comme par exemple l'instrument, sont moins fréquents et nous les ignorerons dans la suite. Il n'y a cependant pas unanimité à propos de cette hiérarchie. Dans des travaux plus récents et basés sur des expérimentations, (Stevenson *et al.* 1994) montrent qu'entre agent et patient, la préférence est significativement pour le patient, du moins dans certaines phrases. Ainsi, pour les phrases qui décrivent un événement, les conséquences de l'événement sont plus présentes dans les représentations mentales que les conditions initiales. Ces conséquences s'appliquant au patient, celui-ci en devient plus saillant que l'agent. Pour les phrases qui ne décrivent pas d'événement, tout dépend des composants de la phrase et rien ne peut être conclu. Pour déterminer quelle est l'entité saillante entre agent et patient, il faudrait détailler chaque type d'action, donc chaque type de verbe, en étudiant sa sémantique. Le problème s'avère complexe, du fait de la multiplicité des critères qui entrent en jeu. (Stevenson *et al.* 1994) ainsi que (Pearson *et al.* 2001) détaillent par exemple le cas des verbes de transfert (donner quelque chose à quelqu'un) et montrent que le receveur est plus saillant que le donneur et que l'objet transféré. Nous retiendrons de tout cela qu'un calcul rigoureux de la saillance thématique est pour l'instant quasiment impossible.

- La saillance liée à la sémantique de l'énoncé :
  Au niveau à la fois de la syntaxe et de la sémantique, le thème correspond à ce dont l'énoncé parle (souvent le sujet syntaxique, souvent en position initiale) et le rhème à ce qui en est dit (souvent le prédicat). Le mot « thème » est ici employé dans un sens très différent des rôles thématiques. Contrairement à la distinction entre connu et nouveau que nous verrons plus loin, la distinction entre thème et rhème concerne le repérage de l'énoncé par le locuteur et non par l'interlocuteur. Selon Caron (1989), la séquence la plus naturelle de deux phrases est celle où le rhème de la première phrase est repris comme thème de la seconde. Le rhème est repris et est donc saillant. Quant au thème de la nouvelle phrase, il est également saillant. Caron conclut qu'il est impossible de savoir si c'est le thème ou le rhème

qui est l'élément le plus saillant dans un énoncé. Une autre distinction concernant la prise en charge de l'information par le locuteur est la distinction entre présupposé et posé : une information présupposée, donc implicite, peut s'avérer aussi saillante qu'une information explicite. Encore une fois, dans l'état actuel des recherches, il n'est pas possible de trancher.

- La saillance liée à la sémantique de la conversation :
  Les équivalents du thème et du rhème au niveau de la conversation sont les notions de topic et de commentaire : à un niveau stylistique, le topic est l'entité dont parle la conversation et le commentaire est ce qui en est dit. Cette différence entre thème et topic est classiquement admise, comme le montre Wolters (2001). Il n'en reste pas moins que les concepts sont vagues et peuvent donner lieu à différentes formulations et formalisations. Wolters évoque par exemple les structures complexes de Chafe, comprenant des « supertopics » et des « subtopics ». Le topic peut recouvrir plusieurs entités et il s'avère ainsi difficile de faire un lien entre topic et saillance. Une autre distinction au niveau de la conversation est celle entre donné (*given*) et nouveau (*new*). Nous nous plaçons cette fois du point de vue de l'interlocuteur. Le donné est déjà connu par l'interlocuteur. D'un point de vue psychologique, il correspond ainsi au point de départ dans le processus d'interprétation (Wolters 2001). Il peut également désigner ce que l'interlocuteur juge pertinent dans le message communiqué, ou encore ce qui permet de lier l'énoncé courant au sujet de la discussion. Ici aussi rien ne permet de conclure quant à la saillance : le donné (ou connu) est saillant parce qu'il est bien présent dans l'esprit des interlocuteurs, le nouveau est saillant justement parce qu'il est nouveau, parce qu'il peut orienter la conversation dans une nouvelle voie.

- La saillance indirecte ou transfert sémantique de saillance :
  Pour compléter cette liste de critères de saillance sémantique, il nous reste à mentionner la saillance indirecte qui s'applique aussi au niveau du sens, à propos des thèmes et des topics. Stevenson (2002) note ainsi qu'un référent très lié au topic est plus saillant qu'un référent qui ne lui est pas apparenté. Cette idée se trouve par exemple dans (Garrod & Sanford 1988) et dans (Marslen-Wilson *et al.* 1993).

Pour récapituler, tout reste très flou et non opérationnel. Les distinctions classiques ne donnent pas de critère permettant d'identifier de manière automatique et systématique l'élément saillant dans un énoncé ou au cours de la conversation. Même la sémantique des mots, *a priori* la plus simple et la plus formalisable, aboutit à deux hiérarchies contraires pour les rôles thématiques.

Des critères liés à la subjectivité. Nous avons mentionné les notions de récence, ainsi que de connu et de nouveau, sans considérer les facteurs cognitifs concernés par ces notions. Parmi ces facteurs, la mémoire joue un rôle important. Comme nous l'avons vu dans le chapitre 3, la mémoire à court terme se limite à sept éléments, ce qui diminue d'autant les possibilités de saillance : nous en déduisons qu'il est inutile de chercher un focus, un thème ou un topic au-delà de sept entités du discours. La familiarité intervient également dans l'affectation de saillance aux entités du discours. Une entité peut être nouvelle, ou au contraire déjà évoquée ou déjà traitée et donc familière. Chacun de ces cas a des conséquences sur la saillance linguistique des expressions référentielles. Mémoire et familiarité dépendent de l'utilisateur et montrent en quoi la saillance est subjective. Du fait de cette subjectivité, il s'avèrerait utile, pour une modélisation de la saillance dans la communication, de distinguer saillance pour le locuteur et saillance pour l'interlocuteur. Pour l'instant, nous ne suivrons pas cette voie, préférant nous limiter aux aspects

structurels liés à la forme de l'énoncé.

Le calcul de la saillance linguistique dans notre approche. Nous avons vu en § 5.1.3 que, dans notre approche, la saillance constitue un critère de différenciation supplémentaire dans une partition d'un domaine de référence. Son calcul se fait donc au moment de la construction d'un domaine. Pour pouvoir être calculée par le système de dialogue, elle doit s'appuyer sur des données formelles, d'une part le résultat de l'analyse syntaxique de l'énoncé courant, d'autre part l'historique du dialogue et sa structuration en domaines de référence. Nous avons détaillé dans le chapitre 4 deux types de facteurs de groupement pour la construction de domaines de référence linguistiques. Il nous reste ici à spécifier l'intervention de la saillance pour ces deux types.

Dans le cas d'un groupement sur le nombre, l'énumération ou la coordination, nous considérons qu'aucun des éléments ne reçoit une focalisation particulière. Aucune saillance linguistique n'a besoin d'être calculée, et le domaine de référence obtenu ne comprend pas de critère de différenciation. Nous nous appuyons sur l'argument de Kleiber (1986) à propos des possibilités de reprise des éléments d'un tel groupement : autant la reprise de tous les éléments du groupement est possible, autant la reprise d'un seul élément, comme « *mets-le sur la droite* » après « *prend un triangle et un carré* », est impossible. Si l'un des éléments était saillant, une telle reprise serait possible. Aucun élément n'est donc plus saillant qu'un autre.

Dans le cas d'un groupement sur la participation à un même événement, nous faisons intervenir les critères liés à la forme de l'énoncé pour calculer la saillance des actants et constituer une partition focalisant le plus saillant. Le principal critère est la fonction grammaticale, et nous suivons la hiérarchie de la Théorie du Centrage pour privilégier le sujet au complément d'objet direct. Nous considérons également la place stratégique dans l'énoncé, en privilégiant le début par rapport au reste. Nous tenons compte des constructions dédiées à la saillance, ainsi qu'à la présence d'un accent tonique. Nous suivons le même principe que pour le calcul de la saillance visuelle : pour chaque expression référentielle, tous les paramètres retenus sont évalués avec des 0 et des 1. Les vecteurs normés obtenus ont des valeurs entre 0 et 1, valeurs que nous comparons pour identifier l'entité saillante. Comparons par exemple « *le triangle rouge se met à côté du bleu* »

A.	« *le triangle rouge* » dans « *le triangle rouge se met à côté du bleu* »	
B.	« *le triangle rouge* » dans « *c'est le triangle rouge que tu dois mettre à côté du bleu* »	
	cas A	cas B
accent tonique :	1	0
construction syntaxique dédiée :	0	1
place stratégique en début d'énoncé :	1	1
fonction grammaticale sujet :	1	0
saillance :	**0.75**	**0.5**

Figure 5.5 – *Scores numériques pour un premier calcul opérationnel de la saillance linguistique.*

avec « *c'est le triangle rouge que tu dois mettre à côté du bleu* ». Les deux énoncés conduisent à la construction d'un domaine de référence regroupant un triangle rouge et un triangle bleu, le facteur de groupement étant la participation à un même événement. En supposant que « *rouge* » est accentué dans le premier énoncé et en retenant les quatre paramètres présentés dans la figure 5.5, le vecteur pour « *le triangle rouge* » s'avère de 0,75 alors que celui correspondant au

triangle bleu est nul. L'écart n'est que de 0,5 dans le deuxième énoncé. Notons que cette méthode fondée sur l'identification d'un ensemble de paramètres et sur l'attribution de 0 et de 1 (ou de + et de −) est souvent utilisée en linguistique computationnelle. Un exemple remarquable pour la résolution des anaphores est celui de (Beaver 2003).

Avec les classifications de facteurs proposées, nous avons montré que la saillance est un concept complexe, difficile à cerner et à formaliser. Nous considérons que nos ébauches à base de scores numériques constituent un point de départ formel pour le calcul de la saillance dans le cadre du modèle des domaines de référence. Nous sommes néanmoins conscient que ce point de départ est simple et que les recherches dans ce domaine sont loin d'être terminées. Nous espérons que notre apport, et en particulier le rapprochement entre caractérisation de la saillance visuelle et caractérisation de la saillance linguistique, incitera à de nouvelles directions de recherche. Dans le chapitre 10 où nous présenterons des spécifications de protocoles expérimentaux, nous reviendrons sur les possibilités de validation de notre conception de la saillance.

RÉCAPITULATIF
Ce chapitre s'est attaché à clarifier la notion de saillance et à en proposer des caractérisations formelles. Compte tenu du flou dans les travaux existants et de la difficulté dans la représentation de cette notion, notre principal apport a été de nous situer à un niveau plus global que la saillance gestuelle, visuelle ou linguistique, pour appréhender les différents aspects structuraux et sémantiques de cette notion, et pour les appliquer ensuite aux saillances spécifiques aux modalités. C'est ainsi que nous avons utilisé les critères identifiés lors de la caractérisation de la saillance visuelle pour celle de la saillance linguistique et inversement. Cette méthodologie et les résultats obtenus, ainsi que les rapports que nous avons identifié entre saillance et geste et entre saillance et domaine de référence, nous permettent de conclure que la saillance est non seulement un phénomène global permettant d'intégrer les modalités, mais également un point de départ pour l'interprétation.

Chapitre 6

Le parcours de domaines

Nos représentations mentales sont-elles ordonnées ? Les objets visuellement perçus, les demonstrata et les entités du discours sont-ils toujours ordonnés ? Quels sont les facteurs d'ordonnancement ? Ces facteurs sont-ils formalisables ? Un système de dialogue peut-il les calculer dans le but d'améliorer ses capacités d'interprétation ?

Nous nous intéressons dans ce chapitre au critère d'ordonnancement caractérisant certaines partitions de domaines de référence, le but étant d'interpréter les actions de référence fondées sur un ordonnancement. Nous analysons tout d'abord (§ 6.1) la nature de ces actions et les conséquences dans notre modèle d'interprétation. Nous détaillons ensuite (§ 6.2) les facteurs d'ordonnancement pour la modalité de support qu'est la perception visuelle et pour les modalités d'interaction que sont la parole et le geste.

6.1 Ordonnancement et référence

6.1.1 Recours à l'ordonnancement lors de la référence aux objets

Ordonnancement et objet de l'interaction. Dans notre cadre de la référence aux objets, un ordonnancement intervient lorsqu'on se focalise sur les objets de l'application selon un ordre particulier. Cet ordre peut être induit par la tâche applicative, ou dépendre des caractéristiques de l'intention et de l'attention de l'utilisateur. Nous avons vu que l'attention s'appliquait à un ensemble d'objets ou contexte. Le traitement de ces objets peut suivre un certain ordre, et des changements de contexte peuvent survenir. Elargissement à un sur-contexte, focalisation dans un sous-contexte, passage à un autre contexte totalement disjoint : les phénomènes sont larges. Ils le sont encore plus lorsque nous considérons les possibilités en termes de domaines de référence : changement de critère de différenciation, changement de partition, changement de domaine de référence dans une même source contextuelle, changement de domaine de référence avec passage dans une autre source contextuelle. Ces phénomènes semblent surtout chaotiques et donc imprévisibles. Le sont-ils vraiment ? Beaucoup de travaux ont tenté de répondre à ce problème en proposant des modèles de contraintes pour restreindre les changements possibles. Certains explorent les contraintes provenant de mécanismes cognitifs généraux, d'autres de connaissances encyclopédiques et de connaissances liées à la tâche applicative.

Dans les mécanismes cognitifs généraux, nous décrirons brièvement les *schémas* et les *cadres*, d'après la présentation et l'analyse qui en est faite dans (Caron 1989). Les schémas sont des organisations très générales en fonction desquelles se structure la mémoire. Le but est de modéliser les régularités dans les activités cognitives. Des tentatives ont été faites pour les formaliser, du moins dans un cadre limité comme celui de la narration, mais ils restent très vagues et nous semblent trop éloignés de notre approche. Les cadres (ou frame) proposés par Minsky représentent également certaines régularités cognitives, en incluant cette fois celles qui gouvernent l'analyse d'une scène visuelle. Un frame est une structure représentant une situation habituelle avec des cases pour chacun de ses éléments. Les cases sont remplies par les données perceptives disponibles. Si celles-ci sont absentes, des valeurs par défaut sont assignées. A chaque frame sont attachées des règles d'utilisation et de transformation.

Dans les contraintes liées aux connaissances encyclopédiques et aux connaissances liées à la tâche applicative, nous évoquerons les *scripts* et les *plans*. Les scripts, proposés par Schank et Abelson (cités par Caron 1989) sont des structures décrivant une séquence stéréotypée d'événements, l'exemple classique étant la succession d'actions que l'on fait quand on va au restaurant. Un script est activé dès la mention d'un élément, comme la table ou un serveur. Dans le cadre de la compréhension automatique, le but est de prédire l'action que l'utilisateur va entamer, du moins lorsque la tâche est suffisamment contrainte pour que ces successions d'actions implicites soient saillantes. La notion plus générale de plan fait intervenir un but avec un ensemble hiérarchisé de sous-buts. Un plan pourra entraîner la mise en œuvre d'un ou plusieurs scripts.

Dans le cadre de l'expérimentation Magnét'Oz et de sa tâche applicative consistant à ranger des objets dans des boîtes appropriées, nous pouvons identifier des régularités qui ressemblent à des scripts et que nous désignerons par le terme « ordonnancement ». Nous remarquons que les sujets suivent deux grandes stratégies : celle qui consiste à considérer un type d'objets et à ranger tous les objets de ce type, et celle qui consiste à ranger les objets en commençant par une zone spatiale (par exemple la gauche de l'écran) et à se décaler progressivement vers les autres zones. Ces deux stratégies induisent un ordonnancement, et prendre en compte cet ordonnancement facilite grandement le processus d'interprétation. Par exemple, dans le cas du traitement des objets de gauche à droite, il est ainsi possible d'interpréter des expressions référentielles telles que « *le suivant* ». Dans notre exemple d'aménagement d'un intérieur, nous pouvons également identifier des ordonnancements. Ainsi, on commence généralement par placer la table avant de placer les chaises qui lui sont associées. Nous faisons l'hypothèse que dans la majorité des cas la spécification d'une application de dialogue homme-machine inclut la définition de tels ordonnancements. L'exemple de la figure 7.3 page 147 viendra confirmer cette hypothèse dans le cadre d'un sous-but précis, de même que les définitions des scénarios dans le projet MIAMM.

Ordonnancement et expression référentielle. Un ordonnancement dans les actions de l'application correspond à un ordre dans les objets concernés par ces actions, et donc à un ordre dans les références aux objets. Si cet ordonnancement peut être détectable, notre modèle doit pouvoir en tenir compte. Il en acquerra une meilleure prédictibilité. La modélisation que nous proposons consiste à déclencher une recherche d'ordonnancements possibles lors de l'interprétation d'expressions dénotant un rang : « *le premier* », « *le second* », « *le suivant* », « *le précédent* », « *l'avant-dernier* », « *le dernier* ». L'ordonnancement est cherché en priorité dans un domaine de référence stable, déjà délimité. Il est cherché ensuite dans le contexte correspondant à la scène visuelle, qui nous semble privilégié pour une interaction basée sur ce support visuel. Nous verrons en effet que les critères d'ordonnancement sont nombreux dans la perception visuelle, alors qu'ils sont plus réduits dans le langage et le geste.

6.1.2 L'ordonnancement dans le modèle des domaines de référence

Ordonnancement et facteur de groupement. Nos domaines de référence présentent des similitudes aussi bien avec les frames qu'avec les scripts : il s'agit de structures avec des cases qui se remplissent au fur et à mesure de l'interaction, et ils constituent un moyen de formaliser des contraintes applicatives. La justification d'un domaine de référence résidant dans le facteur de groupement, nous passons ici en revue les facteurs de groupement possibles pour déterminer ceux qui entretiennent un lien quelconque avec un ordonnancement.

Ainsi, lors de la construction de domaines visuels, il apparaît que seul le critère de continuité induit un ordonnancement. La détection d'un alignement ou d'une disposition régulière d'objets doit donc s'accompagner d'un marquage d'ordonnancement. Ainsi, cinq objets disposés selon une ligne régulière seront marqués de 1 à 5, en commençant par exemple par la gauche qui correspond au sens de lecture dans la culture européenne. Dans le domaine visuel considéré, ces objets seront réunis dans une partition marquée du critère d'ordonnancement « sens de lecture ». Si les objets sont disposés selon un cercle parfait et que cette régularité est détectée pour constituer un domaine sur le critère de bonne continuité, le premier rang est attribué à n'importe lequel de ces objets (au plus saillant si l'un se démarque d'entre les autres malgré la régularité) et le critère d'ordonnancement inclut un indice pour conserver cette particularité. Ainsi, un tel ordonnancement ne sera exploité que pour l'interprétation de « *le suivant* » ou de « *le précédent* », et en aucun cas pour celle de « *le premier* » ou « *le second* ».

Lors de la construction de domaines linguistiques, aucun des critères que nous avons étudiés ne semble fondamentalement lié à un ordonnancement. Nous avons vu en effet qu'une coordination, par nature, ne conduit pas à un ordonnancement des éléments coordonnés. Nous constatons également que la hiérarchie selon la fonction grammaticale n'a aucun rapport avec un ordonnancement autorisant l'utilisation d'expressions dénotant un rang. Seul l'ordre d'énonciation conduit directement à l'identification d'un rang. C'est sur ce critère que s'appuient les références mentionnelles : dans « *prends le triangle rouge et le triangle bleu* » suivi de « *mets le premier ici et efface le second* », « *le premier* » réfère clairement au triangle rouge et « *le second* » au triangle bleu. Nous en déduisons pour notre modèle que les expressions référentielles de chacun des énoncés peuvent être marquées d'un rang si besoin est.

Ordonnancement et saillance. Il n'y a pas équivalence entre un ordonnancement tel que nous le concevons désormais et les hiérarchies de saillance dont nous avons parlé dans le chapitre précédent. Même si nous avons vu comment évaluer la saillance de chaque objet et donc comment classer les objets du plus au moins saillant, des contraintes supplémentaires font que ce classement ne peut pas constituer un ordonnancement. Un exemple de telles contraintes supplémentaires est une ligne de force dans l'image. En effet, une ligne de force contraint le parcourt du regard dans la scène. Même si elle débute sur l'objet le plus saillant de la scène, elle possède ses propres caractéristiques qui la font parcourir certains objets dans un certain ordre indépendant du classement par saillance. Une ligne de force peut se terminer sur un objet très saillant, même après être passée par des objets insignifiants. Au niveau du langage, l'ordre d'énonciation peut de même commencer avec le sujet grammatical saillant, sans pour autant suivre obligatoirement les composants selon leur classement dégressif de saillance. C'est juste une possibilité.

6.2 Caractérisation des critères d'ordonnancement

Comme nous l'avons déjà fait pour la saillance, nous présentons dans cette section les critères que nous avons retenus pour une caractérisation formelle des phénomènes d'ordonnancement, le

but étant qu'un système de dialogue homme-machine puisse calculer ces critères à partir des données dont il dispose sur la scène visuelle et sur l'historique de l'interaction.

6.2.1 Ordonnancement par la perception visuelle

Considérations méthodologiques. A la suite de (Vettraino-Soulard 1993), nous faisons l'hypothèse que l'ordre de perception des objets dans une scène a des répercussions sur la représentation mentale de cette scène et sur l'énonciation qui en découle. Ainsi, si le regard tend spontanément à passer du triangle rouge au triangle bleu pour finir sur le vert, des expressions référentielles telles que « *le premier* », « *le second* » ou « *le troisième* » pourront se fonder sur un tel ordonnancement. Reste à prouver que le parcours du regard peut être si prévisible, et à le caractériser.

Il peut sembler aléatoire de se lancer dans de telles recherches. Nous tenons cependant à rappeler l'importance du chemin parcouru par le regard dans tout le processus de perception visuelle. Les exemples du vase de Rubin, du canard–lapin de Wittgenstein et de toutes les images ambiguës classiques qui ont suivi nous montrent que même la reconnaissance des objets dépend du parcours du regard : si l'on regarde l'image en commençant par telle zone visuelle plutôt que par telle autre, on perçoit tel motif plutôt que tel autre. Nous avons distingué dans le chapitre 3, à propos de notre positionnement face à la sémiotique, l'aspect syntagmatique et l'aspect paradigmatique intervenant lors de la lecture d'image, donc lors de la perception ordonnée des objets. Ces deux axes interviennent même dans nos exemples de scènes épurées comportant uniquement des formes géométriques. En effet, même des triangles, si abstraits soient-ils, peuvent avoir une configuration spatiale rappelant une configuration classique et donc susceptible d'un certain effet dans le parcours du regard. Autrement dit, il semble que le parcours du regard fasse systématiquement invervenir un aspect syntagmatique et un aspect paradigmatique. De ce fait, ajouté au fait que les principaux travaux sur ce sujet partent de l'étude d'une seule image, il paraît difficile de caractériser le parcours du regard.

Certains résultats des études en lecture d'image sont cependant intéressants, en particulier ceux de (Vettraino-Soulard 1993). A partir de l'hypothèse que l'ordre d'énonciation correspond à l'ordre de perception, l'auteur met en œuvre une expérimentation avec un protocole basé sur la perception d'une image de publicité et l'énonciation de ses composants. Suite à une analyse très détaillée des données enregistrées, Vettraino-Soulard conclut qu'en considérant l'objet comme unité, les parcours du regard suivent une trop grande diversité et ne peuvent être modélisés. Quelques régularités sont observées sur les trois premiers objets, mais rien de significatif. Par contre, en considérant le groupe perceptif d'objets comme unité, les parcours du regard suivent certaines régularités. Un tel résultat nous intéresse car il donne une nouvelle dimension aux domaines de référence visuels, et en particulier à ceux construits à l'aide du critère de proximité de la Gestalt. En effet, ceux-ci correspondent à des groupes perceptifs et il devient alors envisageable de modéliser le parcours du regard en termes d'ordonnancement de domaines visuels. Si Vettraino-Soulard se limite aux aspects morphologiques, c'est-à-dire ceux fondés sur la répartition des éléments du message visuel, à savoir l'alternance de lignes, de surfaces, de volumes, de pleins et de vides, nous tenterons d'y ajouter quelques aspects paradigmatiques.

Des critères syntagmatiques (ou morphologiques). Dans la disposition des objets, une scène peut présenter des symétries, des régularités, des alignements qui vont contraindre les parcours possibles du regard. Plus un alignement d'objets est régulier, plus il y a de chances pour qu'un ordre soit perçu par l'utilisateur. Dans les scènes visuelles de Magnét'Oz, le bas de l'écran est occupé par une série de boîtes (se reporter à la figure 1.3 page 24). Les boîtes sont disposées de

manière parfaitement régulière et les sujets n'hésitent pas à utiliser des expressions référentielles telles que « *la première boîte* », « *la deuxième boîte* » ou « *la dernière boîte* ». En quelque sorte, cet alignement de boîtes porte le regard de l'utilisateur, l'empêchant par exemple de s'arrêter à la troisième boîte si ce n'est pas intentionnel. Il constitue donc une ligne de force (appelée aussi ligne de guidage ou ligne directrice) dans la scène.

Quand elles sont fortes, les lignes directrices prennent le pas sur le sens de lecture par défaut, c'est-à-dire de gauche à droite (et de haut en bas) dans la culture européenne. Le regard parcourt l'image en suivant ces lignes : il s'arrête sur les objets placés sur la ligne mais pas sur les autres sujets qui nécessiteraient un détour du regard. D'après les études faites dans le domaine des arts picturaux, par exemple par Itten, Kandinsky ou Klee (cf. Cocula & Peyroutet 1989), les lignes les plus fortes sont celles qui partent, passent et aboutissent sur des points forts, et, s'il n'y en a pas suffisamment, sur les quatre emplacements forts classiques (c'est-à-dire les intersections des tiers du cadre, comme nous l'avons vu dans le chapitre 5). Parmi ces lignes, les diagonales sont plus fortes que les horizontales et les verticales. Les critères qui interviennent peuvent être hiérarchisés de la manière suivante :

1. Le parcours du regard est suggéré par le sujet même. Dans les scènes comportant un être humain ou un avatar, le regard de l'observateur suit la direction du regard de la personne observée. C'est de ce constat qui provient le conseil en photographie consistant à dégager le regard du sujet. De même, dans les scènes comportant un objet caractérisé par une direction de visée, comme par exemple un fusil, le regard suit la direction de cet objet.
2. Le regard passe d'un objet saillant à un autre objet saillant. Les lignes directrices correspondantes sont avant tout les diagonales qui passent par le maximum d'objets, saillants ou non.
3. Le regard suit la perspective lorsque celle-ci est particulièrement importante dans la scène. Il part des objets situés au premier plan pour aboutir au point de fuite.
4. Le regard suit le sens de lecture privilégié, c'est-à-dire celui de la lecture de texte.

Le parcours du regard peut également être cyclique, c'est-à-dire revenir sur une ligne directrice déjà parcourue. Lorsque la scène comprend deux objets particulièrement saillants, on peut assister à un phénomène d'aller-retour entre ces deux points. Le regard peut également s'arrêter soit sur un point fort, soit en sortant du cadre de l'image si une ligne directrice n'aboutit à aucun objet. L'arrêt sur un point fort est particulièrement exploité dans l'image publicitaire, compte tenu du fait que l'objet sur lequel s'arrête le regard est celui sur lequel l'attention va se porter. Cet objet a donc tout intérêt à être le produit à vendre. Dans nos exemples de scènes, il s'agit de l'objet le plus saillant.

Des critères paradigmatiques. Notre perception visuelle se construit au cours de notre vie : nous nous sensibilisons aux constructions fréquemment utilisées dans les arts picturaux ou dans les images publicitaires. Par exemple, à force de voir des tableaux et des photographies comportant le sujet principal non pas au milieu du cadre mais à un tiers, nous associons à cette place une importance particulière. De même, à force de voir des images qui dirigent notre regard en le faisant partir d'un objet saillant situé par exemple en haut à gauche pour l'amener jusqu'au sujet principal, nous acquérons cette particularité dans notre vision des choses. Dans le cadre du dialogue homme-machine, il s'avère ainsi intéressant d'exploiter ces régularités de perception. Le principe est de doter le système d'une connaissance des constructions classiques, en termes de structures de lignes directrices classiques, et d'un algorithme d'identification de ces constructions pour n'importe quelle scène visuelle. Ainsi, dans le cas où la scène courante repose sur une

construction classique liée à des lignes directrices précises, le système peut exploiter sa connaissance de ces phénomènes pour déterminer un ordonnancement possible d'objets ou de domaines de référence visuels. Dans le cas où la scène courante ne présente pas une telle structure, le système n'identifie pas d'ordonnancement privilégié.

Des critères liés à la subjectivité. La sensibilisation de l'individu aux constructions picturales classiques a un corrolaire : la perception d'une image varie d'une personne à l'autre. Comme le conclut Vettraino-Soulard (1993), un modèle prédictif universel du parcours du regard est impossible, compte tenu de la multitude de facteurs personnels qui interviennent. La majorité de ces facteurs, à savoir les facteurs culturels (éducation, histoire, religion), oniriques, émotionnels, etc., ne sont tout simplement pas formalisables. Dans ce qui précède, nous nous sommes limité aux facteurs les plus importants et les plus objectifs. Beaucoup de recherches restent à faire avant de pouvoir aller plus loin dans la modélisation.

La détection des lignes directrices dans notre approche. En attendant un élargissement des recherches et le développement d'outils informatiques pour cette problématique[1], nous posons ici les bases de notre modélisation, sachant qu'il s'agit de propositions non implantées. Pour la détection par le système des lignes directrices à l'aide de critères morphologiques, nous proposons la modélisation suivante :

1. Retenir les lignes identifiées lors de la construction des groupes perceptifs sur le critère de bonne continuité (il s'agit d'un ordonnancement sur des objets) ;
2. Après avoir repéré les objets saillants et, si ceux-ci sont trop nombreux, retenu les trois ou quatre plus saillants, après avoir repéré les groupes perceptifs construits sur la proximité qui contiennent ces objets, retenir les lignes joignant ces groupes perceptifs (il s'agit d'un ordonnancement sur des domaines de référence visuels).

Pour la détection de lignes directrices à l'aide de critères paradigmatiques, nous proposons une modélisation consistant à comparer les coordonnées des objets de la scène avec celles de configurations classiques préalablement enregistrées dans une bibliothèque de scènes, et à utiliser un algorithme classique de reconnaissance de formes pour l'identification de la configuration de la scène courante.

6.2.2 Ordonnancement par les modalités d'interaction

L'ordre d'énonciation. Un ordonnancement immédiat dans les énoncés oraux et multimodaux est celui de l'énonciation. Pour un énoncé oral, il s'agit de l'ordre d'apparition des constituants. Pour un geste ostensif, il s'agit de l'ordre dans lequel les demonstrata sont désignés, du moins si un tel ordre existe. En effet, seuls les entourages et les ciblages semblent pouvoir ordonner leurs demonstrata. Si le fait est clair pour un ciblage, il s'avère plus délicat pour un entourage : un entourage est fait dans un sens précis, et les objets proches de la trajectoire peuvent être ordonnés selon ce sens. Dans la figure 2.6-B page 53, en considérant que le sens de la trajectoire est le sens horaire, les objets `d`, `c` et `b` peuvent être classés selon cet ordre. Si les objets inclus dans l'entourage sont plus nombreux, déterminer un ordre semble très aléatoire. Par conséquent, nous

1. Plusieurs pistes nous semblent intéressantes à moyen et long terme : premièrement l'expérimentation à l'aide d'appareils captant le parcours du regard (*eye-trackers*), deuxièmement une coopération interdisciplinaire entre notre approche et celle de la vision artificielle pour confronter nos hypothèses théoriques avec des algorithmes analytiques, statistiques ou connexionnistes, troisièmement l'élargissement des études en psychologie cognitive (et computationnelle) sur la perception visuelle et la mémoire de travail pour une meilleure considération des phénomènes de parcours du regard.

considérons que l'utilisateur qui veut ordonner les demonstrata utilisera préférentiellement un ciblage. Nous n'attribuons donc un ordre que pour les ciblages. Rappelons que nous avions déjà exploité un tel ordre page 53 pour l'identification des demonstrata d'un ciblage ambigu. Bien entendu, dans le cas de plusieurs gestes successifs, l'ordre des gestes, qui correspond d'ailleurs à l'ordre d'énonciation des expressions référentielles associées, est également retenu.

Un cas particulier. Parmi les expressions référentielles que nous avons considérées, c'est-à-dire les expressions fondées sur une marque d'ordonnancement comme « *le premier* », « *le second* », « *le suivant* » ou « *le dernier* », la seule contrainte était d'identifier un ordonnancement préalable pour interpréter la référence dans cet ensemble ordonné. Un cas particulier est donné par les expressions incomplètes telles que « *l'un ... l'autre* ». Il s'agit dans ce cas du marquage explicite d'un ordonnancement pour une liste déjà énumérée. Dans notre modélisation, rien ne change dans l'interprétation car l'ordonnancement est identifié de la même manière. Par contre, un repère sera conservé après l'interprétation de « *l'un* » pour indiquer la nécessité d'un parcours ultérieur. Ce n'est pas le cas pour l'interprétation de « *le premier* », expression qui n'est pas forcément suivie par une autre telle que « *le second* ».

Nous avons exploré dans ce chapitre le traitement d'expressions référentielles particulières, celles faisant appel à un ordonnancement des référents. Nous avons montré que de tels ordonnancements peuvent être implicites et reposer sur des contraintes perceptives, linguistiques ou applicatives difficiles à identifier. Nous nous sommes particulièrement focalisé sur l'ordonnancement des objets visibles lors de la perception visuelle de la scène. Nous proposons ainsi de modéliser l'appel à un ordonnancement par la prise en compte des caractéristiques suivantes, classées par ordre de priorité :

1. L'ordre d'énonciation des expressions référentielles verbales. Cet ordre est calculable directement par le système lors de la réception de l'analyse syntaxique de l'énoncé oral.
2. L'ordre de désignation des demonstrata. Cet ordre se détermine lors de l'analyse des trajectoires gestuelles.
3. Les lignes directrices identifiées lors du groupage selon la continuité. Leur détermination ne requiert pas d'algorithme particulier puisqu'elle s'appuie sur les résultats du groupage tel que décrit dans le chapitre 4.
4. Les lignes directrices fondées sur les domaines de référence saillants. Ici aussi, la détermination de ces lignes ne requiert pas d'algorithme particulier, puisqu'elle s'appuie sur les résultats du chapitre 4 pour la délimitation des groupes et sur ceux du chapitre 5 pour l'identification des objets et groupes saillants.
5. Les lignes directrices fondées sur le rappel d'une construction classique. Un algorithme particulier est cette fois nécessaire. Il s'agit d'un algorithme classique de reconnaissance de formes à partir d'une bibliothèque de situations préalablement enregistrées.

Comme nous l'avons remarqué pour la notion de saillance, la notion de ligne directrice reste très floue et peu étudiée dans les travaux en linguistique ou en psychologie computationnelle. Notre objectif est ici de proposer des bases à partir desquelles une modélisation globale du processus de communication puisse être envisagée. Bien que les notions de saillance et de ligne directrice ne soient pas complètement cernées, leur nature et leur rôle dans la communication nous semblent suffisamment clairs pour qu'une exploitation puisse en être faite dans notre proposition de modélisation, proposition que nous allons détailler dans le chapitre suivant.

RÉCAPITULATIF
Nous avions montré dans le chapitre 4 en quoi les domaines de référence s'avéraient importants lors de l'interprétation ; nous avions montré dans le chapitre 5 comment la saillance constitue le point de départ de cette interprétation ; nous avons montré ici qu'il existe des contraintes permettant de favoriser l'interprétation correspondant à un certain parcours des éléments d'un domaine de référence. Ces contraintes, lignes dirigeant le regard pour la perception visuelle et ordre d'énonciation pour les modalités d'interaction, sont exploitées à la demande, c'est-à-dire lors de l'interprétation d'expressions dénotant un rang. Nous avons montré qu'il était possible de formaliser la notion très vague de ligne directrice. Nous l'avons rattachée aux domaines visuels construits selon les critères de la Gestalt, adaptant en cela un résultat de Vettraino-Soulard.

Chapitre 7

La pertinence, un critère d'exploitation de domaines

Comment exploiter les notions que nous avons analysées dans les chapitres précédents? Comment exploiter les domaines de référence dont nous avons donné les principes de construction, de focalisation et de parcours? Que peut nous apporter la Théorie de la Pertinence? Pourquoi s'applique-t-elle à notre approche et en quoi ses notions d'effets contextuels et d'effort de traitement constituent des critères formalisables pour la communication multimodale?

Après trois chapitres détaillant la nature des domaines de référence, ce chapitre aboutit à la modélisation de leur fonctionnement. Nous avons vu comment plusieurs hypothèses de domaines étaient générées à partir de l'expression référentielle verbale et du geste produits; nous voyons ici comment exploiter ces hypothèses. Pour cela, le critère de pertinence avancé par la Théorie de la Pertinence nous semble particulièrement adéquat et intéressant. Nous commençons par détailler ce critère et son adaptation à notre approche (§ 7.1), pour présenter ensuite notre modèle d'interprétation de la référence aux objets (§ 7.2) et en tirer des propositions pour une formalisation générale de la pertinence dans la communication multimodale (§ 7.3).

7.1 Définition et adaptation à la multimodalité

7.1.1 Effets contextuels, effort de traitement et pertinence

Le contexte dans la Théorie de la Pertinence. A la base de la Théorie de la Pertinence, présentée dans (Sperber & Wilson 1995), se trouve la notion de double intentionnalité: le locuteur n'a pas seulement l'intention de transmettre un message, il a aussi l'intention de le transmettre de telle façon que son intention soit reconnue. Dans son interprétation, l'interlocuteur s'appuie ainsi sur l'hypothèse que le message est pertinent. Autrement dit il se fonde sur une présomption de pertinence. L'idée principale de la Théorie de la Pertinence est que l'interprétation fait intervenir deux types de mécanismes, des mécanismes de décodage du message et des mécanismes d'inférence. Ces inférences consistent à déduire du message décodé et du contexte des propositions nouvelles qui constituent l'objet de la communication. Le problème principal est de déterminer l'étendue et le contenu du contexte. En effet, c'est de ces deux paramètres que va dépendre la

production de propositions nouvelles.

Face au problème de la délimitation du contexte, Sperber et Wilson proposent de sélectionner dans l'ensemble des informations contextuelles, celles auxquelles le sens de l'énoncé donne accès. Il s'agit des informations liées aux composants de l'énoncé, ainsi que des connaissances encyclopédiques les plus accessibles. Si les propositions nouvelles inférées s'avèrent nulles ou insuffisantes, des informations contextuelles moins accessibles sont exploitées. Avec cette notion d'accessibilité, les auteurs définissent le contexte comme un équivalent de la mémoire à court terme (ou mémoire de travail).

Face au problème de la nature des informations contextuelles, ils proposent de définir le contexte comme un ensemble de propositions, ces propositions pouvant être vraies, probablement vraies, plutôt vraies, plutôt fausses, probablement fausses, ou fausses. Ces degrés sont appelés forces des propositions. Les informations portées par ces propositions sont connues du locuteur et de l'interlocuteur. Plus que cela, les deux interlocuteurs savent que ces informations leur sont connues, leur sont mutuellement manifestes. Comme le montrent Reboul et Moeschler (1998b), le contexte ainsi défini permet d'expliquer la réussite et l'échec de la communication, et permet de mettre en avant le critère de pertinence.

Deux définitions permettent de mieux appréhender les mécanismes d'inférence : celle d'une *contextualisation* et celle d'une *implication contextuelle*. Une contextualisation (de la proposition P dans le contexte C) est une déduction utilisant comme prémisses l'union d'informations nouvelles P et d'informations anciennes C. Une implication contextuelle (de la proposition P dans le contexte C) est un conclusion nouvelle qui ne serait pas dérivable à partir de P seul ou de C seul.

Les notions d'effets contextuels et d'effort de traitement. L'intérêt d'une inférence est de produire des effets. Sperber et Wilson définissent ainsi un effet contextuel comme un résultat du processus d'interprétation : si une contextualisation ne fait qu'ajouter au contexte l'information nouvelle (en totalité ou en partie) sans entraîner d'autres modifications du contexte, alors cette contextualisation n'a pas d'effet contextuel. Dans le cas contraire, l'un au moins des effets contextuels suivants est obtenu :

- effacement de certaines hypothèses du contexte,
- modification de la force de certaines hypothèses du contexte,
- dérivation d'implications contextuelles.

Les effets contextuels sont le produit de processus mentaux qui demandent un certain effort. L'interprétation nécessite ainsi un certain effort de traitement, qui correspond à la dépense d'énergie des processus mentaux activés. Selon les auteurs, l'effort de traitement dépend de la longueur de l'énoncé, de la facilité d'accès aux informations encyclopédiques, ou encore du nombre de règles logiques impliquées par le mécanisme déductif.

Avec ces deux notions, Sperber et Wilson (1995) définissent la pertinence de manière comparative. Ainsi, « une hypothèse est d'autant plus pertinente dans un contexte donné que ses effets contextuels y sont plus importants ». D'autre part, « une hypothèse est d'autant plus pertinente dans un contexte donné que l'effort nécessaire pour l'y traiter est moindre ». La pertinence peut ainsi être vue comme le rapport des effets contextuels sur l'effort de traitement. Avec nos préoccupations computationnelles, c'est de cette manière que nous l'aborderons et que nous tenterons de l'évaluer. Chaque effet contextuel nécessitant un supplément d'effort, les auteurs définissent également l'*effort impliqué par un effet*. Ce supplément d'effort n'est pas coûteux au point d'annuler la contribution de l'effet à la pertinence. Puisqu'il est toujours proportionné aux effets qui le rendent nécessaire, on peut l'ignorer dans l'évaluation de la pertinence.

Vers une évaluation de la pertinence. La Théorie de la Pertinence ne s'attache pas vraiment à évaluer les effets contextuels et l'effort de traitement, mais plutôt à décrire comment l'esprit évalue lui-même ses propres résultats et ses propres efforts, et comment il décide en conséquence de poursuivre ses efforts dans la même direction, ou au contraire de les diriger ailleurs. Il s'avèrerait pourtant très intéressant de pouvoir quantifier la pertinence d'un énoncé. On pourrait alors comparer les pertinences de plusieurs propositions, ou encore déterminer la forme la plus pertinente pour une intention communicative. Il s'agit cependant d'un problème psychologique général faisant intervenir de multiples facteurs.

On ne sait pas quelles sont les opérations élémentaires constitutives des processus intellectuels complexes, et on ne sait donc pas comment évaluer l'effort de traitement d'un énoncé. On sait que certains paramètres comme la durée d'un processus mental ne sont pas de bons indices : du fait de l'impossibilité de détecter une réflexion intense d'une réflexion détendue beaucoup plus lente, la durée d'interprétation n'apporte rien. Gazdar et Good, cités par Sperber et Wilson (1995), proposent d'évaluer les effets contextuels en comptant le nombre des implications contextuelles, et l'effort de traitement en comptant le nombre des opérations de déduction. Mais Sperber et Wilson montrent que compter chaque opération revient à ajouter une opération de comptage dans l'effort, et que si l'évaluation résultait d'un comptage, les sujets devraient être capables de porter des jugements absolus sur la quantité d'effet obtenu ou d'effort dépensé. Ce n'est pas le cas. Certains indices semblent cependant intervenir très fortement. Par exemple, plus une hypothèse est forte et plus ses effets contextuels risquent d'être importants. En ce qui concerne l'effort, plus le contexte comprend d'informations et plus l'effort de traitement est important.

Un autre argument œuvrant contre la calculabilité de la pertinence est son aspect subjectif : effets et effort dépendent de l'individu, de ses dispositions, de ses expériences, du contexte. Par exemple, le contexte visuel peut différer pour les deux interlocuteurs : lorsqu'une bouteille en cache une autre pour le locuteur, « *la bouteille* » sera une expression référentielle pertinente pour lui, alors que cette expression ne le sera pas pour l'interlocuteur qui, de sa position dans la pièce, voit les deux bouteilles.

Comme l'argumentent Sperber et Wilson, la pertinence s'avère pour le moins difficilement calculable. Quand elle est représentée mentalement, elle l'est sous la forme de jugements comparatifs ou éventuellement de jugements absolus vagues et généraux, et non sous la forme de jugements absolus fins et précis tels que le sont les jugements quantitatifs. Nous suivrons par conséquent une approche comparative dans notre exploitation de la pertinence.

7.1.2 Pertinence et référence dans le dialogue homme-machine

Du langage à la référence dans la communication multimodale. Nous venons d'évoquer un exemple de pertinence sur une expression référentielle. C'est effectivement l'approche que nous adoptons ici : calculer les effets contextuels et l'effort de traitement d'une expression référentielle, langagière ou multimodale, pour déterminer sa pertinence dans le contexte de communication.

Un premier problème réside dans la notion de contexte : y a-t-il équivalence entre le contexte tel que défini par Sperber et Wilson, et les sous-ensembles contextuels que sont les domaines de référence ? Il semble que non : le contexte dans la Théorie de la Pertinence se limite à des propositions traduisant des croyances ou des affirmations qui s'expriment par le langage. Le contexte tel que nous le concevons s'avère beaucoup plus large et hétérogène. Nous verrons que notre focalisation au problème de la référence suffit à relativiser ce problème : comme nous ne nous intéressons qu'aux actes de référence, nous n'avons pas besoin des contenus propositionnels des énoncés. Nous verrons également que tenir compte de connaissances extra-linguistiques lors de la détermination des effets contextuels et de l'effort de traitement, non seulement ne remet

pas en cause ces deux notions, mais de plus s'adapte parfaitement aux principes de la Théorie de la Pertinence. Si celle-ci a été élaborée pour un cadre linguistique, les principes qu'elle propose s'avèrent tout à fait valables pour un geste et pour une action de référence compte tenu du contexte visuel. En effet, on peut considérer qu'une trajectoire gestuelle porte en elle-même une présomption de pertinence quant à ses demonstrata. Les effets contextuels et l'effort nécessaire à son traitement se conçoivent aussi bien que pour un énoncé oral.

Un deuxième problème réside dans la confrontation entre saillance et pertinence : puisqu'elle autorise des expressions ambiguës, la notion de saillance ne remet-elle pas en cause la notion même de pertinence ? Nous verrons qu'il n'en est rien : lorsqu'on dit qu'une expression est ambiguë, c'est par rapport à un contexte donné, par exemple au contexte correspondant à la scène visuelle complète dans le cas de l'expression « *le N* » quand plusieurs objets de type N sont visibles. Or cette expression n'est plus ambiguë dans le contexte correspondant à un domaine saillant comprenant un seul objet de type N. Elle peut même être pertinente par rapport à ce contexte, du moins à partir du moment où ce contexte est identifiable facilement.

Un troisième problème est lié à la formalisation de la pertinence pour le dialogue homme-machine : la pertinence ne fait-elle pas intervenir trop d'implicite, impossible à identifier en compréhension automatique, impossible à vérifier ? Le problème reste ici ouvert : en calculant les effets contextuels en fonction des possibilités offertes et l'effort de traitement en fonction de la complexité de l'énoncé et du contexte multimodal, nous interprétons ces notions et nous construisons notre propre modèle de pertinence. Le processus de compréhension de la machine fondé sur ce modèle de pertinence peut s'avérer quelque peu différent du processus de compréhension humain. La différence principale est sans doute que pour calculer la pertinence d'une expression, le système de dialogue doit la comparer aux pertinences d'autres expressions. Autrement dit, la pertinence n'est pas utilisée dans un processus automatique et inconscient comme c'est le cas dans la cognition humaine, mais comme un moyen d'évaluer plusieurs possibilités qui doivent être testées. L'important est moins de copier l'humain que de disposer d'un critère fort permettant de rendre plus efficace la compréhension.

Intérêt de la pertinence pour une action de référence. En compréhension automatique, disposer d'un score numérique pour la pertinence d'une expression référentielle verbale ou multimodale présente plusieurs intérêts :

1. Aider à la résolution de la référence, c'est-à-dire retrouver l'intention référentielle, en mettant en lumière les processus intervenant dans la compréhension : appel pertinent à la saillance, appel pertinent à un sous-ensemble contextuel. L'utilisation de la pertinence va dans le sens d'une compréhension globale, permet d'augmenter les capacités d'interprétation de la machine en l'aidant à retrouver les connaissances partagées sur le critère de présomption de pertinence.
2. Parmi les hypothèses de reconnaissance, favoriser la plus pertinente.
3. Remettre en question le résultat d'une référence quand sa pertinence est faible.
4. Détecter les énoncés incongrus montrant un comportement particulier de l'utilisateur, par exemple quand il n'a pas vu le contexte visuel dans sa globalité, ou quand il montre une intention particulière telle que tester le système.
5. Gérer un stock des pertinences des énoncés précédents, ce qui s'avère utile pour remonter à la source lors d'une incompréhension ou d'un malentendu dans le dialogue. On suppose que cette source correspond à un moment où la pertinence a été mauvaise mais que cette faiblesse n'avait pas empêché la compréhension et la poursuite du dialogue. Nous noterons ici qu'une mauvaise pertinence n'entraîne pas une mauvaise compréhension. Au contraire,

une expression référentielle peut tout à fait être comprise, même si elle demande un effort de traitement très important et présente donc une faible pertinence par rapport aux autres expressions possibles.

En génération automatique, la pertinence d'une expression référentielle verbale ou multimodale s'évalue par rapport aux différents contextes, par rapport à elle-même et à l'intention de référence. Disposer d'une telle évaluation permet d'une part de détecter les éventuelles ambiguïtés entre cette intention et d'autres intentions, entre tel et tel référent, et d'autre part de déterminer parmi les expressions possibles celle qui produit l'effet escompté avec le minimum d'effort de traitement pour l'interlocuteur. La pertinence constitue à ce titre un critère idéal pour un choix parmi plusieurs possibilités. Nous reléguons dans le chapitre 10 nos propositions concernant l'exploitation d'un critère de pertinence en génération, et nous nous focalisons dans ce chapitre sur la compréhension.

7.1.3 La pertinence dans le modèle des domaines de référence

En quoi la pertinence sous-tend le modèle des domaines de référence. Nous voulons montrer ici que la notion de pertinence et en particulier d'effort de traitement est présente dans les concepts que nous avons développés jusqu'ici. La notion de saillance est particulièrement liée à la pertinence. Nous avions en effet caractérisé dans le chapitre 5 la saillance comme un point de départ dans le processus de compréhension, donc comme une base pertinente pour l'interprétation. Nous avions également insisté sur un critère de simplicité qui équivaut quasiment à une minimisation de l'effort de traitement.

Plus que cela, la pertinence intervient dans la majorité des morceaux de modélisation que nous avons proposés jusqu'à présent. A la fin du chapitre 2, nous avions présenté les grandes règles des déterminants ainsi qu'un algorithme d'identification des demonstrata d'une trajectoire gestuelle et un algorithme d'appariement du geste et de la parole. Ces trois modélisations bipolaires incluent un critère de pertinence. A propos de la première, nous avons vu par exemple que l'utilisation d'un démonstratif comme dans « *ce N* » associé à un geste ostensif était possible dans un domaine comprenant un seul objet de type N, mais était plus pertinente dans un domaine comprenant plusieurs objets de type N, la raison en étant l'expression par le démonstratif d'un contraste entre un N focalisé par rapport aux N non focalisés. A propos de trajectoire gestuelle et de leur interprétation en contexte visuel, nous avons vu que plusieurs formes étaient possibles, par exemple un entourage ou un ciblage. Selon la proximité des demonstrata intentionnels, entre eux et par rapport aux autres objets de la scène, un entourage peut s'avérer plus pertinent qu'un ciblage. En effet, un entourage cherche à délimiter précisément les demonstrata des autres objets, et, si l'effort de traitement nécessaire peut s'avérer élevé à cause de la lenteur de production, les effets contextuels sont particulièrement forts, du fait de la précision de la délimitation qui inhibe toute ambiguïté de portée. Nous montrons ainsi tout l'intérêt du critère de pertinence. Il en est de même avec la troisième modélisation bipolaire, celle s'intéressant à l'appariement du geste et de la parole. En effet, tout l'algorithme repose sur un seul but : l'identification du critère pertinent pour l'appariement, principalement entre un critère de synchronisation temporelle et un critère de correspondance des cardinaux.

Dans le chapitre 4, nous nous sommes intéressé à la construction de dendrogrammes pour structurer les différents domaines de référence visuels possibles. La lecture de ces dendrogrammes fait elle aussi intervenir un critère de pertinence. Considérons par exemple un dendrogramme permettant de distinguer trois partitions du contexte visuel. Si l'écart en terme d'indice d'agrégation entre deux regroupements est très élevé, la partition correspondant à cet écart significatif sera

caractérisée par un effet élevé. A effort de traitement constant, elle s'avérera plus pertinente que les deux autres partitions.

Dans le chapitre 5, nous nous sommes également intéressé à l'identification du rôle du geste ostensif en termes de domaines de référence. La pertinence intervient alors dans la délimitation du domaine pertinent, en particulier lorsqu'il ne s'agit ni du contexte visuel complet ni de l'ensemble des demonstrata. L'exemple de la figure 5.2 page 104 était particulièrement intéressant à ce point de vue.

De même, nous avons vu dans le chapitre 6 que la pertinence intervenait dans la caractérisation des ordonnancements, par exemple pour évaluer la régularité d'une disposition d'objets et décider si considérer cet ensemble comme un domaine de référence (fondé sur le critère de continuité) est pertinent ou non. Tout ceci fait intervenir la pertinence dans le processus de construction et de délimitation de domaines de référence. Nous voulons aborder maintenant l'exploitation d'un critère de pertinence pour l'intégration et l'évaluation de ces domaines construits.

Pertinence et intégration de domaines de référence. Nous avons vu que la délimitation d'un domaine faisait intervenir la détermination de l'expression référentielle verbale : on privilégiera le domaine dans lequel les particularités du défini ou du démonstratif s'appliquent le mieux. Grossièrement, cela revient à privilégier le domaine le plus restreint (pour minimiser l'effort) dans lequel le maximum de particularités s'appliquent (pour maximiser les effets). Nous avons vu également comment les critères de la Gestalt permettent de délimiter des domaines visuels. Une façon classique de procéder pour la résolution de la référence consiste à intersecter ces domaines. Nous montrons ici en quoi un critère de pertinence peut s'avérer utile dans cette intégration.

Dans l'exemple de la figure 7.1-A, nous considérons l'expression référentielle « *les carrés à gauche* » dans une scène visuelle comprenant deux carrés à gauche et un énorme amas de formes géométriques diverses à droite. Nous pouvons distinguer deux façons de traiter cette référence :

1. Nous appliquons un premier un filtre linguistique consistant à ne tenir compte que des « *carrés* » dans la scène visuelle, puis nous appliquons un filtre visuel consistant à ne tenir compte que des objets « *à gauche* » dans cet ensemble de carrés.
2. Nous appliquons un premier un filtre visuel consistant à ne tenir compte que des objets « *à gauche* », puis nous appliquons un filtre linguistique consistant à ne tenir compte que des « *carrés* » dans cet ensemble délimité.

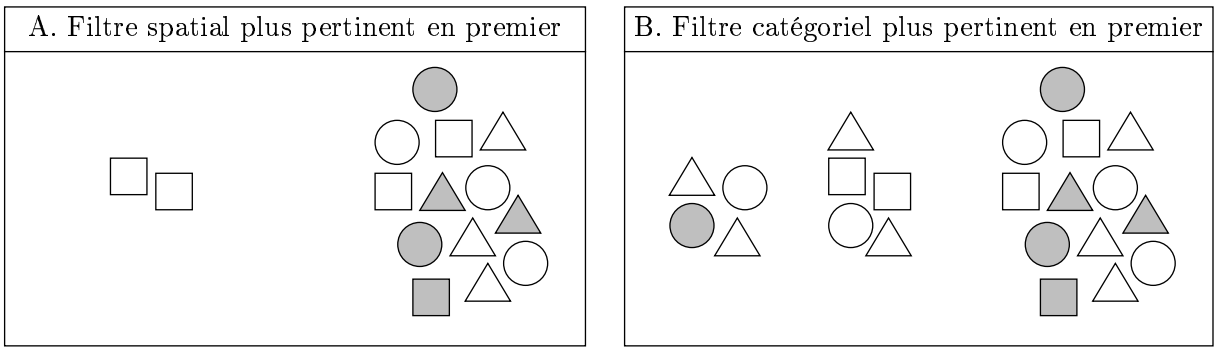


Figure 7.1 – *Pertinence pour l'ordre d'application des filtres catégoriel et spatial.*

La première solution, qui consiste à extraire sur la catégorie avant d'extraire sur la caractéristique spatiale, s'avère beaucoup plus coûteuse d'un point de vue cognitif, car elle nécessite de

distinguer les carrés dans l'amas d'objets à droite. La deuxième solution apparaît en revanche comme immédiate, c'est-à-dire comme nécessitant un effort de traitement beaucoup plus réduit. Dans cette situation particulière, la pertinence permet donc de privilégier l'extraction spatiale à l'extraction sur la catégorie. Une autre situation aurait pu aboutir au résultat inverse. C'est le cas de l'exemple de la figure 7.1-B pour lequel l'application du filtre spatial est délicate compte tenu de la présence de deux groupes plus ou moins à gauche. Pour ce filtre, l'effort de traitement s'avère important et la pertinence en diminue d'autant.

L'utilisation de la pertinence que nous faisons ici constitue une réponse possible au débat sur l'importance des modifieurs par rapport au nom. En effet, quelques-une des questions posées par la syntaxe et la sémantique concernent la nature de la tête d'un groupe nominal (est-ce forcément le substantif?) et celle du statut de l'adjectif qualificatif (est-ce un modifieur du nom ou un argument prédicatif?). Le point de vue classique est de considérer la catégorie N comme tête nominale et donc de commencer par le filtrage catégoriel. Un autre point de vue consiste à commencer par le filtrage correspondant au composant énoncé le premier, donc par « *grand* » dans « *un grand carré* » et par « *gris* » dans « *un carré gris* ». Nous considérons de notre côté que tout dépend du contexte visuel, et que la pertinence avec son critère de minimisation de l'effort de traitement donnera la réponse adaptée à la situation.

Pertinence et choix d'un domaine. Plusieurs points de vue se confrontent quant au choix d'un ou de plusieurs domaines de référence quand le calcul des référents aboutit à plusieurs hypothèses de domaines. Salmon-Alt (2001b) suppose que le coût d'interprétation est lié à la distance entre des domaines activés successivement, et donc que l'interprétation pertinente est celle qui demande le moins d'efforts en termes de changement de domaine. Un autre point de vue consiste à privilégier le domaine le plus simple, c'est-à-dire celui comportant le moins d'objets ou les objets les plus homogènes. Ceci nous amène à la pertinence : il faut faire jouer à la fois un critère de simplicité et un critère d'informativité, donc un critère de pertinence.

Un autre point de vue est le point de vue intensionnel : un domaine de référence constitue une liste de procédures (facteur de groupement, critère de différenciation, critère d'ordonnancement) partant d'un ensemble d'objets et aboutissant à un sous-ensemble. La pertinence d'un domaine peut être vue comme le degré d'optimisation de cette liste. Ainsi, le domaine le plus pertinent est celui qui aboutit au sous-ensemble avec la liste de procédures la plus efficace, c'est-à-dire impliquant le minimum d'effort de traitement. Bien que notre modèle que nous allons présenter maintenant ne gère pas encore la pertinence d'une telle façon, nous retiendrons l'importance du critère de pertinence, ainsi que sa présence constante dans les notions que nous avons exploré et que nous allons intégrer. Nous pourrons ensuite (§ 7.3) faire le point sur la quantification de la pertinence pour l'identification des domaines et des référents pertinents, et proposer des pistes pour la conception d'un modèle totalement fondé sur la pertinence.

7.2 Modèle de résolution de la référence multimodale

Nous intégrons dans cette section les propositions que nous avons faites jusqu'à présent, suivant le schéma de la figure 7.2. Dans ce schéma, nous avons regroupé toutes les prémices à notre modèle et nous décomposons celui-ci en trois phases, premièrement l'interprétation du geste par intégration des dendrogrammes visuels, deuxièmement la détermination de domaines de référence sous-spécifiés à partir de l'expression référentielle langagière, et troisièmement l'appariement de ces domaines sous-spécifiés avec les domaines disponibles. Cette dernière phase comporte plusieurs étapes selon qu'on fait appel aux historiques (du dialogue et de l'interaction), à de

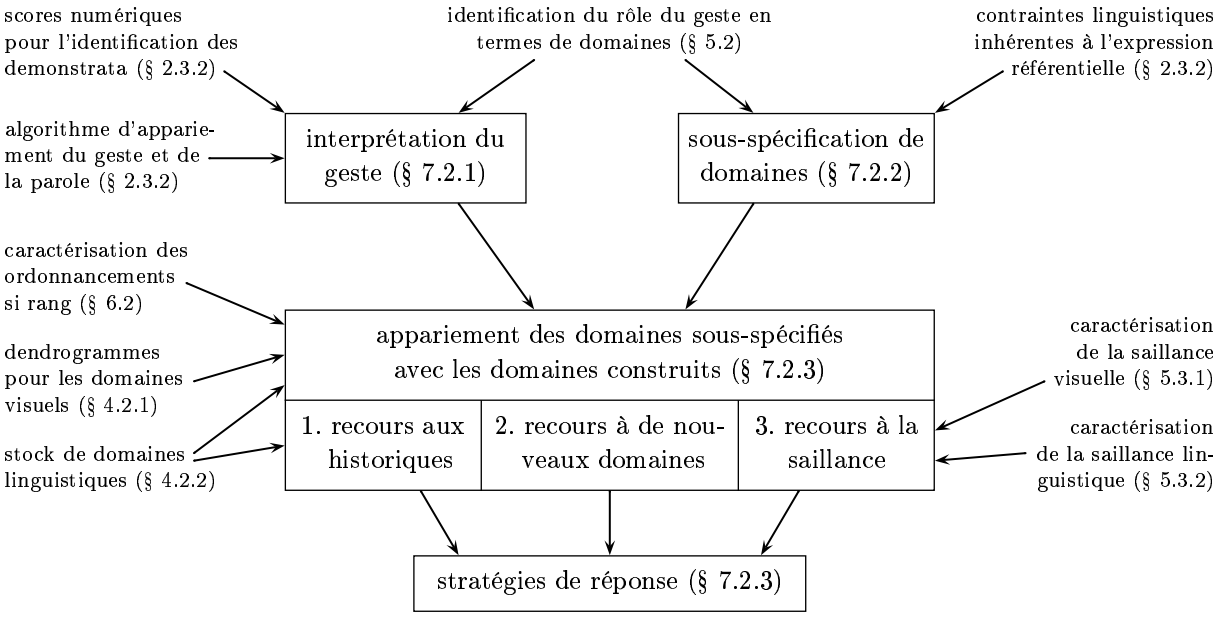


Figure 7.2 – *Schématisation globale des étapes de notre modélisation.*

nouveaux domaines tels que ceux construits sur la perception visuelle, ou à des domaines liés à la saillance. Selon les résultats obtenus lors de chacune de ces étapes, le système peut choisir entre prendre une décision ou demander une précision à l'utilisateur.

7.2.1 Intégration des dendrogrammes pour l'interprétation du geste

Un algorithme pour la modélisation de la focalisation spatiale. Il s'agit ici d'une part de spécifier un modèle de la focalisation spatiale, d'autre part de commencer véritablement l'intégration tripolaire. L'intégration des trois dendrogrammes issus des trois critères de la Gestalt pour la structuration de la scène visuelle en domaines, est une première étape significative dans cette prise en compte de phénomènes tripolaires. Elle se décompose en deux phases : la première part d'un objet ou d'un groupe d'objets focalisé pour aboutir à une liste d'hypothèses de groupes d'objets ; la seconde consiste en procédures d'exploitation de ces résultats. Pour l'interprétation du geste, la première phase part ainsi des demonstrata pour aboutir à une liste d'hypothèses de référents, la seconde phase intégrant les expressions référentielles verbales pour déterminer quelles sont les hypothèses pertinentes dans cette liste. Ce modèle de la focalisation spatiale peut également être utilisé pour l'interprétation langagière fondée sur la saillance. Dans ce cas, la première phase part de l'ensemble des objets les plus saillants pour aboutir à une liste d'hypothèses de domaines de référence, la seconde phase intégrant l'expression référentielle pour déterminer une hypothèse de référents. Dans cette section, nous présentons l'algorithme d'intégration pour l'interprétation du geste, sachant que la transposition pour l'interprétation fondée sur la saillance est immédiate. Deux méthodes sont possibles selon que l'on dispose ou non d'une correspondance entre les mesures des trois dendrogrammes.

Intégration des dendrogrammes sans correspondance des mesures. Cette première méthode est utilisée lorsqu'aucune mesure commune ne permet de confronter les niveaux d'agrégation des différents dendrogrammes, autrement dit lorsqu'il est impossible de comparer une unité de proximité avec une unité de similarité et une unité de continuité. 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 de cette méthode 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 lorsque l'on remonte d'un niveau dans son dendrogramme. Nous ne procédons pas ici à ce processus car il augmente considérablement le nombre de résultats.

Intégration des dendrogrammes 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 fondées sur la 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 que l'on étend aux groupes qui en incluent une partie (plusieurs groupes de la partition considérée peuvent ainsi être réunis) ;
- 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 à un niveau supérieur dans le dendrogramme) ;
- 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 (plusieurs milliers), c'est tout le module perceptif avec le choix de l'algorithme de classification automatique qui est à revoir.

Application à l'exemple de la figure 4.3. Avec cette deuxième méthode, nous procédons à l'intégration des dendrogrammes de la figure 4.3 page 93, en partant d'un ensemble d'objets comprenant T_1 et T_2, c'est-à-dire en imaginant qu'une trajectoire a été effectuée sur les triangles gris T_1 et T_2. Le but de l'intégration est de montrer que ce geste peut également être compris comme englobant le groupe des trois triangles très proches $\{T_1, T_2, T_3\}$, et ce selon plusieurs combinaisons des critères de la Gestalt.

On fait correspondre les trois dendrogrammes avec une même mesure, ce qui donne comme ordre de prise en compte de critères : `sim niv 1` ; `prox niv 1` ; `cont niv 1` ; `sim niv 2` ; `prox niv`

`2` ; `cont niv 2` ; `prox niv 3` et ainsi de suite. L'ordre des opérations est le suivant :

1. `sim niv 1` : on étend la liste des deux demonstrata au premier groupe possible, donc celui des trois triangles gris. Ce résultat est particulier car il fait intervenir la similarité seule, ce qui n'est pas très pertinent : la similarité se combine généralement avec un autre critère. Nous ne le considérerons donc pas.
2. `sim niv 1 + prox niv 1` : on intersecte avec le groupe de cinq objets de `prox niv 1`, ce qui donne le nouveau résultat : $\text{result}_1 = \{T_1, T_2\}$.
3. `sim niv 1 + prox niv 1 + cont niv 1` : on intersecte avec le groupe des trois triangles de `cont niv 1`, ce qui donne toujours : $\text{result}_1 = \{T_1, T_2\}$.
4. `sim niv 2 + prox niv 1 + cont niv 1` : cette fois on repart du groupe des quatre triangles correspondant à `sim niv 2`, pour l'intersecter avec le groupe de cinq objets de `prox niv 1` et le groupe des trois triangles de `cont niv 1`. Un nouveau résultat est alors identifié : $\text{result}_2 = \{T_1, T_2, T_3\}$.
5. `sim 2 + prox 2 + cont 1` : cette phase ajoute juste le carré qui n'est de toute façon pas pris dans l'intersection. Le résultat est le même que précédemment : $\text{result}_2 = \{T_1, T_2, T_3\}$.
6. `sim 2 + prox 2 + cont 2` : $\text{result}_2 = \{T_1, T_2, T_3\}$.
7. `sim 2 + prox 3 + cont 2` : cette fois T_4 est dans les trois groupes à intersecter. Le nouveau résultat est alors : $\text{result}_3 = \{T_1, T_2, T_3, T_4\}$.
8. etc.

Cette liste met l'accent sur result_2, et, dans une moindre mesure, sur result_3 qui apparaît à trois reprises ensuite. Nous remarquerons que si nous avions pris en compte `prox niv 1` avant `sim niv 1`, le premier résultat aurait été le groupe des cinq objets à gauche, et le reste n'aurait pas beaucoup changé. Nous remarquerons également que result_2 correspond à l'interprétation stricte de la trajectoire gestuelle, alors que result_3 correspond à l'interprétation non stricte la plus plausible, compte tenu des liens forts (les trois critères de la Gestalt s'appuyant mutuellement) entre les trois triangles. Si l'expression référentielle va dans le sens de cette deuxième interprétation (comme « *ces trois triangles* » dans l'exemple de l'introduction), le système pourra éjecter la première. C'est à ce tri dans les hypothèses que nous nous intéressons maintenant.

Exploitation des résultats à l'aide de l'énoncé oral. Cette phase consiste essentiellement en une combinaison de critères venant des composants de l'expression référentielle verbale, du geste et des résultats obtenus par la phase précédente. Nous avons vu page 106 que le geste pouvait désigner aussi bien un ensemble de référents qu'un domaine de référence. Les résultats de l'intégration peuvent donc correspondre à ces deux possibilités, pour lesquelles ils ne font que limiter les hypothèses.

Le principal critère pour l'identification du rôle du geste est la présence d'un adjectif numéral dans l'expression référentielle verbale. Ainsi, dans l'exemple de la figure 4.3, « *ces deux objets* » activera le résultat result_1 alors que « *ces trois objets* » activera le résultat result_2. Si l'expression référentielle est « *ces objets* », le seul critère permettant de choisir entre result_1 et result_2 est donné par leur pertinence, c'est-à-dire l'ordre dans lequel ils sont apparus au cours de l'intégration des dendrogrammes. C'est ainsi que result_1 est retenu. Plus précisément, c'est-à-dire en tenant compte des catégories et de la possibilité qu'a le geste de désigner un domaine, nous proposons l'algorithme suivant pour l'interprétation :

1. Nous rapprochons le nombre apparaissant dans l'expression référentielle et le nombre d'objets de chacun des ensembles obtenus par l'intégration des dendrogrammes, quelles que

soient les catégories. Trois cas apparaissent : soit le nombre venant de l'expression verbale est inférieur à celui venant de l'intégration (c'est le cas de l'exemple de la figure 5.1-B page 104) ; soit il lui est supérieur (c'est le cas de l'exemple de la figure de l'introduction) ; soit les deux nombres ne présentent pas d'incompatibilité (c'est le cas lorsque les deux nombres sont égaux, lorsque l'expression verbale est « *ça* », ou lorsqu'un pluriel sans adjectif numéral est utilisé en association avec un geste désignant plusieurs demonstrata).

2. Dans le premier cas, nous identifions le rôle du geste comme désignant un domaine de référence, celui-ci correspondant à l'ensemble obtenu par l'intégration des dendrogrammes (pour l'hypothèse considérée). Nous cherchons alors les contraintes linguistiques qui permettent d'extraire les référents de ce domaine. Si l'expression référentielle ne contient pas de catégorie (« *ces objets* » par exemple), nous considérons qu'il y a incompréhension et nous rejetons l'hypothèse. Si des filtres liés à la catégorie ou aux modifieurs peuvent être identifiés, nous les appliquons et en déduisons les référents. L'hypothèse est alors validée.
3. Dans le deuxième cas, nous identifions le rôle du geste comme désignant une partie des référents, et nous cherchons un critère dans l'expression référentielle qui permette d'étendre l'ensemble obtenu par intégration des dendrogrammes (pour l'hypothèse considérée) à un ensemble de référents valable. Par exemple, si l'expression référentielle est pertinente pour une référence générique, une telle interprétation sera retenue (« *j'aime bien ces fauteuils* » associé à un geste sur un fauteuil, par exemple). Si ce n'est pas le cas, notamment si l'expression référentielle contient un adjectif numéral (ce qui exclut l'interprétation générique), nous procédons à l'extension avec le critère de similarité. L'ensemble obtenu est alors identifié comme l'ensemble des référents. Pour déterminer le domaine, nous passons à l'étape 5.
4. Dans le troisième cas, nous identifions le rôle du geste comme désignant un ou plusieurs référents, à savoir l'ensemble obtenu par l'intégration des dendrogrammes. Si les filtres liés à la catégorie et aux modifieurs ne remettent pas en cause ce résultat, nous passons à l'étape 5 pour déterminer le domaine de référence. Sinon, nous considérons qu'il y a incompréhension et nous rejetons l'hypothèse.
5. Si l'ensemble obtenu par l'intégration des dendrogrammes ne convient pas, pour des raisons d'application des règles liées aux déterminants ou à des composants linguistiques particuliers comme le superlatif, nous excluons l'hypothèse courante et nous passons à la suivante.

Cet algorithme, présenté dans (Landragin 2002), permet de tenir compte de tous les phénomènes que nous avons observés à propos de rapport entre demonstrata et référents. Il aboutit à la bonne interprétation, aussi bien pour l'exemple de l'introduction que pour les exemples des figures 5.1 et 5.2. Il ne prend néanmoins pas encore en compte toutes les particularités du langage, notamment celles liées à une continuité dans le dialogue, et c'est sur ce point que nous allons maintenant axer notre modélisation. D'autre part, cet algorithme reste à améliorer et devrait en particulier inclure une phase préliminaire chargée de déterminer si la trajectoire gestuelle est ambiguë compte tenu du contexte visuel. Un tel test permettrait en effet de ne pas faire d'hypothèses non pertinentes, comme celle consistant à étendre selon la proximité un ensemble de demonstrata clairement délimité par un geste d'entourage lent et précis.

7.2.2 Sous-spécification de domaines pour l'interprétation du langage

Rôles des composants de l'expression référentielle verbale. Le but de cette section est de proposer une classification des phénomènes liés à chacun des composants linguistiques potentiels, pour déterminer quelles structures de domaines peuvent être liées à quels composants. Cette

classification regroupe tous les phénomènes dont nous avons parlé, en termes de référents et de domaines de référence. Elle part des déterminants et intègre ensuite, dans la mesure du possible, les autres composants linguistiques. Avant de la présenter, il s'avère utile de préciser les principaux rôles de ces composants. Comme nous l'avons fait dans (Landragin *et al.* 2002c), nous considérons le déterminant comme fournissant des contraintes sur les partitions d'un domaine de référence, et les autres composants comme fournissant des contraintes sur le type d'un domaine et sur le critère de différenciation d'une partition. Etudions tout d'abord la détermination :

- Groupes nominaux indéfinis :
  Selon Corblin (1987), un indéfini de forme « *n N* » extrait n éléments d'un domaine comprenant des éléments de type N. L'expression « *un triangle* » cherchera donc à s'interpréter dans un domaine formé par des objets de type « triangle ». 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.

- Groupes nominaux définis :
  Corblin (*Ibid.*) montre qu'un défini s'interprète dans un domaine à l'intérieur duquel son contenu constitue un signalement singularisant. Une expression telle que « *le triangle de droite* » s'interprète alors dans un domaine partitionné selon au moins un critère de différenciation. Ici, on cherchera par exemple un domaine comprenant des objets de type « triangle », 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.

- Groupes nominaux démonstratifs :
  Un démonstratif tel que « *ce triangle* » cherchera à s'interpréter dans un domaine comportant obligatoirement un élément identifiable autrement que par la désignation linguistique elle-même. C'est évidement le rôle d'un geste ostensif, une autre possibilité étant une focalisation discursive, par exemple lorsqu'une entité du discours correspondant à un triangle se trouve en position de focus ou de thème. Par ailleurs, il doit être possible de reclassifier l'élément. Un « triangle » peut ainsi être reclassifié comme « pyramide », comme nous le verrons dans l'exemple de la figure 7.3 page 147. 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 au demonstratum de la figure 5.1-A page 104 par « *ce triangle* » associé à un geste, ou par « *cette figure* » associé au même geste, aurait en effet des conséquences différentes sur l'interprétation d'une expression telle que « *les autres* » dans la suite du dialogue : dans le premier cas, celle-ci désignerait tous les triangles (un seul dans la scène visuelle considérée), 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).

A partir de ces mécanismes, nous intégrons maintenant les autres composants linguistiques, c'est-à-dire essentiellement la catégorie et les modifieurs. Comme les mécanismes de ces derniers

dépendent de la détermination de l'expression, nous reprenons la liste précédente :

- Groupes nominaux indéfinis :
  Le domaine dont un indéfini « *n N P* » (« *un grand triangle* » par exemple) extrait des éléments doit être un domaine comprenant des éléments de type N ayant, le cas échéant, la propriété P.

- Groupes nominaux définis :
  Pour un défini nu (« *le N* »), 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 « *le triangle* » s'extrait d'un domaine de formes géométriques), soit le type correspondant à un objet dont le référent est une partie (le référent de « *la fenêtre* » s'extrait d'un domaine « mur » ou « pièce »). 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 la base des objets.
  Pour un défini déterminé par un adjectif (« *le N P* »), une première singularisation passe par N et une seconde par la valeur P pour un attribut particulier. Cette dernière reposant sur la première, le critère de différenciation est donc l'attribut particulier, le type commun des éléments du domaine étant N.

- Groupes nominaux démonstratifs :
  Corblin (1987) fait l'hypothèse qu'un démonstratif « *ce N* » 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 la base des objets : il s'agit soit de relations hiérarchiques de typage (un type est compatible avec ses sous-types) ; soit de connaissances spécifiques à l'application (dans le corpus Ozkan et par exemple dans l'extrait de la figure 7.3, on observe des reclassifications entre éléments géométriques et figuratifs telles que « *rond* » puis « *soleil* » ou « *triangle* » puis « *pyramide* ») ; soit de connaissances encyclopédiques de façon générale.

- 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 l'une des désignations possibles pour le référent.

Parmi les autres possibilités linguistiques se trouvent les expressions d'altérité (« *les autres triangles* »), les expressions dénotant un rang (« *le premier triangle* ») ou encore l'emploi du superlatif (« *les formes les plus claires* »). Or le point commun de ces expressions est de présupposer explicitement un domaine de référence partitionné selon des critères de différenciation spécifiques : « *les autres triangles* » présuppose un domaine de type « triangle » ayant préalablement été partitionné selon un critère quelconque ; « *le premier triangle* » présuppose un domaine de type « triangle » dont les éléments peuvent être distingués selon un ordre ; « *les formes les plus claires* » présuppose un domaine de type « forme » partitionné selon la couleur de ses éléments. Nous montrons ainsi un aspect intéressant de notre modélisation, à savoir l'intégration aisée des

contraintes liées à des expressions souvent considérées comme particulières, c'est-à-dire comme nécessitant habituellement des traitements adéquats.

Classification des mécanismes de référence en termes de domaines de référence. Dans la liste suivante, nous énumérons tous les emplois possibles de chacun des types fondamentaux de détermination pour des actions de référence aux objets. L'éventail des phénomènes repose sur les classifications données dans le chapitre 1, et leur caractérisation en termes de domaines de référence sur les résultats des chapitres 4, 5 et 6. Comme nous l'avons fait ci-dessus, nous partons d'une expression du type « *le N* » ou « *ce N* » pour considérer ensuite chaque composant supplémentaire, adjectif numéral, adjectif qualilificatif ou composant particulier.

- **Groupes nominaux indéfinis :**
  1. Introduction langagière d'un nouveau référent. Dans « *crée un carré* » par exemple, « *un carré* » ne réfère pas encore à un objet particulier mais pourra être repris ultérieurement. Le pluriel et le remplacement par un déterminant numéral cardinal ou par un déterminant indéfini quantifiant sont possibles (« *ajoute deux triangles et quelques cercles* »). Un geste coréférent simultané n'est pas possible. Les opérations effectuées par le système sont l'ajout du ou des nouveaux objets dans le domaine de référence correpondant au contexte visuel global, et la création d'une nouvelle partition dans ce domaine avec un critère de différenciation correspondant au prédicat, partition dans laquelle le nouveau référent est focalisé.

  2. Extraction d'un élément quelconque d'un ensemble focalisé. L'ensemble doit être plus restreint que celui de la classe entière. Le pluriel et le remplacement par un déterminant numéral ou quantifiant sont possibles. L'ensemble est un domaine de référence qui peut être :
     - délimité par une référence langagière dans l'énoncé précédent (« *sélectionne les carrés et les triangles* » puis « *enlève un carré* » ou « *enlève quelques triangles* », qui se comprennent comme « *enlève un des carrés sélectionnés* » ou « *enlève quelques-uns de ces triangles* ») ;
     - délimité un geste ostensif associé à l'énoncé courant (« *enlève un carré* » associé à un geste délimitant un ensemble de carrés) ;
     - non précisé, auquel cas on considère le contexte visuel courant (« *enlève deux carrés* » qui se comprend comme « *enlève deux des carrés visibles* » et éventuellement comme « *enlève deux des carrés visuellement saillants* »).

     Dans les trois cas, les opérations effectuées par le système sont la création d'une nouvelle partition dans le domaine focalisé (le critère de différenciation étant fourni par le prédicat) et la focalisation du référent choisi dans cette partition.

  3. Désignation d'un référent particulier qui est focalisé par ailleurs. Ce cas constitue une utilisation déviante par rapport aux théories comme celle de Corblin (1987), mais néanmoins présente dans notre corpus de référence. Le pluriel et le remplacement par un déterminant numéral ou quantifiant sont possibles. C'est le geste ostensif qui propose un choix de référent (« *enlève un carré* » associé à un geste désignant un carré particulier). Les opérations effectuées par le système sont les mêmes que précédemment, le choix du référent étant cette fois contraint.

  4. Référence générique. Exemple : « *un carré a quatre côtés* ». Le pluriel, le remplacement par un déterminant quantifiant et le geste coréférent ne sont pas possibles. Le

remplacement par un déterminant numéral est possible : « *deux triangles ayant un côté commun forment un quadrilatère* ». L'opération effectuée par le système est l'activation d'un domaine de référence générique, c'est-à-dire d'un domaine désignant la classe entière.

- **Groupes nominaux définis :**

  1. Extraction d'un élément particulier d'un ensemble focalisé. Le pluriel et l'adjectif numéral cardinal sont possibles. L'ensemble est un domaine de référence qui peut être :
     - délimité préalablement par une référence langagière dans l'énoncé précédent (anaphore comme « *déplace le triangle bleu et le carré vert* » suivi de « *enlève le triangle* » ; « *déplace les triangles* » suivi de « *le triangle rouge* » ; « *affiche les meubles* » suivi de « *la table* » ; référence mentionnelle comme « *le triangle rouge, le vert et le bleu* » suivi de « *le premier* » ; anaphore associative comme « *la chaise* » suivi de « *le dossier* ») ;
     - délimité préalablement par une focalisation à un sous-espace visuel (exemple de l'introduction) ;
     - délimité par une précision langagière dans l'expression référentielle même (« *le triangle en haut à gauche de la scène* ») ;
     - délimité par un geste ostensif dans l'énoncé même (« *le triangle* » associé à un geste désignant un triangle et un carré) ;
     - non précisé, auquel cas on considère le contexte visuel courant, ou éventuellement un sous-ensemble saillant de celui-ci.

     Les opérations effectuées par le système sont l'extraction et la focalisation d'un élément du domaine focalisé. L'élément doit être isolable selon le critère de différenciation fourni par les composants tels que la catégorie et les modifieurs.

  2. Désignation d'un référent particulier qui est focalisé par ailleurs. Le pluriel et l'adjectif numéral sont possibles. Le référent peut être :
     - donné par une référence langagière dans l'énoncé précédent (« *sélectionne un triangle rouge* » repris par « *le triangle* » ; « *le triangle et le carré* » repris par « *les formes* ») ;
     - montré par un geste ostensif dans l'énoncé même (« *le carré* » associé à un geste désignant un carré ; « *la forme-ci et la forme-là* » associé à un ou deux gestes).

     L'opération effectuée par le système est la construction, autour du référent focalisé, d'un nouveau domaine dont le type est un sur-type de N, le critère de différenciation de la partition créée conjointement étant N.

  3. Désignation d'un ensemble particulier dont un élément est focalisé. Le pluriel et l'adjectif numéral sont possibles. L'élément peut être :
     - donné par une référence langagière dans l'énoncé précédent (« *le carré avec des cercles qui l'entourent* » suivi de « *le groupe* » ; « *le dossier* » suivi de « *la chaise* ») ;
     - montré par un geste ostensif dans l'énoncé même (« *les fauteuils* » associé à un geste désignant un seul fauteuil).

     L'opération effectuée par le système est la même que précédemment.

4. Référence générique. Exemples : « *le triangle est une figure géométrique simple* » ; « *les chaises ont quatre pieds* ». Le pluriel est possible. L'adjectif numéral et le geste coréférent ne le sont pas. L'opération effectuée par le système est l'activation d'un domaine de référence générique.

- **Groupes nominaux démonstratifs :**
  1. Extraction d'un élément particulier d'un ensemble focalisé. Le pluriel et l'adjectif numéral sont possibles. La focalisation est forcément due à une référence langagière antérieure (« *déplace le triangle bleu et le carré vert* » suivi de « *enlève ce carré* » ou de la référence mentionnelle « *ce dernier* »). Le geste simultané n'est pas possible. Les opérations effectuées par le système sont l'extraction et la focalisation d'un élément du domaine focalisé. L'élément doit être isolable selon le critère de différenciation fourni par les composants tels que la catégorie et les modifieurs.

  2. Désignation d'un référent particulier qui est focalisé par ailleurs. Le pluriel et l'adjectif numéral sont possibles. La focalisation du référent peut provenir :
     - d'une référence langagière dans l'énoncé précédent (anaphore comme « *déplace le triangle bleu* » repris par « *enlève ce triangle* » ou comme « *la chaise* » repris par « *ce meuble* » ; reclassification comme « *le triangle* » repris par « *cette pyramide* ») (le geste ostensif est alors impossible) ;
     - d'un geste ostensif dans l'énoncé même (« *ce triangle* » ou « *cette forme-ci* » associé à un geste désignant un triangle).

     L'opération effectuée par le système est la construction, autour du référent focalisé, d'un nouveau domaine de type N, le critère de différenciation de la partition créée conjointement étant fourni par le prédicat ou par l'intervention d'un geste coréférent.

  3. Désignation d'un ensemble particulier dont un élément est focalisé. Le pluriel et l'adjectif numéral sont possibles. L'élément peut être :
     - donné par une référence langagière dans l'énoncé précédent (« *le carré avec des cercles qui l'entourent* » suivi de « *ce groupe* ») ;
     - montré par un geste ostensif dans l'énoncé même (« *ces fauteuils* » associé à un geste désignant un seul fauteuil).

     L'opération effectuée par le système est la même que précédemment.

  4. Référence générique. Le pluriel est possible (et plus fréquent que le singulier). L'adjectif numéral n'est pas possible. La référence peut prendre trois formes :
     - le passage au générique à partir d'un antécédent langagier (« *cette forme bizarre* » suivi de « *ces formes* ») ;
     - le passage au générique à partir d'un antécédent gestuel (« *j'aime bien ces fauteuils* » associé à un geste désignant un seul fauteuil) ;
     - le générique direct (« *ce fauteuil est confortable* » associé à un geste désignant un fauteuil, se comprenant comme « *ce type de fauteuils* », et entraînant de ce fait une ambiguïté totale avec la référence au fauteuil spécifique désigné).

     L'opération effectuée par le système est l'activation d'un domaine générique.

- **Pronoms personnels :**
  1. Désignation d'un référent particulier qui est focalisé par ailleurs. Le pluriel est possible, le geste coréférent ne l'est pas. En dehors du cas particulier caractérisé par une

intention manifeste (« *attention,* ***il*** *risque de te mordre* », exemple décrit page 99), la focalisation du référent provient forcément d'une référence langagière dans l'énoncé précédent. Il s'agit donc de l'anaphore, comme dans « *déplace la chaise bleue* » suivi de « *supprime-la* ». Nous mettrons ici aussi le cas particulier où le pronom réfère à un autre exemplaire de l'objet désigné par l'antécédent (« *j'ai effacé le triangle vert mais il est revenu* »). Le référent étant déjà focalisé, et la nature de cette focalisation ne changeant pas, aucune opération n'est effectuée par le système.

2. Désignation d'un ensemble particulier dont un élément est focalisé. Le geste coréférent est impossible. L'élément est donné par une référence langagière préalable. Pour la reprise anaphorique, le pluriel est obligatoire puisqu'il permet de construire l'ensemble : « *ajoute un triangle vert* » suivi de « *supprime-les* ». L'opération effectuée par le système est la construction, autour de l'élément focalisé, d'un nouveau domaine dont le type est un sur-type de N, le critère de différenciation de la partition créée conjointement étant N (N et P pour l'exemple donné).

3. Référence générique. Le pluriel est obligatoire. Quant au geste coréférent, il est encore une fois impossible. Ce cas correspond au glissement au générique dans l'anaphore : « *j'ai acheté une Toyota parce qu'elles sont robustes et bon marché* » (Gaiffe 1992). L'opération effectuée par le système est l'activation d'un domaine générique.

- **Pronoms démonstratifs :**

1. Extraction d'un élément particulier d'un ensemble focalisé. Le pluriel est possible. La focalisation est forcément due à une référence langagière antérieure (« *la table, la chaise et le fauteuil* » suivi de la référence mentionnelle « *celui-ci* » pour désigner le dernier élément cité). Le geste coréférent est impossible. Les opérations effectuées par le système sont l'extraction et la focalisation d'un élément du domaine focalisé. L'élément doit être isolable selon le critère de différenciation fourni par l'ordre d'énonciation : est retenu le dernier élément cité (le genre intervient en plus dans l'exemple donné).

2. Désignation d'un référent particulier qui est focalisé par ailleurs. Les pronoms démonstratifs combinent une référence démonstrative et une anaphore : ils sont associés à un geste ostensif pour désigner un nouvel objet ayant les caractéristiques d'un objet désigné dans la partie précédente du discours. La focalisation du référent provient donc forcément d'un geste dans l'énoncé même. Dans « *déplace cette chaise bleue* » suivi de « *supprime celle-ci* », « *celle-ci* » désigne avec le geste coréférent une autre « *chaise bleue* ». Le pluriel est possible. L'opération effectuée par le système est la construction, autour du référent focalisé, d'un nouveau domaine de type N, le critère de différenciation de la partition créée conjointement étant fourni par l'intervention du geste coréférent.

3. Référence générique. Comme pour le pronom personnel, le pluriel est indispensable et le geste coréférent impossible (« *J'ai acheté une Lada et une Toyota. Celles-ci sont robustes et bon marché* »). L'opération effectuée par le système est l'activation d'un domaine générique.

Résultats et commentaires. Nous remarquons dans cette classification que les rôles d'un groupe nominal en terme de mécanisme de référence sont en nombre limité. Nous trouvons les possibilités

suivantes :

- introduction langagière d'un nouveau référent ;
- extraction d'un élément quelconque d'un ensemble focalisé ;
- extraction d'un élément particulier d'un ensemble focalisé ;
- désignation d'un référent particulier qui est focalisé par ailleurs ;
- désignation d'un ensemble particulier dont un élément est focalisé ;
- référence générique.

Cette liste illustre parfaitement l'intérêt des domaines de référence avec la focalisation possible d'un de leurs éléments au travers d'une partition. Comme nous nous sommes limité aux actions de référence aux objets, nous n'avons pas inclu l'utilisation attributive et non référentielle (« *il y a toujours un triangle en haut à gauche de la scène* » ; « *les deux triangles en haut à gauche, quels qu'ils soient* »). Une remarque que nous pouvons faire est que l'introduction multimodale d'un nouveau référent n'existe pas : du fait de la nécessaire coréférence, il s'agit d'une désignation langagière d'un référent particulier qui est focalisé par le geste. Pour le pronom personnel, nous n'avons pas mentionné l'extraction d'un élément particulier d'un ensemble focalisé telle qu'elle peut apparaître avec des êtres humains : « *un homme et une femme entrèrent* » suivi de « *elle était belle* ». Nous considérons en effet cette exploitation du genre comme un cas extrême. Une autre remarque concerne le mot « *autre* » : sans entrer dans les détails de l'utilisation de ce mot, il n'y a jamais contradiction entre son fonctionnement et la détermination du groupe nominal dans lequel il se trouve. L'important, c'est que le domaine soit calculable dans son extension, de façon à ce que la partition concernée puisse être épuisée lors de l'interprétation.

Le résultat principal de cette analyse est que les opérations effectuées par le système lors de l'interprétation de l'expression référentielle dépendent non seulement du déterminant, mais également d'un des six rôles identifiés ci-dessus. Ainsi, l'opération consistant à activer un domaine générique est propre à cette interprétation particulière, que l'on retrouve dans les possibilités offertes par tous les groupes nominaux à l'exception des pronoms démonstratifs. Les opérations consistant à créer un nouveau domaine ou à focaliser un élément dans une partition existante ne sont pas propres à un déterminant particulier. Les exemples les plus importants de fonctionnements indépendants du déterminant (entre défini et démonstratif) sont les suivants : l'extraction d'un élément particulier d'un ensemble délimité préalablement par une référence langagière antérieure (« *déplace le triangle bleu et le carré vert* » suivi aussi bien de « *enlève le carré* » que de « *enlève ce carré* ») ; la désignation d'un référent particulier qui est focalisé par ailleurs, que ce soit par une référence langagière antérieure (« *sélectionne un triangle rouge* » repris aussi bien par « *le triangle* » que par « *ce triangle* ») ou par un geste coréférent (« *le triangle-ci* » ou « *ce triangle* » associé à un geste désignant un triangle) ; ou encore la désignation d'un ensemble particulier dont un élément est focalisé (« *le carré avec des cercles qui l'entourent* » suivi aussi bien de « *le groupe* » que de « *ce groupe* »). En revanche, lors de la création d'un nouveau domaine par exemple, le choix du type du domaine (N ou sur-type de N) dépend du déterminant.

C'est en ce point que notre approche diffère fortement de celle de Salmon-Alt (2001b), qui considère que le déterminant seul suffit quasiment à spécifier les opérations effectuées par le système. Notre analyse de la référence multimodale, en montrant que dans certaines circonstances un groupe nominal défini pouvait être utilisé de la même façon qu'un groupe nominal démonstratif, relativise l'importance du déterminant pour mettre l'accent sur la combinaison des rôles de ce déterminant et du geste en termes de désignation de référents ou de domaines. C'est l'identification d'une telle combinaison qui entraîne la spécification des opérations que le système doit effectuer.

Ces opérations consistent en la construction d'un domaine de référence sous-spécifié et en la formalisation de contraintes sur ce domaine. Lorsque le système reçoit une expression référentielle verbale, il identifie les rôles possibles du groupe nominal en terme de mécanisme de référence. Pour chaque possibilité, il génère un domaine sous-spécifié avec ses contraintes. Lorsque plusieurs hypothèses restent valables après l'analyse de l'éventuel geste ostensif, les domaines sous-spécifiés correspondants sont conservés. Contrairement à l'approche de Salmon-Alt qui ne conserve qu'un seul domaine sous-spécifié dans lequel le minimum de contraintes sont formalisées (ou pour lequel des contraintes peuvent ensuite être relâchées), nous gardons plusieurs hypothèses de domaines sous-spécifiés. Le but est maintenant de montrer comment nous allons exploiter, sans les remettre en cause, ces hypothèses.

7.2.3 Appariement de domaines pour la compréhension globale

Recherche d'un domaine compatible. Nous proposons de procéder à l'appariement des domaines sous-spécifiés avec des domaines construits en suivant trois phases (cf. figure 7.2 page 134). La première, c'est-à-dire celle que nous considérons comme la plus naturelle et comme devant testée en priorité, consiste à ne considérer que les domaines construits lors des énoncés précédents. Ces domaines constituent l'historique du dialogue et sont indispensables, notamment pour la résolution des anaphores. Cette phase correspond à tester une interprétation dénotant une continuité dans le dialogue. Elle s'oppose aux interprétations fondées sur un nouveau contexte ou sur une rupture.

La deuxième phase correspond à tester de telles interprétations : elle consiste en effet à considérer des domaines construits pour le traitement de l'énoncé courant. C'est le cas lorsque l'utilisateur reconsidère la scène visuelle dans sa globalité ou lorsqu'il commence une nouvelle phase du dialogue en oubliant les situations précédentes. Le système passe à cette deuxième phase quand la première n'a donné aucun résultat pertinent. Contrairement à cette première phase, nous disposons ici de critères de détection. En effet, l'interprétation correspondant à cette deuxième phase est très probable lorsque tous les objets du domaine courants ont été traités. Un autre critère est celui de redondance : selon Beun et Cremers (1998), la présence de redondance dans l'énoncé courant est le signe que l'utilisateur a reconsidéré la situation et veut changer de domaine. Il arrive maintenant que cette phase aboutit à une ambiguïté.

La troisième phase consiste à recourir à la saillance face à de telles ambiguïtés. En simplifiant, il s'agit de faire appel aux objets et aux domaines saillants quand la considération des contextes complets génère trop d'hypothèses pour la résolution de la référence. L'interprétation selon cette saillance permettra peut-être de mettre en avant certaines hypothèses, idéalement une seule. Le système devra ensuite adapter sa stratégie de réponse.

Sélection de domaines et résolution de la référence. Le but est d'aboutir à une liste de couples comprenant un domaine et un ensemble de référents, ces couples étant ordonnés selon leur pertinence. Le principe est celui de l'unification : le système fait correspondre chaque hypothèse de domaine sous-spécifié avec chaque domaine construit. La meilleure unification, c'est-à-dire celle où le maximum de contraintes correspondent parfaitement, est retenue. Pour l'instant, nous ne décrirons pas plus le mécanisme d'unification, renvoyant la description et la formalisation de ce principe au chapitre 9 (§ 9.3).

Le principal problème est que de nombreuses hypothèses de domaines construits peuvent s'unifier à l'un des domaines sous-spécifiés. Il faut pouvoir évaluer ces hypothèses pour les ordonner. Un modèle local de pertinence s'avère ici utile : en calculant la pertinence de chacune des hypothèses, on dispose d'un critère numérique permettant de déterminer un ordre. Nous propo-

sons de faire le calcul de cette pertinence en appliquant les notions d'effet contextuel et d'effort de traitement à l'unification : plus le domaine sous-spécifié et le domaine construit considérés sont similaires, et plus l'effet est important ; plus le nombre d'opérations élémentaires impliquées dans l'unification est élevé, et plus l'effort est important. Nous assimilons ainsi l'effet au nombre d'attributs similaires sur le nombre total d'attributs, les attributs étant tous les paramètres caractérisant un domaine, donc le critère de différenciation, le facteur de groupement, le type des objets contenus dans la partie focalisée de la partition, etc. De même, nous assimilons l'effort au nombre d'opérations nécessaires à l'appariement.

Le système doit maintenant exploiter cette liste ordonnée d'hypothèses. En étudiant pour chacune d'entre elles l'identité des référents et l'étendue du domaine de référence, la lecture des résultats se fait de la manière suivante :

1. 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.
2. S'il n'y a pas ambiguïté sur l'identité des référents, il peut néanmoins 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 différentes possibilités sont sauvegardées et conduiront éventuellement à détecter une ambiguïté lors de l'analyse de l'énoncé ultérieur.
3. 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.
4. Si, après ce relâchement de contrainte, aucun résultat n'est encore trouvé, cela privilégie l'hypothèse de l'erreur. Celle-ci peut provenir soit de l'utilisateur (comme le lapsus transformant « *horizontale* » en « *verticale* » dans l'exemple de la figure 7.3), soit du système (généralement 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).

7.2.4 Déroulement sur un exemple

Une situation de dialogue avec support visuel médiatisé. La figure 7.3 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 qui est ici 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 7.3.

Parmi les expressions référentielles à interpréter au cours de cet extrait consistant à construire une ligne d'horizon dans une scène représentant les grandes pyramides d'Egypte, nous nous concentrerons sur celles, en italique dans l'extrait, qui correspondent aux références dans la zone de manipulation. L'intérêt de la suite : « *les deux grands triangles* » (I_2), « *la pyramide de gauche* » (I_5), « *la pyramide de droite* » (I_7) et « *la petite pyramide* » (I_{10}), réside dans l'interprétation de « *la pyramide de droite* » (I_7) qui ne désigne pas, comme nous pourrions nous y attendre hors

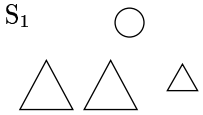
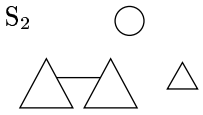
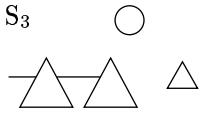
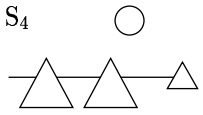
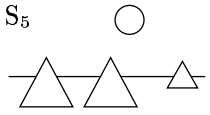
I_1	« Il faut prendre une grande horizontale. »	S_1 
I_2	« Et la placer à la pointe *des deux grands triangles* [...] »	
M_1	*geste de manipulation directe menant à la scène* S_2	
I_3	« Voilà comme ça... et tu en prends une deuxième. »	S_2 
I_4	« Une petite [...] »	
I_5	« Tu la places à gauche de *la pyramide de gauche* [...] »	
M_2	*geste de manipulation directe menant à la scène* S_3	
I_6	« Voilà comme ça [...] et t'en prends une autre petite [...] »	S_3 
I_7	« Et tu la places à droite de *la pyramide de droite* [...] »	
M_3	*geste de manipulation directe menant à la scène* S_4	
I_8	« Voilà comme ça... et tu prends une autre petite verticale. »	S_4 
M_4	« Une autre petite verticale? »	
I_9	« Euh horizontale pardon. »	S_5 
I_{10}	« Et puis tu la places dans la même lignée à droite de *la petite pyramide.* »	
M_5	*geste de manipulation directe menant à la scène* S_5	

Figure 7.3 – *Extrait du corpus Ozkan (transcription et scènes visuelles).*

contexte, le petit triangle le plus à droite de la scène, mais le plus à droite parmi les deux grands triangles, et ceci sans ambiguïté pour les interlocuteurs.

Cette interprétation échappe 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 *et al.* 2001) 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* ». Comme nous allons le montrer, le modèle des domaines de référence est parfaitement adapté à la compréhension d'une telle expression.

Illustration de l'interaction des modalités. L'extrait illustre bien la collaboration des critères perceptifs, linguistiques et applicatifs : dans les scènes visuelles, les trois triangles sont regroupés sur les critères de similarité (ce sont les trois triangles de la scène), de proximité et de continuité (ils sont disposés quasiment selon une même horizontale). La similarité et la proximité jouent surtout pour regrouper les deux grands triangles, de même qu'ils sont regroupés par l'expression « *les deux grands triangles* » conservée dans l'historique du dialogue. La tâche applicative a également un rôle important : la ligne d'horizon est construite par morceaux en s'aidant de supports, ces supports étant les trois pyramides. Or la pose du premier segment horizontal s'appuie sur les deux grandes pyramides, puis la pose du deuxième segment s'appuie sur la plus à gauche de ces deux pyramides. Il est presque prévisible que la pose du troisième segment s'appuie sur l'autre. D'autres arguments viennent renforcer l'interprétation de « *la pyramide de droite* » comme « *la pyramide de droite parmi les deux grandes pyramides* ». Au niveau de la perception

visuelle, c'est la symétrie de positionnement par rapport aux deux grandes pyramides. A un niveau linguistique, c'est la symétrie dans les expressions référentielles « *à gauche de la pyramide de gauche* » et « *à droite de la pyramide de droite* »). Tout concourt donc à rendre cette dernière expression non ambiguë.

Calcul des domaines de référence. La figure 7.4 montre la seule hypothèse pertinente de domaine sous-spécifié élaborée lors de l'interprétation de « *la pyramide de droite* » dans l'exemple de la figure 7.3. Elle montre d'autre part quelques exemples de domaines construits. 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 DR_2 sur le schéma). L'unification de ce domaine avec celui correspondant à la requête se fait grâce à l'assimilation de « *pyramide* » et « *triangle* » tout en acceptant le modifieur « *grand* », de même que sont assimilés « *gauche* » et « $\neg$*droite* » (et « $\neg$*gauche* » à « *droite* »). Le module visuel renvoie quant à lui deux domaines DR_4 et DR_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 O_2.

L'unification de ces domaines conduit bien à l'identification du référent O_2 dans le domaine constitué par O_1 et O_2 : le module linguistique met en avant ce résultat ; le module visuel classe également ce résultat (correspondant à DR_4) avant l'autre possibilité (DR_5), car la similarité au niveau le plus immédiat (forme et taille) intervient avant la combinaison de la bonne continuité et de la similarité à un niveau plus faible (uniquement la forme) ; le module tâche conduit également à ce résultat avec une focalisation directe sur O_2. Si l'ambiguïté est détectée, la prise en compte de multiples critères permet d'interpréter comme l'a fait intuitivement le manipulateur.

7.3 Perspectives pour un modèle formel de la pertinence

Maintenant que nous avons exploré les caractéristiques référentielles de la perception visuelle, du langage et du geste, maintenant que nous avons, en spécifiant un modèle d'interprétation de la référence, exploré les composantes, les étapes, les besoins de critères d'évaluation impliqués dans le processus de compréhension, nous pouvons tenter d'aller plus loin dans l'adaptation de la Théorie de la Pertinence. Nous proposons dans cette section quelques pistes pour une formalisation de la pertinence, c'est-à-dire pour un modèle évaluant directement la pertinence d'une expression référentielle, en tenant compte de tous les paramètres contextuels que nous avons détaillés. Après avoir montré l'intérêt de cette démarche, nous analysons en détail les composantes des effets contextuels et de l'effort de traitement que nous proposons pour une expression référentielle multimodale. Cette proposition, même si elle reste à l'état à la fois d'ébauche de modélisation et de projet de recherche, réunit les principaux apports des sections § 7.1 et § 7.2.

7.3.1 Intérêts et approches pour une quantification de la pertinence

Les effets contextuels d'une action de référence. Sperber et Wilson (1995) considèrent le contexte C comme un ensemble d'informations et définissent une contextualisation de la proposition P dans C comme une déduction ayant pour prémisses les informations de P et de C

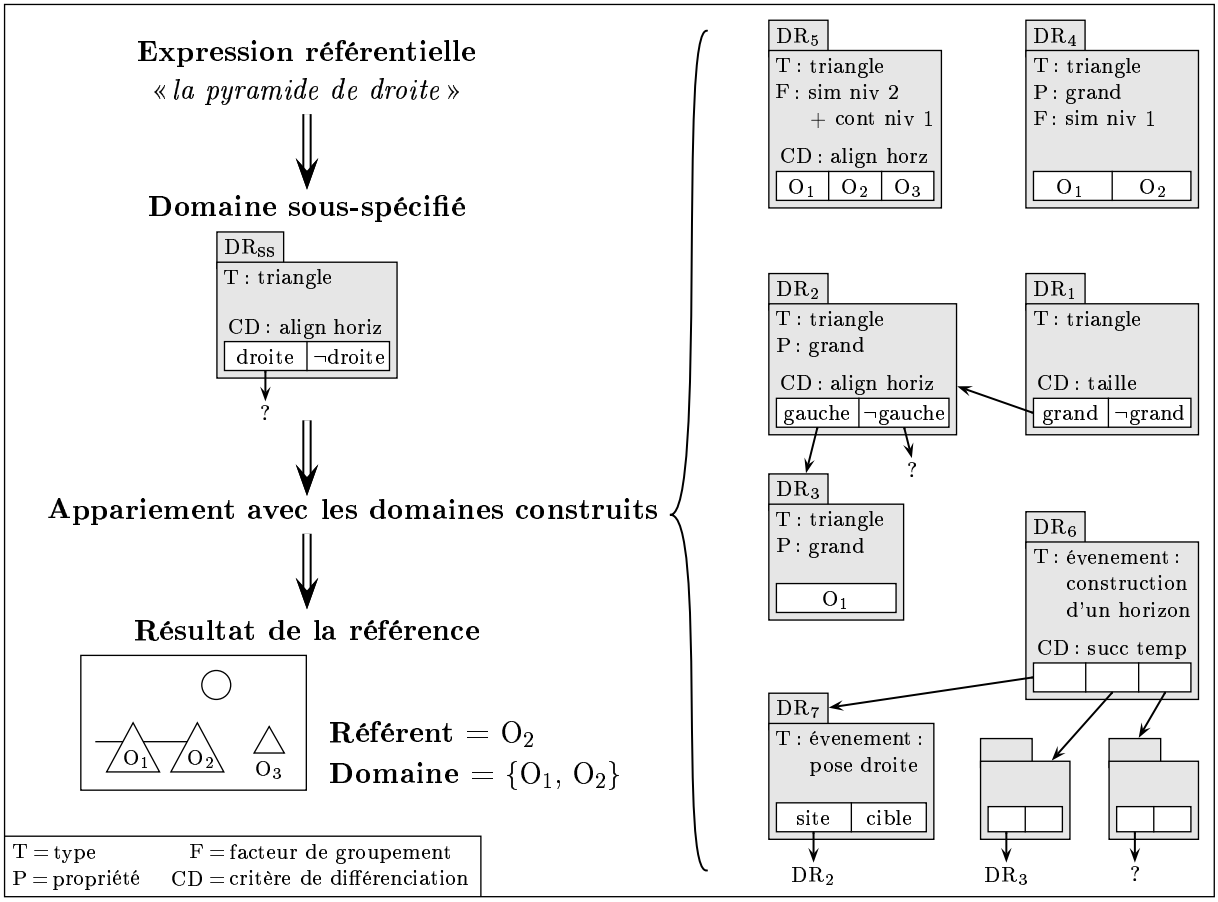


Figure 7.4 – *Quelques domaines de référence pour l'exemple de la figure 7.3.*

(cf. page 128). Il y a effet contextuel quand une contextualisation ne fait pas qu'ajouter à C les informations nouvelles de P mais entraîne en plus d'autres modifications de C (par exemple l'effacement de certaines hypothèses de C). Notre but ici est de déterminer quels sont les effets contextuels non pas d'une proposition issue d'un énoncé complet, mais d'une action de référence. En considérant que C contient des propositions telles que « *deux triangles rouges sont visibles dans la scène* », « *le triangle le plus saillant visuellement est le triangle vert* », ou encore « *aucun objet n'est visuellement focalisé* », nous pouvons avancer l'idée que les effets contextuels de l'expression référentielle « *le grand triangle rouge* » sont les suivants :

- il y a référence ;
- cette référence porte sur l'objet O_{42} (le seul grand triangle rouge dans le domaine de référence impliqué par la référence) ;
- l'objet O_{42} est focalisé (visuellement et linguistiquement).

Cette dernière proposition suffit à montrer qu'en plus des informations nouvelles liées à la référence, C a bien été modifié par la suppression de la proposition « *aucun objet n'est visuellement focalisé* ». Nous remarquons d'autre part que ces trois propositions ne dépendent pas de la forme prise par l'expression référentielle. Compte tenu du fait que ces propositions sont liées au résultat de la référence et non à son processus de résolution, nous n'irons pas plus loin dans l'étude des effets contextuels. Dans notre cadre de recherche, il est encore trop tôt pour rendre possible une évaluation des effets contextuels : la nature exacte des informations et des hypothèses contenues

dans C doit tout d'abord être clairement spécifiée, ainsi que les façons dont elles sont déterminées. Nous laissons de côté ce problème pour nous intéresser à l'effort de traitement.

L'effort de traitement d'une expression référentielle. En ce qui concerne l'évaluation de l'effort de traitement d'une référence, la méconnaissance des opérations élémentaires constituant les processus mentaux activés lors d'un tel traitement nous conduit à des hypothèses. D'une manière générale, nous supposerons que l'effort est proportionnel à la complexité des différentes informations intervenant lors du traitement, à leur accessibilité, et à l'importance des interactions entre ces informations.

Pour pouvoir évaluer numériquement l'effort de traitement, il s'avère nécessaire de le décomposer en traits. Nous parlions par exemple de complexité des informations intervenant lors de la résolution de la référence. Parmi ces informations se trouvent les composants de l'expression référentielle elle-même. Deux exemples de traits sont donc la complexité de l'expression verbale et la complexité du geste ostensif. Chaque d'eux peut de même se décomposer en d'autres traits. Ce qu'il est intéressant de constater, c'est que chacun de ces traits fait appel à une notion que nous avons étudié. Ainsi, la complexité du geste fait intervenir le nombre de singularités présentes dans une trajectoire, ce qui donne un intérêt supplémentaire à cette notion de singularité. La modélisation en traits que nous allons proposer repose ainsi sur le travail que nous avons fait jusqu'à présent. Avant de nous lancer dans des considérations quantitatives, nous nous attacherons à identifier le maximum de traits. Pour cela, nous procéderons en quatre étapes. La première ne fera intervenir que les informations provenant de la perception visuelle. La seconde intégrera les informations apportées par l'expression référentielle verbale, dans le but de proposer une base pour la modélisation de l'interprétation de la référence langagière. La troisième étape intégrera les informations provenant de la perception et d'une trajectoire gestuelle. Pour l'interprétation de la référence multimodale, la quatrième et dernière étape intégrera les informations apportées par les trois modalités.

7.3.2 La pertinence d'une expression référentielle multimodale

Informations liées à la perception. Les informations nouvelles qui viennent s'ajouter dans C lors de la perception visuelle sont liées à la structuration de la scène. Il s'agit de la construction d'une représentation cognitive du contexte visuel. Quant à l'effort de traitement, il est lié à la complexité des dendrogrammes construits avec les critères de la Gestalt. Cette complexité s'évalue avec la profondeur des dendrogrammes (nombre de niveaux de lecture, c'est-à-dire nombre de partitions), éventuellement avec une pondération selon la distance qui sépare ces niveaux. On peut également considérer qu'elle fait intervenir le nombre de nœuds des dendrogrammes.

Informations liées à la perception et à l'expression référentielle verbale. Les informations nouvelles qui viennent s'ajouter dans C lors de l'interprétation en contexte visuel d'une expression référentielle verbale sont les suivantes : il y a eu une référence verbale ; cette référence a porté sur tel(s) objet(s). Quant à l'effort de traitement, il se décompose selon les traits suivants :

1. adéquation de l'expression verbale avec l'intention de référence en contexte applicatif ;
2. adéquation de l'expression avec l'intention de référence en contexte dialogique ;
3. complexité de la perception visuelle ;
4. complexité de l'expression verbale ;
5. effort nécessaire pour isoler le(s) référent(s) à l'aide de la perception, de l'expression verbale et de l'historique de l'interaction.

Le premier trait correspond à l'effort nécessaire pour intégrer l'expression référentielle verbale dans le contexte de tâche. Par exemple, si la tâche impose le traitement des objets par type, comme c'est le cas de la tâche de rangement dans la simulation Magnét'Oz, « *sélectionner un triangle et un carré* » demande un effort particulier du fait de son incongruité (aucune action ne peut être appliquée à cet ensemble hétérogène).

Le deuxième trait correspond à l'effort nécessaire pour contextualiser l'expression verbale dans le contexte dialogique. Par exemple, « *le rouge* » après « *le triangle vert* » demande un effort pour la résolution de l'ellipse.

Le troisième trait correspond à l'effort nécessaire pour prendre connaissance de la scène. Il s'évalue à l'aide de critères tels que le nombre d'objets présents dans la scène, le nombre de groupements perceptifs, la diversité des objets (par exemple le nombre de couleurs différentes).

Le quatrième trait s'évalue à l'aide de critères syntaxiques tels que le nombre de mots et la complexité de la structure syntaxique, par exemple la profondeur ou le nombre de nœuds de l'arbre syntaxique. Les mots ayant plus ou moins d'importance, des pondérations peuvent être envisagées. Le cas particulier des connecteurs pragmatiques est intéressant. Selon Moeschler dans (Reboul & Moeschler 1998b), les connecteurs (conjonctions de coordination, « *parce que* », etc.) jouent un rôle au niveau de la facilitation du traitement de l'information : leur fonction est de minimiser les efforts cognitifs. Ce sont des guides pour l'interprétation. Bien que Moeschler analyse les connecteurs dans une proposition et non dans un simple groupe nominal, nous pouvons tenir compte de cette facilitation, par exemple en ne comptant pas le mot « *et* » dans une coordination, ou en le pondérant de manière très faible.

Le cinquième trait s'évalue à l'aide de critères tels que la difficulté à repérer dans la scène les objets vérifiant les propriétés données dans l'expression verbale. Par exemple, « *le triangle rouge* » demandera moins d'effort que « *le petit triangle* » si le référent est le seul objet rouge de la scène alors qu'il n'est pas le seul petit objet (les autres petits objets n'étant pas des triangles). L'historique de l'interaction est pris en compte de la manière suivante : d'une part en ajoutant un effort lié au rappel d'un contexte visuel antérieur, par exemple pour l'interprétation de « *remets le triangle vert* » après « *efface les triangles* » ; d'autre part en ajoutant un effort lié au rappel d'informations linguistiques utiles pour la résolution de la référence (par exemple pour l'interprétation de « *l'autre* »).

Informations liées à la perception et au geste. Les informations nouvelles qui viennent s'ajouter dans C lors de la perception d'un geste ostensif en contexte visuel sont les suivantes : une partie du contexte perceptif a été rendue saillante ; cette saillance s'est appliquée sur tel(s) objet(s). Quant à l'effort de traitement, il se décompose selon les traits suivants :

1. adéquation du geste avec l'intention de démonstration en contexte applicatif ;
2. complexité de la perception visuelle ;
3. complexité de la trajectoire gestuelle ;
4. effort nécessaire pour isoler le(s) référent(s) à l'aide de la perception, du geste et de l'historique de l'interaction.

Le premier trait correspond à l'effort nécessaire pour intégrer le geste dans le contexte de tâche. Par exemple, si la tâche impose le traitement des objets par type, un geste entourant deux objets de type différent demande un effort particulier du fait de son incongruité.

Le deuxième trait est identique à celui correspondant lors de la première étape.

Le troisième trait s'évalue à l'aide de critères tels que la longueur et la complexité de la trajectoire gestuelle. La longueur s'évalue par rapport au gabarit des objets, et la complexité par rapport au nombre de singularités, comme nous l'avons déjà évoqué.

Le quatrième trait s'évalue à l'aide de critères tels que la difficulté à repérer dans la scène les objets désignés par le geste. Par exemple, un geste imprécis demande un effort particulier, même s'il n'est pas ambigu. Un autre exemple important est le suivant : lorsque les référents sont proches de distracteurs, un geste élictif demande plus d'effort qu'un geste séparateur. D'autre part, l'appel à l'historique de l'interaction consiste à ajouter un effort supplémentaire lorsque par exemple le geste effectué est nouveau en termes de type de trajectoire et de type d'accès aux demonstrata. Au contraire, lorsque un geste complexe est effectué plusieurs fois de suite sans variation, l'effort nécessaire à son traitement diminue à chaque fois.

Informations liées à la perception et à l'expression référentielle multimodale. Les informations nouvelles qui viennent s'ajouter dans C lors de l'interprétation en contexte visuel d'une expression référentielle multimodale sont les suivantes : il y a eu une référence multimodale ; cette référence a porté sur tel(s) objet(s). Quant à l'effort de traitement, il se décompose selon les traits suivants :

1. adéquation de l'expression multimodale avec l'intention de référence en contexte applicatif ;
2. adéquation de l'expression avec l'intention de référence en contexte dialogique ;
3. complexité de la perception visuelle ;
4. complexité de la ou des expressions référentielles verbales ;
5. complexité du ou des gestes ;
6. effort nécessaire pour associer le ou les gestes à la ou aux expressions verbales ;
7. effort nécessaire pour isoler le(s) référent(s) à l'aide de la perception, de l'expression multimodale et de l'historique de l'interaction.

Le premier trait correspond à l'effort nécessaire pour intégrer l'expression référentielle multimodale dans le contexte de tâche. Il s'évalue à l'aide des efforts correspondants dans les étapes précédentes.

Le deuxième trait correspond à l'effort nécessaire pour intégrer l'expression multimodale dans le contexte dialogique. Par exemple, « *celui-là* » après « *celui-ci* » demande moins d'effort que « *celui-ci* » dans la même situation.

Les troisième, quatrième et cinquième traits ne diffèrent pas de ceux des étapes précédentes.

Le sixième trait s'évalue à l'aide de critères tels que la composition de l'expression multimodale (nombre de gestes, nombre d'expressions verbales), la qualité de la synchronisation entre geste(s) et expressions(s) verbale(s), la présence d'une ambiguïté dans l'une des modalité (ambiguïté résolue par l'autre modalité au prix d'un certain effort), ainsi que le choix du déterminant, d'éventuels marqueurs déictiques et d'éventuels adjectifs numéraux. Par exemple, du fait de son association avec un geste, un démonstratif est plus naturel qu'un défini et demande moins d'effort.

Le dernier trait s'évalue à l'aide de critères tels que la difficulté à repérer dans la scène les objets désignés par l'expression multimodale. Les évaluations de cette partie de l'effort lors des étapes précédentes n'ont ici aucune utilité. Par exemple, l'expression multimodale composée d'un geste désignant un triangle et associé à « *ce triangle* » demande moins d'effort que l'expression multimodale composée du même geste associé à « *cet objet* », pour laquelle intervient une récupération visuelle de la catégorie. L'historique de l'interaction est pris en compte en ajoutant un effort lié au rappel d'un contexte visuel antérieur ou au rappel d'informations linguistiques utiles pour la résolution de la référence (comme nous l'avons vu lors de la deuxième étape). L'historique des trajectoires gestuelles déjà effectuées intervient également pour diminuer l'effort de traitement d'une trajectoire souvent utilisée.

Critique de cette approche et perspectives. Il manque beaucoup de choses à cette première proposition. Nous n'avons par exemple considéré l'intention de référence (ou but communicatif)

qu'au moment précis de la référence. Or celle-ci peut être effectuée dans un certain but qui n'apparaîtra que plus tard, lors d'une référence ultérieure. Évaluer la pertinence sans tenir compte de ce but s'avère insuffisant. En effet, même s'il est impossible pour le système de deviner le but communicatif de l'interlocuteur au sens large, il est toujours intéressant de calculer la pertinence *a posteriori*, dans le but par exemple de revenir à la source d'un malentendu. Ce calcul peut alors s'appuyer sur les résultats des énoncés les plus récents, et donc sur les intentions référentielles identifiées. En lien avec cet aspect de continuité d'une référence à l'autre, il manque à notre proposition une meilleure prise en compte des phénomènes liés aux domaines de référence et aux partitions. Par exemple, il semble probable qu'un changement de partition dans un même domaine demandera plus d'effort de traitement qu'un changement d'élément focal dans la même partition du même domaine.

La saillance visuelle n'a pas non plus été prise en compte dans la caractérisation de l'effort de traitement. Son rôle est pourtant double. D'une part elle intervient en tant que facteur de réussite de la référence : elle peut rendre non ambiguë une référence telle que « *le N* » dans un environnement comprenant plusieurs N dont un fortement saillant. Les évaluations doivent être révisées en conséquence. D'autre part elle intervient en tant que facteur de diminution de l'effort de traitement.

Pour la caractérisation des effets contextuels, il semble que la contextualisation ne se limite pas à « il y a eu référence qui a porté sur tels objets ». Par exemple, l'emploi de « *l'un* » a un effet linguistique particulier dû à ce qu'on attend ensuite « *l'autre* ». De même, l'emploi de « *celui-ci* » dans une intention mentionnelle a un effet particulier, dû à ce qu'on peut éventuellement attendre ensuite « *celui-là* ». La tâche pourrait également intervenir dans le calcul des effets dans le sens que si l'utilisateur s'occupe d'un triangle parmi deux, le deuxième va sans doute être traité bientôt, du moins si la tâche incite à traiter les objets par catégorie. Dans un même ordre d'idée, le contexte visuel intervient avec la notion de ligne directrice : une référence à un objet situé au départ de la ligne entraîne l'hypothèse d'une référence ultérieure à l'objet suivant selon l'ordonnancement amorcé.

Avec ces ébauches de caractérisations des effets contextuels et de l'effort de traitement, nous montrons que la formalisation de la pertinence est encore un objectif à long terme, car faisant intervenir tous les paramètres intervenant dans la communication d'une intention. Nous espérons que les traits proposés, ainsi que les pistes données pour leur quantification, constitue une base intéressante pour l'élaboration de futures caractérisations. Bien que nous nous soyons limité aux actions de référence aux objets, nous considérons qu'une extension au problème de la compréhension en général est possible : en partant du calcul des effets et de l'effort des composants de l'énoncé, il sera plus facile ensuite d'appréhender les effets et l'effort de l'énoncé complet.

RÉCAPITULATIF
L'idée principale de ce chapitre est que le critère de pertinence avancé par Sperber et Wilson, non seulement peut s'appliquer à la communication multimodale telle que nous la concevons, mais de plus sous-tend les morceaux de modélisation que nous avons proposés jusqu'à présent. Le modèle que nous présentons ici se fonde donc fortement sur la pertinence. Les trois parties qui le composent, c'est-à-dire la modélisation de la focalisation spatiale, la sous-spécification de domaines de référence et leur unification avec des domaines disponibles, constituent notre réponse au problème de la référence aux objets dans un contexte regroupant trois modalités. Nous montrons que c'est lors de la phase de détermination d'un domaine sous-spécifié que sont émises les hypothèses relatives aux composants linguistiques de l'expression référentielle. Cette phase s'appuie sur tous les phénomènes que nous avons retenus et constitue notre principal apport. Suite à cette modélisation qui constitue un aboutissement des chapitres précédents, nous pouvons explorer en profondeur la notion de pertinence pour proposer des pistes dans l'élaboration d'un modèle de la communication entièrement fondé sur elle. Bien que ces pistes prennent pour le moment la forme de simples ensembles de traits, nous montrons quels sont les intérêts et les difficultés ouvertes par cette voie de recherche.

Troisième partie

Applications du modèle

Chapitre 8

Une architecture pour la gestion de domaines

Comment le modèle des domaines de référence peut-il se traduire en recommandations d'implantation ? Quelles sont les données gérées lors du processus d'interprétation ? Comment les différentes facettes de ce processus peuvent-elles se répartir dans des modules ? Comment se déclenchent les appels aux différents modules? Quelles sont les données qui transitent ?

Nous abordons dans ce chapitre certains aspects logiciels liés aux domaines de référence et à la façon dont ils permettent d'intégrer des informations provenant de la perception visuelle, du langage et du geste. Compte tenu des particularités de la structuration de domaines visuels en dendrogrammes, du calcul de la saillance visuelle ou de la saillance linguistique, nous serons amené à spécifier un module particulier pour chacun de ces aspects de notre modèle. Le but de ce chapitre est de détailler ces modules et leur fonctionnement. Pour cela, nous commençons (§ 8.1) par spécifier les types de données sur lesquelles repose l'architecture, pour enchaîner (§ 8.2) sur les conditions d'exploitation de ces données, c'est-à-dire sur les algorithmes caractérisant le fonctionnement global du système, et terminer (§ 8.3) sur la spécification des modules qui regroupent ces algorithmes.

8.1 Les données gérées par le système

Nous avons déjà parlé de la base des objets et de la base des fonctions caractérisant la tâche applicative. Nous avons évoqué dans le chapitre 1 le modèle du domaine regroupant ces deux bases, ainsi que le modèle de la langue regroupant les connaissances qu'a le système du fonctionnement du langage. Nous détaillons dans cette section quelques aspects de ces éléments, ainsi que la nature du modèle de l'utilisateur et celle de deux modèles propres à notre système, le modèle de la saillance et le modèle de la pertinence. Nous analysons également de manière approfondie la nature et les composants de l'historique de l'interaction.

8.1.1 Données statiques

Le modèle du langage. D'une manière générale, il regroupe les connaissances qu'a le système des composantes lexicales, syntaxiques et sémantiques du langage courant (cf. par exemple Pierrel 1987). Ces connaissances sont statiques, c'est-à-dire déterminées *a priori* et inchangées au cours du dialogue. Il s'agit donc du vocabulaire usuel qui comprend par exemple les déterminants présentés dans le chapitre 1, des structures syntaxiques et des caractéristiques sémantiques courantes. Ces dernières se représentent généralement sous la forme de traits (trait « animé » pour un être vivant, par exemple). Toutes ces données permettent de traduire le signal vocal en une forme logique. Comme nous nous préoccupons du processus d'interprétation et de la pragmatique en général, nous n'entrerons pas dans les détails de représentation et d'étendue de ces connaissances.

Le modèle du dialogue. Il contient la description de diverses situations de dialogues, pour permettre au système de réagir correctement à telle ou telle situation. Dans sa modélisation de ce modèle, Bilange (1992) place les règles de gestion de l'historique du dialogue ; des règles de dialogue telles que les règles de contrôle (permettant de déterminer à quel moment le système peut prendre l'initiative) ; les règles de gestion des tours de parole ; les grandes règles de calcul des interventions du système, de prédiction des interventions de l'utilisateur et d'interprétation de ces interventions. Ces données permettent d'assurer le bon déroulement du dialogue. Comme nous nous préoccupons des actions de référence, nous n'entrerons pas plus dans les détails.

Le modèle du domaine (ou modèle de la tâche ou de l'application). En complément aux modèles du langage et du dialogue, le modèle de la tâche vient ajouter les connaissances spécifiques à la tâche applicative concernée. Il s'agit ainsi de la base des objets et des fonctions de l'application ; de la définition des buts et des sous-buts qui spécifient l'accès aux données de l'application (comprenant en particulier les valeurs par défaut qui permettent d'interpréter des énoncés qui ne précisent qu'un ensemble minimal de paramètres) ; de la spécification des stratégies de gestion du dialogue spécifiques à la tâche (précisant par exemple les liaisons entre les diverses situations). L'instanciation du contexte de tâche pour une application donnée ne doit pas modifier la structure du modèle de la tâche. Nous reviendrons sur ce modèle et sur la nature du contexte de tâche dans le chapitre 9 page 177 (§ 9.2).

Le modèle de la saillance. Il comprend la liste des propriétés sur lesquelles baser les calculs de saillance. Il comprend également les pondérations propres à chacune de ces propriétés, traduisant ainsi leur importance relative. A titre d'exemple, nous étudierons en § 9.1.1 l'adaptation du modèle de la saillance visuelle à une tâche applicative particulière dans le cadre du projet COVEN traitant d'interaction multimodale dans un environnement virtuel.

Le modèle de la pertinence. Il comprend les principes de calcul des effets contextuels et de l'effort de traitement, que ce soit pour les expressions référentielles, pour les trajectoires gestuelles, pour les domaines de référence, pour les contenus propositionnels des énoncé, ou, d'une manière générale, pour l'évaluation d'hypothèses émises dans un cadre permettant le calcul des effets et de l'effort, et nécessitant la sélection des hypothèses les plus pertinentes.

8.1.2 Données dynamiques

L'état de l'environnement applicatif. Les caractéristiques des objets changent au cours du dialogue : les coordonnées changent suite à une action de déplacement, le nombre d'objets évolue

suite aux créations et aux suppressions. Il s'agit de données dynamiques, gérées dans une base de données. Ce sont les différents stades par lesquels passe l'état de l'environnement qui sont enregistrés dans un corpus multimodal. Dans le corpus Magnét'Oz, chaque état est conservé et mis en rapport avec les énoncés et les trajectoires produites.

Les historiques. Nous avons déjà beaucoup parlé de l'historique de l'interaction et de l'historique du dialogue. A la fin du chapitre 7, nous avons également évoqué un historique des trajectoires gestuelles effectuées par l'utilisateur. Nous voulons ici faire le point sur la nature des différents historiques et sur les données qui y sont conservées. A quoi sert un historique ? Que représente-t-il ? Qu'y stocker et dans quelles limites ? Faut-il un historique pour la perception visuelle, pour le geste, pour la parole, pour les référents, pour les domaines ? Nous allons passer en revue les types d'historique possibles.

L'historique du dialogue (ou historique langagier) a pour but principal de garder la trace des expressions utilisées pour référer à tel ou tel objet. Plus précisément, sont ainsi conservées les chaînes de référence complètes, c'est-à-dire les expressions référentielles, les référents, les domaines de référence, et surtout leur structuration arborescente. L'intérêt de cet historique est double : d'une part, il permet la réutilisation des algorithmes s'appuyant sur le contexte linguistique, en particulier pour la résolution des pronoms personnels ; 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 dans l'exemple suivant : « *un triangle rouge et un triangle vert* » suivi de « *mets le triangle rouge sur la droite et supprime l'autre* » (possible et effectivement autorisé par l'historique), ou de « *mets-le sur la droite* » (impossible et effectivement interdit par l'historique).

Comme nous l'avons vu dans le chapitre 4, l'historique du dialogue consiste en une structuration de domaines de référence, certains éléments identifiés étant focalisés, d'autres n'étant pas encore identifiés mais supposés exister. Suite à une référence, sont conservés à la fois l'expression référentielle et le référent. En effet, ne conserver que l'expression référentielle s'avère insuffisant, non seulement pour ne pas avoir à refaire le calcul de la référence lors d'un appel à une référence antérieure, mais aussi pour un traitement adéquat des référents évolutifs (cf. exemple page 20) ou des suppressions d'objets (dans « *supprime le grand bureau ; remplace-**le** par une table basse* », la reprise de l'objet est impossible puisque l'objet n'existe plus). Conserver l'expression référentielle en plus du référent est indispensable pour une exploitation ultérieure de la façon dont l'utilisateur a accédé à ce référent.

L'historique visuel conserve les états et les structurations successives de la scène. Un exemple d'utilisation 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. A l'image de l'exemple précédent, le système peut en effet avoir à traiter l'énoncé suivant : « *supprime le grand bureau et mets une table basse un peu plus vers le mur* ». Dans le corpus Magnét'Oz, on assiste de même à la situation suivante : « *enlève cet objet* » associé à un geste et suivi de « *enlève aussi celui au-dessous* ».

L'historique de tâche conserve les actions effectuées, les référents sur lesquels elles ont porté, et surtout un repérage de ces objets et de ces actions par rapport aux buts et sous-buts courants. Sa description n'entre pas dans la problématique qui nous intéresse ici, car elle fait intervenir les stratégies de communication des interlocuteurs. Nous y reviendrons partiellement dans le chapitre 9, lors de l'adaptation de notre modèle à un type de tâche particulier (§ 9.2).

L'historique global regroupe dans une certaine mesure toutes ces informations. Il ne s'agit pas seulement de pointeurs vers les historiques locaux, mais il contient également des repères à propos des différentes actions et des différentes phases du dialogue : résultats des références, évaluations *a posteriori*, réponses du système, repères temporels. Il inclut en ceci ce que nous avons appelé jusqu'à présent l'historique de l'interaction.

Un aspect fondamental de l'historique global est de gérer sa propre capacité de stockage. D'un point de vue cognitif, si nous ne retenons pas ici le critère de Miller (1956) consistant à ne conserver que les sept éléments remplissant la mémoire de travail de l'utilisateur, nous retiendrons un critère lié à l'attention de celui-ci. En effet, plus une entité est ou a été l'objet de l'attention de l'utilisateur, et plus il nous semble important de la repérer comme telle dans l'historique, qu'il s'agisse de l'historique visuel ou de l'historique de l'interaction. Nous proposons ainsi la gestion de scores d'attention selon le principe suivant : plus l'utilisateur travaille sur un objet, plus le score attentionnel de cet objet augmente. Il est alors possible d'étendre ce principe à la catégorie, voire à chaque propriété de l'objet : plus l'utilisateur travaille sur un type d'objet particulier (ou sur une couleur particulière, par exemple), plus le score attentionnel de cette entité (classe d'objets, propriété) augmente. De même, plus l'utilisateur effectue un type d'action, plus le score attentionnel de la fonction applicative correspondante augmente. Tous ces scores sont retenus, de façon à ce que la résolution de la référence traite en priorité les objets ayant les scores les plus élevés. Avec ces considérations, le score d'un objet se calcule selon le score courant de sa catégorie et de chacune de ses propriétés. Au cours du temps, tous les scores diminuent régulièrement pour rendre compte des phénomènes de diminution de l'attention. Cette gestion de l'attention s'inspire directement du modèle des logogènes de Morton (1982) que nous avons présenté page 42. Nous illustrons en cela l'intérêt de certains modèles issus de la psycholinguistique, bien que ces modèles ne soient initialement pas prévus pour un traitement automatique.

Le modèle de l'utilisateur. Une autre ensemble de données dynamiques regroupe tout ce qui concerne l'utilisateur. En plus des informations propres à la gestion des droits (droits de certains utilisateurs par rapport à d'autres, en particulier pour les applications d'interrogation de bases de données), nous y mettons tout ce qui peut simplifier l'interprétation et particulièrement le calcul de la référence, c'est-à-dire :

1. Des informations relatives à la reconnaissance de la parole : règles phonologiques et prosodiques propres à un locuteur (acquises au cours de la reconnaissance).
2. Des informations sur la façon dont l'utilisateur semble percevoir la scène visuelle : plutôt en groupes perceptifs fondés sur le critère de proximité ; avec un sens de lecture préférentiellement de gauche à droite ; en fonction de son propre référentiel ou du référentiel intrinsèque de l'objet, par exemple pour l'interprétation de « *à gauche de la chaise* » pour une chaise qui lui fait face (la gauche de la chaise correspond dans ce cas à la droite de l'utilisateur).
3. Des informations sur la façon dont il utilise le geste : plutôt élictif ou au contraire souvent séparateur ; généralement précis ; se traduisant par des trajectoires toujours complètes.
4. Des informations sur son vocabulaire et ses constructions syntaxiques privilégiées : fréquence d'utilisation des différentes constructions ; emploi privilégié de certains mots ambigus (le système retiendra par exemple qu'il arrive à l'utilisateur d'employer « *cet objet* » pour désigner un groupe d'objets).
5. Des informations sur la façon dont il associe la parole et le geste : synchronisation généralement parfaite (ou au contraire tendance à produire le geste avant l'expression référentielle coréférente) ; tendance à utiliser un groupe nominal défini avec le geste ; etc.

8.1.3 Données pour le calcul de la référence

En partant des données statiques et dynamiques, le calcul de la référence implique aussi des données spécifiques, créées au cours de l'interprétation et détruites juste après. Parmi ces données se trouvent les domaines de référence calculés, sous-spécifiés ou construits à l'aide des différentes sources contextuelles. En nous plaçant du point de vue du module chargé de l'interprétation d'un énoncé multimodal, nous considérons l'énoncé en entrée et la spécification des fonctions de l'application à exécuter en sortie. Du point de vue du module chargé de la résolution de la référence aux objets, nous considérons les gestes et expressions référentielles verbales en entrée,

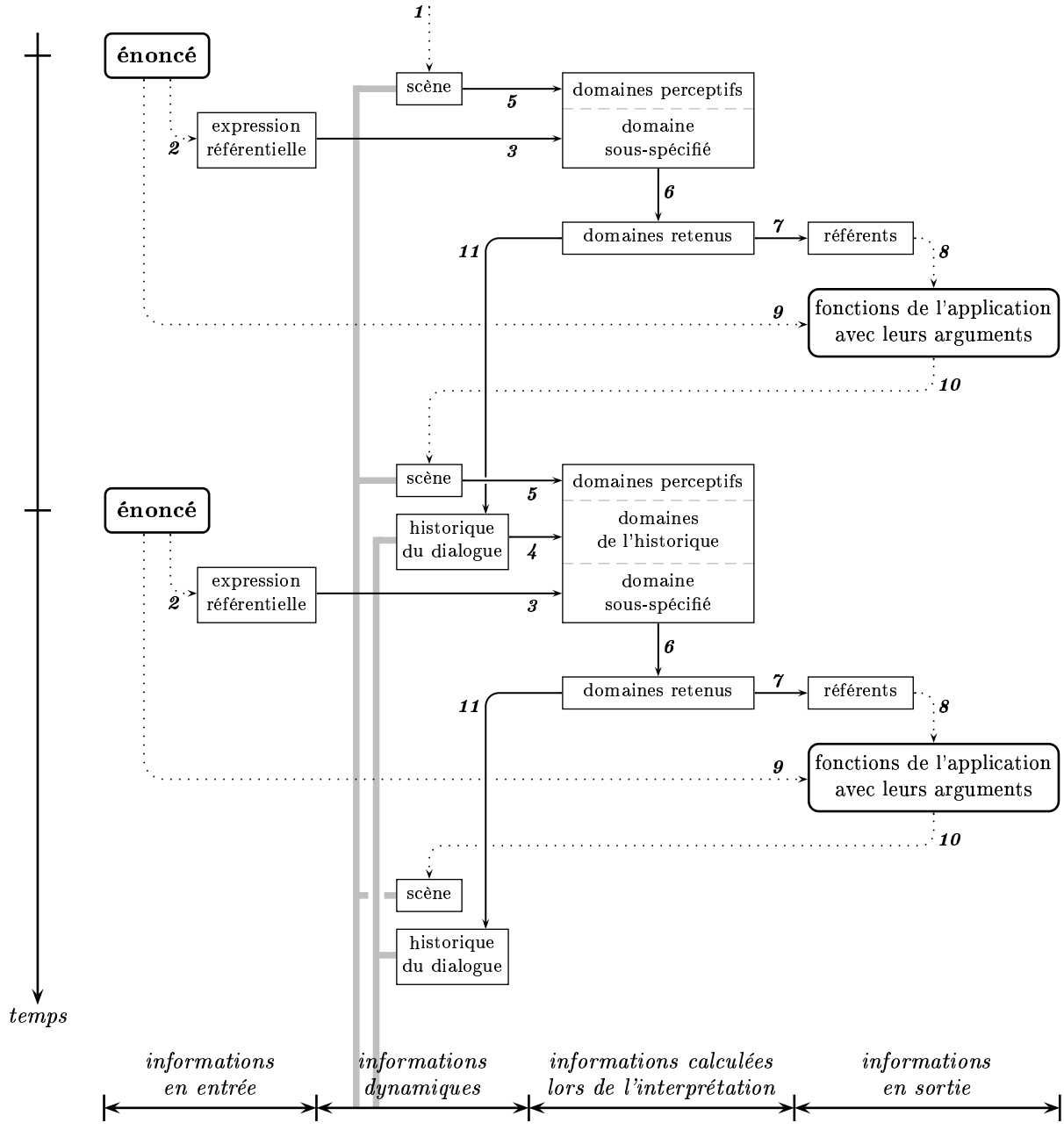


Figure 8.1 – *Fonctionnement simplifié d'un système de dialogue sur deux énoncés oraux.*

et les référents qui seront les paramètres des fonctions à appliquer en sortie. Les quatre types de données mis ainsi en jeu sont présentés dans la figure 8.1 pour l'interprétation d'un énoncé oral en première mention et d'un deuxième énoncé oral qui lui fait suite.

Dans cette figure, les flèches entre les boîtes représentent des étapes du traitement. Celles représentées par un trait continu concernent la résolution de la référence et sont au cœur de notre problématique. Celles représentées en pointillé concernent d'autres compétences d'un système de dialogue. Ces flèches sont identifiées par des nombres qui renvoient aux processus suivants :

1. Initialisation de la scène visuelle, c'est-à-dire constitution d'une base de données correspondant à l'état de la scène : liste des objets et de leurs caractéristiques.
2. Analyse linguistique de l'énoncé et extraction de l'expression référentielle (qui, dans un souci de clarté, est ici uniquement verbale).
3. Détermination du domaine de référence sous-spécifié correspondant à l'expression référentielle. Cette détermination fait appel au contexte applicatif : elle nécessite une mise en correspondance des composants de l'expression avec les catégories et les propriétés définies dans l'application.
4. Récupération dans l'historique du dialogue des différents domaines possibles, dans lesquels l'utilisateur peut se placer.
5. A partir des données contenues dans la base décrivant l'état de la scène, identification des domaines de référence liés à la perception visuelle.
6. Parmi les domaines possibles qui ont été identifiés, choix d'un ou de plusieurs domaines considérés comme plus pertinents que les autres.
7. Dans le cadre de ces domaines, extraction du ou des référents.
8. Mise en correspondance des référents avec les identifiants des objets de l'application pour déterminer les paramètres de la ou des fonctions applicatives à exécuter.
9. Résolution de la référence aux actions pour déterminer la ou les fonctions applicatives à exécuter.
10. Mise à jour de la scène en fonction des actions effectuées sur les référents.
11. Mise à jour (ou initialisation s'il s'agit du premier énoncé) de l'historique du dialogue.

8.2 Comment le système gère les notions étudiées

8.2.1 Conditions d'appel aux domaines de référence

Quand faut-il construire les domaines de référence ? Nous avons décrit les processus de construction de domaines et d'appel à ceux-ci pour l'appariement avec les domaines sous-spécifiés émis suite à la réception de l'énoncé oral. Nous voulons maintenant considérer les aspects algorithmiques de ces processus. Il peut en effet s'avérer très coûteux de construire tous les domaines possibles. A propos des domaines visuels, par exemple, les critères de similarité, de proximité et de continuité permettent de générer un grand nombre de domaines, surtout lorsque l'on tient compte de plusieurs niveaux de granularité pour chaque critère. Or, si ces domaines permettent d'avancer des hypothèses lors de l'interprétation d'un geste imprécis ou d'une expression faisant appel à la saillance, ils ne suffisent pas à interpréter toutes les expressions référentielles s'appuyant sur le contexte visuel. Des expressions telles que « *le triangle de gauche* » font ainsi intervenir l'abscisse (ou l'ordonnée pour « *le triangle d'en bas* »), information qui n'apparaît pas explicitement dans les domaines construits.

Le problème soulevé ici est celui de l'attribution d'un critère de différenciation : dans le domaine correspondant au contexte visuel complet (ou au sous-contexte délimité par une expression antérieure ou par tout autre indice de focalisation, comme dans l'exemple du corpus Ozkan), une partition avec « positionnement horizontal » comme critère de différenciation permet d'interpréter « *le triangle de gauche* ». La question porte sur le moment de la construction de cette partition. Faut-il envisager tous les critères de différenciation avant même la réception de l'expression référentielle, ou faut-il au contraire utiliser les composants de cette expression pour la génération du critère de différenciation dans un domaine identifié?

Un exemple de scène extrêmement simple, présenté dans la figure 8.2, montre l'explosion possible du nombre de critères de différenciation, et penche pour une génération au moment du traitement de l'expression référentielle. Au contraire, un exemple tel que « *enlève ce triangle* » associé à un geste désignant un triangle dans un groupe de deux (placés l'un au-dessus de l'autre), suivi de « *enlève aussi celui au-dessous* », penche pour la solution consistant à envisager tous les critères possibles avant l'interprétation. En effet, si un domaine comprenant les deux triangles avec le critère de différenciation « positionnement vertical » n'a pas été généré, l'interprétation de « *celui au-dessous* » pose problème : l'objet site n'existe plus, et aucun domaine disponible ne comprend de partition avec un critère lié au positionnement vertical. La seule manière d'interpréter consiste à reprendre l'état de la scène lors de l'énoncé précédent et à recommencer tout le processus. La génération *a priori* de toutes les partitions possibles semble donc utile. D'une manière générale, c'est cette génération systématique qui permet l'identification de possibles ambiguïtés. Or il nous semble important que le système soit conscient des ambiguïtés.

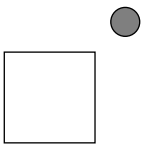
Scène visuelle	Possibilités de critères de différenciation
	Catégorie : critère utilisé pour l'expression « *le carré* » Taille : critère utilisé pour l'expression « *le grand objet* » Couleur : critère utilisé pour l'expression « *l'objet blanc* » Abscisse : critère utilisé pour l'expression « *l'objet à gauche* » Ordonnée : critère utilisé pour l'expression « *l'objet en bas* »

Figure 8.2 – *Combinatoire des critères de différenciation.*

Face à ce dilemme, deux solutions se présentent : la première consiste à construire non pas une multitude de domaines de référence, mais quelques métadomaines à partir desquels toutes les possibilités peuvent se retrouver, en cas de besoin. Ces métadomaines correspondent à des schémas généraux indiquant les procédures à appliquer pour aboutir à des domaines. Cette méthode consiste ainsi à passer de l'extension à l'intension, comme nous l'avons fait pour la traduction d'une expression référentielle en domaines sous-spécifiés. Elle présente des avantages algorithmiques : seuls quelques métadomaines sont gérés, et, parmi ceux-ci, seuls les plus pertinents (compte tenu de l'énoncé oral) sont activés. L'inconvénient majeur est que la structuration en dendrogrammes est perdue. En effet, seule une construction en extension permet de spécifier la hiérarchie entre les éléments instanciés. Pour cette raison, nous abandonnons cette solution.

Des déclencheurs pour la construction de domaines. Nous choisissons la méthode consistant à ne générer qu'un certain nombre de critères de différenciation. Un choix doit ainsi être fait entre les critères retenus et les critères écartés. Ce choix se fonde sur des indices inclus dans l'énoncé oral. Ainsi, la présence d'un pluriel, d'un adjectif numéral ou d'un démonstratif sont autant de

déclencheurs pour la construction de domaines. Dans notre exemple « *le triangle de gauche* », le défini « *le* » déclenche la construction des domaines visuels liés à une partition comprenant un élément isolé, le terme « *triangle* » déclenche la construction des seuls domaines relatifs à cette catégorie et le complément « *de gauche* » déclenche la spécification, dans chacun des domaines construits, du critère de différenciation lié au positionnement horizontal. Avec « *les triangles de gauche* », l'emploi du pluriel dénote en priorité la désignation d'un groupe perceptif fort, et déclenche ainsi de manière prioritaire la construction de tous les domaines liés à la proximité. Avec « *enlève aussi celui au-dessous* » après « *enlève ce triangle* », il est nécessaire de revenir à l'état antérieur de la scène et de refaire la construction de domaines, mais celle-ci est cette fois beaucoup plus rapide du fait du déclencheur supplémentaire « *au-dessous* ».

Cette méthode permet de réduire le nombre de domaines construits, tout en considérant qu'un simple mot comme « *les* » peut être à l'origine d'un grand nombre de possibilités. Elle consiste à généraliser le principe exploité dans le chapitre 6 pour l'appel à un ordonnancement. Elle reste néanmoins à l'état de proposition non validée dans notre travail. Il faudrait en effet tester son efficacité et vérifier que des possibilités ne sont pas rejetées trop rapidement. A terme, le but est de spécifier des combinaisons minimales de déclencheurs fonctionnant conjointement, par exemple la combinaison d'un démonstratif, d'un pluriel et d'un geste ostensif. Nous retiendrons de cette proposition que les modules spécialisés dans la construction de domaines de référence peuvent être appelés de deux façons : avec ou sans l'exploitation de déclencheurs. Cette considération sera d'importance lors de la spécification de l'architecture.

8.2.2 Conditions d'appel à la saillance

Le problème de la confrontation des saillances. Nous avons vu dans les pages précédentes que la saillance avait plusieurs rôles dans la résolution de la référence :

1. Prédire sur quel objet ou groupe d'objets l'utilisateur va se focaliser et axer son énoncé, en particulier s'il s'agit d'une première mention.
2. Interpréter une trajectoire gestuelle en étendant l'ensemble des demonstrata au groupe saillant qui les contient.
3. Interpréter une référence langagière ambiguë dans le contexte visuel : « *le N* » dans une scène affichant plusieurs N ; « *les n N* » pour trouver un groupe saillant comportant le bon nombre d'objets de la catégorie N ; « *les objets* » pour désigner le groupe le plus saillant et non la totalité des objets affichés.
4. Interpréter une référence langagière ambiguë dans le contexte de l'interaction : « *le N* » dans un contexte où un N vient d'être manipulé ou vient d'être mentionné (anaphore).

Pour l'interprétation de l'expression référentielle « *le N* », plusieurs possibilités apparaissent ainsi entre le contexte visuel, le contexte linguistique et l'historique de l'interaction. Par exemple, quelle chaise choisir pour l'interprétation de « *la chaise* » quand une chaise est visiblement saillante et une autre vient d'être manipulée? Cette situation pose le problème de la confrontation des saillances : la saillance visuelle est-elle plus importante que la saillance due à la mémoire de l'interaction? Une réponse positive entraîne l'identification de la chaise visuellement saillante. Une réponse négative entraîne l'identification de la chaise récemment manipulée. Pas de réponse entraîne une ambiguïté. Dans tous les cas, les conséquences algorithmiques sont complexes : pour répondre autrement que de manière figée et simpliste, le système doit disposer d'une échelle de comparaison des différentes saillances, pour accorder la prépondérance de la saillance visuelle sur la mémoire dans telle situation, et l'inverse dans telle autre situation. Les questions qui se

posent alors ont trait à la validité de cette échelle de comparaison, ainsi qu'à son paramétrage par la tâche applicative.

Face à ce problème, nous disposons d'un argument certes valide mais qui nous semble néanmoins trop réducteur : la primauté du visuel sur la mémoire et sur l'audition (cf. page 68 lors de notre positionnement par rapport à la psychologie). Dans l'exemple précédent, nous donnerons ainsi systématiquement l'avantage à la saillance visuelle. Nous retiendrons cependant qu'il est vraisemblable que d'autres facteurs entrent en jeu. Les expérimentations que nous spécifions dans le chapitre 10 vont dans ce sens.

Des déclencheurs pour l'appel à la saillance. Il n'est pas nécessaire de reprendre tous les composants des groupes nominaux étudiés afin de déterminer ceux qui peuvent amener à un appel à la saillance. Avec l'identification des rôles présentée ci-dessus, il apparaît que la saillance intervient presque systématiquement dans le dialogue : en première mention, lors de la production d'un geste ostensif, lors de la production d'une expression référentielle au singulier ou au pluriel, ces cas recouvrant la majorité des actions de référence présentes dans le corpus Magnét'Oz. N'oublions pas cependant que la saillance constitue un moyen de lever une ambiguïté. L'appel à la saillance intervient donc, comme notre modèle le décrit, en dernière phase, c'est-à-dire lorsque l'interprétation sans son recours aboutit à plusieurs possibilités qu'il s'agit de trier.

Nous montrons ainsi que l'appel à la saillance dépend moins des composants de l'énoncé en entrée du processus que des résultats partiels obtenus au cours de ce processus. Il nous semble important que cette caractéristique apparaisse dans l'architecture du système, plus précisément dans la spécification des protocoles d'échanges entre les modules. Nous allons ainsi montrer qu'il existe un modèle d'architecture conçu pour que le résultat d'un premier appel à un module puisse décider d'un deuxième appel à ce module. Un tel modèle nous permettra, à terme, d'implanter efficacement les appels aux domaines de référence et à la saillance.

8.3 Spécification des modules et des échanges entre les modules

8.3.1 Quelques types d'architecture

Les architectures linéaires. Le traitement séquentiel que nous avons décrit page 17 pour l'interprétation d'un énoncé oral ou gestuel se traduit par une architecture linéaire comprenant quatre modules : un module chargé de l'analyse lexicale, un pour l'analyse syntaxique, un troisième pour l'analyse sémantique et le dernier pour la pragmatique. Comme dans toute architecture de système de dialogue, la modularité est surtout une nécessité pratique. Chaque module est expert dans son domaine et consomme des informations que les autres modules sont incapables d'interpréter et qui sont considérées comme indépendantes des autres informations transitant dans l'architecture. Les traitements et leur programmation sont ainsi clairement délimités.

Dans l'architecture linéaire précédente, le module sémantique reçoit des informations du module syntaxique, informations qu'il a produites suite à la réception d'informations du module lexical. Un module n'intervient donc que lorsque le module inférieur a terminé son traitement. Cette approche est dite dirigée par les données (ou *data-driven* ou encore *bottom-up*). Si elle s'avère relativement simple et par conséquent sans difficultés liées à sa structure lors d'une implantation, sa plausibilité cognitive est en revanche mauvaise. Nous ne nous arrêtons pas à un problème lexical (dû par exemple à un bruit dans l'énoncé perçu) avant d'aborder l'analyse syntaxique. Il semble que nous tenons compte du contexte dès le début du traitement, c'est-à-dire que les quatre processus ont lieu en même temps, parallèlement.

C'est donc plutôt une architecture dirigée par les concepts (ou *concept-driven* ou encore

knowledge-driven) qui semble plus plausible. Celle-ci consiste en processus parallèles qui communiquent les uns avec les autres, chacun orientant les autres, éventuellement par l'intermédiaire d'un noyau ou processeur central. C'est à ce type d'architecture que nous allons nous intéresser.

Les architectures dédiées à la multimodalité. Pour modéliser des modules qui fonctionnent en parallèle, on est amené à inclure les protocoles de communication dans la définition même d'un module. On arrive ainsi aux architectures multi-agents : un agent regroupe un ensemble de fonctionnalités, des connaissances sur lui-même et sur les autres agents, et surtout une capacité d'agir de manière autonome (requérir des informations en entrée et faire connaître les informations produites en sortie). Les architectures regroupant plusieurs agents qui communiquent les uns avec les autres sont particulièrement fréquentes dans les systèmes multimodaux. Le nombre et l'indépendance des processus liés au traitement de la parole et du geste amènent en effet directement à la définition d'agents. Bellik (1995) présente ainsi quelques modèles caractérisés par une séparation entre le noyau fonctionnel *a priori* indépendant des modalités d'interaction, et les interfaces liées à une instanciation dans une modalité particulière. Le modèle MVC (Modèle, Vue, Contrôleur) se caractérise par exemple par des agents comportant trois facettes : le « modèle » qui définit les fonctionnalités du module, la « vue » qui correspond à la perception qu'a l'utilisateur du modèle, et le « contrôleur » qui traite les entrées de l'utilisateur. D'autres modèles éclatent les modules liés aux interfaces. Le modèle ARCH, par exemple, distingue en deux modules les abstractions des concepts d'interactions et les instanciations de ces abstractions (Bellik 1995).

Dans l'ensemble, tout ce qui concerne l'interprétation est placé dans un noyau fonctionnel, souvent appelé « contrôleur de dialogue », dont les fonctionnalités restent imprécises, le reste de l'architecture découlant de la nature des modalités prises en compte en entrée et en sortie. Cela s'avère insuffisant pour notre approche. Deux grands principes de communication entre les agents viennent compléter ces considérations : le *tableau noir* et le *carnet d'esquisses*.

Le tableau noir. Dans le modèle du tableau noir, proposé dans le système Hearsay II décrit par exemple dans (Pierrel 1987), les modules s'ignorent mutuellement et ne communiquent que par l'intermédiaire d'une base de donnée externe appelée tableau noir (ou *blackboard*). La caractéristique principale est que les modules peuvent accéder aux données de manière opportuniste. Ces données peuvent être des arbres syntaxiques, des étiquetages sémantiques ou encore des résultats de l'analyse pragmatique. Le dynamisme dans l'accessibilité aux connaissances, ainsi que le parallélisme de l'analyse qui en découle, rendent ce modèle plausible d'un point de vue cognitif.

Néanmoins, Sabah (2000) note deux inconvénients majeurs. Le premier est d'ordre cognitif et concerne le manque de rétroactions : une information écrite sur le tableau noir n'est pas remise en cause à un niveau supérieur dans le traitement. Les seules rétroactions concernent ainsi le choix parmi plusieurs possibilités déjà fixées. Autrement dit, le comportement d'un module n'est pas modifié par un module supérieur. Le second est d'ordre informatique et concerne la nécessité d'un contrôle explicite lourd à réaliser. Sabah illustre de ce point de vue les problèmes dans la réalisation du système Caramel I. Il introduit alors un modèle (le carnet d'esquisses) favorisant les rétroactions et pour lequel le contrôle est implicite.

Le carnet d'esquisses. Dans ce modèle, chaque module est doté d'une capacité à donner l'indice de confiance qu'il a envers sa propre production. Par exemple, le module chargé de l'analyse syntaxique à partir d'une suite de mots est capable d'évaluer la pertinence de l'arbre produit. Si cette pertinence s'avère faible, il peut alors la communiquer au module chargé de l'analyse lexicale, dans le but de lui faire refaire son travail dans une autre direction. Lorsqu'il

reçoit la nouvelle suite de mots et procède à une nouvelle analyse syntaxique, il peut comparer la pertinence de cette nouvelle analyse à celle de la précédente. S'il constate une augmentation, il peut continuer à demander que l'analyse lexicale soit refaite dans la même direction. S'il constate une diminution, il la fait refaire dans une autre direction. Au moyen de ces boucles de rétroaction, le système atteint un état d'équilibre caractérisé par des pertinences maximales. L'architecture du carnet d'esquisses se caractérise également par une mémoire des données qui conserve les informations transitant entre les modules. Dans cette mémoire, les données qui reçoivent les meilleures rétroactions sont étiquetées comme telles. En revanche, les autres données voient leur probabilité diminuer et finissent par être détruites, du moins dès que la place manque. Certaines hypothèses émises par exemple par l'analyse syntaxique sont ainsi rejetées.

Les arguments pour une plausibilité cognitive, outre l'existence de boucles de rétroaction des niveaux supérieurs vers les niveaux inférieurs, sont la modélisation de l'interprétation comme la fusion et l'interaction entre des résultats trouvés parallèlement, l'assimilation du sentiment de compréhension à la stabilité du système, et la modélisation de la mémoire de travail avec ses mécanismes d'oubli avec la mémoire des données et son mécanisme de destruction des données improbables (Sabah 2000).

8.3.2 Une architecture pour le modèle des domaines de référence

La pertinence comme capacité de métacognition. Même si nous lui trouvons quelques inconvénients, nous choisissons le modèle du carnet d'esquisse comme architecture privilégiée pour le modèle des domaines de référence. En effet, la possibilité d'appels simultanés aux modules visuel et linguistique, ainsi que les possibilités de rétroaction semblent convenir à notre approche où des hypothèses sont souvent remises en question. De plus, nous donnons une importance toute particulière à la capacité de chaque module à porter un jugement sur ce qu'il produit. Cette capacité se rapproche de la notion de métacognition (connaissance que l'on a de son propre fonctionnement cognitif) issue de la psychologie cognitive. Si Sabah évoque à son propos la notion de pertinence (ou parfois d'entropie) d'une façon un peu ambiguë, nous avons ici l'avantage de disposer d'un critère de pertinence conséquent.

Les notions d'effets contextuels et d'effort de traitement semblent en effet pouvoir s'adapter au comportement d'un module, à travers les informations traitées et produites par ce module. Les effets contextuels produits par le résultat d'un module s'évaluent à l'aide des réponses des modules traitant ce résultat. Si nous considérons par exemple le module chargé de l'analyse syntaxique et que nous voulons déterminer les effets contextuels de l'arbre syntaxique émis par ce module, il nous faut attendre le jugement émis par le module sémantique : si ce module n'arrive pas à produire une bonne analyse à partir de l'arbre syntaxique, les effets contextuels de celui-ci seront mauvais. Nous déterminons ainsi les effets contextuels du module syntaxique. Quant à l'effort de traitement, il suffit de quantifier la complexité des opérations mises en jeu lors du traitement de l'information prise en entrée, c'est-à-dire le résultat de l'analyse lexicale, pour produire l'arbre syntaxique considéré. Le rapport effets sur effort caractérise alors la pertinence du travail effectué par le module syntaxique et nous semble pouvoir être assimilé au degré de confiance qu'il porte sur ce travail. Nous exploitons ainsi l'analyse que nous avons faite dans le chapitre 7 de la pertinence. Bien que la proposition avancée n'aie pas encore été implantée dans un système, elle nous semble en adéquation parfaite avec le modèle des domaines de référence, particulièrement avec le principe consistant à utiliser un critère de pertinence dans l'exploitation des données gérées par le système. Nous n'irons pas plus loin dans les aspects fonctionnels de l'architecture, pour nous concentrer sur la spécification de ses aspects structurels.

Spécification des modules et des échanges entre les modules. Nous gardons l'idée d'un noyau central qui regroupe les algorithmes proposés pour l'interprétation de la référence aux objets. D'une manière générale, ce module reçoit les analyses partielles des composants de l'expression référentielle multimodale et du contexte dans lequel cette expression a été produite par l'utilisateur, pour émettre une ou plusieurs hypothèses de domaines de référence et de référents. Comme indiqué sur le schéma récapitulatif de la figure 8.3, nous appelons ce module « interpréteur d'expressions référentielles ». Plus précisément, il reçoit du module « langage » un domaine sous-spécifié et du module « perception visuelle » une interprétation de la trajectoire gestuelle en contexte visuel. Il émet alors une requête en parallèle aux trois modules « langage », « perception visuelle » et « tâche », éventuellement à travers les historiques de chacun. Il reçoit en retour trois listes ordonnées de domaines, certains d'entre eux ayant éventuellement un référent focalisé. Il confronte ces résultats et les évalue en s'aidant des informations que lui fournissent les modules « modèle de l'utilisateur » et « modèle de pertinence ». Il retire de cette évaluation le résultat le plus probable, qui se traduit par un ensemble de référents et un ou plusieurs domaines de référence envisageables.

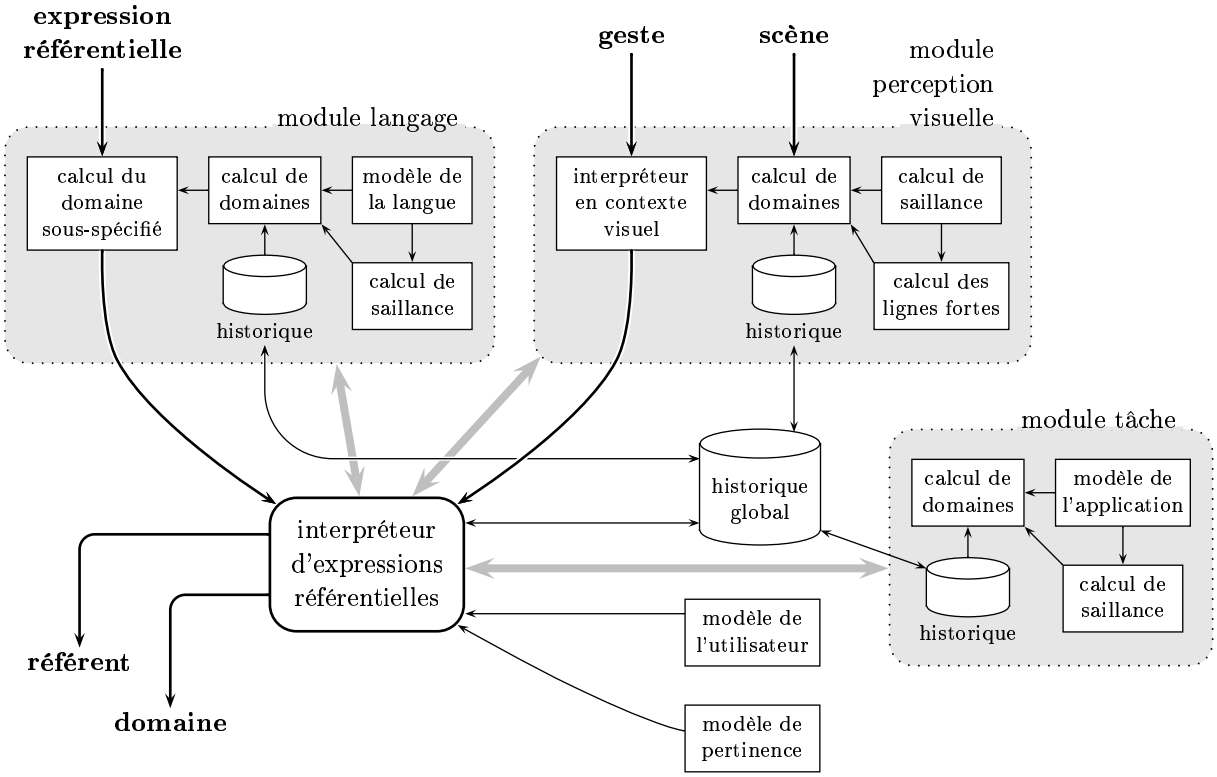


Figure 8.3 – *Architecture logicielle pour la compréhension des références.*

Le schéma de la figure 8.3 met l'accent sur la description des fonctionnalités de chacun des modules « langage », « perception visuelle » et « tâche ». Deux fonctionnalités que l'on retrouve systématiquement sont ainsi le calcul de la saillance et la construction de domaines de référence. Selon les besoins de l'interpréteur central, certaines de ces fonctionnalités sont activées ou désactivées. Ceci permet plusieurs types de requêtes (par exemple : « avec calcul de la saillance », « sans exploitation des déclencheurs »), cette pluralité influant sur les comportements des modules et donc sur la façon d'atteindre la stabilité du système.

RÉCAPITULATIF
Nous décrivons dans ce chapitre les composants d'un système de dialogue qui conduisent à la spécification de modules. Nous mettons particulièrement l'accent sur les données conservées dans des historiques : en plus de l'historique du dialogue tel qu'il apparaît dans la majorité des systèmes existants, nous introduisons l'historique visuel, l'historique de tâche et l'historique global dont le but ultime est de modéliser l'attention de l'utilisateur. Nous proposons pour cela la gestion de scores attentionnels. Nous faisons le point sur les conditions d'appel aux notions étudiées dans les chapitres précédents. Nous précisons ainsi des règles supplémentaires par rapport à celles données dans le chapitre 7, ces règles étant liées non plus à une certaine plausibilité cognitive mais à des préoccupations algorithmiques. Enfin, nous regroupons les aspects architecturaux dans un schéma illustrant les principales caractéristiques d'un système de dialogue fondé sur les domaines de référence. Nous nous appuyons en cela sur le modèle du carnet d'esquisses de Sabah.

Chapitre 9

L'adaptation à une application

Comment les notions composant le modèle des domaines de référence sont-elles affectées par l'adaptation de celui-ci à un type d'interaction, c'est-à-dire à l'utilisation de dispositifs d'acquisition ou de visualisation différents de ceux pour lesquels il a été conçu ? Comment le processus de construction de domaines de référence et le modèle de la saillance sont-ils complétés lors de l'instanciation du modèle à une tâche applicative particulière ? Peut-on adapter le modèle dans sa globalité à un formalisme computationnel existant ?

Nous avons recueilli et analysé des phénomènes de référence essentiellement pour une interaction fondée sur des scènes visuelles en 2D et sur le microphone et l'écran tactile comme dispositifs d'acquisition. Nous montrons ici (§ 9.1) que notre modèle reste compatible avec la visualisation en 3D, le geste en 3D et le geste haptique, c'est-à-dire que son adaptation à de tels types d'interaction s'effectue à moindre coût. Nous nous sommes limité à des tâches applicatives mettant en jeu des objets simples tels que des formes géométriques ou des meubles. Sans prendre en compte de tâches spécifiques supplémentaires mais en nous interrogeant sur les mécanismes dirigés par la tâche en général, nous montrons ici (§ 9.2) comment se traduisent les contraintes du contexte de tâche. Nous nous sommes focalisés sur les aspects théoriques des domaines de référence. Nous tentons ici (§ 9.3) de proposer des pistes pour leur adéquation à des formalismes classiquement utilisés dans le domaine du traitement automatique du langage naturel.

9.1 L'adaptation du modèle à un type d'interaction

9.1.1 Adaptation à la nature du support visuel

Paramètres liés au support visuel. L'adaptation de notre modèle aux particularités des scènes visuelles d'une application consiste d'une part en l'ajustement de la hiérarchie des critères de saillance visuelle, d'autre part en l'ajustement des correspondances des critères de la Gestalt pour la lecture dans les dendrogrammes. L'ajustement de la saillance se fait en déterminant la liste des propriétés ainsi que leurs coefficients pour le calcul des scores numériques présentés dans le chapitre 5. Pour la détermination de ces coefficients comme pour l'ajustement des poids relatifs des critères de groupage, plusieurs solutions sont possibles. Une première méthode consiste à déduire des coefficients théoriques à partir des particularités visuelles. Ainsi, lorsqu'une propriété comme

la couleur fait partie des buts communicatifs (pour une application de jeu vidéo, par exemple), il est clair que le coefficient relatif à la couleur doit être favorisé. Une deuxième méthode, plus rigoureuse mais plus longue, consiste à considérer un ensemble de scènes représentatives et à déterminer manuellement (ou suite à des expérimentations) les coefficients requis. Les moyennes obtenues pourront alors être utilisées comme valeurs propres à l'application. Si les écarts sont trop importants, implanter une bibliothèque et scènes et un algorithme de reconnaissance peut s'avérer utile.

Un exemple d'adaptation : la saillance dans le projet COVEN. Dans le cadre de ce projet (*COllaborative Virtual ENvironments*) et pour le scénario consistant en l'aménagement de bureaux, il s'agit de déterminer, dans la liste des caractéristiques visuelles concourant à la saillance, celles que l'on peut raisonnablement retenir et exploiter lors du calcul de la référence, c'est-à-dire lors de la détermination d'une hypothèse privilégiée parmi les hypothèses de référents. En suivant la première méthode évoquée ci-dessus, nous reprenons les caractéristiques de la classification du chapitre 5.

La saillance par une mise en évidence explicite de l'objet est la première à être traitée car elle est à la base de l'interaction : elle intervient quand l'application veut mettre en avant un objet par rapport aux autres. Dans COVEN, c'est le cas lorsque l'utilisateur a effectué une manipulation directe sans production d'énoncé oral : le résultat de sa manipulation est mis en évidence par un rendu visuel particulier, par exemple l'affichage en traits fins de la boîte englobante de l'objet manipulé. Cet affichage montre que la suite de l'interaction (par exemple une reprise pronominale) s'appliquera sur cet objet. Il n'y a dans ce cas aucune ambiguïté, et aucun traitement basé sur un choix entre plusieurs objets n'est donc à envisager.

La saillance par la catégorie n'intervient pas non plus dans la résolution d'une ambiguïté. Nous avons vu en effet qu'une ambiguïté sur la catégorie n'apparaissait que si l'utilisateur emploie un substantif ambigu tel que « *objet* ». Or nous avons montré que la mention de la catégorie n'était pas plus coûteux. Des tests préliminaires montrent en effet que l'utilisateur ne désigne quasiment jamais les meubles de la scène autrement que par leur catégorie.

La saillance par les caractéristiques physiques soulève la même question : dans un environnement composé de plusieurs chaises rouges et d'une chaise bleue, le contraste des couleurs est tellement évident que l'utilisateur employera sans doute l'expression « *la chaise bleue* » (plutôt que « *la chaise* » avec comme sous-entendu « *la chaise saillante* ») pour désigner celle-ci, même si cette expression est quelque peu coûteuse. De même que la nature des meubles, les choix de couleurs sont importants dans cette tâche d'aménagement d'un intérieur. Les propriétés physiques des objets sont par conséquent explicites dans les énoncés. Nous noterons toutefois qu'il arrive qu'une information à propos d'une caractéristique physique soit sous-entendue par l'action : dans le même environnement que précédemment, l'énoncé « *peins la chaise en rouge* » utilise l'expression référentielle minimale et ambiguë « *la chaise* ». En revanche, le fait de vouloir la peindre en bleu sous-entend qu'elle n'est pas bleue. Ce présupposé permet de résoudre l'ambiguïté sans faire appel à la saillance. L'appel à la saillance par les caractéristiques physiques ne semble donc pas intervenir outre mesure.

La saillance par la localisation dans la scène (ou saillance spatiale) semble intervenir dans la résolution d'une ambiguïté : dans un environnement contenant plusieurs chaises dont une seule est mise en valeur par sa proximité, il semble probable que l'utilisateur ayant une chaise juste devant les yeux ne verra pratiquement que cette chaise et qu'il utilisera l'expression référentielle « *la chaise* » pour la désigner. Il est clair en tout cas qu'il évitera des expressions lourdes telles que « *la chaise la plus proche* » ou « *la chaise juste devant mes yeux* ». Dans cet exemple comme

dans celui d'une chaise isolée par rapport à un groupe de chaises, l'appel à la saillance comme critère de choix d'un référent semble nécessaire, d'autant plus que l'agencement spatial est la principale caractéristique de la tâche. Cet appel fait intervenir notre structuration de l'espace visuel en groupes perceptifs, ainsi que la procédure de lecture dans les dendrogrammes obtenus.

La saillance par l'incongruité ou l'aspect énigmatique consiste à repérer l'objet qui est en situation inhabituelle, par exemple une chaise renversée. Il est nécessaire pour cela de disposer pour chaque caractéristique de chaque type d'objet d'une valeur par défaut correspondant à l'objet dans son état normal. Dans l'exemple de la chaise renversée à même le sol, une base de données doit ainsi contenir le fait que l'état normal d'une chaise est, par exemple, soit les quatre pieds au sol, soit deux pieds au sol et le dossier en appui sur une table, soit posée les pieds en l'air sur une table. Une chaise dans une autre configuration est alors étiquetée comme saillante par incongruïté. La question qui se pose est relative à la possibilité d'une telle situation. En effet, lorsqu'un objet est introduit dans la scène par le système, il est placé dans une configuration standard. Dans COVEN, un objet ne peut donc être mis en situation inhabituelle que par l'utilisateur. Or ceci s'avère contraire au but de tâche, c'est-à-dire l'aménagement d'un intérieur en suivant les lois de gravitation et d'utilisation des différents meubles. Nous en déduisons que l'incongruïté, si elle apparaît, relève de cas exceptionnels, et qu'aucun appel à cette saillance ne semble donc nécessaire.

La saillance par les fonctionnalités doit être prise en compte : dans l'exemple de l'ordinateur allumé parmi plusieurs ordinateurs éteints, il est probable que l'expression « *l'ordinateur* », même en dehors de toute considération liée au prédicat (du moins si celui-ci n'est ni « *éteins* » ni « *allume* »), référera à celui qui est allumé.

La saillance par la dynamique de l'objet consiste à étiqueter comme saillant un objet en mouvement dans un environnement où tous les autres objets sont statiques. Il semble en effet qu'un tel objet retient probablement toute l'attention de l'utilisateur, et que sa désignation par une expression minimale soit suffisante. Dans le cadre de COVEN, tous les objets sont inanimés et immobiles, donc le problème ne se pose pas et l'appel à cette saillance s'avère inutile.

La saillance indirecte, mettant en relation un objet saillant et un objet sur lequel se répercute cette saillance, est plus difficile à repérer. A notre avis, cette saillance ne doit pas être l'objet de traitements car l'utilisateur exprimera probablement le lien entre les deux objets : il employera par exemple une relation fonctionnelle (« *la chaise du bureau* ») ou une locution propositionnelle (« *la chaise à côté du bureau* »), rendant ainsi explicite l'information de transfert. Dans COVEN, nous ne tiendrons donc pas compte de ce dernier type de saillance.

Nous déduisons de tout cela qu'il semble intéressant d'exploiter surtout les saillances fonctionnelle, spatiale, dynamique et incongrue. Dans le cadre de COVEN, nous avons montré que les saillances dynamique et incongrue n'intervenaient pas. En considérant que la fonctionnalité peut être traitée à l'aide d'un test spécifique intervenant de manière prioritaire dans le processus de résolution de la référence, nous mettons l'accent sur la seule saillance spatiale. Nous utilisons pour cela notre modélisation de la focalisation spatiale présentée dans la section 7.2.1 pour le geste ostensif et applicable comme nous l'avons montré à la saillance visuelle en général.

9.1.2 Adaptation à la 3D : le projet COVEN

Quelques conséquences de la 3D. Une autre particularité visuelle du projet COVEN est la visualisation en 3D. Que ce soit à l'aide d'un casque de réalité virtuelle ou d'un écran standard, l'utilisateur perçoit la scène visuelle avec une profondeur. Plus que cela, les manipulations directes et les gestes ostensifs sont réalisés en 3D. Que ce soit à l'aide d'un gant de désignation ou avec

la souris standard, les pointages se font en 3D. Pour traduire visuellement cette particularité, il a été décidé de faire apparaître par un trait la direction d'un pointage.

Avant de détailler cette méthode et les algorithmes qu'elle met en jeu, nous noterons que le geste en 3D est imprécis. Que ce soit dans la communication de face à face ou en dialogue homme-machine, cette imprécision joue à deux niveaux : du point de vue du locuteur et du point de vue de l'interlocuteur. En effet, le locuteur vise un certain point qui peut ne pas être le plus pertinent, ce qui induit une imprécision proportionnelle à la distance entre lui et l'objet. Il calcule pour cela la direction à partir de son œil mais fait le geste à partir de l'épaule, voire plus bas, d'où l'apparition d'un décalage. Du point de vue de l'interlocuteur, nous retrouvons l'imprécision due à la distance entre le locuteur et l'objet, ainsi que l'erreur sur le point de départ supposé du geste (l'œil et non l'épaule du locuteur). S'ajoute le fait que l'interlocuteur aurait peut-être visé un autre point que celui choisi par le locuteur, fait lié à la variabilité inter-locuteur qui augmente d'autant l'imprécision.

Le geste ostensif en 3D. A partir de (Landragin 1998), nous présentons maintenant le travail d'implantation informatique que nous avons fait dans le cadre du projet COVEN pour l'intégration du geste de désignation dans un environnement en 3D. Le retour visuel choisi pour une désignation gestuelle est l'affichage d'une droite matérialisant la direction du pointage. Cette droite part d'un point situé au niveau de l'épaule[1] de l'avatar représentant l'utilisateur, et s'arrête au premier objet rencontré dans son parcours (d'un point de vue algorithmique, nous nous rabattons sur la méthode classique du lancé de rayon). Il ne s'agit pas forcément de l'objet intentionnellement désigné, mais par exemple du sol ou d'un mur. C'est le cas si l'objet est une chaise vue de profil, c'est-à-dire ne présentant que peu de surface visible à l'utilisateur, et si la désignation a été effectuée par exemple entre les pieds de la chaise.

Afin de retrouver l'objet désiré, un moyen consiste à créer une zone de désignation s'étendant autour du rayon tracé. Comme on ne sait pas à quelle distance de l'avatar se trouve le demonstratum, cette zone doit commencer au niveau même de l'avatar et s'étendre au-delà du point d'arrêt du rayon. En outre, plus l'objet se trouve éloigné de l'avatar et plus la désignation peut être imprécise. C'est pourquoi notre choix de forme de zone consiste en un cône ayant l'épaule du participant comme sommet et la direction de désignation comme axe. L'angle au sommet du cône n'est pas déterminé dynamiquement, mais fixé de manière empirique.

L'implantation informatique dans le cadre du projet COVEN. L'implantation de la construction de la zone lorsqu'un geste de désignation est effectué ne pose pas de problème particulier. Il faut quand même tenir compte d'un certain nombre de situations. Si par exemple le pointage aboutit sur le sol, sur un mur ou sur le plafond, c'est sans doute qu'un objet a été raté. Il faut alors faire attention, en particulier si l'élément sur lequel le lancé de rayon s'est arrêté est proche de l'avatar, de tenir compte de l'imprécision supplémentaire. Une augmentation de l'angle au sommet du cône apparaît alors comme une solution simple.

La recherche des objets dans la zone construite est plus complexe à implanter. Nous choisissons de considérer la sphère englobante de chaque objet présent dans l'environnement, et de tester l'inclusion ou l'intersection de cette sphère dans le cône. Si ce test est positif, l'objet est retenu. Reste alors à déterminer les demonstrata dans l'ensemble des objets retenus. Une première

1. En plus de la plausibilité physiologique de ce point de départ, une raison pratique vient justifier ce choix : l'épaule de l'avatar correspond en vue subjective à un point situé en bas de l'image, alors que l'œil correspond au centre de l'image. Or une droite partant de ce point s'avère illisible, surtout si l'objet sur lequel s'arrête le pointage est également au centre de l'image.

approche consiste à considérer les trois cas suivants :

1. Désignation d'un seul objet, c'est-à-dire geste ostensif associé à un groupe nominal démonstratif (éventuellement défini) au singulier : parmi les objets retenus, on choisit celui qui se trouve le plus près de l'axe du cône. Si ce dernier ne contient aucun objet, une reconstruction élargie de la zone ou une stratégie de réponse doit être envisagée.
2. Désignation de plusieurs objets en nombre indéterminé, c'est-à-dire geste associé à un groupe nominal au pluriel : on conserve tous les objets retenus dans le cône. Si celui-ci ne contient aucun objet ou n'en contient qu'un seul, une reconstruction élargie de la zone ou une stratégie de réponse doit être envisagée.
3. Désignation de plusieurs objets en nombre déterminé, c'est-à-dire geste associé à un groupe nominal au pluriel contenant un adjectif numéral (n) : parmi les objets retenus, on choisit les n plus proches de l'axe du cône. De même que précédemment, si la zone ne contient pas suffisamment d'objets, on la reconstruit en l'élargissant ou on envisage une stratégie de réponse fondée sur l'incapacité de l'algorithme à identifier les demonstrata.

Cette approche manque de souplesse vis-à-vis des groupes perceptifs : le geste peut ne désigner qu'une chaise, la référence s'appliquant au groupe de trois chaises dont elle fait partie. Face à cette situation, il arrive que l'algorithme retiennent deux des chaises du groupe (les deux plus proches de l'axe) et une troisième chaise placée elle aussi dans la trajectoire de désignation mais pas sur le même plan que le groupe visé. Modifier l'angle du cône ne change rien au problème et il est alors nécessaire de reconsidérer la méthode de recherche des objets. Une prise en compte des groupements perceptifs et de la distance entre ces groupes et l'avatar constitue alors une deuxième approche. Il s'agit de structurer en groupes la scène en 3D, en ne tenant compte que du critère de proximité. Dans les partitions obtenues, on recherche, en commençant par le sommet du dendrogramme, les groupes contenant le bon nombre d'objets (trois dans notre exemple). Parmi les groupes obtenus, on retient ceux pour lesquels le maximum d'éléments se trouvent dans le cône de désignation. S'il reste encore plusieurs groupes candidats, on retient celui qui contient l'objet le plus proche de l'axe du cône. Compte tenu de la prise en compte des dendrogrammes dont la construction est présentée dans cette thèse, cette deuxième approche n'a pas été implantée dans COVEN. Elle illustre cependant les apports mutuels entre l'implantation réalisée dans le cadre de ce projet et le travail théorique réalisé ensuite.

9.1.3 Adaptation à la modalité haptique : le projet MIAMM

Geste haptique, geste ostensif et geste de manipulation directe. Le geste haptique est le geste effectué à l'aide d'un dispositif générant un retour de force et permettant ainsi une perception tactile de l'environnement. Avec une modalité haptique, geste de désignation et geste de manipulation directe sont fusionnés. Lorsque l'on appuie sur un objet, on effectue deux opérations simultanées, la désignation de l'objet et sa manipulation, une déformation par exemple. Sont également fusionnées les trois fonctions du geste identifiées par Cadoz (1994) en communication homme-machine (cf. § 1.1.1 page 13). En effet, la fonction épistémique caractérise la prise de connaissance de l'environnement, donc ici la perception tactile que l'on peut avoir par exemple de la texture de l'objet. La fonction ergotique caractérise l'action sur l'environnement et apparaît lors de la déformation de l'objet. Quant à la fonction sémiotique, elle correspond à l'intention de désignation.

Ce qui nous intéresse ici, ce n'est pas de définir et de valider un nouveau mode d'interaction spontané à base d'un dispositif haptique, mais de montrer comment notre modèle à base de domaines de référence s'adapte aux caractéristiques d'un tel dispositif. Il s'agit donc de déterminer

si la prise en compte des trois fonctions du geste, et non plus celle de la seule fonction sémiotique, remet en cause certains composants de notre modèle. Nous avançons l'idée que l'intégration du geste haptique ne change rien ni au principe des domaines de référence, ni au processus d'unification d'un domaine sous-spécifié avec des domaines construits à l'aide des différentes sources contextuelles, ni même à l'architecture qui en découle. Nous montrons au contraire que la modalité haptique confirme les places relatives du geste et du module lié à la perception. Cette idée a été présentée dans (Landragin *et al.* 2002a), dans le cadre du projet MIAMM (*Multidimensional Information Access using Multiple Modalities*) et du dispositif PHANToM.

Identification de la fonction de geste. Compte tenu de la possibilité des trois fonctions, le nouveau problème qui se pose est l'identification de la ou des fonctions dans lesquelles l'utilisateur s'est placé. Le geste effectué, et en particulier la force avec laquelle il a été produit, apporte un indice. Il semble en effet que seule la fonction ergotique peut être liée à une force importante. L'expression référentielle, que ce soit « *cet objet* » ou « *cette texture* », n'apporte quasiment aucun indice. En effet, il peut s'agit dans les deux cas d'une intention sémiotique, ergotique (« *écrase cet objet* », « *lisse cette texture* »), ou épistémique (« *cet objet semble fragile* », « *quelle est cette texture ?* »). Les exemples cités mettent l'accent sur l'indice le plus important : le prédicat. En effet, une intention ergotique semble plutôt correspondre à un prédicat utilisé à l'impératif et dénotant une action, et une intention épistémique à une question ou à une assertion sans autre but communicatif. Les autres composants de l'énoncé oral peuvent également apporter des indices. Dans « *déplace cette résistance comme ça* » associé à une pression, « *comme ça* » semble très lié à la façon de procéder à une action ergotique. Dans « *la texture de cet objet* », l'expression référentielle « *cet objet* » s'avère moins utile que « *la texture de* » qui montre l'intention d'extraire quelque chose de l'environnement, et qui dénote donc une intention épistémique. Sans spécifier pour l'instant d'algorithme d'identification, ces éléments nous permettent de supposer qu'un tel algorithme est réalisable et que les cas d'ambiguïté entre deux ou trois fonctions seront marginaux.

Adaptation des mécanismes de résolution de la référence. Si les mécanismes classiques tels que les filtrages réalisés sur les composants de l'énoncé oral s'appliquent également à un énoncé incluant un geste haptique, il n'en est peut-être pas de même des filtrages réalisés sur le contexte perceptif. En effet, les phénomènes de focalisation à un sous-espace perceptif sont sans doute remis en cause par le changement de nature de la perception. Celle-ci n'est en effet plus seulement visuelle mais inclut aussi la perception tactile. Nous étudions ici les conséquences de ce changement.

Du point de vue des phénomènes en entrée, nous aurons des domaines de référence perceptifs plus nombreux, ces domaines se distinguant en deux types *a priori* incompatibles : les domaines visuels et les domaines tactiles. Ces derniers sont spécifiques, dans la mesure où la perception visuelle fonctionne par connaissance immédiate de la répartition spatiale de plusieurs objets, alors que la perception tactile fonctionne par connaissance selon un parcours. Les objets ne sont pas forcément perçus un par un : plusieurs objets peuvent être perçus simultanément avec le dispositif haptique, par exemple des objets agglutinés ou emboîtés (l'utilisateur touche l'ensemble et se réfère à une partie) ou encore une classe complète d'objets. A moins que la tâche applicative les encourage, ces cas sont sans doute peu fréquents.

C'est donc surtout au niveau de l'historique de l'interaction et de l'ordonnancement des éléments dans un domaine tactile que des changements ont lieu. Nous pouvons ainsi considérer un historique épistémique correspondant au stockage des actions successives de prise de connaissance de l'environnement. Nous devons aussi considérer le nouveau critère d'ordonnancement qu'est ce

critère épistémique : au fur et à mesure que les actions épistémiques se succèdent, un domaine tactile se construit avec pour seule partition celle caractérisée par ce critère.

Du point de vue de l'interprétation, les deux sources de construction de domaines perceptifs émettront leurs hypothèses en parallèle, dans un même dendrogramme. Le système rejettera des branches de ce dendrogramme en fonction de l'expression référentielle. Ainsi, une expression telle que « *les objets de cette consistance* », « *les objets ayant cette texture* », ou « *les objets qui vibrent* » fait appel à une caractéristique non visible. Cette caractéristique est forcément liée à un domaine tactile donc seules seront considérées les branches du dendrogramme correspondant à des domaines tactiles. Ces domaines sont également utiles lors de l'interprétation d'expressions faisant appel à un parcours effectué. Ainsi, « *la chaise qui était collée au sol* » ou « *les objets qui vibrent* » s'interprètent dans l'ensemble des objets parcourus, c'est-à-dire dans le domaine tactile activé. La modalité haptique apporte de même de nouvelles possibilités d'interprétation des références mentionnelles telles que « *le premier* » et « *le dernier* ».

Nous avons vu que l'interprétation de la référence aux objets pouvait faire intervenir la notion de saillance. La saillance haptique existe-t-elle ? Non seulement elle existe, mais elle s'avère importante : lors du parcours de plusieurs objets, un objet en particulier peut avoir été manipulé avec plus de force ou pendant plus de temps que les autres. Cette interaction particulière l'a rendu saillant et permet sa reprise ultérieure par une expression du type « *le N* », semblable à celle que nous avons analysée lors du recours à la saillance visuelle.

En conclusion, il apparaît que les mécanismes et les notions du modèle des domaines de référence s'adaptent à l'intégration d'une composante haptique. La principale modification est la transformation du module relatif à la perception visuelle, en module regroupant perception visuelle et perception tactile. Par rapport au schéma de la figure 8.3 page 168, les seuls changements sont ainsi la dénomination du module et la distinction de sous-modules pour le calcul de la saillance visuelle et le calcul de la saillance haptique. Tous les liens entre ce module et le reste de l'architecture sont inchangés. En particulier, le geste arrive toujours en entrée de ce seul module. L'architecture du projet MIAMM suit ce principe, avec un module autonome nommé VisHaptic (cf. `http://www.loria.fr/projets/MIAMM/`).

9.2 L'adaptation du modèle à un type de tâche

Parmi les exemples que nous avons étudiés, se trouvent des situations liées à une tâche d'aménagement d'un intérieur (COVEN) et des situations impliquant le rangement de formes géométriques dans des boîtes appropriées (Magnét'Oz). Nous avons montré que certaines particularités de ces tâches applicatives avaient des répercussions sur la caractérisation de la saillance visuelle ou des ordonnancements. Dans Magnét'Oz, la tâche est tellement contrainte qu'il en devient parfois possible de prédire sur quels objets l'utilisateur va porter son attention. Ainsi, le calcul de la saillance visuelle permet presque d'identifier le ou les objets qui vont être traités en première mention. De plus, la tâche incite fortement à produire des commandes du type « *mets cet objet dans cette boîte* », incluant un geste ostensif pour l'objet et un second pour la destination. Compte tenu de l'obligation de ranger tous les objets ayant la même forme dans une même boîte, la deuxième référence est totalement prévisible et s'avère par conséquent quasiment inutile du point de vue de l'interprétation. Son principal intérêt pour le système réside ainsi dans l'entrainement du modèle de l'utilisateur. Le but de cette section est de préciser les composantes principales du modèle de tâche et de montrer comment la considération d'une application particulière influe sur ces composantes. Nous prendrons pour cela l'exemple du corpus Ozkan que nous avons présenté dans la figure 7.3 page 147.

9.2.1 Nature du modèle de tâche

Importance de la tâche dans la résolution de la référence. A la suite de (Wright 1990), nous considérons que la tâche fournit des contraintes non-linguistiques qui, comme la saillance, facilitent la compréhension de « *le N* » dans un environnement comprenant plusieurs N : « *le N* » peut en effet se comprendre comme « *le N suivant dans la succession des tâches* ». Encore faut-il que le système soit capable de modéliser cette succession des tâches. La section § 6.1.1 avait abordé le problème sous l'angle des ordonnancements d'actions et de références. Nous montrons ici comment une succession peut se représenter sous la forme de domaines de référence. 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 plus finement 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. Les trois notions interviennent dans la construction de domaines de référence. Ainsi, pour ce qui concerne la première notion, 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. Pour ce qui concerne la deuxième notion, l'ontologie des objets, elle repose essentiellement sur une hiérarchie de types, assortie de mécanismes d'inférence permettant par exemple de comprendre des reclassifications telles que « *un chat* » puis « *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* » puis « *les segments* » ; « *la maison* » puis « *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 du corpus Ozkan montre les limites d'une ontologie statique et générique : en effet, 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* ». Elles posent donc le problème de l'acquisition d'ontologies spécifiques, rendue 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. Cette troisième notion est sans doute la notion la plus importante, car elle va permettre de contraindre ou de privilégier une interprétation pour l'énoncé à venir. Le terme « but » regroupe beaucoup de choses : on considère généralement que les buts peuvent être dirigés par les états, les objets, la tâche ou les objectifs. Cette multiplicité ne va pas dans le sens d'une modélisation simple des buts. Nous nous fondons pour cela sur les travaux de Grosz et Sidner (1986), qui ont fait apparaître que la structure de la tâche, c'est-à-dire une hiérarchie intentionnelle, permet de créer des espaces attentionnels limitant l'espace de recherche pour les antécédents d'une expression donnée. Même si Gaiffe (1992) argumente que l'application de cette proposition suppose une tâche fortement hiérarchique, bien définie, connue et exécutée dans un ordre strict par les sujets (ce qui n'est pas forcément le cas pour des tâches tout-venant), nous pensons que faire jouer ce facteur a un intérêt, en tout cas lorsqu'il se combine avec d'autres facteurs. C'est sur cette combinaison que nous avons insisté page 148.

Plus généralement à propos du corpus Ozkan, nous noterons que l'ordre de construction des éléments des dessins a été prédéfini et est le plus souvent respecté. Nous remarquerons également que la clôture des sous-tâches, par exemple la pose d'un élément au bon endroit ou la 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 l'exemple suivant : « *ensuite on a les pyramides dans le désert* », suivi de « *il faut que tu prennes le gros triangle* » puis de « *voilà, et tu* **le** *reprends une deuxième fois un peu décalé par rapport au premier* ». Etant 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 manipulée qu'au type dont relève cette instance. L'exploitation des indices tels que « *ensuite* » et « *voilà* » est une piste importante dans la détection des débuts et fins des buts et sous-buts. Elle fait partie du processus d'identification de l'implicite lié à la tâche que nous allons aborder maintenant.

9.2.2 Identification de l'implicite en tenant compte de la tâche

Restriction des hypothèses liées à l'implicite. D'une manière générale, l'implicite est contraint par la tâche applicative. Sans nous étendre sur ce sujet qui sort de celui concernant la résolution de la référence aux objets, nous montrons ici que l'implicite peut être modélisé comme une liste d'hypothèses complétant l'énoncé émis et à prendre en compte lors de son interprétation. La figure 9.1 montre un exemple d'une telle liste pour l'interprétation de l'énoncé oral « *peins la chaise* » dans notre application d'aménagement d'un intérieur. En fonction du contexte visuel et du déroulement de l'interaction, certaines hypothèses vont être favorisées ou au contraire rejetées.

Enoncé	Implicite sur l'objet	Implicite sur la couleur
« *peins la chaise* »	[celle que je viens de manipuler] [celle à laquelle je n'ai pas encore touché] [la seule visible] [la seule saillante] [la seule à ne pas être peinte] [celle qui a l'air d'une antiquité] [celle que je montre du doigt] [celle que je regarde] [celle juste à côté du meuble que je viens de manipuler] [... enfin le fauteuil, pardon !]	[de la couleur avec laquelle je viens de peindre une autre chaise] [de la couleur utilisée précédemment pour peindre la table] [de la seule couleur qu'ont les autres meubles de la pièce, sauf justement cette chaise] [de la seule couleur que j'ai utilisée jusqu'à présent] [de n'importe quelle couleur, mais fais quelque chose au lieu de refuser toutes mes commandes !] etc.

FIGURE 9.1 – *Exemples d'hypothèses sur l'implicite lorsque l'explicite est incomplet.*

Dans la liste présentée, les hypothèses portent fréquemment sur les actions précédentes ou sur la perception visuelle. Le dialogue homme-machine met en jeu beaucoup d'implicite et le retrouver s'avère nécessaire pour réellement identifier l'intention de l'utilisateur. La détection des indices permettant l'identification de l'implicite est le point de vue que nous adoptons dans (Landragin 2003) et que nous présentons rapidement ici, dans le but de montrer l'importance de la tâche dans ce processus.

Autres composantes de l'implicite. Les aspects implicites qui montrent la prépondérance de la tâche jusque dans l'interprétation de la référence aux objets sont les suivants :

- L'implicite lié au déroulement de l'interaction :
  A force de suivre une même stratégie, l'utilisateur donne des indications sur sa manière de considérer le problème qui lui est soumis. Dans l'exemple de la tâche de Magnét'Oz, la lecture du corpus permet de distinguer deux grandes stratégies dans le rangement des objets : premièrement le rangement guidé par les types d'objets (l'utilisateur se focalise sur un type et range tous les objets de ce type avant de passer au suivant), deuxièmement le rangement guidé par la perception visuelle (l'utilisateur parcourt la scène dans un certain ordre, par exemple celui incité par la saillance et les lignes directrices, et range les objets selon cet ordre sans considération de type). C'est ce type de stratégie que le système devrait être capable de détecter. En effet, un utilisateur qui suit toujours la même stratégie va probablement continuer, et ses actions de référence en seront dirigées implicitement.

- Les implicitations liées à l'énoncé oral courant :
  Le terme implicitation défini page 18 regroupe ce qui est calculé par inférence à partir de la forme propositionnelle de l'énoncé et du contexte. Il s'agit d'une part des présuppositions, d'autre part des implicatures conversationnelles (cf. par exemple Ducrot 1972). Une présupposition est une condition pour qu'une proposition puisse s'évaluer. Elle joue à un niveau linguistique et fait partie du sens de l'énoncé. L'exemple le plus flagrant dans le corpus Magnét'Oz est lié au fait de ranger et d'avoir rangé des objets : dans une scène comportant trois objets dont l'un est visiblement rangé, l'utilisateur va pouvoir dire « *range les deux objets* » sans faire de geste. La présupposition est que « *les deux objets* » réfère à des objets qui ne sont pas encore rangés. Si l'on ne tient pas compte de cette information, l'énoncé est ambigu car le nombre d'« *objets* » présents dans la scène n'est pas celui indiqué. En revanche, si l'on tient compte de cette information, le nombre d'« *objets non encore rangés* » correspond bien à celui indiqué dans l'énoncé. Une implicature conversationnelle est une inférence que l'interlocuteur doit faire lorsque le locuteur viole ouvertement le principe de coopération sur lequel se fondent les maximes de Grice (1975). Elle correspond à la différence entre ce qui est dit (le sens) et ce qui est communiqué (son interprétation). L'utilisateur pourrait par exemple dire « *ces objets sont mal placés* » pour communiquer au système la nécessité de les ranger. Le problème avec les implicitations est qu'elles font intervenir des informations très diverses, aussi bien sur la tâche en cours que sur des connaissances générales sur le monde. Un système face à l'expression précédente doit comprendre qu'il ne s'agit pas seulement d'une constatation mais de l'intention de provoquer une réaction. Ce fait n'est pas évident et doit faire partie du contexte. Or de nombreux faits du même ordre doivent aussi être inclus dans le contexte, qui en devient complexe et hétérogène. En fin de compte, si l'implicite lié aux présuppositions semble formalisable dans le cadre d'une tâche donnée, il n'en est pas de même de l'implicite lié aux implicatures conversationnelles.

- Les effets interprétatifs de l'énoncé oral :
  Dans un même ordre d'idée que l'implicature conversationnelle mais à un niveau extralinguistique, se trouvent les effets interprétatifs comme les sous-entendus. D'une manière générale, un sous-entendu correspond à ce à quoi l'auditeur se réfère pour saisir l'intention (au sens large) du discours qui lui est adressé. Nous n'en avons trouvé aucun exemple dans le corpus Magnét'Oz, si ce n'est peut-être la répétition systématique d'une même commande que l'on peut interpréter comme un signe dénotant que le locuteur se lasse. Un exemple classique est « *tu peux me passer le sel* » qui peut sous-entendre que le plat est mal préparé

ou que le locuteur veut changer de sujet de conversation. Encore une fois, l'identification d'une telle information implicite est délicate pour un système et nous espérons que la tâche contient suffisamment de contraintes pour que ces cas de figure n'apparaissent pas.

9.3 L'adaptation du modèle à un formalisme computationnel

Plusieurs possibilités d'implantation se présentent selon la partie du modèle que l'on veut mettre en valeur. Ainsi, l'appariement de domaines incite fortement à implanter un mécanisme, classique en traitement automatique des langues, d'unification de structures de traits typées. Nous commencerons par présenter cet aspect à la fois représentatif et algorithmique. En revanche, la description du contexte visuel avec les critères de la Gestalt et la notion de la saillance s'avère efficace avec un formalisme permettant des inférences. Nous montrerons ainsi comment les logiques de description permettent certains calculs, pour lesquels nous partirons également d'une représentation de l'énoncé sous une forme logique similaire à celles utilisées couramment en sémantique formelle. Explorer ces trois types de formalisation permettra d'appréhender ultérieurement le travail d'implantation proprement dit, travail nécessitant l'intégration dans une plate-forme incluant reconnaissance de la parole et du geste, et relégué par conséquent en perspectives.

9.3.1 Adaptation au mécanisme d'unification de structures de traits

Notes préliminaires sur le mécanisme d'unification. L'appariement de domaines de référence sous-spécifiés avec des domaines construits à partir des caractéristiques du contexte consiste à sélectionner les couples de domaines proches au niveau des caractéristiques spécifiées : un domaine sous-spécifié de type « triangle » et pour lequel une partition met en jeu le critère de différenciation « couleur » ne pourra s'apparier qu'avec des domaines de type « triangle » et contenant au moins une partition avec le critère de différenciation « couleur », toutes les autres propriétés étant libres. Dans ce mécanisme d'appariement (ou unification), l'ordre des propriétés (ou attributs) testées n'est pas neutre. La première propriété testée est ainsi favorisée, et, d'une manière générale, il doit exister une relation d'ordre sur l'ensemble des propriétés. De plus, les structures regroupant les attributs ne doivent pas contenir de cycle. Enfin, l'unification nécessite une hiérarchie de types, pour pouvoir déterminer le plus grand unificateur lors de l'unification de deux types. Ce sont de telles préoccupations que nous retenons dans la modélisation qui suit.

Modélisation des domaines de référence sous la forme de 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 :

1. Un identifiant garantissant l'unicité de cette structure pour l'ensemble des espaces de représentation du système de dialogue considéré.

2. Un facteur de groupement indiquant la nature du module qui justifie l'existence de ce domaine de référence.
3. Un type qui, par référence à une ontologie du domaine concerné, subsume l'ensemble des types des entités composant le domaine.
4. Un modifieur ou ensemble de propriétés pouvant préciser le type.
5. Une cardinalité exprimée soit sous la forme d'un entier, soit par un code de numération : simple, pluriel, massif, inconnu, etc.
6. Une séquence éventuellement vide de partitions.

La relation d'ordre entre les attributs est la suivante : facteur de groupement, type, modifieur du type, cardinalité. Pour l'unification par exemple de « triangle » et de « carré », la hiérarchie des types correspond à celle de l'ontologie. Une partition est une sous-structure d'un domaine de référence, et est pour sa part représentée à l'aide de quatre caractéristiques principales :

1. Un critère de différenciation, représentant la ou les caractéristiques justifiant la discrimination des différentes entités composant le domaine pour la partition considérée. Ce critère est pris au sein d'une ontologie spécifique regroupant les critères linguistiques, les différentes combinaisons des critères de la Gestalt, et les critères spécifiques à l'application comme ceux qui sont liés à la structuration en buts et sous-buts.
2. Une marque d'ordonnancement à valeur binaire (oui ou 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.
3. Un contenu, formé d'une suite de références à des domaines de référence ou à des entités individuelles.
4. Une marque de focalisation, éventuellement vide, correspondant à un index sur le contenu de la partition.

La relation d'ordre est cette fois : critère de différenciation, critère d'ordonnancement, marque de focalisation. La combinaison des structures de traits s'effectue à l'aide d'un mécanisme 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. Du fait des possibles cycles (un domaine renvoie à une partition qui peut renvoyer à des domaines), un test d'arrêt s'avère de plus nécessaire. Enfin, nous remarquerons 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 ». Nous noterons qu'une telle représentation permet d'intégrer aisément le traitement des anaphores associatives par décomposition (« *le triangle* » puis « *les segments* ») à l'aide de partitions particulières des entités élémentaires. Les structures de traits et les caractéristiques du mécanismes d'unification ont été présentées dans (Landragin 2002c). L'implantation informatique correspondante n'est cependant pas encore réalisée et constitue l'une de nos principales perspectives à court terme.

9.3.2 Adaptation aux logiques de description

Notes préliminaires sur les logiques de description. Les logiques de description constituent une famille de langages formels spécifiant syntaxe et sémantique, et dotés de mécanismes d'inférence

spécialisés. Le principe de base est la décomposition des connaissances en deux ensembles : un niveau terminologique (*TBox* pour *Terminological Box*) et un niveau factuel (*ABox* pour *Assertional Box*). Le premier regroupe la description des concepts (entités génériques) et de leurs rôles relatifs, le second regroupe celle des individus (instances particulières). A partir du moteur d'inférences, les requêtes possibles sont par exemple : le test de subsomption (est-ce qu'un concept donné est plus spécifique qu'un autre ?) ; le test de satisfiabilité d'un concept (est-ce qu'un concept donné admet des instances ?) ; le test d'instanciation (est-ce qu'un individu donné est bien une instance d'un concept donné ?).

Modélisation des domaines de référence visuels. Dans le cadre de la construction des domaines visuels et de la gestion de la saillance visuelle pour l'interprétation, il s'avère intéressant de décrire dans la *TBox* tout ce qui concerne les entités visuelles et leur fonctionnement, et dans la *ABox* tout ce qui concerne l'instanciation du contexte visuel à un instant donné. Le but est de construire une base répondant aux requêtes suivantes : quel objet est saillant ? quels sont les objets rouges ? à quel groupe appartient tel objet ? quels sont les autres éléments de tel groupe ? etc. Nous définissons donc dans la *TBox* les types d'objets, les types de propriétés que peuvent prendre les objets, ainsi que les principes de fonctionnement d'un groupe perceptif. Une illustration très rapide est la suivante :

$triangle \sqsubseteq object$
$object : has\_attribute\ X\ Y$
$object \sqsubseteq \exists has\_colour.Colour$
$object \sqsubseteq \exists has\_size.Size$
...
$salient\_group = group \sqcap \exists has\_member.salient$
$focussed\_group = group \sqcap \exists has\_member.focus$

Dans la *ABox*, nous mettons la description d'une scène visuelle à un instant donné, c'est-à-dire la description des objets, de leurs propriétés, de leur éventuelle saillance, et des groupes perceptifs. Ainsi, l'algorithme de construction de domaines visuels vient modifier cette *ABox* en spécifiant des groupes et leurs éléments. Le calcul de la saillance visuelle aboutit dans notre exemple à $salient(t_1)$:

$triangle(t_1)$
$has\_colour(t_1, green)$
$attribute(X, t_1, 640)$
...
$salient(t_1)$
$focus(t_1)$
...
$group(g_1)$
$has\_member(g_1, t_1)$
...
$group(g_2)$
$has\_member(g_2, g_1)$

Un traitement rapide de l'exemple de l'introduction pourrait alors être le suivant : l'expression

orale « *ces trois triangles* » se traduit par la forme logique :

$$\exists E\ (salient\_group(E)) \wedge (card(E) = 3) \wedge (\forall x \in E\ triangle(x)).$$

Cette forme logique correspond à une interprétation sémantique de l'expression référentielle, sans considérations pragmatiques. De son côté, la trajectoire gestuelle, après avoir été interprétée comme ciblant un groupe d'objets éventuellement plus large que les deux objets recouverts par elle, se traduit également par une forme logique :

$$(\exists E\ pointing(E)) \vee (\exists E\ \exists X \subset E\ pointing(X) \wedge group(E)).$$

Comme nous l'avons montré dans (Landragin *et al.* 2000), on peut voir l'interprétation de la coréférence comme la fusion de ces deux formes logiques en s'aidant des caractéristiques du contexte visuel. Ces caractéristiques étant ici incluses dans la spécification des $TBox$ et $ABox$, nous pouvons concevoir que le système retrouve le groupe perceptif contenant les deux triangles recouverts par la trajectoire. Si cette illustration reste simpliste, elle permet néanmoins de montrer l'intérêt des logiques de description et des mécanismes d'inférence associés pour le traitement d'une partie des problèmes soulevés dans l'introduction.

RÉCAPITULATIF
Nous montrons dans ce chapitre comment le modèle générique des domaines de référence peut être instancié pour un type d'interaction particulier, pour une tâche applicative particulière, et dans un formalisme computationnel particulier. Réalisé à partir d'exemples d'interaction dans un espace en 2D avec écran tactile, le modèle s'avère parfaitement compatible avec une interaction en 3D et pour le geste haptique. Nous montrons ainsi comment les seules modifications ont trait au réglage de quelques paramètres, et non à la structure même des domaines de référence ou de l'architecture décrite dans le chapitre précédent. Nous explorons les facettes de la tâche applicative, et nous mettons l'accent sur la difficulté pour le système de tenir compte de toutes ces facettes. Nous reformulons ainsi le problème du point de vue de l'identification de l'implicite lié à une tâche particulière, problème pour lequel nous identifions des perspectives de recherche. Enfin, nous donnons deux exemples de formalisations qui montrent que certaines composantes du modèle s'adaptent directement à des formalismes existants, mais que l'ensemble du modèle est encore trop complexe et trop hétérogène pour être implanté facilement.

Chapitre 10

Exploitations connexes du modèle

En plus de l'interprétation des expressions référentielles multimodales, le modèle des domaines de référence permet-il la modélisation d'autres processus cognitifs ? Comment peut se faire son « retournement » pour la génération automatique d'expressions référentielles ? Quels problèmes autres que la référence peuvent être traités ? Quelles peuvent être les ambitions de ce modèle ?

Nous avons parfois évoqué dans les chapitres précédents le problème de la génération d'expressions référentielles. Nous voulons ici (§ 10.1) faire le point sur ce problème et sur l'adaptation du modèle des domaines de référence pour son traitement. D'autre part, nous avons souvent mis en avant notre souci de plausibilité cognitive dans la formalisation des concepts et dans la spécification des processus liés aux domaines de référence. S'il nous apparaît bien trop tôt pour intégrer les domaines de référence à un modèle général de la cognition, nous voulons étudier ici (§ 10.2) les bases nécessaires à une telle ambition. Nous montrerons en particulier comment valider à l'aide d'outils issus de la psychologie les notions que notre modèle met en œuvre.

10.1 Génération automatique d'expressions référentielles

10.1.1 Notes préliminaires sur la génération

Dissemblance cognitive et similitudes computationnelles. D'un point de vue cognitif, comprendre et produire du langage sont deux choses bien différentes. Il est classiquement admis qu'ils font appel à deux systèmes distincts. Cette hypothèse s'appuie sur le fait qu'un auditeur peut se contenter d'une analyse partielle pour comprendre un énoncé (les indices sémantiques et situationnels peuvent le dispenser d'une analyse syntaxique complète), alors que le locuteur doit expliciter complètement tous les aspects de l'énoncé qu'il produit (Caron 1989). Cette hypothèse s'appuie aussi sur la distinction classique entre aphasies d'expression et aphasies de réception : un malade atteint de l'aphasie de Broca reste capable de dénomination mais a perdu la capacité à combiner les mots en phrases ; un malade atteint de l'aphasie de Wernicke échoue à retrouver les mots, qu'il remplace par des néologismes.

Même si les mécanismes sont différents, on peut supposer qu'ils reposent sur des structures et des types de procédures similaires. D'un point de vue informatique, les échanges entre les deux branches de recherche que sont la compréhension automatique et la génération automatique

s'avèrent régulièrement fructueux. De nombreuses considérations, aussi bien théoriques qu'algorithmiques, sont communes, et de nombreux algorithmes sont conçus pour être réversibles. Nous pouvons donc tenter d'appliquer les concepts de notre modèle à la génération, en particulier la notion de domaine de référence.

Exploration des expressions référentielles possibles. Dans le cadre de la génération d'expressions référentielles verbales ou multimodales, le principal problème est d'ordre combinatoire. Il est impossible de générer toutes les expressions référentielles possibles puis d'en choisir une, sur un critère ou un autre. Il s'avère au contraire nécessaire de spécifier des heuristiques pour n'émettre que des hypothèses intéressantes. Pour une même intention référentielle, les mots et les constructions syntaxiques possibles s'avèrent en effet très nombreux. Les possibilités pour référer à un objet sont multiples, surtout dans un contexte multimodal où la combinatoire explose avec la prise en compte des paramètres visuels et gestuels.

Une stratégie consiste donc à faire des choix dans l'exploration des possibilités, par exemple de restreindre une première mention à l'utilisation d'un groupe nominal défini, ou de restreindre l'association d'un geste ostensif à un démonstratif. Ces choix peuvent éventuellement être remis en cause, par exemple si toutes les descriptions définies pour une première mention s'avèrent très complexes. La stratégie se caractérise alors par des retours en arrière qu'il s'agit de minimiser. Enfin, lorsque les hypothèses ont été émises, il s'agit de choisir parmi elle. Que ce soit pour ce choix ou pour la minimisation des retours en arrière, les critères que nous avons présentés dans notre travail, et particulièrement le critère de pertinence, constituent des réponses que nous allons développer.

10.1.2 Apports du modèle des domaines de référence

Rôles de la gestion de groupes d'objets. Appliquer notre structuration des objets en domaines de référence à la génération présente quelques intérêts. Premièrement, c'est une façon de favoriser la cohérence de la communication en la restreignant à des domaines correspondant à des buts et des sous-buts, ou à des espaces attentionnels. Deuxièmement, cela permet de prendre en compte la continuité dans le dialogue, en autorisant des suites de références internes à un domaine. Troisièmement, la structuration de la scène visuelle en groupes perceptifs permet de réduire la complexité des gestes et des expressions référentielles verbales. Ainsi, si un geste ostensif doit désigner un groupe perceptif, ce geste pourra être simplifié sans ambiguïté. La minimisation de la fatigue physique est un critère important pour la génération réaliste de gestes. Au niveau de l'expression référentielle, certaines constructions comme le pluriel sans adjectif numéral s'avéreront suffisantes.

Rôles de la saillance. Dans un même ordre d'idée, une prise en compte de la notion de saillance permet de réduire la complexité des expressions verbales, ou encore de remettre en question la nécessité de produire un geste. Une remarque importante ici est que si la saillance permet de trouver une interprétation privilégiée à une expression *a priori* ambiguë, elle ne doit pas pour autant être utilisée pour produire de telles expressions. La saillance est ainsi exploitée de manière très différente en interprétation et en génération. L'exemple que nous avons souvent discuté est celui d'un environnement comprenant plusieurs N, un seul d'entre eux étant saillant. Si l'on veut référer à ce dernier, la saillance fait que l'expression « *le N* » ne sera pas ambiguë en compréhension, ou en tout cas trouvera une interprétation très privilégiée. En reprenant cet exemple en génération, la question qui se pose est la suivante : la production de « *le N* » dans le même environnement est-elle pertinente ? La possible ambiguïté de cette expression (si l'utilisateur ne voit

pas la saillance ou ne la considère pas comme suffisante) va dans le sens d'une réponse négative. La simplicité de l'expression « *le N* » par rapport à une expression telle que « *le N* P_1 P_2 », et donc la diminution de l'effort de traitement, va dans le sens d'une réponse positive. Notre but n'est pas ici de trancher, mais de montrer qu'un modèle en génération d'expressions référentielles a tout intérêt à se poser ce problème et à étudier la notion de saillance. Nous espérons que la caractérisation que nous en faisons dans le chapitre 5 permettra d'apporter quelques éléments à de telles études que nous n'avons pas encore entreprises.

Rôles de la mesure de la pertinence. Une fois que le système a déterminé sur quel(s) objet(s) va porter son énoncé, c'est-à-dire une fois qu'il a fixé son intention référentielle, le problème est de choisir l'information à générer. Il s'agit premièrement de choisir entre produire une expression référentielle verbale et une expression référentielle multimodale, deuxièmement de sélectionner parmi les informations possibles celles à générer. Pour l'éventuel geste ostensif, il s'agit ainsi de déterminer non seulement la forme de la trajectoire, mais aussi les demonstrata, sachant que l'ostension peut s'appliquer aux seuls référents ou à un domaine dans lequel ceux-ci seront extraits par l'expression verbale. Pour celle-ci, il s'agit tout d'abord de choisir un déterminant, en fonction de l'historique de l'interaction (par exemple si l'intention référentielle est une anaphore) ou en fonction de la présence d'un geste coréférent. Il s'agit ensuite de choisir les propriétés discriminantes qui permettent à l'interlocuteur de distinguer les référents intentionnels dans l'ensemble des référents potentiels. Ces propriétés sont prioritairement la catégorie (qui se traduit généralement par un substantif), et ensuite les propriétés de couleur, de taille ou de positionnement qui se traduisent par des adjectifs qualificatifs ou des compléments du nom. Une telle approche est illustrée de manière très claire par l'algorithme incrémental de Reiter et Dale (1992), dont nous avons déjà parlé dans le chapitre 5.

Une mesure de la pertinence telle que nous l'avons envisagée (du moins sous la forme de traits) dans la section 7.3, s'avère très utile, en particulier pour la phase de détermination des informations à générer. La pertinence sert tout d'abord à identifier un contexte ou domaine de référence pertinent. Ainsi, lorsque les derniers énoncés se sont placés dans un contexte précis et que l'intention de l'énoncé courant est de s'y tenir, le choix des propriétés discriminantes se fait dans le sous-ensemble d'objets correspondant. C'est là tout l'intérêt des domaines de référence pour la génération d'expressions référentielles. La pertinence sert ensuite à évaluer les différentes hypothèses d'expressions référentielles verbales ou multimodales. A l'aide des ensembles de traits que nous avons proposés, nous pouvons évaluer les effets contextuels et l'effort de traitement de chaque hypothèse dans le domaine. La confrontation des scores obtenus pour chaque hypothèse permet alors de choisir la plus pertinente.

Nous avons également évoqué une autre utilisation de la pertinence : celle consistant à minimiser les retours en arrière lors de l'émission des hypothèses. Nous montrons ici comment combiner les avantages de l'exploitation de la pertinence en génération avec ceux de son exploitation en compréhension. En effet, il se peut qu'une évaluation *a posteriori* des hypothèses soit nécessaire, c'est-à-dire du point de vue de l'interlocuteur et non du système. Cette évaluation consiste à oublier l'intention référentielle à partir de laquelle ont été construites les hypothèses, et à utiliser la pertinence comme critère d'interprétation. En fonction des résultats obtenus, par exemple l'identification claire de l'intention référentielle ou au contraire la détection d'une possible ambiguïté, un retour en arrière s'avérera peut-être nécessaire, dans le but d'émettre une autre série d'hypothèses et de recommencer le processus de leur évaluation. Un tel retour en arrière est coûteux et doit être évité dans la mesure du possible. Une réponse partielle à ce problème réside dans la notion de domaine de référence : c'est la pertinence d'un domaine dans une situation de dialogue donnée qui garantit la pertinence de tout le processus.

10.1.3 Vers un modèle de génération multimodale

Génération d'un geste ostensif. van der Sluis (2001) propose une extension de l'algorithme de génération d'expressions référentielles de Dale et Reiter (1995) pour la génération d'expressions référentielles multimodales. Elle identifie deux facteurs permettant de prendre la décision de générer un geste ostensif ou non : d'une part l'efficacité d'un tel geste, d'autre part l'inefficacité d'une longue description langagière. L'efficacité du geste se base sur la classique loi de Fitts (1954). Il s'agit d'une formule quantifiant la difficulté (id) à désigner en fonction de la distance (d) et de la taille (w) du demonstratum : $id = \log_2(2d/w)$. Un seuil déterminé de manière empirique vient placer une limite dans la difficulté : au-dessous de cette limite, on considère qu'un geste peut être produit facilement et efficacement. Quant au facteur lié à l'inefficacité d'une description langagière, il s'appuie sur la méthode incrémentale de Dale et Reiter : si la liste des propriétés nécessaires pour distinguer le référent des autres objets s'avère plus élevée qu'un certain seuil (déterminé également de manière empirique), on recourt au geste ostensif. L'expression verbale associée peut alors être réduite aux seules propriétés discriminantes dans l'espace visuel focalisé par le geste.

Nous proposons d'ajouter à cette approche une prise en considération du contexte visuel plus fine que la seule taille du demonstratum. Nous en avons en effet montré dans les chapitres 4 et 5 en quoi les considérations visuelles étaient relatives. Ainsi, il nous apparaît insuffisant de parler de la taille d'un objet sans considérer les tailles des objets qui lui sont proches. La saillance visuelle nous semble ainsi un facteur important dans la décision de générer un geste : si le demonstratum est plus saillant que les objets qui l'entourent, le geste en est d'autant plus efficace. De même, la notion de domaine de référence visuel apporte un facteur supplémentaire : si le demonstratum est isolé de tout groupe perceptif, un geste imprécis peut le désigner aussi facilement qu'un geste précis ; au contraire, si le demonstratum fait partie d'un groupe perceptif, un geste doit détruire la cohésion du groupe pour l'en extraire, requiert donc une précision particulière et s'avère en cela bien plus difficile à générer. C'est donc à partir d'une combinaison des facteurs de distance, de taille, de saillance et d'appartenance à un domaine visuel que nous proposons d'appréhender la génération d'un geste ostensif. Des mesures empiriques restent à réaliser pour déterminer les poids relatifs de chacun de ces facteurs.

Génération de la partie verbale de l'expression référentielle. Parmi les propriétés se trouve le type ou catégorie. A la suite de Dale et Reiter (1995) et de van der Sluis (2001), nous proposons de générer la catégorie, même si elle n'est pas discriminante. Elle correspond généralement au substantif qui s'avère souvent indispensable, un groupe nominal sans nom ne s'utilisant que dans certains cas particuliers. Or un substantif sans précision de catégorie, tel que « *forme géométrique* » ou « *objet* », s'avère aussi coûteux à générer que « *triangle* » ou « *carré* ». A coût égal, autant générer la catégorie.

Considérons maintenant la génération d'un adjectif numéral en plus de l'information de nombre portée par le déterminant. Lorsque le geste à lui seul ne suffit pas à indiquer le nombre de demonstrata, l'information de cardinalité peut s'avérer indispensable. C'est le cas si le geste désigne par exemple un domaine de référence dans lequel les référents devront être extraits, ou s'il s'agit d'un pointage vers un groupe perceptif. Dans l'exemple de la figure 7.1-B page 132, imaginons que l'on veuille désigner les triangles contenus dans le groupe perceptif hétérogène placé tout à droite de la scène. L'association d'un geste désignant globalement le groupe, par exemple un entourage, et de l'expression verbale « *les triangles* », peut amener à une extraction incomplète des référents. En les cherchant bien, ceux-ci sont au nombre de cinq, deux d'entre eux étant de couleur grise et les trois autres de couleur blanche. Dans un même contexte, l'ex-

pression « *les cinq triangles* » s'avère plus pertinente à générer, car elle permet à l'interlocuteur d'être sûr de ne pas oublier un référent. Si l'on considère maintenant le groupe perceptif situé au milieu de la même figure, l'association d'un geste désignant globalement le groupe et de l'expression « *les triangles* » s'avère aussi pertinente à générer que le même geste avec « *les deux triangles* ». Le groupe ne comporte en effet que cinq objets, parmi lesquels les deux triangles se distinguent facilement. Or le pluriel porté par « *les* » suffit à déterminer que le résultat de la référence concerne bien les deux triangles possibles. L'adjectif numéral « *deux* » n'apporte aucune information supplémentaire, il ne fait qu'augmenter la taille de l'expression référentielle. Nous proposons d'appliquer ces règles à la génération de l'adjectif numéral, non seulement dans les cas d'ostension non coréférente, mais également dans tous les cas où le geste peut être interprété de manière imprécise (typiquement les cas du pointage et du ciblage).

Enfin, lors de la génération des propriétés discriminantes, nous proposons de suivre le principe de l'algorithme de Dale et Reiter, en ajoutant la règle suivante : lorsque plusieurs possibilités s'avèrent équivalentes en terme de coût, c'est-à-dire en nombre de propriétés à générer, nous privilégions la propriété qui apparaît comme le critère de différenciation ou d'ordonnancement de la partition activée dans le domaine de référence incluant les référents intentionnels. Nous exploitons avec cette règle l'intérêt des domaines de référence en terme d'organisation des informations. Encore une fois, il ne s'agit ici que de propositions qui restent à évaluer dans un système ou par des expérimentations. Nous allons maintenant nous intéresser à l'élaboration de telles expérimentations.

10.2 Vers un modèle cognitif de la communication

Les rapports entre ce que nous faisons et la psychologie cognitive sont multiples : pour la validation de ses modèles, la psychologie utilise non seulement des méthodes statistiques, mais également des méthodes de simulation, en particulier à l'aide de l'informatique. L'idée généralement admise est que, si le système reproduit correctement le comportement observé avant modélisation, alors le système est considéré comme satisfaisant. Cela nous semble insuffisant : on ne peut valider que certaines situations, qui ne correspondent pas à toutes les situations possibles et qui ne correspondent généralement pas à l'apport du modèle par rapport aux approches classiques. Pour le moment, notre objectif n'est pas de valider un modèle complet de la cognition, mais de nous focaliser sur les notions de domaines, de saillance et de pertinence sur lesquelles s'axe notre modèle. Ces notions peuvent-elles être l'objet d'un modèle cognitif général de la communication humaine? Pour ce faire, il s'avère nécessaire de valider chacune de ces notions, paramètre par paramètre. Cette section est ainsi axée sur la définition et la spécification, étape par étape, situation par situation, de protocoles expérimentaux de validation.

10.2.1 Elaboration de protocoles expérimentaux

De la nécessité d'expérimentations. Pourquoi procéder à des expérimentations? Parce que les corpus ne suffisent pas. Comme nous l'avons déjà évoqué dans le chapitre 3, la majorité d'entre eux ne contiennent en effet aucune description détaillée du contexte visuel, et ne permettent pas de valider des détails aussi précis que, par exemple, l'intérêt ou l'inutilité de l'adjectif numéral « *deux* » dans la situation décrite ci-dessus. De plus, et ceci est le principal argument pour des expérimentations avec un protocole différent de celui du magicien d'Oz, les corpus ne suffisent pas parce qu'ils sont généralement assez courts et contiennent peu de situations mettant en avant certaines finesses du processus d'interprétation. Autrement dit, les cas les plus fréquents ne permettent généralement pas de valider un point précis du modèle. Les situations du corpus

Magnét'Oz comprennent ainsi un grand nombre de références directes telles que « *cet objet* » associé à un geste désignant un objet, mais quasiment aucune situation illustrant les possibilités d'évolution des domaines de référence au cours du dialogue. Nous devons donc nous tourner vers des données obtenues par une mise en situation très précise.

Les expérimentations que nous proposons consistent ainsi à placer le sujet dans un rôle d'interprétation dans une situation correspondant à un aspect précis de notre modèle. Le but est d'enregistrer l'interprétation que font les sujets d'expressions référentielles fixées dans un contexte linguistique, visuel et mémoriel maîtrisé. Elles se rapprochent en cela des méthodes classiques de la psycholinguistique. L'avantage de cette méthode est que les conclusions sur le comportement d'un système de dialogue homme-machine sont faciles à déduire. Par exemple, si certaines interprétations sont systématiquement rejetées, le modèle pourra ne pas en tenir compte. Si, dans le cadre d'une ambiguïté, une interprétation particulière est très souvent privilégiée, le modèle pourra la favoriser.

Un protocole expérimental. Le contexte linguistique, visuel et mémoriel se caractérise par l'affichage de scènes successives (projetées sur un mur ou tout simplement sur un écran d'ordinateur) et par l'émission de messages vocaux pré-enregistrés, au nombre d'au moins un par scène. L'enregistrement des réactions du sujet se fait soit à l'aide d'une caméra vidéo, soit, de manière plus directe, par l'enregistrement des trajectoires (si l'expérimentation inclut un écran tactile) ou des clics effectués avec la souris. Dans tous les cas, des précautions particulières seront prises pour que tous les paramètres soient maîtrisés et pour que les résultats soient exploitables. Ainsi, pour rendre l'expérimentation facile à mettre en œuvre et donc facile à reproduire et à analyser, nous proposons les quelques règles suivantes :

- simplicité et aspect ludique de l'expérience, la rendant ainsi réalisable par des enfants ;
- souplesse : pauses autorisées ; possibilité d'ignorer une situation, de ne pas répondre ;
- qualité de l'interface : bonne synchronisation temporelle, audibilité des messages, « *bip* » clair à chaque clic de la souris, retours visuels saillants, etc., le but étant que le sujet ne se demande jamais ce qu'il doit faire ou si sa réponse a bien été prise en compte.

Nous mettons d'autre part l'accent sur la nécessité de spécifier les situations les plus neutres et les plus banales possible. L'important est de trouver des situations banales dont notre modèle prédit une interprétation qui ne rentre pas dans le cadre des théories classiques. La neutralité est obtenue en attribuant des valeurs insignifiantes à tous les composants d'une situation qui ne constituent pas des paramètres à tester. Ainsi, comme nous nous focalisons sur l'interprétation des expressions référentielles, nous inclurons dans nos énoncés les actions les plus neutres possibles. Une action de déplacement d'objets n'est par exemple pas neutre puisqu'elle entraîne une modification des groupes perceptifs de la scène visuelle. De même pour une action de suppression. Ces deux actions ne pourront donc être utilisées qu'à la fin de l'exploitation d'une scène visuelle, c'est-à-dire juste avant le passage à une autre scène. Les actions suivantes nous semblent en revanche être neutres et mettre l'accent sur la façon dont les objets sont introduits : « *voyez-vous les N ?* » ; « *cette scène contient des N ; montrez-moi les N P* » ; « *sélectionnez les deux N* » ; « *désignez le N* » ; « *marquez le N* » ; « *cliquez sur le N* ».

Une attention toute particulière doit être donnée à la spécification d'une succession d'actions destinée à tester l'intérêt du modèle lors d'une continuité dans le dialogue. Ainsi, la succession « *sélectionnez les N* » puis « *supprimez les P* » incite à chercher les P dans l'ensemble des N. La succession « *montrez-moi les formes de grande taille* » puis « *montrez-moi les formes rouges* » incite plutôt à oublier le résultat de la première action lors de l'exécution de la seconde. Nous ne proposons pas ici de règles d'interprétation, nous voulons juste montrer que le choix du prédicat

peut influer sur elle, et que ce choix doit donc être réfléchi.

Des mesures pour exploiter les résultats. Pour que les résultats soient valides, certaines règles doivent être suivies, premièrement dans le nombre de sujets (20, 40, ou plus selon les variations des conditions expérimentales), deuxièmement dans la mesure des temps de réaction si celle-ci s'avère d'importance (calcul d'un seuil maximal au-delà duquel le résultat ne sera pas pris en compte, pondération par la distance entre la main du sujet et l'écran où se fait la désignation, ce qui peut leur imposer un placement fixe), troisièmement dans l'analyse des résultats. Pour cette dernière phase, de nombreux tests statistiques sont disponibles dans la littérature (cf. Blalock 1979). Ils correspondent à des caractéristiques précises des données. Dans (Landragin *et al.* 2002b), nous avons par exemple utilisé le test de Mann-Whitney qui correspond à deux échantillons de variables non mesurables sur une échelle d'intervalle. L'indice de significativité permet alors d'interpréter une observation : une significativité de 0,05 est suffisante, l'idéal étant de 0,0001. N'ayant pas procédé à des expérimentations dans le cadre de cette thèse, nous n'entrerons pas plus dans les détails, pour nous concentrer sur la description de situations.

10.2.2 Spécification de situations expérimentales

Validation expérimentale de la gestion de groupes d'objets. Pour vérifier l'hypothèse que l'interprétation se fait dans tel domaine de référence et non dans tel autre, il faudrait tester l'interprétation d'une succession d'énoncés en montrant, d'une part que toutes les réponses correspondent à des domaines calculés, et d'autre part qu'aucune réponse ne peut correspondre à un domaine non calculé. Ceci nécessite l'élaboration de scènes avec suffisamment de successions du même genre que celle de l'exemple de l'introduction pour tester tous les domaines possibles. Nous proposons ici quelques situations expérimentales prototypiques centrées sur le test des domaines de référence :

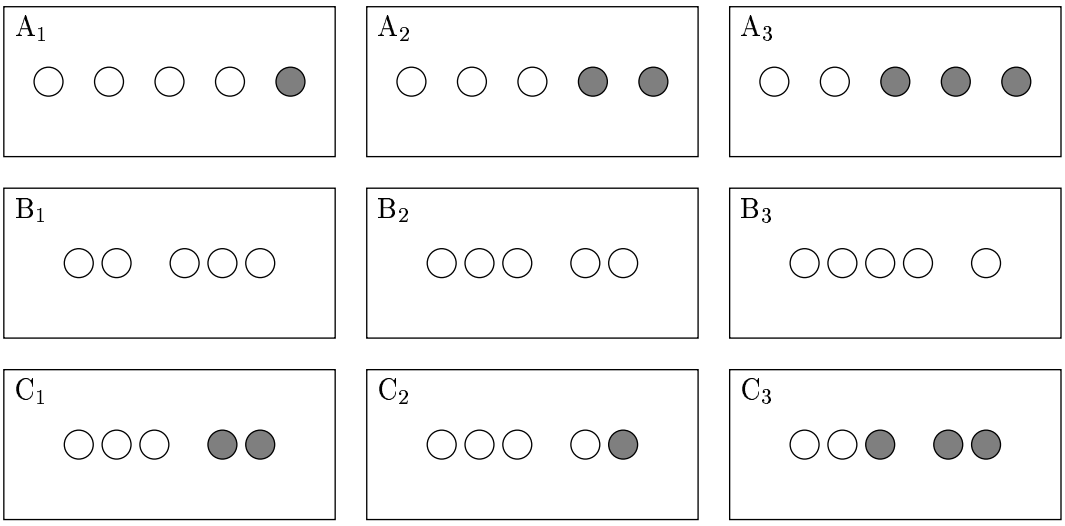


Figure 10.1 – *Quelques scènes pour tester la séparation entre la gauche et la droite.*

- Validation de l'existence des domaines de référence visuels à l'aide des critères de la Gestalt (toutes les scènes évoquées pour ce point sont regroupées dans la figure 10.1) :
  1. Situations où la similarité permet de délimiter des domaines. Dans la série des fi-

gures 10.1-A qui contiennent cinq cercles alignés et équidistants, la similarité de couleur permet de construire deux groupes. L'interprétation donnée à « *montrez-moi les cercles de gauche* » permet de vérifier l'existence de ces groupes : si le sujet montre quatre cercles dans la figure 10.1-A_1, trois dans la figure 10.1-A_2 et deux dans la figure 10.1-A_3, il est clair que c'est le critère de similarité qui partitionne la scène. Dans le cas contraire, si par exemple le sujet assimile le milieu de la scène au troisième cercle, les cercles de droite aux deux cercles les plus à droite et les cercles de gauche aux deux les plus à gauche, alors ce n'est pas la similarité qui intervient dans le partitionnement mais seulement la position horizontale.

2. Situations où la proximité permet de délimiter des domaines. Dans la série des figures 10.1-B, le paramètre utilisé dans la présentation des cercles n'est plus la similarité mais la proximité. L'interprétation que fait le sujet de « *les cercles de gauche* » permet de vérifier la prédominance de ce paramètre dans le partitionnement de la scène en deux groupes perceptifs.
3. Situations où la proximité et la similarité se renforcent ou entrent en conflit. Le principe et l'expression testée sont toujours les mêmes, les paramètres intervenant dans la présentation des objets étant cette fois au nombre de deux. Soit ces deux paramètres se renforcent (figure 10.1-C_1), ce qui incite fortement à interpréter « *les cercles de gauche* » comme les trois cercles blancs ; soit ils donnent chacun lieu à un partitionnement différent (figures 10.1-C_2 et 10.1-C_3). Le but est alors de tester si l'un des deux paramètres est prépondérant. Si par exemple la majorité des réponses des sujets correspondent aux trois cercles les plus à gauche, nous en déduiront que le critère de proximité est plus important que celui de similarité. Les conséquences sur notre modèle seront dans ce cas l'attribution d'une plus grande importance aux domaines visuels construits sur la proximité, ainsi que l'éjection des résultats fondés sur la seule similarité lors de l'intégration des dendrogrammes.

- Validation de l'existence des domaines visuels par la focalisation dans un sous-espace visuel : le but est ici de tester l'interprétation de « *le N* » dans un environnement comprenant plusieurs N, après focalisation dans un sous-ensemble de l'environnement qui ne contient qu'un seul N, cette focalisation étant réalisé par une première référence. L'exemple typique est celui de l'introduction, un exemple similaire étant présenté dans (Kessler *et al.* 1996). Suite aux expérimentations présentées dans ce rapport, la conclusion des auteurs va dans notre sens.

- Validation de l'existence des domaines linguistiques à l'aide de l'ellipse nominale :
  1. Dans la scène de la figure 10.2 comportant un grand nombre d'objets aux tailles et couleurs variées, la suite d'expressions référentielles « *les carrés* », « *les gris* », « *les petits* », éventuellement complétée par un très ambigu « *les autres* », doit pouvoir s'interpréter comme une incrémentation de contraintes, traduisant la continuité de l'interprétation dans un domaine linguistique de plus en plus restreint.
  2. Dans la suite « *les N P_1* » puis « *les P_2* », l'interprétation de cette dernière expression en « *les N P_1 P_2* » est-elle possible? Notre but est de montrer que cette interprétation dénote une continuité dans un domaine et est donc possible. D'autre part, l'interprétation en « *les formes P_2* » est-elle possible? Nous voulons montrer ici aussi que cette interprétation correspond à un domaine correspondant à la scène visuelle complète et qu'elle est donc possible. L'interprétation en « *les N P_2* » est-elle possible? Là, nous voulons montrer que ce n'est pas le cas, cette interprétation ne correspondant à aucun

domaine construit. Enfin, l'interprétation en « *les formes* P_1 P_2 » est-elle possible? Là non plus, nous considérons que cette interprétation ne correspond à aucun domaine construit et qu'elle ne doit donc apparaître chez aucun sujet.

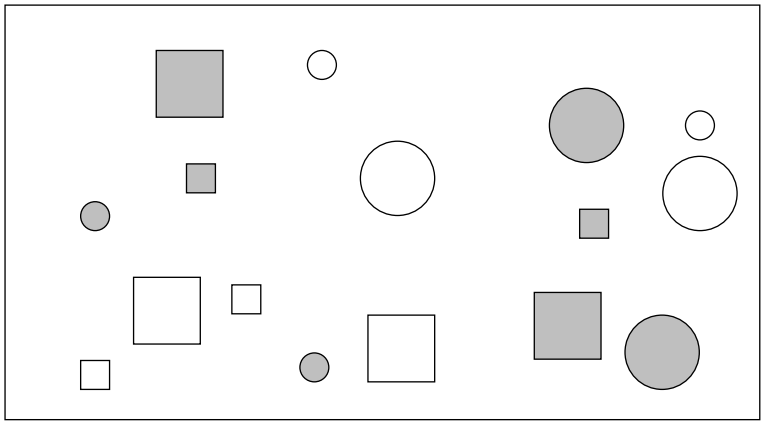


Figure 10.2 – *Scène pour tester une succession de critères de différenciation.*

- Validation de l'existence de domaines de référence à l'aide du mécanisme d'épuisement de la partition de l'association du défini avec « *autre* », quelque soit le facteur de groupement (visuel ou linguistique) :
  1. Vérification que « *autre* » épuise la partition dans un domaine visuel : dans la scène de la figure 10.3-A, nous soumettons le sujet à un pointage sur un cercle associé à l'expression référentielle « *ce triangle* », puis à « *les deux autres* ». Nous vérifions alors que l'interprétation qu'il donne de cette dernière référence correspond à l'épuisement du groupe perceptif focalisé par le pointage.
  2. Vérification que les différentes interprétations de « *autre* » en association avec une ellipse correspondent à des domaines linguistiques construits : dans la suite « *un N P* » puis « *les autres* », cette dernière expression se comprend soit comme « *les autres N P* », soit comme « *les autres formes* ». Dans la suite « *le N P* » puis « *les autres* », les interprétations possibles pour cette dernière expression sont « *les autres formes* » ou éventuellement « *les autres N* ». Dans la suite « *ce N P* » puis « *les autres* », les interprétations possibles sont « *les autres N P* » s'il y en a encore plus d'un, et éventuellement « *les autres N* » s'il n'y a plus qu'un autre N P.
- Validation de la difficulté des opérations cognitives relatives aux domaines de référence :
  1. Mesure du temps de réaction lors d'un changement d'élément dans une partition (« *le triangle blanc* » puis « *le triangle noir* »).
  2. Comparaison avec le temps de réaction lors d'un changement de partition (« *le triangle blanc* » puis « *le triangle de gauche* »).
  3. Comparaison avec le temps de réaction lors d'un changement de domaine (« *le triangle blanc* » puis « *le carré blanc* »).

Validation expérimentale de l'importance des phénomènes de saillance. Nous avons fait l'hypothèse que si la scène visuelle comportait un objet particulièrement saillant, une référence à cet

objet pouvait se faire d'une manière très simple, c'est-à-dire à l'aide de peu de propriétés discriminantes, voire d'une manière ambiguë (« *le N* » dans un environnement comportant plusieurs N, l'un d'entre eux étant saillant). Sans entrer dans les détails de situations faisant intervenir tel ou tel critère de saillance, nous nous intéressons aux protocoles permettant de tester la saillance, quelque soit la nature de celle-ci (visuelle, linguistique, ou liée à la tâche applicative). Le principe général est de spécifier des scènes comportant des objets saillants et d'autres non saillants, les calculs de saillance s'appuyant sur les scores numériques que nous avons présentés dans le chapitre 5. Du point de vue de la génération, le but est de montrer que toutes les références aux objets saillants sont facilitées, c'est-à-dire que les expressions référentielles peuvent être simplifiées sans entraver la compréhension, et qu'aucune référence à un objet non saillant ne fait l'impasse sur une propriété discriminante. Du point de vue de l'interprétation, deux protocoles peuvent être utilisés :

1. Validation de la saillance comme choix privilégié lorsqu'aucun référent n'est imposé. C'est le cas de l'interprétation du groupe nominal indéfini : dans la scène de la figure 10.3-B, « *montre-moi un cercle blanc* » devrait privilégier le cercle blanc saillant, c'est-à-dire celui se trouvant le plus à gauche, séparé des autres cercles blancs par un cercle gris.
2. Validation de la saillance comme interprétation possible de « *le N* » dans une scène comprenant plusieurs N : l'interprétation privilégiée concerne le N le plus saillant.

D'autre part, nous avons soulevé dans le chapitre 8 le problème de la confrontation des saillances. Une situation faisant intervenir simultanément la saillance visuelle et la saillance due à une manipulation récente permet de tester la prépondérance de l'une par rapport à l'autre, du moins dans certaines conditions. Le résultat est fonction des réponses obtenues à l'expression « *la chaise* » lorsque deux chaises sont présentes, l'une d'entre elles étant visuellement saillante et l'autre venant d'être manipulée. Le même principe permet de confronter la saillance visuelle et la saillance linguistique : des constatations sur les conditions de prédominance de l'une par rapport à l'autre peuvent être déduites en faisant varier les critères de saillance et en testant l'interprétation de « *la chaise* » dans une situation où une chaise est saillante visuellement et une autre linguistiquement (par exemple parce qu'elle se trouve au début de l'énoncé précédent).

Validation expérimentale de l'intérêt de l'ordonnancement des éléments dans un domaine. Nous avons fait l'hypothèse que la présence d'un ordonnancement d'objets, par exemple dû à une disposition spatiale régulière, peut entraîner l'utilisation d'expressions référentielles dénotant un rang. Sans entrer dans les détails, le protocole consiste ici à montrer au sujet une scène comportant une telle disposition d'objets (figure 10.3-C), et à tester son interprétation de « *le premier cercle* » en première mention, c'est-à-dire dans un contexte linguistique vide. Dans la scène de la figure 10.3-C, nous testerons de plus la prédominance du sens de lecture de gauche à droite : si la majorité des sujets choisissent le cercle le plus à gauche, ce sens de lecture sera privilégié dans notre modèle pour le type de scène considéré.

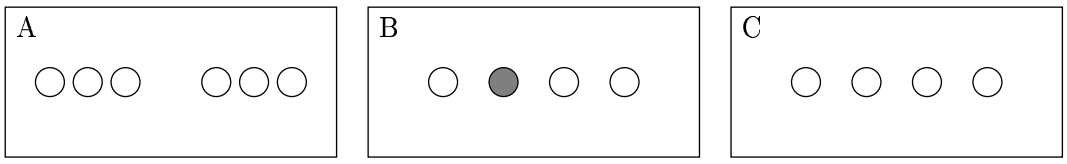


Figure 10.3 – *Scènes diverses.*

Validation expérimentale de la mesure de la pertinence. Suite aux pistes que nous avons données en § 7.3 pour une évaluation de la pertinence d'une expression référentielle donnée dans un domaine de référence donné, nous évoquons ici rapidement un protocole lié à la validation d'une telle approche. Comme pour la validation des scores de saillance, il s'agit de spécifier une situation de dialogue et une intention référentielle, puis d'évaluer la pertinence de toutes les expressions référentielles possibles. Du point de vue de la génération, cette évaluation est validée si les sujets choisissent systématiquement les expressions les plus pertinentes, voire la plus pertinente. Du point de vue de l'interprétation, aucun protocole ne permet de valider la pertinence. En effet, même une expression référentielle très peu pertinente, comme par exemple une expression incluant un grand nombre de propriétés (en particulier celles qui ne sont pas discriminantes), peut rester compréhensible et ne pas choquer le sujet au point de l'inciter à ne pas répondre.

La validation de notre modèle à travers les notions sur lesquelles il repose implique donc un très grand nombre d'expérimentations. A raison de 40 sujets par groupe de 10 situations (par exemple), cela conduit à un processus lourd, très coûteux en temps et en moyens humains. C'est pour cette raison que nous n'avons pas pu procéder à ce travail dans le cadre de cette thèse, et que nous le renvoyons dans nos perspectives de recherche, perspectives que nous allons détailler maintenant, après une synthèse des apports de notre travail.

RÉCAPITULATIF
Deux exploitations connexes sont présentées dans ce chapitre : le « retournement » du modèle pour son utilisation en génération automatique d'expressions référentielles multimodales, et sa validation et son extension pour une future modélisation de la cognition humaine sur le problème de la référence aux objets. L'idée principale de la première exploitation est que la notion de domaine de référence s'avère parfaitement compatible avec les modèles existants en génération, et que leur prise en compte peut améliorer les performances de ces modèles. L'idée principale de la deuxième exploitation est que la proposition d'un modèle de la cognition passe par la validation psychologique des notions étudiées. Le meilleur moyen de valider compte tenu de la nature de ces notions se trouve dans l'expérimentation. Les protocoles que nous spécifions pour cela sont fondés sur l'interprétation que font des sujet de certaines expressions référentielles bien choisies dans des situations maîtrisées. Après cette nécessaire étape, le modèle des domaines de référence nous semblera pouvoir constituer un modèle plausible de la référence d'un point de vue cognitif, illustrant en cela un exemple de retour de l'informatique vers la psychologie cognitive.

Conclusion et perspectives

Récapitulatif. Tout ce que nous avons exploré au cours de ce travail tend vers le même objectif, montrer que la communication multimodale dans sa globalité est ostensive et inférentielle. La perception visuelle, le geste et le langage ne donnent que des indices ostensifs, complétés ensuite par des inférences. La Théorie de la Pertinence (Sperber et Wilson 1995) montre que la communication langagière est ostensive et inférentielle. Nous ne nous contentons pas d'y ajouter la perception visuelle et le geste pour l'adapter à la communication multimodale. Nous montrons en plus comment la notion de contexte doit être repensée pour appréhender la multimodalité dans sa globalité. Nous explorons ainsi les multiples interactions entre les trois pôles cités et nous décrivons les phénomènes de dialogue dans leur complexité.

Dans le processus d'interprétation de ces phénomènes, qui inclut l'exploitation des indices ostensifs et la formulation d'inférences, nous nous focalisons sur le problème de la référence aux objets. Nous montrons la prédominance de ce problème d'un point de vue à la fois cognitif et computationnel : l'identification des référents met en jeu aussi bien les trois modalités que des facteurs cognitifs comme la mémoire, l'attention ou l'intention. Nous montrons ainsi comment tout acte référentiel passe par l'activation d'un domaine de référence. Cette étape intermédiaire entre l'énoncé et les objets de l'application permet, d'une part de décrire et d'exploiter formellement ces facteurs, d'autre part de proposer une solution au problème de la nature du contexte.

Au niveau de la perception visuelle, les indices ostensifs sont les caractéristiques visuelles des objets et les propriétés structurelles de leur disposition. Ces indices contribuent à la construction de structures cognitives que nous modélisons dans des domaines de référence visuels. Au niveau du geste ostensif, nous montrons comment un grand nombre de gestes n'indiquent pas directement les référents mais constituent au contraire des indices ostensifs. Ces indices forment une base pour des inférences aboutissant à l'identification des référents, éventuellement par l'intermédiaire d'un domaine de référence gestuel. Au niveau du langage, tout mot, toute structure syntaxique est un indice concourant à l'identification de procédures pour l'identification des référents. Nous montrons en particulier comment la détermination se trouve à la source d'inférences exploitables pour la spécification de contraintes sur le domaine de référence linguistique, domaine essentiel non seulement pour l'interprétation courante mais aussi pour celle des expressions ultérieures.

Ces domaines de référence, complétés par d'éventuels domaines liés à la tâche applicative, sont modélisés sous une même forme incluant des partitions et des caractéristiques fondamentales telles qu'un facteur de groupement, un critère de différenciation ou un critère d'ordonnancement. Nous montrons en quoi ces caractéristiques permettent d'intégrer les divers domaines et de diriger l'interprétation des références : toute expression référentielle multimodale peut se traduire en requêtes sur des domaines, requêtes incluant ou non des précisions sur le facteur de groupement, le critère de différenciation, le critère d'ordonnancement, ou encore la nature de la partie non focalisée d'une partition. L'utilisation d'un même cadre formel pour l'intégration des modalités et des facteurs cognitifs intervenant lors de l'interprétation nous semble être le principal intérêt

et le principal apport de notre travail.

L'unification de domaines sous-spécifiés (déterminés à partir de l'énoncé) avec des domaines apportés par les différentes facettes du contexte est le moyen que nous proposons pour accroître les capacités de compréhension de la machine. Nous montrons l'utilité d'un critère de pertinence pour l'évaluation des résultats de cette unification. Nous montrons également que ces domaines de référence sont plausibles d'un point de vue cognitif. Pour prouver leur validité dans le cadre d'un système de dialogue, il reste à mettre en œuvre des expérimentations que nous spécifions à défaut de les avoir menées (nous avons montré que le cadre d'une thèse s'avérait insuffisant pour procéder à ce travail dans les règles). Les quelques résultats que nous avons obtenus suite à des pré-expérimentations nous amènent à penser que notre modèle a toutes les caractéristiques d'un modèle de la communication multimodale spontanée pour le problème de la référence aux objets. Nos perspectives de recherche vont dans ce sens.

En plus des questions posées dans l'introduction et auxquelles nous venons de répondre, le déroulement de notre travail nous a amené à nous interroger sur d'autres points et à proposer des pistes pour la continuation de ce travail. Nous nous sommes ainsi intéressé à la transposition de critères de saillance pour les contextes visuels et linguistiques. Les questions méthodologiques que nous nous sommes alors posées nous ont grandement aidé dans la proposition de classifications, le résultat montrant tout l'intérêt d'une telle méthode. La question fondamentale qui ressort particulièrement du chapitre 7 a trait à la pertinence : comment peut-on exploiter et étendre la Théorie de la Pertinence au dialogue homme-machine ? Est-il possible de quantifier la pertinence ? Nous avons montré que la gestion d'un critère quantitatif de pertinence était réalisable, et nous avons proposé quelques pistes de recherche dans ce domaine très peu exploré.

Parmi les apports de notre travail, nous citerons également les caractérisations de phénomènes que nous avons proposées avant de les utiliser pour l'élaboration de notre modèle. L'étude de corpus que nous avons faite nous a alors été indispensable, et le temps passé à cette étude s'est révélé fructueux. Si l'aspect théorique reste prépondérant, nous avons également mis l'accent sur l'aspect pratique avec les scores numériques que nous avons proposés pour l'interprétation des trajectoires gestuelles, pour la formalisation des critères de la Gestalt, pour la quantification des saillances visuelle et linguistique, ou encore pour la quantification de la pertinence. Ces scores utilisables par l'ensemble des composants de notre modèle peuvent se confronter facilement et contribuent à tenir compte des liens entre les modalités. Ils constituent une solution provisoire satisfaisante, et les réflexions qu'ils ont permis d'initier aboutissent sur diverses perspectives de recherche.

Perspectives pour l'amélioration et la validation du modèle proposé. Sans entrer dans les détails comme nous l'avons fait dans le chapitre 10 pour certains aspects, nous énumérons ici nos perspectives à court et moyen terme :

- Ré-exploration de certains aspects du modèle traités un peu rapidement, en particulier la formalisation de la Gestalt et celle de la saillance. Une perspective consiste ainsi à intégrer le quatrième critère de groupage de la Gestalt, celui de la fermeture. Il serait en effet intéressant de pouvoir exploiter ce critère. Par exemple, un groupe de trois objets formant un triangle équilatéral est mieux perçu qu'un groupe de trois objets formant un triangle quelconque. La fermeture apparaît dans le corpus Magnét'Oz, implicitement et explicitement, un exemple de ce dernier cas étant « *ces objets qui forment un triangle* ». Pour la formalisation, nous pourrons nous inspirer des travaux de Zeevat et Wang (1996) qui partent de la reconnaissance de points, traits et courbes, pour reconnaître à l'aide d'inférences des triangles ou d'autres configurations géométriques. En ce qui concerne la

formalisation de la saillance, une perspective consiste à revenir à la psychologie cognitive (et à la psycholinguistique pour la saillance linguistique), pour intégrer nos propositions dans une théorie plausible d'un point de vue psychologique.

- Elargissement théorique du modèle aux phénomènes d'indexicalité tels qu'ils sont étudiés dans les travaux en philosophie du langage, par exemple dans ceux de Kaplan (1989). Cette perspective inclut une étude plus fine du démonstratif et de ses rapports avec les demonstrata, la démonstration et la deixis. Le but est de formaliser l'indexicalité, peut-être sous une forme logique adaptée.

- Elargissement théorique du modèle aux références spatiales et à tout ce qui concerne les images mentales liées au contexte visuel. Il s'agit par exemple de tenir compte des caractéristiques des images mentales telles qu'analysées par Denis (1989), ou encore d'étudier les rapports entre cible et site à la suite du travail de Vandeloise (1986). La formalisation pourra alors s'inspirer des travaux de Vieu (1991).

- Elargissement théorique du modèle aux événements, avec la gestion de domaines de référence événementiels et d'une saillance événementielle. En effet, les domaines sont construits essentiellement à partir des propriétés des objets, alors que les informations concernant les événements auxquels les objets participent devraient être prises en compte. Ce type d'information permettrait par exemple de traiter les coréférences entre indéfinis, ou encore de formuler des contraintes supplémentaires sur l'interprétation des pronoms. Par ailleurs, ces connaissances sont nécessaires afin de résoudre les anaphores associatives à des participants d'une éventualité (« *un meurtre* » puis « *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 les domaines de référence constituent un cadre de modélisation adapté à une telle extension.

- Elargissement théorique du modèle aux spécificités pragmatiques du dialogue finalisé : actes de langage (étudiés également par Grisvard dans le cadre des domaines de référence), gestion du dialogue (stratégies pertinentes de réponse et de réparation), nature et étendue de l'intervention de la tâche applicative.

- Comparaisons théoriques poussées du modèle avec des formalismes existants et en évolution constante, par exemple les logiques de description ou la DRT.

- Test de la prédictibilité du modèle sur des phénomènes impossibles, sur des accidents référentiels ou des incompréhensions tirés de situations de dialogue réelles. Il s'agit ainsi de vérifier que le modèle refuse de traiter ces phénomènes ou qu'il les trouve très peu pertinents. Le corpus Magnét'Oz s'est révélé insuffisant pour le faire, contenant trop peu de telles situations du fait de la tâche applicative fortement contrainte.

- Continuation de la boucle caractérisée par une simulation selon le principe du magicien d'Oz, puis une modélisation à partir des données recueillies, puis à nouveau une simulation dirigée par cette modélisation, puis une nouvelle modélisation intégrant les nouvelles données, etc. Une perspective consiste ainsi à monter une simulation similaire à celle de Magnét'Oz, mais en spécifiant la tâche de manière à ce qu'elle mette l'accent sur certains aspects de notre modèle, en particulier l'intervention de la mémoire lors de la référence aux objets. Le but est d'obtenir un corpus qui corresponde à ce qui intéresse notre modèle, pour d'une part en valider certains aspects et d'autre part tenter de détecter des phénomènes

que nous n'avons pas encore pris en compte.

- Validation des aspects non encore validés du modèle, par la spécification de protocoles expérimentaux en suivant les règles de la psycholinguistique. En plus des expériences présentées dans le dernier chapitre, il serait intéressant de monter des expériences plus en rapport avec le geste et la multimodalité. L'utilisation d'un *eye-tracker* pour observer les éventuels phénomènes d'anticipation du regard par rapport au geste, ainsi que pour étudier le parcours du regard dans l'image, permettrait de valider certains aspects du modèle.
- Instanciation du modèle pour des applications particulières variées. L'aménagement d'intérieur ou l'interrogation de bases de données multimedia ne sont que deux exemples de telles applications. Il serait également utile de tester le modèle avec d'autres applications du même type, voire avec des applications sans support visuel. Une façon intéressante de valider le modèle consisterait par exemple à tester les domaines de référence visuels dans une situation où les interlocuteurs parlent d'une scène visuelle qu'ils ne voient pas ou dont ils ne partagent pas la perception.

Perspectives pour un nouveau modèle. Nous présentons ici des perspectives de recherche à plus long terme. Compte tenu du fait que nous n'avons pas pris en compte un grand nombre de phénomènes tels que les actes de langage ou les fonctions dialectiques pouvant intervenir dans le dialogue, ainsi que les gestes communicatifs qui peuvent être produits dans leur diversité, il nous semble qu'une refonte du modèle s'impose pour la modélisation de l'interaction autrement qu'avec un écran tactile. Cette refonte n'est cependant pas un retour à zéro : non seulement nous garderons la méthodologie et les caractérisations de phénomènes acquises, mais également le même appui sur la Théorie de la Pertinence. Nos perspectives de recherche suivent en cela la voie décrite par Feyereisen :

> « *Les recherches futures sur la compréhension gestuelle devraient s'orienter dans deux directions complémentaires, à partir du présupposé que plusieurs opérations interviennent entre la présentation d'une information et la décision qu'elle entraîne. Ces directions correspondent à l'étude des deux mécanismes fondamentaux qui sous-tendent la communication : l'ostension et l'inférence* (Sperber & Wilson 1995). *D'un côté, il existe des traitements spécifiques, propres aux différentes catégories de signaux disponibles : reconnaissance de la parole, reconnaissance visuelle d'un objet dans un dessin, analyse d'un geste de désignation, d'un geste descriptif ou d'un mime d'action. Ces traitements ont pour effet de rendre plus ou moins manifeste un élément de l'environnement physique ou social. A cet égard, il serait nécessaire d'examiner de manière plus approfondie dans quelle mesure la présentation d'un geste attire l'attention de l'interlocuteur. D'un autre côté, interviennent des mécanismes centraux pour mettre en relation des informations lexicales, visuelles, spatiales et motrices. Celles-ci activent en outre des connaissances en mémoire sémantique et épisodique. Des éléments du contexte influencent également l'élaboration et le test de l'hypothèse émise à propos de l'intention du locuteur. Dans ce processus d'inférence, les informations gestuelles ne sont pas les seules utilisées ; elles ne sont pas pour autant totalement négligées.* » (Feyereisen 1997)

D'une manière générale, le travail présenté dans cette thèse constitue une première étape dans la modélisation de la communication multimodale, dont les deux objectifs fondamentaux sont premièrement l'élaboration d'un modèle cognitif exhaustif de la communication, et deuxièmement

la conception d'un système de dialogue homme-machine dont les capacités de compréhension seraient comparables à celles d'un humain. Nous présentons maintenant le problème sous l'angle des informations sémantiques communiquées.

Les informations mises en jeu au cours du dialogue peuvent être distinguées en informations nouvelles et informations connues et partagées par les deux interlocuteurs. Pour avoir une utilité communicative, un énoncé apporte à l'interlocuteur des informations qu'il ne possède pas déjà. Ces informations nouvelles peuvent ne pas être toutes incluses dans l'énoncé : une partie implicite complète souvent la partie explicitée et justifie ainsi les inférences. S'il veut maintenir une certaine cohérence dans le dialogue, cet énoncé s'appuie sur des informations partagées par les deux interlocuteurs, certaines de ces informations se déduisant de l'espace visuel partagé, d'autres se déduisant des échanges précédents dans le dialogue, d'autres encore provenant de connaissances générales sur le monde et sur l'utilisation du langage et du geste pour communiquer.

Une voie dans l'analyse des informations nouvelles est ouverte par la Théorie de la Pertinence avec la notion d'effets contextuels dénotant l'intérêt de ces informations compte tenu de la situation. Cet intérêt est confronté à l'effort nécessaire au traitement des informations, le ratio effets sur effort constituant l'indice de pertinence. Les pistes que nous avons données permettent de mieux aborder ces notions dans un contexte multimodal, plus proche de la communication humaine que l'interaction purement langagière telle qu'apréhendée par Sperber et Wilson (1995). L'identification de la partie implicite d'un énoncé constitue une voie de recherche à part entière. Si nous avons donné dans ce travail des pistes pour l'identification de l'implicite dans une action de référence multimodale, reste à élargir la problématique aux autres phénomènes de communication, et à intégrer ainsi les implicitations, les sous-entendus, les effets interprétatifs de l'énoncé complet. Il est vraisemblable que, comme nous l'avons dit à propos d'adaptation de la Théorie de la Pertinence, toute la notion de contexte soit à repenser pour appréhender ces aspects dans leur globalité. Nous espérons que ce que nous avons tiré de ce travail nous permettra d'appréhender efficacement les étapes de cette refonte.

Les informations partagées par les interlocuteurs ont été abordées dans ce travail à travers la notion de domaine de référence, sous-ensemble contextuel se déduisant de l'espace visuel partagé, de l'historique du dialogue, ou encore de contraintes liées à la tâche applicative. Un prolongement en serait une modélisation plus formelle de la notion de coopération dans le dialogue. Parmi les informations partagées se trouvent également des connaissances générales sur le fonctionnement du langage et du geste, sur les grands principes de leur utilisation coopérative. Cet aspect métacommunicatif n'est pas négligeable : ce n'est que parce que l'on sait que telle expression linguistique est liée à tel type d'implicitations que l'on comprend des énoncés *a priori* ambigus. Sperber et Wilson ont exploré cette voie avec la présomption de pertinence qu'a tout locuteur à propos de l'énoncé qu'il produit. Cette notion nous semble pouvoir être étudiée de manière à en déduire des principes opérationnels dans la communication homme-machine.

La modélisation de ces deux grands types d'information, avec des préoccupations multimodales et pas seulement linguistiques, nous apparaît nécessaire pour l'élaboration de systèmes de dialogue homme-machine capables d'exploiter les capacités communicatives de l'utilisateur humain. Elle nous semble de plus capable d'apporter à la psychologie cognitive un point de vue non dénué d'intérêt, à la fois large au niveau des phénomènes et formel au niveau de leur traitement.

Bibliographie

ALQUIER, L. (1998), *Analyse et représentation de scènes complexes par groupement perceptuel*, Thèse de doctorat, Université de Montpellier 2.

ALSHAWI, H. (1987), *Memory and Context for Language Interpretation*, Cambridge University Press (Studies in Natural Language Processing).

ARIEL, M. (1988), Referring and Accessibility, *Journal of Linguistics* 24, pp. 65–87.

ASHER, N. (1993), *Reference to Abstract Objects in Discourse*, Kluwer, Dordrecht.

BARTHES, R. (1964), Rhétorique de l'image, *Communications* 4, Seuil, Paris, pp. 40–51.

BATICLE, Y. (1985), *Clés et codes de l'image*, Magnard Université.

BEAVER, D. (2003), The Optimization of Discourse, *Linguistics and Philosophy* ?, pp. .

BELAÏD, A. & BELAÏD, Y. (1992), *Reconnaissance des formes : méthodes et applications*, Inter-Editions.

BELLALEM, N. (1995), *Etude du mode de désignation dans un dialogue homme-machine finalisé à forte composante langagière : analyse structurelle et interprétation*, Thèse de doctorat, Université Henri Poincaré de Nancy.

BELLALEM, N. & ROMARY, L. (1995), Reference Interpretation in a Multimodal Environment Combining Speech and Gesture, In: *Proceedings of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces*, Edinburgh.

BELLENGER, L. (1979), *L'expression orale*, PUF (collection Que sais-je?), Paris.

BELLIK, Y. (1995), *Interfaces multimodales : concepts, modèles et architectures*, Thèse de doctorat, Université de Paris 11, Orsay.

BEUN, R.-J. & CREMERS, A.H.M. (1998), Object Reference in a Shared Domain of Conversation, *Pragmatics and Cognition* 6(1/2), pp. 121–152.

BILANGE, E. (1992), *Dialogue personne-machine : modélisation et réalisation informatique*, Hermès, Paris.

BLALOCK, H.M. (1979), *Social Statistics*, McGraw-Hill, New York.

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

BONNET, C., GHIGLIONE, R. & RICHARD, J.-F. (1989), *Traité de psychologie cognitive* (vol. 1 : *perception, action, langage*), Dunod, Paris.

BONNET, C., GHIGLIONE, R. & RICHARD, J.-F. (1990), *Traité de psychologie cognitive* (vol. 2 : *le traitement de l'information symbolique*), Dunod, Paris.

BORDRON, J.-F. (1991), Les objets en partie (esquisse d'ontologie matérielle), *Langages* 103, pp. 51–65.

BOS, E., HULS, C. et CLAASSEN, W. (1994), EDWARD: Full Integration of Language and Action in a Multimodal User Interface, *International Journal of Human-Computer Studies* 40, pp. 473–495.

Boucart, M. (1996), *La reconnaissance des objets*, Presses Universitaires de Grenoble (collection La Psychologie en Plus).

Braffort, A. (1996), *Reconnaissance et compréhension de gestes, application à la langue des signes*, Thèse de doctorat, Université de Paris 11, Orsay.

Briffault, X. (1992), *Modélisation informatique de l'expression de la localisation en langage naturel*, Thèse de doctorat, Université de Paris 6.

Brison, E. (1997), *Stratégies de compréhension dans l'interaction multimodale*, Thèse de doctorat, Université Paul Sabatier de Toulouse.

Cadoz, C. (1994), Le geste canal de communication homme-machine. La communication instrumentale, *Techniques et Sciences Informatiques* 13(1), pp. 31–61.

Carbonell, N., Valot, C., Mignot, C. & Dauchy, P. (1997), Etude empirique de l'usage du geste et de la parole en situation de communication homme-machine, *Le travail humain* 60(2), pp. 155–184.

Caron, J. (1989), *Précis de psycholinguistique*, PUF, Paris.

Carré, R., Dégremont, J.-F., Gross, M., Pierrel, J.-M. & Sabah, G. (1991), *Langage humain et machine*, Presses du CNRS, Paris.

Cassell, J., McNeill, D. et McCullough, K.-E. (1999), Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information, *Pragmatics and Cognition* 7(1), pp. 1–33.

Chen, L., Harper, M. & Quek, F. (2002), Gesture Patterns during Speech Repairs, In: *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces* (icmi'02, Pittsburgh, PA), IEEE Computer Society, Los Alamitos, CA, pp. 155–160.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Clark, H.H., Schreuder, R. et Buttrick, S. (1983), Common Ground and the Understanding of Demonstrative Reference, *Journal of Verbal Learning and Verbal Behavior* 22, pp. 245–258.

Clark, H.H. & Wilkes-Gibbs, D. (1986), Referring as a Collaborative Process, *Cognition* 22, pp. 1–39.

Cocula, B. & Peyroutet, C. (1989), *Sémantique de l'image. Pour une approche méthodique des messages visuels*, Delagrave.

Cohen, P.R. (1984), The Pragmatics of Referring and the Modality of Communication, *American Journal of Computational Linguistics* 10, pp. 97–146.

Cohen, P.R. (1992), The Role of Natural Language in a Multimodal Interface, In: *Proceedings of User Interface Software and Technology Conference* (uist'92), Academic Press, Monterey, CA, pp. 143–149.

Corazza, E. (1995), *Référence, Contexte et Attitudes*, Bellarmin/Vrin, Montréal/Paris.

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

Corblin, F. (2002), *Représentation du discours et sémantique formelle*, PUF (collection Linguistique nouvelle), Paris.

Cosnier, J. & Vaysse, J. (1997), Sémiotique des gestes communicatifs, *Nouveaux actes sémiotiques* 52-53-54, pp. 7–28.

Coutaz, J. & Caelen, J. (1991), A Taxonomy for Multimedia and Multimodal User Interfaces, In: *Proceedings of the First ERCIM Workshop on Multimodal Human-Computer Interaction*, Lisbon.

Cremers, A.H.M. (1996), *Reference to Objects. An Empirically Based Study of Task-Oriented Dialogues*, Ph.D. Thesis, Technische Universiteit Eindhoven.

CYRULNIK, B. (1995), *La naissance du sens*, Hachette Littératures, Paris.

DALE, R. (1992), *Generating Referring Expressions*, MIT Press, Cambridge, MA.

DALE, R. & REITER, E. (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions, *Cognitive Science* 19, pp. 233–263.

DANON-BOILEAU, L. (1987), *Enonciation et référence*, Ophrys, Paris.

DEKKER, P. (1998), Speaker's Reference, Descriptions and Information Structure, *Journal of Semantics* 15(4), pp. 305–334.

DENIS, M. (1989), *Image et cognition*, PUF, Paris.

DONNELLAN, K. (1966), Reference and Definite Descriptions, *Philosophical Review* 75, pp. 281–304.

DUCROT, O. (1972), *Dire et ne pas dire*, Hermann, Paris.

ECO, U. (1972), *La structure absente*, Mercure de France, Paris.

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

EDMONDS, P.G. (1994), Collaboration on Reference to Objects that are not Mutually Known, In: *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94, Kyoto), pp. 1118–1122.

EVANS, G. (1985), *The Varieties of Reference*, Oxford University Press, Oxford.

FAUCONNIER, G. (1984), *Espaces mentaux*, Editions de Minuit, Paris.

FELDMAN, J. (1999), The Role of Objects in Perceptual Grouping, *Acta Psychologica* 102, pp. 137–163.

FEYEREISEN, P. (1997), La compréhension des gestes référentiels, *Nouveaux actes sémiotiques* 52-53-54, pp. 29–48.

FISCHER, M. (1998), *Automatic Generation of Spatial Configurations in User Interfaces*, Ph.D. Thesis, University of Brighton (ITRI).

FITTS, P.M. (1954), The Information Capacity of the Human Motor System in Controlling Amplitude of Movement, *Journal of Experimental Psychology* 47, pp. 381–391.

FLORÈS, C. (1972), *La mémoire*, PUF (collection Que sais-je?), Paris.

FODOR, J. (1986), *La modularité de l'esprit*, Minuit, Paris.

FORD, W. & OLSON, D. (1975), The Elaboration of the Noun Phrase in Children's Description of Objects, *Journal of Experimental Child Psychology* 19, pp. 371–382.

FRANCÈS, R. (1963), *La perception*, PUF (collection Que sais-je?), Paris.

FREGE, G. (1892), Über Sinn und Bedeutung, *Zeitschrift für Philosophie und philosophische Kritik* 100, pp. 25–50.

FUCHS, C., DANLOS, L., LACHERET-DUJOUR, A., LUZZATI, L. & VICTORRI, B. (1993), *Linguistique et Traitements Automatiques des Langues*, Hachette.

GAIFFE, B. (1992), *Référence et dialogue homme-machine : vers un modèle adapté au multimodal*, Thèse de doctorat, Université Henri Poincaré de Nancy.

GARROD, S. & SANFORD, A.J. (1988), Thematic Subjecthood and Cognitive Constraints on Discourse Structure, *Journal of Pragmatics* 12, pp. 519–534.

GIBSON, J.J. (1979), *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston.

GLENBERG, A.M. & KRULEY, P. (1992), Pictures and Anaphora: Evidence for Independent Processes, *Memory and Cognition* 20(5), pp. 461–471.

GOODMAN, B.A. (1986), Reference Identification and Reference Identification Failures, *Computational Linguistics* 12, pp. 273–305.

Grice, H.P. (1975), Logic and Conversation, In: Cole, P. & Morgan, J. (Eds.), *Syntax and Semantics*, vol. 3, Academic Press, pp. 41–58.

Grisvard, O. (2000), *Modélisation et gestion du dialogue oral homme-machine de commande*, Thèse de doctorat, Université Henri Poincaré de Nancy.

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

Grosz, B.J., Joshi, A.K. et Weinstein, S. (1995), Centering: A Framework for Modelling the Local Coherence of Discourse, *Computational Linguistics* 21(2), pp. 203–225.

Guillaume, P. (1979), *La psychologie de la forme*, Flammarion (collection Champs), Paris.

Gundel, J.K., Hedberg, N. & Zacharski, R. (1993), Cognitive Status and the Form of Referring Expressions in Discourse, *Language* 69(2), pp. 274–307.

Habert, B., Nazarenko, A. & Salem, A. (1997), *Les Linguistiques de corpus*, Armand Colin, Paris.

Hauptmann, A.G. & McAvinney, P. (1993), Gestures with Speech for Graphic Manipulation, *International Journal of Man-Machine Studies* 38(2), pp. 231–249.

Heeman, P.A. & Hirst, G. (1995), Collaborating on Referring Expressions, *Computational Linguistics* 21(3), pp. 351–382.

Heilig, M. (1992), El Cine del Futuro: The Cinema of the Future, *Presence: Teleoperators and Virtual Environments* 1(3), pp. 279–294.

Huls, C., Claassen, W. et Bos, E. (1995), Automatic Referent Resolution of Deictic and Anaphoric Expressions, *Computational Linguistics* 21(1), pp. 59–79.

Isard, S. (1975), Changing the Context, In: Keenan, E.L. (Ed.) *Formal Semantics of Natural Language*, Cambridge University Press, London & New York, pp. 287–296.

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

Jakobson, R. (1963) *Essais de linguistique générale, tome 1*, Minuit, Paris.

Johnston, M., Cohen, P.R., McGee, D.R., Oviatt, S.L., Pittman, J.A. & Smith, I. (1997), Unification-based Multimodal Integration, In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (acl'97), pp. 281–288.

Joly, M. (1993), *Introduction à l'analyse de l'image*, Nathan Université, Paris.

Jönsson, A. & Dählback, N. (1988), Talking to a Computer is not like Talking to your Best Friend, In: *Proceedings of the Scandinavian Conference on Artificial Intelligence*, Tromsø.

Jørgensen, S.W. (2000), Computational Reference. An Investigation, Development and Implementation of Kronfeld's Theory of Reference, Ph.D. Thesis, Copenhague Business School.

Kamp, H. & Reyle, U. (1993), *From Discourse to Logic*, Kluwer, Dordrecht.

Kandinsky, W. (1979), *Point and Line to Plane*, Dover Publications Inc., New York.

Kaplan, D. (1989), Demonstratives, In: Almog, J., Perry, J. & Wettstein, H. (Eds.), *Themes from Kaplan*, Oxford University Press, New York.

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

Karttunen, L. (1976), Discourse referents, In: McCawley, J. (Ed.), *Syntax and Semantics*, vol. 7, Academic Press, New York, pp. 363–385.

Kendon, A. (1980), Gesticulation and Speech: Two Aspects of the Process of Utterance, In: Key, M.R. (Ed.) *The Relation between Verbal and Nonverbal Communication*, Mouton, La Hague, pp. 207–227.

KENDON, A. (1994), Do Gestures Communicate? A Review, *Research on Language and Social Interaction* 27, pp. 175–200.

KENNEDY, A., WILKES, A., ELDER, L. & MURRAY, W. (1988), Dialogue with machines, *Cognition* 30, pp. 73–105.

KERBRAT-ORECCHIONI, C. (1996), *La conversation*, Seuil (collection Mémo), Paris.

KESSLER, K., DUWE, I. & STROHNER, H. (1996), *Sprachliche Objektidentifikation in ambigen Situationen: Empirische Befunde*, SFB 360 Situierte künstliche Kommunikation, Report 96/1, Universität Bielefeld.

KIEVIT, L., PIWEK, P., BEUN, R.J. & BUNT, H. (2001), Multimodal Cooperative Resolution of Referential Expressions in the DENK System, In: BUNT, H. & BEUN, R.J. (Eds.), *Cooperative Multimodal Communication*, Springer, Berlin & Heidelberg, pp. 197–214.

KITA, S. (2000), How Representational Gestures Help Speaking, In: MCNEILL, D. (Ed.), *Language and Gesture*, Cambridge University Press, New York.

KLEIBER, G. (1986), Pour une explication du paradoxe de la reprise immédiate, *Langue française* 72, pp. 54–79.

KLEIBER, G. (1991), Anaphore-deixis: où en sommes-nous?, *L'information grammaticale* 51, pp. 3–16.

KLEIBER, G. (1994), *Anaphores et pronoms*, Duculot (collection Champs linguistiques), Louvain-La-Neuve.

KLINKENBERG, J.-M. (1996), *Précis de sémiotique générale*, De Boeck Université, Bruxelles.

KLIPPLE, E. & GURNEY, J. (2002), Some Observations on Deixis to Properties, In: VAN DEEMTER, K. & KIBBLE, R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, pp. 355–390.

KOBSA, A., ALLGAYER, J., REDDIG, C., REITHINGER, N., SCHMAUKS D., HARBUSCH, K. & WAHLSTER, W. (1986), Combining Deictic Gestures and Natural Language for Referent Identification, In: *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, West Germany, pp. 356–361.

KÖHLER, W. (1947), *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*, Liveright Publishing Corporation, New York.

KRAHMER, E. & PIWEK, P. (2000), Varieties of Anaphora: Introduction, In: KRAHMER, E. & PIWEK, P. (Eds.), *Varieties of Anaphora*, Reader ESSLLI 2000, Birmingham, pp. 1–15.

KRAHMER, E. & THEUNE, M. (2002), Efficient Context-Sensitive Generation of Referring Expressions, In: VAN DEEMTER, K. & KIBBLE, R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, pp. 223–264.

KUBOVY, M. & WAGEMANS, J. (1995), Grouping by Proximity and Multistability in Dot Lattices: A Quantitative Gestalt Theory, *Psychological Science* 6(4), pp. 225–234.

KUBOVY, M., HOLCOMBE, A.O. & WAGEMANS, J. (1998), On the Lawfulness of Grouping by Proximity, *Cognitive Psychology* 35(1), pp. 71–98.

LAMBRECHT, K. (1996), On the Formal and Functional Relationship between Topics and Vocatives. Evidence from French, In: GOLDBERG, A.E. (Ed.), *Conceptual Structure, Discourse, and Language*, CSLI Publications, Stanford, CA, pp. 267–288.

LANDRAGIN, F. (1998), *Interaction multimodale dans un environnement virtuel*, Mémoire présenté en vue d'obtenir le diplôme d'ingénieur IIE, Thomson-CSF & Institut d'Informatique d'Entreprise.

LANDRAGIN, F. (1999), *Analyse de l'articulation entre parole et geste dans un corpus multimodal*, Mémoire de DEA, LORIA & Université de Marne-La-Vallée.

LANDRAGIN, F., GAIFFE, B., BELLALEM, N. et ROMARY, L. (2000), Fusion de contraintes pour la synchronisation des modalités et pour la résolution des références dans un énoncé multimodal, dans : *Colloque sur les Interfaces Multimodales (10 ans de multimodalité)*, Grenoble, pp. 17–20.

LANDRAGIN, F., BELLALEM, N. & ROMARY, L. (2001a), Compréhension automatique du geste et de la parole spontanés en communication homme-machine : apport de la théorie de la pertinence, dans : *Oralité et gestualité. Interactions et comportements multimodaux dans la communication* (*actes du colloque* ORAGE 2001, Aix-en-Provence), L'Harmattan, pp. 390–393.

LANDRAGIN, F., BELLALEM, N. & ROMARY, L. (2001b), Visual Salience and Perceptual Grouping in Multimodal Interactivity, In: *Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 151–155.

LANDRAGIN, F. (2002), The Role of Gesture in Multimodal Referring Actions, In: *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces* (ICMI'02, Pittsburgh, PA), IEEE Computer Society, Los Alamitos, CA, pp. 173–178.

LANDRAGIN, F., BELLALEM, N. & ROMARY, L. (2002a), Referring to Objects with Spoken and Haptic Modalities, In: *Proceedings of the fourth IEEE International Conference on Multimodal Interfaces* (ICMI'02, Pittsburgh, PA), IEEE Computer Society, Los Alamitos, CA, pp. 99–104.

LANDRAGIN, F., DE ANGELI, A., WOLFF, F., LOPEZ, P. & ROMARY, L. (2002b), 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*, CSLI Publications, Stanford, CA, pp. 391–409.

LANDRAGIN, F., SALMON-ALT, S. & ROMARY, L. (2002c), Ancrage référentiel en situation de dialogue, *Traitement Automatique des Langues* 43(2), pp. 99–129.

LANDRAGIN, F. (2003), Clues for the Identification of Implicit in Multimodal Referring Actions, In: *Proceedings of the tenth International Conference on Human-Computer Interaction* (HCI International 2003, Heraklion, Crete, Greece), Lawrence Erlbaum Associates, Mahwah, NJ.

LAPPIN, S. & LEASS, H.J. (1994), A Syntactically Based Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics* 20(4), pp. 535–561.

LASSWELL, H.D. (1948), The Structure and Function of Communication in Society, In: BRYSON, L. (Ed.), *The Communication of Ideas: A Series of Addresses*, Institute for Religious and Social Studies, New York, pp. 37–51.

LESTER, J.C., VOERMAN, J.L., TOWNS, S.G. & CALLAWAY, C.B. (1999), Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents, *Applied Artificial Intelligence* 13(4-5), pp. 383–414.

LINSKY, L. (1967), *Referring*, Routledge & Kegan Paul Humanities Press, New York.

LOFTUS, G.R. & MACKWORTH, N.H. (1978), Cognitive Determinants of Fixation Location during Picture Viewing, *Journal of Experimental Psychology: Human Perception and performance* 4, pp. 565–572.

LOPEZ, P. (1999), *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*, Thèse de doctorat, Université Henri Poincaré de Nancy.

MARSLEN-WILSON, W.D., TYLER, L.K. & KOSTER, J. (1993), Integrative Processes in Utterance Resolution, *Journal of Memory and Language* 32, pp. 647–666.

MATHIEU, F.-A. (1997), *Prise en compte de contraintes pragmatiques pour guider un système de reconnaissance de la parole : le système COMPPA*, Thèse de doctorat, Université Henri Poincaré de Nancy.

MCGINN, C. (1981), The Mechanism of Reference, *Synthese* 49, pp. 157–186.

MCNAMARA, T.P. (1986), Mental Representation of Spatial Relations, *Cognitive Psychology* 18(1), pp. 87–118.

MCNEILL, D. (1992), *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, Chicago.

MIGNOT, C. (1995), *Usage de la parole et du geste dans les interfaces multimodales, étude expérimentale et modélisation*, Thèse de doctorat, Université Henri Poincaré de Nancy.

MILLER, G.A. (1956), The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information, *Psychological Review* 63, pp. 81–97.

MILNER, J.-C. (1982), *Ordres et raisons de langue*, Seuil, Paris.

MILNER, J.-C. (1989), *Introduction à une science du langage*, Seuil (collection Travaux linguistiques), Paris.

MOESCHLER, J. & REBOUL, A. (1994), *Dictionnaire encyclopédique de pragmatique*, Seuil, Paris.

MOREL, M.-A. & DANON-BOILEAU, L. (Eds.) (1990), *La deixis, colloque en Sorbonne 8–9 juin 1990*, PUF, Paris.

MORRIS, C.W. (1974), Fondements de la théorie des signes, *Langages* 35, pp. 15–21.

MORTON, J. (1982), Disintegrating the Lexicon: An Information Processing Approach, In: MEHLER, J., WALKER, E.C.T. & GARRETT, M.F. (Eds.), *On Mental Representation*, Erlbaum, Hillsdale, NJ, pp. 89–109.

MOULTON, J. & ROBERTS, L.D. (1994), An AI Module for Reference Based on Perception, In: *Proceedings of the AAAI Workshop on Integration of Natural Language and Vision Processing*, Seattle.

NEAL, J.G. & SHAPIRO, S.C. (1991), Intelligent Multimedia Interface Technology, In: SULLIVAN, J.W. & TYLER, S.W. (Eds.), *Intelligent User Interfaces*, ACM Press, New York, pp. 11–43.

NIGAY, L. (1994), *Conception et modélisation logicielles des systèmes interactifs : application aux interfaces multimodales*, Thèse de doctorat, Université Joseph Fourier de Grenoble.

NORMAND, V., PERNEL, D. & BACCONNET, B. (1997), Speech-based Multimodal Interaction in Virtual Environments: Research at the Thomson-CSF Corporate Research Laboratories, *Presence: Teleoperators and Virtual Environments* 6(6), pp. 687–700.

OLÉRON, P. (1974), *L'intelligence*, PUF (collection Que sais-je?), Paris.

OLSON, D.R. (1970), Language and Thought: Aspects of a Cognitive Theory of Semantics, *Psychological Review* 77, pp. 257–273.

OSGOOD, C.E. & BOCK, J.K. (1977), Salience and Sentencing: Some Production Principles, In: ROSENBERG, S. (Ed.), *Sentence Production: Developments in Research and Theory*, Erlbaum, Hillsdale, NJ, pp. 89–140.

OVIATT, S.L. (1997), Multimodal Interactive Maps: Designing for Human Performance, *Human-Computer Interaction* 12, pp. 93–129.

OVIATT, S.L., DE ANGELI, A. et KUHN, K. (1997), Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction, In: *Proceedings of Conference on Human Factors in Computing Systems* (CHI'97), ACM Press, New York. Also in ANDRÉ, E. (Ed.), *Proceedings of the ACL Workshop on Referring Phenomena in a Multimedia Context and their Computational Treatment* (ACL/EACL-97), Madrid.

OVIATT, S.L. (1999), Ten Myths of Multimodal Interaction, *Communications of the ACM* 42(11), pp. 74–81.

OZKAN, N. (1994), *Vers un modèle dynamique du dialogue : analyse de dialogues finalisés dans une perspective communicationnelle*, Thèse de doctorat, Institut National Polytechnique de Grenoble.

PEARSON, J., POESIO, M. & STEVENSON, R. (2001), The Effects of Animacy, Thematic Role and Surface Position on the Focusing of Entities in Discourse, In: *Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processing*, Edinburgh.

PEIRCE, C.S. (1960), *Collected Papers*, Harvard University Press, Cambridge.

PÉNINOU, G. (1970), Physique et métaphysique de l'image publicitaire, *Communications* 15, Seuil, Paris, pp. 96–109.

PIERREL, J.-M. (1987), *Dialogue oral homme-machine*, Hermès, Paris.

POESIO, M. (2000), Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms: Issues and Preliminary Results, In: *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC-2000), Athens.

POPESCU-BELIS, A. (1999), *Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation*, Thèse de doctorat, Université de Paris 11, Orsay.

POUTEAU, X. (1995), *Dialogue de commande multimodal en milieu opérationnel : une communication naturelle pour l'utilisateur ?*, Thèse de doctorat, Université Henri Poincaré de Nancy.

QUINE, W.v.O. (1971), The Inscrutability of Reference, In: STEINBERG, D.D. & JAKOBOVITS, L.A. (Eds.), *Semantics*, Cambridge University Press, pp. 142–154.

RASTIER, F. (1991), *Sémantique et recherches cognitives*, PUF, Paris.

REBOUL, A., BALKANSKI, C., BRIFFAULT, X., GAIFFE, B., POPESCU-BELIS, A., ROBBA, I., ROMARY, L., & SABAH, G. (1997), *Le projet CERVICAL : représentations mentales, référence aux objets et aux événements*, rapport de recherche LORIA-LIMSI.

REBOUL, A. (1998), A Relevance Theoretic Approach to Reference, In: *Proceedings of Relevance Theory Workshop*, Luton.

REBOUL, A. & MOESCHLER, J. (1998a), *La pragmatique aujourd'hui. Une nouvelle science de la communication*, Seuil, Paris.

REBOUL, A. & MOESCHLER, J. (1998b), *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*, Armand Colin, Paris.

RÉCANATI, F. (1993), *Direct Reference: From Language to Thought*, Blackwell, Oxford.

REIMER, M. (1991), Demonstratives, Demonstrations, and Demonstrata, *Philosophical Studies* 63, pp. 187–202.

REITER, E. & DALE, R. (1992), A Fast Algorithm for the Generation of Referring Expressions, In: *Proceedings of the 14th International Conference on Computational Linguistics* (COLING'92), Nantes, pp. 232–238.

REITER, E. & DALE, R. (1997), Building Applied Natural-Language Generation Systems, *Journal of Natural-Language Engineering* 3, pp. 57–87.

REITHINGER, N. (1987), Generating Referring Expressions and Pointing Gestures, In: KEMPEN, G. (Ed.) *Natural language generation: New Results in Artificial Intelligence, Psychology and Linguistics*, Nijhoff, Dordrecht, pp. 71–81.

RIST, T., ANDRÉ, E. & MÜLLER, J. (1997), Adding Animated Presentation Agents to the Interface, In: *Proceedings of the Second International Conference on Intelligent User Interfaces* (IUI'97), Orlando, pp. 79–86.

ROBBE, S. (1998), *Etude ergonomique de contraintes d'expression orales et gestuelles dans un environnement multimodal d'interaction homme-machine*, Thèse de doctorat, Université Henri Poincaré de Nancy.

ROBERTS, C. (2002), Demonstratives as Definites, In: VAN DEEMTER, K. & KIBBLE, R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, pp. 89–136.

ROCK, I. & BROSGOLE, L. (1964), Grouping Based on Phenomenal Proximity, *Journal of Experimental Psychology* 67, pp. 531–538.

ROMARY, L. (1993), L'interprétation de ici dans des énoncés de positionnement, dans : VIVIER, J. (Ed.) *Le dialogue homme-robot en langage naturel : problèmes psychologiques*, Presses Universitaires de Caen.

ROUSSEL, D. (1999), *Intégration de l'analyse du langage naturel parlé avec la reconnaissance de la parole*, Thèse de doctorat, Université Joseph Fourier de Grenoble.

RUSSELL, B. (1905), On Denoting, *Mind* 14, pp. 479–493.

SABAH, G. (1989), *L'intelligence artificielle et le langage : processus de compréhension* (vol. 2), Hermès, Paris.

SABAH, G., VIVIER, J., PIERREL, J.-M., ROMARY, L., VILNAT, A. & NICOLLE, A. (1997), *Machine, langue et dialogue*, L'Harmattan, Paris.

SABAH, G. (2000), The Fundamental Role of Pragmatics in Natural Language Understanding and its Implications for Modular, Cognitively Motivated Architectures, In: BUNT, H. & BLACK, B. (Eds.), *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics*, John Benjamins, Amsterdam, pp. 151–188.

SALMON-ALT, S. (2001a), Reference Resolution within the Framework of Cognitive Grammar, In: *Proceedings of the Seventh International Colloquium on Cognitive Science* (ICCS'01), San Sebastián, Spain.

SALMON-ALT, S. (2001b), *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*, Thèse de doctorat, Université Henri Poincaré de Nancy.

SANMIGUEL, D. (2000), *Perspective et composition*, Gründ.

SAUSSURE, F. de (1916), *Cours de linguistique générale*, Payot, Paris.

SCHANG, D. (1997), *Représentation et interprétation de connaissances spatiales dans un système de dialogue homme-machine*, Thèse de doctorat, Université Henri Poincaré de Nancy.

SHANNON, C.E. & WEAVER, W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

SIDNER, C.L. (1979), *Towards a Computational Theory of Definite Anaphora in English Discourse*, Ph.D. Thesis, MIT.

SPERBER, D. & WILSON, D. (1995), *Relevance. Communication and Cognition* (2nd edition), Blackwell, Oxford UK and Cambridge USA.

STEVENSON, R.J., CRAWLEY, R.A. & KLEINMAN, D. (1994), Thematic Roles, Focus and the Representation of Events, *Language and Cognitive Processes* 9(4), pp. 519–548.

STEVENSON, R.J. (2002), The Role of Salience in the Production of Referring Expressions, In: VAN DEEMTER, K. & KIBBLE, R. (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, pp. 167–192.

STRAWSON, P.F. (1950), On Referring, *Mind* 59, pp. 320–344.

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

Thórisson, K.R. (1996), *Communicative Humanoids. A Computational Model of Psychosocial Dialogue Skills*, Ph.D. Thesis, MIT.

Treisman, A.M. & Gelade, G. (1980), A Feature-Integration Theory of Perception, *Cognitive Psychology* 12, pp. 97–136.

van der Sluis, I.F. (2001), An Empirically Motivated Algorithm for the Generation of Multimodal Referring Expressions, In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, pp. 67–72.

Vandeloise, C. (1986), *L'espace en français : sémantique des prépositions spatiales*, Seuil, Paris.

Vettraino-Soulard, M.-C. (1993), *Lire une image*, Armand Colin, Paris.

Vieira, R. (1998), A Review of the Linguistic Research on Definite Descriptions, *Acta Semiotica et Linguistica* 7, pp. 219–258.

Vieu, L. (1991), *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en langage naturel*, Thèse de doctorat, Université Paul Sabatier de Toulouse.

Vivier, J. (2001), Introduction : la psycholinguistique au secours de l'informatique, *Langages* 144, pp. 3–19.

von Heusinger, K. (2000), The Reference of Indefinites, In: von Heusinger, K. & Egli, U. (Eds.), *Reference and Anaphoric Relations*, Kluwer Academic Publishers, Dordrecht, pp. 247–265.

Wahlster, W., André, E., Finkler, W., Profitlich, H.J. & Rist, T. (1993), Plan-Based Integration of Natural Language and Graphics Generation, *Artificial Intelligence* 63(1-2), pp. 387–427.

Wahlster, W. (Ed.) (2000), *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer.

Weil-Barais, A. (Ed.) (1993), *L'homme cognitif*, PUF, Paris.

Wertheimer, M. (1923), Untersuchungen zur Lehre von der Gestalt II, *Psychologische Forschung* 4, pp. 301–350.

Westerståhl, D. (1985), Determiners and Context Sets, In: van Benthem, J. & ter Meulen, A. (Eds.), *Generalized Quantifiers in Natural Language*, Foris Publications, Dordrecht, pp. 45–71.

Wilson, D. (1992), Reference and Relevance, *UCL Working Papers in Linguistics* 4, pp. 165–191.

Winograd, T. (1972), *Understanding Natural Language*, Academic Press, San Diego, CA.

Wolff, F., De Angeli, A. et Romary, L. (1998), Acting on a Visual World: The Role of Perception in Multimodal HCI, In: *Proceedings of AAAI Workshop on Multimodal Representation*, Madison.

Wolff, F. (1999), *Analyse contextuelle des gestes de désignation en dialogue homme-machine*, Thèse de doctorat, Université Henri Poincaré de Nancy.

Wolters, M.K. (2001), *Towards Entity Status*, Ph.D. Thesis, Bonn University.

Wright, P. (1990), Using Constraints and Reference in Task-Oriented Dialogue, *Journal of Semantics* 7, pp. 65–79.

Zeevat, H. & Wang, D. (1996), Minimal Sorts for Interpreting Pictures, CWI Report of Department of Interactive Systems, Amsterdam.

Zeevat, H. (1999), Demonstratives in Discourse, *Journal of Semantics* 16(4), pp. 279–313.

Index

Table des matières

Liste des figures :

Résumé

Notre manière de percevoir les objets qui nous entourent détermine nos choix langagiers et gestuels pour les désigner. Les gestes que nous produisons structurent notre espace visuel, les mots que nous utilisons modifient à leur tour notre manière de percevoir. Perception visuelle, langage et geste entretiennent ainsi de multiples interactions. Il s'agit bien d'une seule problématique qui doit être appréhendée globalement, premièrement pour comprendre la complexité des phénomènes de référence, deuxièmement pour en déduire une modélisation informatique exploitable dans tout système de dialogue homme-machine qui se veut un tant soit peu compréhensif.

Nous montrons comment tout acte de référence se produit dans un sous-ensemble d'objets, ce sous-ensemble appelé domaine de référence étant implicite et pouvant découler de multiples indices. Parmi ces indices, certains proviennent du contexte visuel et de l'énoncé émis, d'autres proviennent de l'intention, de l'attention et de la mémoire de l'utilisateur. Nous proposons une formalisation des domaines de référence en tenant compte de ces critères et en nous axant sur la notion de saillance dont nous proposons une caractérisation formelle. Il nous apparaît en effet que l'implicite se retrouve en priorité à l'aide des indices saillants. Nous montrons comment un système de dialogue peut exploiter les hypothèses obtenues en s'aidant d'un critère de pertinence. Nous posons quelques pistes pour une calculabilité de ce critère. Notre contribution s'attache ainsi à identifier l'implicite dans la communication multimodale, en termes de structurations d'objets et de formalisation de critères cognitifs.

Mots-clés : communication multimodale spontanée, perception visuelle, traitement automatique des langues, architecture logicielle, pragmatique, modélisation cognitive, référence aux objets, contexte, saillance, pertinence.

Abstract

The way we see the objects around us determines speech and gestures we use to refer to them. The gestures we produce structure our visual perception. The words we use have an influence on the way we see. In this manner, visual perception, language and gesture present multiple interactions between each other. The problem is global and has to be tackled as a whole in order to understand the complexity of reference phenomena and to deduce a formal model. This model may be useful for any kind of man-machine dialogue system that focuses on deep comprehension.

We show how a referring act takes place into a subset of objects. This subset is called reference domain and is implicit. It can be deduced from a lot of clues. Among these clues are those which come from the visual context and from the utterance, and those from the user's intention, attention and memory. We propose a formalization of reference domains taking these parameters into account. We focus on the notion of salience for which we propose a formal characterization. In fact, it seems that implicit information can most readily be retrieved from salient clues. We show how a dialogue system can exploit the resulting hypotheses with the help from a relevance criterion. We lay the foundations of the computation of this criterion. Our contribution is then directing along the identification of implicit information in multimodal communication, in terms of objects structures and of cognitive criteria formalizations.

Keywords: spontaneous multimodal communication, visual perception, natural language processing, dialogue system architecture, pragmatics, cognitive modelling, reference to objects, context, salience, relevance.