MÉMOIRE

présenté en vue d'obtenir

LE DIPLÔME D'INGÉNIEUR I.I.E.

Frédéric LANDRAGIN

Interaction Multimodale dans un Environnement Virtuel

Directeur du Mémoire : M. Célestin SÉDOGBO, Responsable du Laboratoire "Communication Homme-Machine" du Groupe Informatique du Laboratoire Central de Recherches de THOMSON-CSF.

Soutenu le 2 octobre 1998 devant le jury : M. M. MALLEM, Maître de conférences.
M. A. CORNUÉJOLS, Maître de conférences.
M. C. SÉDOGBO.

Conservatoire National des Arts et Métiers

INSTITUT D'INFORMATIQUE D'ENTREPRISE

FICHE SIGNALÉTIQUE

Mémoire d'Ingénieur I.I.E.

Interaction Multimodale dans un Environnement Virtuel

Auteur : Frédéric LANDRAGIN

Directeur de recherche : M. Célestin SÉDOGBO, Responsable du Laboratoire “Communication Homme-Machine” du Groupe Informatique du Laboratoire Central de Recherches de THOMSON-CSF.

Descriptif : Dans le cadre d'une interaction multimodale associant la commande vocale en langage naturel avec le geste de désignation dans un environnement virtuel, le système conçu permet la désignation floue et à distance d'objets (par le geste), ainsi que la prise en compte des caractéristiques spatiales saillantes des objets lors des références (par la commande vocale).

Résumé

Une des finalités de l'interaction dans un environnement virtuel consiste à interagir avec les objets présents dans la scène virtuelle. L'interaction multimodale associant geste de désignation et commande vocale offre pour cela une grande puissance d'expression. Afin de minimiser les efforts d'adaptation de l'utilisateur, elle se doit d'utiliser au mieux le caractère naturel de ses deux modes d'interaction. Ce mémoire traite certains aspects de la mise en œuvre d'une interaction multimodale dans une application de réalité virtuelle.

Nous étudions dans un premier temps les problèmes liés aux gestes de désignation imprécis, dont l'interprétation s'appuie sur les éléments contenus dans l'énoncé oral. Nous nous intéressons à la notion de désignation floue et nous proposons une mise en œuvre de zones de désignation floue.

Nous étudions dans un second temps les problèmes liés aux références en langage naturel. Afin de résoudre une référence, c'est-à-dire de faire le lien entre les groupes de mots utilisés et les objets virtuels, il est nécessaire d'identifier le contexte à l'intérieur duquel la référence a été produite par l'utilisateur. Ce contexte peut être linguistique, lié à la tâche ou perceptif. Dans le cas d'un contexte perceptif, l'utilisateur peut être amené à référencer un objet en utilisant une caractéristique saillante de cet objet, c'est-à-dire une caractéristique de l'objet qui le distingue des autres objets. Cette caractéristique peut être soit explicite dans l'énoncé de la commande vocale, soit implicite. Le problème revient alors à déterminer les circonstances dans lesquelles la saillance peut être utilisée, et, lorsque c'est le cas, à retrouver le type de saillance et à le prendre en compte lors du calcul de référent. Dans la section 3.3 dédiée à la saillance qui constitue l'objet principal de ce mémoire, nous étudions tout d'abord les structures linguistiques pouvant conduire à une recherche de saillance. Afin de déterminer les types possibles de saillance, nous étudions ensuite les caractéristiques d'un objet pouvant amener à sa saillance. Nous proposons une hiérarchie de ces caractéristiques et nous déterminons celles que l'on peut retenir et exploiter lors du calcul de référents. Nous nous intéressons alors à la saillance spatiale, c'est-à-dire à la saillance découlant des regroupements des objets et de leur répartition dans l'espace selon le point de vue du participant. Nous décrivons l'implémentation réalisée ainsi que les résultats obtenus, et nous proposons finalement quelques possibilités de prolongements de l'étude.

Remerciements

Je tiens à exprimer tous mes remerciements à Béatrice Bacconnet pour m'avoir accueilli dans le groupe Informatique, et à Célestin Sédogbo pour m'avoir accueilli dans le laboratoire Communication Homme-Machine, m'ayant ainsi fait entrer dans le monde du langage naturel.

Je remercie vivement Véronique Normand qui, malgré ses longs mois d'absence, a dirigé et relu mon travail avec attention, ainsi que Nicolas Farcet, François-Arnould Mathieu, Frédéric Meunier, Thierry Poibeau, Rodrigo Reyes, David Roussel dont les explications, les conseils et les relectures m'ont été utiles.

Merci aussi à Pascal Bisson, Nathalie Colineau, Ariane Halber ainsi qu'à Laurence Griffon pour leur accueil, de même qu'aux membres du laboratoire Architectures et Technologies Systèmes et du laboratoire Modélisation et Analyse Numérique. Merci également à Jean-Franck Olivier-André pour sa disponibilité.

Sommaire

1. Environnement

1.1. Présentation de l'entreprise

Thomson-CSF LCR :

Thomson-CSF est l'un des premiers groupes mondiaux d'électronique de défense. Il dispose d'un spectre de compétences très large, de l'acoustique sous-marine aux systèmes de missile et de communication, de l'optronique au traitement du signal.

Les activités de recherches sont centralisées au Laboratoire Central de Recherches (LCR). Cette unité d'environ 300 personnes est située à 25 kilomètres de Paris, sur le site de Corbeville, à proximité de la ville universitaire d'Orsay, au cœur de l'une des plus importantes communautés scientifiques d'Europe. Le LCR est l'un des plus grands laboratoires mondiaux de recherches avancées en physico-chimie et en électronique. Ces dernières années, il a joué un rôle de pionnier dans les domaines des matériaux, de l'optique, de la visualisation, de l'enregistrement magnétique et des composants à semi-conducteurs rapides. Sa vocation principale est l'exploration des technologies nouvelles dont la maîtrise sera cruciale pour les systèmes de demain.

le laboratoire CHM :

Le LCR est organisé en groupes, chaque groupe contenant un certain nombre de laboratoires. Le groupe "Informatique et Systèmes" représente environ 30 personnes et se divise en deux laboratoires : le laboratoire "Architectures et Technologies Systèmes" et le laboratoire "Communication Homme-Machine" (CHM). Le mémoire s'est déroulé dans ce dernier.

Le laboratoire CHM accueille une quinzaine de personnes, dont quatre thésards et deux stagiaires. Son activité consiste en recherche et développement, essentiellement dans le cadre de projets européens. Un certain nombre de contrats sont signés également avec la Direction Générale de l'Armement. Le thème principal de recherches est le traitement automatique du langage naturel (extraction, filtrage, génération, oral, etc.). Le deuxième thème abordé est celui de la réalité virtuelle et de la réalité augmentée. Initialement thème de recherches, il est aujourd'hui surtout l'objet de développements et sert en particulier de cadre applicatif pour les recherches du premier thème.

1.2. Contexte du mémoire

le projet COVEN :

En octobre 1995, le laboratoire CHM a mis en place avec douze partenaires européens un projet d'une durée de quatre ans relatif aux environnements virtuels distribués multi-utilisateurs en tant que nouvelle classe d'outils de support au travail coopératif : le projet ACTS COVEN (COllaborative Virtual ENvironments). Plusieurs scénarios d'application sont étudiés dans le cadre de ce projet : business

(conférences virtuelles, arrangement d'intérieur, etc.) et grand public (parcours routiers, agence de voyage virtuelle, etc.).

Le projet d'arrangement d'intérieur, développé au laboratoire CHM, consiste en manipulations d'objets d'ameublement divers dans une scène représentant un ensemble de bureaux. Le laboratoire CHM apporte ses compétences en langage naturel afin d'intégrer la commande vocale à l'interaction.

l’encadrement :

Outre le chef de projet, un ingénieur travaillait à temps plein sur ce projet au laboratoire CHM, avant son départ du LCR. Pendant les premiers mois de ce mémoire, j'ai pris sa suite dans les travaux de développement, ce qui m'a permis de me former sur l'interaction dans les environnements virtuels en général et sur la plate-forme utilisée dans le laboratoire. J'ai ensuite abordé mon travail de recherche puis de développement sur quelques problèmes posés par l'usage du geste de désignation et de la commande vocale dans les environnements virtuels.

Le développement s'est effectué sur des stations Silicon Graphics de type Impact, en langage C++ avec les librairies de fonctions dVS/dVISE de la société DIVISION. Les périphériques utilisés sont les périphériques 2D classiques : écran, clavier et souris. Le projet d'aménagement d'intérieur représente environ 20.000 lignes de codes. Un module de 3.500 lignes est écrit en LISP.

2. Objet de la recherche

2.1. Problématique

présentation globale :

L'interaction dans un environnement virtuel consiste avant tout à interagir avec les objets présents dans la scène virtuelle. Des mécanismes de sélection et de désignation doivent ainsi être proposés à l'utilisateur, lui permettant de choisir sur quel(s) objet(s) il va effectuer un traitement. Ce traitement peut consister en une manipulation directe classique dans les environnements virtuels (telle qu'un déplacement de la représentation graphique de l'objet) ou en une action spécifique à la logique de l'application gérant la scène. L'interaction à base de pointeur 2D ou 3D constitue classiquement le mode d'interaction privilégié dans un environnement virtuel. Les limites de ce mode d'interaction sont atteintes dès que l'on veut agir sur des ensembles d'objets, sur des objets qui n'existent pas (pour les créer), ou qui sont hors du champ de visibilité et de manipulation.

L'interaction multimodale associant interaction vocale et gestuelle apparaît comme une alternative offrant une grande puissance d'expression tout en préservant la simplicité de l'interaction : ces deux modes sont naturels et ne nécessitent donc que très peu d'efforts de la part de l'utilisateur. L'intégration de la commande vocale à l'interaction dans un environnement virtuel pose différents problèmes liés au traitement automatique du langage naturel. En particulier, l'emploi que nous faisons des groupes nominaux et des pronoms pose le problème des références : à quels objets réfèrent-ils, autrement dit comment faire le lien entre les groupes de mots employés et les objets de l'environnement virtuel.

Le calcul de référents, c'est-à-dire le traitement consistant à faire ce lien, doit tenir compte d'un grand nombre de paramètres : les commandes vocales précédentes, l'utilisation éventuelle d'un geste, la prise en compte des seuls objets visibles ou de tous les objets de l'environnement, etc. De façon générale, il est nécessaire d'identifier le contexte à l'intérieur duquel la référence a été produite par l'utilisateur ; ce contexte peut être linguistique, lié à la tâche ou perceptif. Dans le cas d'un contexte perceptif, l'utilisateur peut être amené à référencer un objet en utilisant une caractéristique saillante de cet objet, c'est-à-dire une caractéristique de l'objet qui le distingue des autres objets. Cette caractéristique peut être soit explicite dans l'énoncé de la commande vocale, soit implicite. Le problème revient alors à déterminer les circonstances dans lesquelles la saillance peut être utilisée, et, lorsque c'est le cas, à retrouver le type de saillance et à le prendre en compte lors du calcul de référents.

objectifs du mémoire :

Nous nous intéresserons dans un premier temps aux mécanismes de désignation : comment désigner un objet à distance (avec un périphérique 2D ou 3D) ? Quels sont les problèmes liés à ce type de désignation et comment les résoudre ? D'autre part, comment désigner plusieurs objets simulta-

nément, c'est-à-dire de manière plus efficace et plus rapide que la désignation successive de chacun des objets ?

Nous introduirons alors la désignation par la commande vocale et nous étudierons les mécanismes liés à l'utilisation de commandes vocales en langage naturel. Nous nous focaliserons sur les problèmes posés par le calcul de référents, en particulier lorsque l'on souhaite prendre en compte la saillance des objets. Nous serons amenés pour cela à nous poser les questions suivantes : qu'est-ce qu'un objet saillant ? Quels sont les types de saillance ? Quand et comment prendre en compte la saillance pour les références ?

2.2. Introduction à l'interaction dans un environnement virtuel

Les recherches effectuées dans le cadre de ce mémoire concernant essentiellement les mécanismes de désignation par le geste et les principes d'une interface en langage naturel dans un environnement virtuel, il est nécessaire de commencer par une introduction présentant les concepts de base d'un environnement virtuel et les principes de base d'une interaction multimodale dans un environnement virtuel.

La section 2.2.1 présente les environnements virtuels avec pour principales sources bibliographiques les travaux de Grigore Burdea et Philippe Coiffet [Burdea Coiffet 1993], et ceux de Bernard Jolivalt [Jolivalt 1995]. La section 2.2.2 présente les interactions multimodales avec pour principales sources bibliographiques la thèse de Xavier Pouteau [Pouteau 1995] et de celle de Christophe Mignot [Mignot 1995]. Nous nous placerons au cours de cette section dans le cadre d'une interaction bimodale associant le geste de désignation et la commande vocale.

2.2.1. Environnement virtuel

définitions :

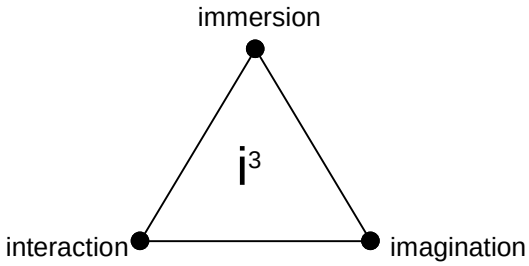
Grigore Burdea [Burdea Coiffet 1993] donne la définition suivante : "Un système de réalité virtuelle est une interface qui implique de la simulation en temps réel et des interfaces via de multiples canaux sensoriels. Ces canaux sensoriels sont ceux de l'homme : vision, audition, odorat, toucher, goût."

Au terme "réalité virtuelle" entré dans le langage courant, nous préférerons l'expression "environnement virtuel" qui met davantage l'accent sur l'implication de l'homme immergé dans le monde artificiel. Le terme "environnement" élude l'ambiguïté attachée au mot "réalité" ; il marque ses distances avec la simulation qui demeure résolument ancrée dans la reproduction exacte du réel. Il évoque l'immersion, la position de l'opérateur au centre de l'univers qui se construit autour de lui. L'environnement virtuel devient ainsi une expérience multidimensionnelle générée totalement ou partiellement par un ordinateur et susceptible d'être validée par le participant sur le plan cognitif. [Jolivalt 1995].

Mel Slater[1] et Martin Usoh, chercheurs à University College London, définissent [Slater Usoh 1993] l'environnement virtuel comme "un environnement créé par l'interaction d'un participant humain avec un monde généré par l'ordinateur. Il produit des informations d'ordres visuels, auditifs, kinesthétiques (y compris les retours tactiles et d'effort)".

La réalité virtuelle n'a pas inventé les mondes virtuels. Elle les a emprunté à la simulation. La simulation, qui est à l'origine de la réalité virtuelle, en est également le produit. La réalité virtuelle est non seulement capable de simuler le monde, mais de proposer une approche différente, inédite, fondée sur l'imagination et la créativité. Les programmeurs de jeux vidéo n'ont pas manqué d'exploiter cette ouverture sur des mondes mythiques. L'industrie et les armées s'intéressent de près aux potentialités de la réalité virtuelle, non seulement parce qu'elle propose l'immersion dans des simulations, mais parce qu'elle donne à voir des phénomènes qui échappent aux sens, comme la mise en évidence des contraintes subies par un matériau ou la vision thermique qui permet de voir dans la frange de l'infrarouge.

Grigore Burdea, professeur à l'Université Rutgers, où il dirige le département de recherche sur les interfaces homme-machine, inscrit la réalité virtuelle [Burdea Coiffet 1993] dans un triangle dont chaque sommet comporte la lettre "I". Elles se rapportent aux trois fondements de la réalité virtuelle : l'immersion, l'interaction et l'imagination :



- **Immersion :** Afin qu'il se sente pleinement présent dans le monde virtuel qui l'environne de toute part, l'opérateur doit être physiquement plongé dans le monde virtuel, ou du moins en éprouver la sensation, grâce à la faculté d'orienter le regard dans toutes les directions et d'entendre un son provenant de n'importe quel point de l'espace. Il doit pouvoir se déplacer et explorer l'univers qui l'entoure. Dans les applications les plus sophistiquées, il peut effleurer un objet et en percevoir la texture, le saisir délicatement avec ses doigts ou le soulever à pleine main en ressentant son poids.

  Le concept d'immersion n'est pas une retombée de la réalité virtuelle. Il est sous-jacent à toute entreprise de spectacle, dont l'une des finalités est de proposer une autre vision du monde. Afin d'éviter toute interférence fâcheuse avec le réel, le spectateur est isolé du monde extérieur. La recherche d'un spectacle total, qui donnerait au spectateur l'illusion d'être plongé dans l'image, d'en faire partie intégrante, n'est pas récente. A la fin des années 1950, le producteur Morton Heilig [Heilig 1992] conçoit le Sensorama, un cinéma expérimental qui solliciterait la plupart des sens : la vue et l'ouïe bien sûr, auxquelles s'ajouteraient la perception de la température et du souffle du vent, et même l'odorat (faute de financement, le Sensorama ne fût jamais construit). Selon Morton Heilig, qui avait

[1] Mel Slater fait partie des partenaires du projet COVEN.

longuement étudié les techniques de perception, les sens accaparaient l'attention selon l'échelle suivante :

1. la vue : 70%
2. l'ouïe : 20%
3. l'odorat : 5%
4. le toucher : 4%
5. le goût : 1%

Bien qu'il ne s'agisse ici que de réception d'informations, on peut en déduire que la vue puis l'ouïe sont à la base de la communication, non seulement homme-homme, mais aussi homme-machine. Le graphique fournit l'un des moyens les plus naturels pour communiquer avec un ordinateur, car nos capacités de reconnaissance de formes 2D et 3D, grandement développées, nous permettent de percevoir et traiter des données picturales rapidement et efficacement. [Foley et al. 1990].

- **Interaction :** L'interaction donne à l'opérateur un pouvoir sur le monde virtuel. Elle lui permet de s'y mouvoir à sa guise, de saisir des objets ou de communiquer des ordres au système informatique ou de converser avec les êtres de synthèse qu'il rencontre. Sans interaction, l'opérateur ne serait que le spectateur passif d'un univers sur lequel il n'aurait aucune prise. L'interaction lui permet non seulement de se déplacer librement et d'agir mais aussi de recevoir en retour des stimuli visuels (mise à jour du visuel d'après la direction du regard), auditifs ou haptiques (choc, élasticité d'un objet, poids ou résistance à la poussée).

  L'interaction est obtenue par un échange de données bidirectionnel entre l'opérateur et le monde virtuel. Elle implique, de la part du système informatique, des temps de réponse très brefs. Le déplacement du point de vue, par exemple, s'accompagne aussitôt d'une remise à jour du visuel. Les actions sont liées à des effets gérés par l'ordinateur : toucher un bouton virtuel ouvre une porte virtuelle ou allume la lumière, heurter le mur d'un appartement dessiné en image de synthèse se répercute, grâce à un mécanisme de retour d'effort, dans la manette tenue à la main ou sur l'exosquelette qui la gante. L'opérateur perçoit ainsi le choc contre l'obstacle.

  L'interaction ne doit pas être confondue avec l'interactivité, qui est une interaction à sens unique : l'opérateur commande et la machine exécute l'ordre.

- **Imagination :** L'imagination laisse le concepteur libre de définir les lois qui régissent l'univers virtuel. Il peut, suivant le cahier des charges, modéliser un monde et des objets strictement conformes à la réalité (applications apparentées à la simulation), ou s'affranchir de tout ou partie des lois physiques (applications ludiques, facilités de déplacement afin de mieux appréhender un projet).

  Le terme de "réalité virtuelle altérée" désigne des applications qui prennent des libertés avec le réel en introduisant des règles et des lois qui défient celles du monde qui nous entoure. Réalité virtuelle et réalité virtuelle altérée sont deux concepts fréquemment confondus. Une distinction est néanmoins toujours faite entre la réalité virtuelle et la réalité augmentée, qui superpose les objets de la réalité virtuelle par-dessus le monde réel.

Il existe bien d'autres définitions des environnements virtuels, en particulier la quantification de David Zeltzer, chercheur au MIT, qui situe la réalité virtuelle en un point localisé à l'intérieur d'un cube

tridimensionnel dont les axes quantifient l'autonomie, l'interaction et la sensation de présence. Ces définitions sont intéressantes car elles donnent une idée de l'importance et de l'étendue de l'interaction.

fonctionnement général :

Il existe maintenant sur le marché et dans les universités un certain nombre de systèmes de réalité virtuelle. Certains d'entre eux abordent le problème du fonctionnement distribué dans le cadre d'environnements virtuels multi-utilisateurs. Parmi ceux-ci, des prototypes de recherche (notamment DIVE, SPLINE et MASSIVE) et des produits commerciaux (par exemple dVS™ et WorldToolkit™). Selon les cas, ces plates-formes peuvent offrir des fonctions relativement sophistiquées, assurant :

- la distribution d'une base de données graphique 3D, avec mise à jour en temps réel des copies à travers le réseau, selon des schémas d'architecture réseau variés autorisant ou non une mise à l'échelle (scalability) de ces sytèmes,
- des fonctions de détection de collision entre les objets 3D, et de gestion d'événements d'interaction (en provenance de périphériques d'entrée divers), d'événements de collision, et d'événements divers, liés à la session (connexion de participants), à l'affichage ou au temps,
- des fonctions de visualisation graphique 3D, de rendu sonore spatialisé, de rendu vidéo parfois (dans DIVE notamment),
- des fonctions de représentation des participants à l'intérieur de la scène 3D (corps virtuels ou clone), et la navigation de ces participants à l'intérieur de la scène,
- des fonctions de gestion autorisant la connexion et la déconnexion de participants au cours d'une session,
- des langages de script permettant le prototypage rapide d'applications (par exemple, TCL/TK dans DIVE, ou VDI dans dVS).

Certaines de ces plates-formes (par exemple dVS) ont pour ambition de constituer des environnements de développement offrant notamment des outils graphiques (2D et 3D) de prototypage rapide.

La plupart d'entre elles gèrent la majorité des périphériques disponibles sur le marché. Ces périphériques sont présentés par catégorie dans la liste suivante :

- Les capteurs de position tridimensionnels permettent de connaître la position et l'orientation absolues d'un objet mobile (généralement la tête de l'utilisateur).
- Les systèmes de contrôle à boule (trackballs) permettent la navigation dans l'environnement virtuel ou le contrôle incrémental de la position d'un objet virtuel. Le support de la boule comprend généralement une série de boutons poussoirs qui permettent par exemple le couplage de la boule avec l'objet virtuel.
- Les gants sensitifs mesurent les valeurs angulaires des articulations et permettent, lorsqu'ils sont associés à un capteur de position, de manipuler des objets de manière naturelle.
- Les systèmes de vision stéréoscopiques permettent à l'utilisateur d'avoir la sensation de profondeur et donc d'immersion dans l'environnement virtuel. Les casques de visualisation et les lunettes stéréoscopiques sont les plus répandus.

- L'écran, la souris et le clavier, c'est-à-dire les périphériques 2D classiques, sont parfois suffisants. Ce sont en tout cas les plus utilisés dans la mesure où les périphériques plus complexes restent très chers.

La vue représentant 70% de l'utilisation des sens, l'algorithme chargé du rendu visuel doit être bien étudié. Comme nous verrons dans la suite de ce rapport quelques mécanismes liés à cet algorithme, il est nécessaire d'énoncer les grands principes de son fonctionnement.

Le monde virtuel 3D se compose de la scène et des objets sur lesquels une interaction est possible. Il contient en outre une répresentation de l'utilisateur (ou participant) que l'on nomme avatar. Les objets graphiques sont décrits en termes de géométrie, de matériau, de couleur et de texture. Une caméra est liée au participant et permet ainsi, à partir de la position et de l'orientation de celui-ci, de calculer l'image perçue. Ce calcul consiste en la projection des objets dans un plan perpendiculaire à la direction de visée du participant. Un certain nombre d'algorithmes de rendu accompagne cette projection. Dans le cas d'un écran 2D classique, le résultat de la projection est directement affichée à l'écran. Dans le cas d'une vision stéréophonique, deux projections sont effectuées avec pour origine chacun des deux yeux du participant puis transmises au système de vision. Les projections sont perspectives.

Pour que l'animation soit fluide, un minimum de 20 images par secondes est nécessaire. Tous les calculs sont effectués en temps réel et doivent donc être très rapides. C'est pourquoi dans les plates-formes de réalité virtuelle classiques, la géométrie des objets est surfacique et se compose uniquement de facettes polygonales. Les algorithmes de rendu proprement dits (éclairage de Gouraud, placage de textures, anti-crénelage) ne sont pas effectués de manière logicielle mais sont pris en charge par la carte graphique accélératrice 3D de la station de travail.

2.2.2. Interaction multimodale

définition et généralités :

La multimodalité dans l'interaction consiste à utiliser simultanément plusieurs modes de communication. Une interaction permettant d'utiliser simultanément le geste et la voix est un exemple d'interaction bimodale. La multimodalité n'est pas une spécificité des systèmes d'information, bien au contraire, la communication humaine est par nature essentiellement multimodale [Gaiffe 1992]. Dans le cadre d'un dialogue homme-homme, le geste est un deuxième mode de communication utilisé comme complément au premier mode qu'est la communication parlée.

Les modes de communication se font généralement par l'intermédiaire de médias différents. Or les temps de réponse des divers médias sont différents. Ceci pose des problèmes sur la gestion et l'interprétation des événements multimodaux. D'autre part, combiner plusieurs modes semble poser un problème du fait qu'il est possible d'adopter plusieurs points de vue sur un même acte. Dans sa thèse, Christophe Mignot [Mignot 1995] définit les différents points de vue possibles sur un énoncé multimodal et distingue un point de vue en deux catégories : point de vue discret et point de vue continu. Les énoncés multimodaux sont des énoncés dont l'effet ne dépend pas d'un seul mode mais de plusieurs. Il est possible de construire trois points de vue : un sur chacun des modes et un troisième sur leur combinaison. Les combinaisons possibles de deux points de vue selon leurs catégories sont les suivantes :

- **Combinaison de deux points de vue continus :** l'intérêt est de simuler une indépendance entre les deux modes afin de permettre à l'utilisateur le développement de ses capacités de coordination. L'exemple que donne Mignot est celui d'un système de simulation du pilotage d'un tank : à l'aide de deux manettes indépendantes contrôlant chacune la vitesse de rotation d'une chenille, l'utilisateur peut faire avancer et tourner le char.
- **Combinaison d'un point de vue discret et d'un point de vue continu :** par exemple dans une interface capable de combiner un geste de dessin avec une commande vocale, le geste est utilisé pour tracer une forme continue et la commande vocale pour donner l'épaisseur du trait. Les deux modes étant indépendants, la combinaison entre les deux ne peut être que temporelle[2].
- **Combinaison de deux points de vue discrets :** le geste peut par exemple être utilisé pour désigner un objet et la commande vocale pour effectuer une action sur cet objet. Cette combinaison peut être temporelle ou sémantique[3]. On remarque que le geste, selon son utilisation, peut être considéré comme un point de vue discret ou continu.

interaction bimodale associant geste de désignation et commande vocale :

Cette approche permet d'aborder les problèmes liés à l'association du geste et de la voix qui nous intéressent. Le geste est utilisé pour la désignation et la voix est utilisée pour la commande. L'association de ces deux modes est intéressante car la langue naturelle et le geste sont les deux principaux moyens spontanés d'expression ou d'action d'une personne sur son environnement. Ceci a été confirmé par des expériences du type magicien d'Oz[4] comme celle présentée par Mignot dans sa thèse.

Lorsqu'il est utilisé dans un but de désignation dans un environnement virtuel, le geste permet d'effectuer une action discrète telle que la désignation d'un objet dans la scène ou la désignation d'un endroit particulier de cette scène. La commande vocale permet également d'effectuer une action discrète telle que la création ou la modification d'une propriété d'un objet. Nous sommes donc dans le cas d'une combinaison de deux points de vue discrets.

La combinaison temporelle du geste de désignation et de la commande vocale est alternée ou parallèle. La combinaison peut être sémantique dans le sens où un même geste de désignation peut vouloir désigner un objet ou un endroit de la scène. Le choix est effectué lors de l'interprétation de la commande vocale (par exemple selon la présence de l'expression "cet objet" ou de l'expression "ici"). Le geste a ainsi un rôle de support sur lequel s'appuie la commande vocale.

---

[2] Une combinaison temporelle de deux modes peut se faire de plusieurs manières : elle peut être parallèle (l'intervalle de temps pris par un mode est inclus dans l'intervalle pris par l'autre), alternée (un mode précède l'autre) ou synchronisée (la prise en compte des deux modes se fait à la demande de l'utilisateur, par exemple lorsque celui-ci dit : "top").

[3] Une combinaison sémantique des modes signifie que l'interprétation de chaque expression contribue à définir le sens de l'énoncé multimodal. [Romary 1993] montre par exemple que la combinaison du geste et de la commande vocale peut être sémantique : le sens du geste peut dépendre de l'énoncé oral.

[4] Une expérience de type magicien d'Oz place un utilisateur en situation d'utilisation d'un système dont les réactions sont simulées par un homme (le compère).

commande vocale en langage naturel :

Après une phase où les développements informatiques ont été guidés par les avancées scientifiques et technologiques et ont conduit à des systèmes auxquels l'homme a dû s'adapter, il convient aujourd'hui de redonner la priorité à l'homme en adaptant les systèmes informatiques aux besoins et aux potentialités humains [Schang 1997]. Dès lors, si l'on se fixe comme objectif de minimiser l'effort d'adaptation de l'utilisateur à la machine, le choix de modes d'interaction naturels semble s'imposer. Les périphériques 3D tels que les gants sensitifs dont nous avons parlé dans la section 2.2.1 vont dans ce sens. La langue naturelle[5] semble également un choix intéressant. Elle présente en effet l'avantage d'être une compétence déjà acquise par l'utilisateur et un moyen de communication particulièrement puissant.

La commande vocale permet par exemple d'effectuer plusieurs opérations de natures différentes en un seul énoncé. Elle permet aussi de prendre du recul par rapport aux manipulations manuelles directes. L'opposition entre la manipulation et la commande vocale rejoint la distinction faite entre *faire* et *faire-faire* explicitée par Xavier Pouteau ([Pouteau et al. 1994] puis [Pouteau 1995]) : la machine peut se différencier en tant qu'outil de travail et en tant qu'interlocuteur. Deux relations se posent alors entre l'opérateur et la machine :

- Une logique de *faire*, où les objets de l'interface réagissent aux opérations de manipulation virtuelle effectuées par le canal gestuel (prolongé par les outils employés). Il est alors nécessaire de gérer le fonctionnement des outils pour donner à l'utilisateur l'illusion d'une vraie relation directe aux objets manipulés.
- Une logique de *faire-faire*, où la manipulation n'est plus la logique globale de fonctionnement de la machine, et où l'opérateur indique à l'exécutant informatique les opérations à effectuer.

Si l'on se place par exemple dans un environnement dans lequel l'interaction est fondée sur des menus, et que l'on veuille ajouter trois chaises en plastique bleu dans la scène, cela peut nécessiter la série d'opérations suivantes : dans le menu correspondant à la création d'un objet, choisir l'objet "chaise" (ce qui a pour effet de créer une chaise avec des attributs par défaut) ; recommencer deux fois l'opération ; sélectionner les trois chaises ainsi créées ; dans le menu correspondant à la modification de matériau, choisir "plastique" (en supposant que la modification s'applique sur l'ensemble de sélection) ; dans le menu correspondant à la modification de couleur, choisir "bleu". On obtient bien alors trois chaises en plastique bleu et on a l'impression d'avoir *fait* les actions[6].

Or toutes ces opérations manuelles sont efficacement remplacées par la simple et unique commande vocale : "ajoute trois chaises en plastique bleu". Le résultat est alors le même et on a l'impression d'avoir *fait faire* le travail par la machine, de n'avoir eu qu'un rôle de dirigeant et d'observateur.

---

[5] Une commande vocale en langage naturel suit les principes grammaticaux de la langue. Une commande vocale en langage artificiel serait par exemple une suite de mots dans leur forme canonique et sans liens grammaticaux (verbe à l'infinitif, pas d'article). L'emploi d'un langage artificiel impose donc un certain effort à l'utilisateur.

[6] L'interaction par menus et l'interaction par la commande vocale reposent en fait sur la même logique de *faire-faire*, l'interaction par menus étant cependant plus proche du *faire*. Une interaction basée véritablement sur une logique de *faire* obligerait par exemple l'utilisateur à prendre un pinceau pour changer la couleur des chaises.

On peut reprendre le même exemple en apportant la multimodalité avec le geste de désignation : si l'on veut ajouter trois chaises en plastique bleu dans un coin particulier de la scène, une manipulation simple consisterait à créer les trois chaises et à modifier leurs propriétés comme précédemment, puis à les sélectionner et à les déplacer dans la zone désirée. Outre le fait qu'il s'agit d'une opération manuelle, il est à noter que le déplacement d'un objet en 3D est difficile même avec un système de contrôle à boule (une direction est difficile à garder).

Or un geste de désignation vers la zone désirée associé à la commande vocale "ajoute trois chaises en plastique bleu ici" est bien plus efficace et rapide mais potentiellement moins précis. Comme nous le verrons dans la section 2.4.1, l'adverbe de lieu "ici" est un déictique et il réfère à la position désignée. Avant d'étudier ces mécanismes linguistiques, il est toutefois nécessaire d'étudier le geste de désignation.

2.3. Mécanismes de désignation

Nous considérons le geste de désignation comme une alternative à la sélection directe d'un objet. Il est effectué à distance de l'objet, avec un degré d'imprécision, et nécessite donc un traitement afin de déterminer l'objet désigné. Il peut de plus s'effectuer sur un groupe d'objets, ce qui procure un avantage par rapport à la sélection successive d'objets.

La section 2.3.1 présente les concepts classiques de la désignation à distance. La section 2.3.2 présente les concepts classiques de la désignation multiple.

2.3.1. Désignation approximative

sélection et désignation approximative :

Sélectionner un objet dans un environnement virtuel, c'est effectuer une manipulation directe sur la représentation de cet objet afin qu'il soit mis en valeur par rapport aux objets non sélectionnés et qu'il soit utilisé pour une opération future (notion classique de sélection dans les interfaces 2D ou 3D). Afin d'appliquer une opération sur un ensemble d'objets, plusieurs objets peuvent être successivement sélectionnés, ce qui nécessite la gestion d'un ensemble de sélection. La mise en valeur consiste généralement en un rendu visuel spécial. Ce rendu visuel permet à l'utilisateur de savoir immédiatement si un objet est sélectionné ou non. En 2D, un rendu visuel classiquement utilisé est par exemple l'inverse vidéo. En 3D, plusieurs solutions sont possibles : affichage de l'objet en mode transparent, clignotant, en surbrillance, en fil de fer, etc. (Le choix effectué dans COVEN consiste à afficher la boîte englobante de l'objet sélectionné en utilisant une couleur caractérisant le participant à l'origine de la sélection).

La sélection classique dans un environnement en 2D consiste à placer le pointeur lié à la souris sur l'objet et à cliquer. Comme les pixels occupés par l'objet sont connus à tout moment, il suffit de récupérer les coordonnées 2D du pointeur au moment du clic pour savoir quel objet vient d'être sélectionné. En effet, un objet 2D classique tel qu'une icône ou une fenêtre occupe une zone rectangulaire dans le plan de l'écran, axée selon les axes horizontal et vertical de l'écran ; il suffit donc

de quatre tests (deux sur les abcisses et deux sur les ordonnées) pour savoir si le pointeur est placé sur l'objet. La notion de distance n'intervient donc pas en 2D.

Plusieurs modes de sélection sont classiquement proposés en 3D. Un premier mode de sélection consiste en la mise en contact de la main de l'avatar avec l'objet. Un second mode de sélection consiste à déplacer un pointeur 2D sur l'image projetée, et à effectuer un lancé de rayon dans la scène 3D lorsque l'utilisateur effectue la sélection proprement dite (par un clic souris).

On oppose sélection et désignation sur un critère de précision : que ce soit en 2D ou en 3D, la sélection est précise et ne conduit jamais à une ambiguïté sur l'objet choisi ; par contre, désigner un objet dans un environnement 3D consiste à effectuer un geste dans la direction approximative de cet objet. Ce geste est effectué à distance et reste vague, pouvant ainsi conduire à une ambiguïté sur l'objet choisi. Le problème revient à gérer des zones de désignation floues et à en extraire un objet. Dans le cadre d'une interaction multimodale associant la commande vocale à ce mode de désignation, l'énoncé peut aider à la résolution de l'ambiguïté.

mécanismes de désignation approximative :

En supposant que les périphériques utilisés sont les classiques écran, clavier et souris ; et en supposant que la désignation se fait par un clic de la souris comme pour la sélection en 2D, on ne peut pas savoir à quel objet appartient le pixel sur lequel se trouve le pointeur lors du clic. En effet, ce pixel appartient à la projection d'un objet et il n'existe pas de lien direct de la projection vers l'objet source. Ceci est du au fait que l'algorithme de projection est à un seul sens : à partir des identifiants des objets, l'algorithme récupère leur géométrie et les projette sur le plan correspondant à l'écran. Seule la profondeur des objets est gardée au cours de l'algorithme (dans le Z-buffer). Lorsqu'un objet en cache un autre, c'est la profondeur du plus proche qui est retenue. Mais les identifiants des objets ne sont pas retenus. Ceci est du en particulier au fait que ces algorithmes sont effectués par le hard et que le hard ne peut pas faire des renvois à des variables du soft.

Le seul moyen permettant de déterminer à quel objet appartient un pixel donné de l'image projetée est le suivant : à partir de l'origine de la projection, c'est-à-dire l'œil du participant, et des coordonnées 2D du pixel, on calcule la direction de désignation. Dans la scène 3D, on effectue un lancé de rayon à partir des coordonnées 3D de l'œil du participant et dans la direction calculée. On s'arrête au premier objet rencontré dont on retient l'identifiant. L'inconvénient de cette méthode est le temps relativement élevé de calcul dans le cadre d'un système en temps réel. Cette méthode peut être utilisée aussi bien pour la désignation que pour la sélection.

Dans le cadre d'un environnement virtuel sur une machine dépourvue de carte graphique 3D, une deuxième méthode consisterait à spécifier les algorithmes de rendu en tenant compte de ce besoin d'identification d'un objet par un pixel. On pourrait pas exemple construire un buffer équivalent au Z-buffer, contenant non pas des réels correspondant aux profondeurs, mais des pointeurs correspondant aux objets. Pour chaque pixel de l'écran, on disposerait ainsi d'un pointeur sur l'objet visible en ce pixel, ce qui faciliterait grandement les sélections et les désignations.

2.3.2. Désignation multiple

sélection multiple et désignation multiple :

La sélection multiple consiste à sélectionner plusieurs objets de façon à effectuer ensuite une opération globale sur l'ensemble de ces objets. La sélection multiple classique se fait en sélectionnant successivement chacun des objets. Ce mécanisme est contraignant dans le cas d'un grand nombre d'objets mais n'a pas d'inconvénients majeurs : étant donné que les objets sélectionnés sont affichés avec un rendu visuel particulier, l'utilisateur sait à tout moment quels sont les objets appartenant à l'ensemble de sélection. Il lui est par exemple possible de procéder à une vérification ou à une correction[7] de l'ensemble de sélection avant de lui appliquer une opération.

La désignation multiple est effectuée dans le même but que la sélection multiple. La différence est qu'un seul geste est effectué pour désigner plusieurs objets. Ces objets doivent alors être groupés. Comme on l'a vu à la section 2.3.1, le geste peut ainsi conduire à une ambiguïté sur les objets choisis. Le problème revient à extraire des zones de désignation floues un groupe d'objets.

mécanismes de désignation multiple :

En supposant que les périphériques utilisés sont les classiques écran, clavier et souris, on peut utiliser la possibilité offerte par la souris de tracer des formes sur l'écran : la désignation multiple d'un groupe d'objets peut se faire en entourant les objets par un geste en forme de cercle. Cette méthode a été employée dans plusieurs interactions en 2D, dans le cadre de la sélection multiple : l'utilisateur entoure un certain nombre d'objets qui, une fois le geste terminé, apparaissent avec un rendu montrant qu'ils appartiennent à l'ensemble de sélection. Si l'utilisateur veut extraire un objet de cet ensemble, il l'entoure à nouveau.

Les mécanismes d'interprétation du geste en 2D sont relativement complexes [Bellalem Romary 1993] : à partir de quelques points, la courbe est modélisée par des B-splines, les courbures sont déterminées, de même que les points de rebroussement et les points d'inflexion ; puis la courbe est simplifiée en fusionnant les éléments consécutifs de même type ; des indices de segmentation sont ensuite cherchés : points d'arrêt dus à une pression ou à un relachement d'un bouton, points d'intersection ; la représentation finale de la courbe est enfin confrontée avec les indices de segmentation afin de déterminer l'ensemble des gestes candidats.

Ce principe d'interprétation du geste est surtout utilisé lorsque la forme du geste contient une information relative à l'action à effectuer : par exemple, un trait ou une croix sur un objet est classiquement un ordre de destruction de l'objet. Lorsque le but est la sélection, les rectangles que l'on étire par un clic continu de la souris (dans Windows par exemple) sont généralement suffisants.

Ces mécanismes sont difficilement implémentables en 3D : l'utilisateur aurait à créer avec le gant un volume entourant les objets, ce qui semble difficile. D'autre part, la solution consistant à interpréter le geste dans la projection 2D n'est pas envisageable du fait du nombre relativement important de lancés de rayons nécessaires.

---

[7] Dans le cas où la sélection d'un objet se fait par un clic de la souris, le clavier est généralement utilisé en tant que modifieur pour la sélection multiple : un clic associé à l'appui de la touche Ctrl permet d'ajouter ou de retirer l'objet de l'ensemble de sélection.

2.4. Interface en langage naturel

La commande vocale apporte une certaine efficacité à l'interaction, comme nous l'avons vu dans la section 2.2.2. Elle entraîne par contre un certain nombre de difficultés dans son traitement : lorsque l'on autorise l'utilisateur à employer un langage naturel, il est nécessaire de traiter les constructions grammaticales qu'il peut être amené à utiliser. Dans une interaction homme-machine, les commandes vocales sont le plus souvent des ordres, c'est-à-dire des phrases relativement courtes à l'impératif, contenant généralement un verbe, un complément du verbe et un complément circonstanciel de lieu (par exemple). Afin de manipuler les objets de l'environnement, l'utilisateur va utiliser des références à ces objets, c'est-à-dire des groupes nominaux liés aux entités physiques que sont ces objets. La principale difficulté d'un système de traitement des commandes vocales est de faire le lien entre ces groupes de mots et les objets. Ce traitement, nommé calcul de référents, sera présenté dans la section 2.4.1. Son application à l'interaction multimodale dans un environnement virtuel sera étudiée dans la section 2.4.2. Notre cadre de travail sera alors défini et nous nous intéressons à la principale problématique de ce mémoire dans la section 2.4.3 : comment les caractéristiques saillantes des objets peuvent-elles être prises en compte dans les mécanismes de référence ? Les principales sources bibliographiques utilisées dans cette section sont la thèse de Bertrand Gaiffe [Gaiffe 1992] et celle de Philip Edmonds [Edmonds 1993].

2.4.1. Mécanismes de calcul de référents

étude des groupes nominaux :

Comme l'écrit Bertrand Gaiffe dans sa thèse sur la référence et le dialogue homme-machine [Gaiffe 1992], il faut partir des possibilités de la langue et non des possibilités d'implémentation : "Notre objectif est bien de déterminer quels modèles et quelles implémentations sont nécessaires pour traiter le langage et non pas quel langage nous pouvons traiter compte tenu de nos connaissances actuelles. Peut-être faut-il voir là la différence entre l'informatique qui utilise du langage et une véritable informatique-linguistique." Gaiffe part ainsi de l'étude des différents groupes nominaux pour déterminer les figures de style propres à chacun d'eux et la nature des références associées à ces figures. Nous reprendrons une partie de son étude en y ajoutant des cas relatifs à l'interaction multimodale que Gaiffe n'a pas pris en compte étant donné leur maladresse voire leur incorrection du point de vue de la linguistique.

Trois catégories intéressantes se distinguent : les pronoms définis, les groupes nominaux définis et les groupes nominaux démonstratifs. Le but de cette étude est de montrer un certain nombre de possibilités offertes par la commande vocale en langage naturel, bien que ces possibilités ne seront pas toutes exploitées dans COVEN.

1. les pronoms définis

La fonction principale du pronom est l'*anaphore*, c'est-à-dire la référence aux objets dont on vient de parler. Dans l'exemple suivant, le pronom réfère à la chaise verte :

"déplace la chaise verte"
"mets-la ici"

Se pose alors la question de l'antécédent repris par le pronom : est-ce le groupe de mots, la représentation sémantique ou l'objet lui-même ? Reprendre le groupe de mots nécessite de recalculer la référence ; reprendre l'objet lui-même consiste à reprendre directement le résultat de la référence précédente. Comme nous allons le voir sur les exemples suivants, chacune de ces méthodes a des inconvénients :

"peints la chaise verte en bleu"
"déplace-la ici"
La reprise du groupe de mots donne : "la" = "la chaise verte" qui justement n'est plus verte.

"détruit le grand bureau"
"ajoute une table aussi grande que lui"
La reprise de l'objet est impossible puisque l'objet n'existe plus.

Une solution consiste à reprendre la représentation sémantique, c'est-à-dire la structure du groupe nominal (en particulier s'il contient une coordination) et les propriétés des objets. Dans le cadre de COVEN, seules des commandes relativement simples sont utilisées et le cas de figure du deuxième exemple ne se produira pas. Comme les coordinations ne sont pas acceptées, la méthode choisie est la reprise de l'objet.

L'anaphore sous forme d'ensemble est un cas plus complexe d'anaphore, comme le montre l'exemple suivant :

"déplace la chaise verte"
"déplace la chaise rouge"
"mets-les ici"

L'anaphore peut enfin être *divergente*, c'est-à-dire que le pronom ne réfère pas au même objet, comme le montre l'exemple suivant :

"Ne lui achète pas ce livre, il l'a déjà"

Le *changement d'état* est une figure de style assez proche :

"On a rasé la chevelure de Samson, mais elle a repoussé"

Enfin, le pronom défini anaphorique peut impliquer un *glissement au générique* :

"J'ai acheté une Toyota parce qu'elles sont robustes"

2. les groupes nominaux définis

La première fonction du groupe nominal défini est la *désignation directe* :

"déplace la chaise bleue"

Cette désignation s'accompagne d'un filtrage lorsque par exemple il existe plusieurs chaises et qu'une seule de ces chaises est bleue.

La deuxième fonction du groupe nominal défini est l'*anaphore* :

"déplace la chaise verte et le bureau bleu"
"supprime la chaise"

L'emploi d'un groupe nominal défini pose donc une ambiguïté : est-ce une désignation directe ou une anaphore ? Comme nous le verrons dans la section 2.4.2, il est nécessaire de tenir compte du contexte pour résoudre cette ambiguïté.

L'anaphore peut être *associative*, c'est-à-dire que la deuxième référence est associée à la première par une relation de composition comme dans l'exemple suivant :

"j'ai acheté un stylo mais la plume est cassée"

3. les groupes nominaux démonstratifs

Ils semblent avant tout dédiés à l'association avec le geste dans les systèmes multimodaux. Pourtant, la reprise *anaphorique* par le groupe démonstratif existe :

"déplace la chaise verte"
"supprime cette chaise"

L'anaphore peut s'associer à une *hyperonymie*[8] comme dans l'exemple suivant :

"déplace le bureau vert"
"supprime ce meuble"

Les pronoms démonstratifs sont traités de la même façon que les groupes nominaux démonstratifs : ils désignent un ou plusieurs objets dont les caractéristiques sont indiquées par l'anaphore. Les groupes nominaux indéfinis ne font l'objet d'aucune étude puisqu'ils n'impliquent aucune référence. Ils sont essentiellement utilisés lors de la création d'objets. Les pronoms indéfinis ont le même rôle que les groupes nominaux indéfinis, les caractéristiques de l'objet ou des objets à créer étant indiquées par l'anaphore.

conséquences sur les énoncés :

Lorsque l'emploi d'un mot tel qu'un pronom ou un adjectif démonstratif est autorisé, toutes les figures de style relatives à ce mot devraient être traitées. En effet, l'utilisateur ne doit pas se poser la question, à chaque fois qu'il utilise un mot dans un emploi particulier, de savoir s'il va être compris ou non. Dans le cadre d'une interaction homme-machine, certaines figures de style ne seront jamais utilisées par l'utilisateur car les situations pouvant amener à leur utilisation ne se produiront jamais. C'est le cas par exemple du changement d'état dans un environnement composé d'objets d'ameublement. Ceci permet donc de ne traiter que certains emplois sans pour autant obliger l'utilisateur à faire un effort particulier.

D'autre part, certaines règles classiques, lorsqu'elles sont suivies, simplifient la tâche du calcul de référents. Il s'agit par exemple du principe d'économie qui dit que l'utilisateur ne décrit que les caractéristiques pertinentes d'un objet. Dans un environnement contenant par exemple une seule chaise bleue, l'utilisateur emploiera le groupe nominal défini "la chaise" et non "la chaise bleue", économisant ainsi un mot inutile. Lorsque cette règle n'est pas respectée, le calcul de référents se demande pourquoi l'utilisateur emploie l'adjectif bleu alors que la seule chaise visible est bleue. Il peut envisager les explications suivantes : erreur de reconnaissance ou erreur de l'utilisateur qui a par exemple peut-être cru voir une autre chaise non bleue dans l'environnement.

De manière plus générale, on peut considérer les maximes de Grice [Grice 1975], en particulier les maximes de quantité, comme des règles que tout utilisateur devrait suivre. Les maximes de Grice sont les suivantes :

[8] meuble est un hyperonyme de chaise ; table et chaise sont des hyponymes de meuble.

maximes de quantité	Apporter suffisamment d'information. Ne pas apporter plus d'information que ce qui est nécessaire.
maximes de qualité	Ne rien dire que l'on croit faux. Ne rien dire que l'on ne puisse démontrer.
maxime de pertinence	Etre pertinent.
maximes de manière	Eviter d'utiliser des expressions obscures. Eviter d'utiliser des expressions ambiguës. Etre bref. Donner les informations dans le bon ordre.

Il s'avère que dans la plupart des cas, ces règles sont respectées de manière plus ou moins inconsciente par les utilisateurs.

2.4.2. Application à l'interaction multimodale dans un environnement virtuel

notion de contexte référentiel :

La résolution d'une référence dépend de l'historique du dialogue et de son association éventuelle à un geste de désignation. Plus généralement, elle dépend de l'environnement, en particulier dans le cadre d'une interaction multimodale avec un environnement virtuel.

Nous avons vu qu'un groupe nominal démonstratif par exemple pouvait impliquer une anaphore ou une désignation directe. Gaiffe propose de supposer en premier lieu qu'il s'agit d'une anaphore puis, si cette supposition ne permet pas de résoudre la référence, de supposer en second lieu qu'il s'agit d'une désignation directe. Cela revient à tester dans la référence dans un contexte anaphorique puis dans un contexte lié à l'environnement. Nous introduisons ainsi la notion de contexte référentiel : les différents contextes permettant de résoudre la référence sont traités l'un après l'autre, dans un ordre précis, le passage au contexte suivant ne se faisant que si la référence n'a pu être résolue. Dans chaque contexte n'est prise en compte qu'une partie de l'environnement.

Dans le cadre d'une interaction dans un environnement virtuel, et plus généralement d'une interaction dans un environnement graphique 3D ou 2D, deux types de contextes sont à distinguer : les contextes liés au discours et les contextes liés à l'environnement graphique. Les contextes liés au discours sont forcément anaphoriques et peuvent se décomposer en un contexte correspondant à la référence de la dernière commande et en un contexte correspondant aux références des dernières commandes (ou historique des commandes). Dans le cadre d'un environnement virtuel, les contextes liés à l'environnement graphiques peuvent se décomposer en un contexte correspondant aux objets dans le champ de vision du participant et en un contexte correspondant à tous les objets de l'environnement. Ces quatre contextes sont testés dans l'ordre suivant :

- les références de la dernière commande,
- les références des commandes précédentes,
- les objets visibles dans l'environnement,
- tous les objets de l'environnement.

Typiquement, dans le cas d'une expression telle que "la chaise", le calcul de référents commence par considérer la commande précédente : si on avait manipulé une chaise juste avant, la chaise qui avait été retenue lors du calcul de référents précédent sera testée en premier. Si cette chaise ne peut être celle référencée (par exemple si elle vient d'être détruite), on passe au contexte suivant, c'est-à-dire que l'on teste les chaises précédemment référencées. Si aucune de ces chaises ne correspond, ou si l'historique du dialogue est vide, on passe au contexte des objets visibles, c'est-à-dire que l'on teste chacune des chaises visibles. Si on trouve une chaise visible, le référent est trouvé ; si on trouve plusieurs chaises visibles, une ambiguïté est annoncée ; si on ne trouve aucune chaise visible, on passe au contexte suivant. On cherche alors une chaise dans l'ensemble des objets de l'environnement : si on en trouve une, le référent est trouvé, sinon la référence n'est pas résolue.

Tous les contextes référentiels ne sont pas toujours testés : dans le cas d'un groupe nominal indéfini, par exemple "ajoute une chaise bleue ici", aucun des contextes n'est exploré ; dans le cas d'un pronom indéfini, par exemple "ajoutes-en une ici", les contextes anaphoriques ne sont utilisés que pour récupérer la catégorie et les propriétés de l'objet. Par contre, dans le cas d'un groupe nominal défini, d'un pronom défini, d'un groupe nominal démonstratif ou d'un pronom démonstratif, tous les contextes référentiels peuvent être testés.

association de la commande vocale avec le geste de désignation :

Lorsqu'un ou plusieurs objets sont sélectionnés et qu'ils apparaissent avec un rendu particulier les mettant en valeur, il est logique de les considérer comme étant prioritaires pour le calcul de référents. Or la sélection pouvant être effectué manuellement, c'est-à-dire par un mode autre que la commande vocale, les objets sélectionnés ne sont a priori présents dans aucun contexte anaphorique. Si on manipule une chaise, que l'on sélectionne ensuite une autre chaise et que l'on dit "peints-la en bleu", le pronom défini référera la première chaise, ce qui est correct d'un point de vue linguistique (c'est-à-dire si l'on ne considère que les commandes vocales) mais ne correspond sans doute pas à ce que l'utilisateur attendait : le fait qu'il mette la deuxième chaise en valeur sous-entend qu'il va appliquer une action sur cette chaise.

La notion d'ensemble de sélection est donc indépendante dans le sens qu'elle doit être traitée avant le premier contexte référentiel. Ce n'est que dans le cas d'une ambiguïté que l'appel aux contextes référentiels est effectué. D'autre part, nous remarquons que l'association de la commande vocale et des manipulations directes peut donner lieu à des structures linguistiques incorrectes : autant une commande vocale seule reste linguistiquement correcte, autant une suite de commandes vocales alternées avec des manipulations directes fait apparaître des structures linguistiques aberrantes au niveau des anaphores.

L'association du geste de désignation tel que nous l'avons étudié dans la section 2.3 et de la commande vocale pose le problème des co-références. De façon générale, on parle de co-référence chaque fois que deux désignations (gestuelles ou vocales) aboutissent à un même objet. Dans un contexte multimodal, on peut alors distinguer deux types de co-références [Gaiffe 1992] :

- les **co-références synchrones** qui par nature exigent l'utilisation de deux modes, dans notre cas une référence est effectuée par le discours et l'autre par le geste,
- les **co-références asynchrones** qui consistent en la désignation d'un même objet plusieurs fois dans un même mode, dans notre cas par le discours exclusivement.

Nous pouvons alors distinguer trois situations : soit il y a référence simple, c'est-à-dire qu'aucun geste de désignation n'a été effectué et que l'historique des commandes (ou historique du dialogue) est vide ; soit il y a co-référence synchrone, c'est-à-dire qu'un geste de désignation a été effectué et qu'on ne tient pas compte de l'historique du dialogue ; soit il y a co-référence asynchrone, c'est-à-dire qu'aucun geste de désignation n'a été effectué mais que l'historique du dialogue contient déjà une référence qui est prise en compte. L'étude des groupes nominaux nous permet de dresser la liste des figures de style possibles dans chacune de ces situations :

1. **référence simple :**

   Seul le groupe nominal défini est possible. Il correspond à une désignation directe, avec ou sans filtrage.

2. **co-référence synchrone :**

   groupe nominal défini : cet usage est maladroit, il correspond à une désignation directe, avec ou sans filtrage.

   groupe nominal démonstratif : désignation directe sans filtrage.

3. **co-référence asynchrone :**

   groupe nominal défini : désignation directe (avec filtrage), anaphore, anaphore associative.

   pronom défini : anaphore, anaphore divergente, changement d'état, glissement au générique.

   groupe nominal indéfini : anaphore, anaphore hyperonyme, glissement au générique.

Cette liste sera reprise dans la section 3.3.1 afin d'étudier les situations de co-référence et les structures linguistiques liées au phénomène de saillance.

2.4.3. Prise en compte de la saillance

circonstances d'appel à la saillance :

Un objet saillant est un objet qui se distingue des autres objets, qui est mis en valeur. La notion de saillance dénote le degré de cette mise en valeur. Ce nom commun n'existe pas dans le vocabulaire français mais son utilisation et son sens se déduisent facilement de l'adjectif "saillant".

Classiquement, la saillance en langage naturel est utilisée lors de la description d'objets : pour désigner un objet, on indiquera quelques unes de ses caractéristiques, et en particulier sa ou ses caractéristiques saillantes. La principale application de la saillance se trouve dans le dialogue lorsque les deux interlocuteurs n'ont pas de connaissance commune des référents utilisés par l'un ou par l'autre. La thèse de Philip Edmonds [Edmonds 1993] contient la description d'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. Afin que l'itinéraire soit clair, le premier (le "speaker") 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 veras un immeuble très haut". Le second (le "hearer") accepte la description ou demande des précisions lorsque la caractéristique ne lui semble pas pertinente : "haut comment ?". Lorsque le "hearer" accepte la description de l'itinéraire, c'est qu'il a confiance dans sa validité.

A première vue, plus le "speaker" emploie de descripteurs, plus il y a de chances pour que le "hearer" 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 (cf. les interprétations des maximes de Grice dans [Dale Reiter 1996]). L'idéal est de ne laisser dans le message que les informations pertinentes, c'est-à-dire saillantes. L'homme fait des desciptions courtes non pas en minimisant la longueur des expressions mais en utilisant des informations saillantes [Reiter Dale 1992].

La saillance dans une situation de communication homme-machine n'a été l'objet que de très peu de recherches. Un objet devrait a priori être décrit sans ambiguïté pour que la communication soit efficace. Lorsqu'il y a ambiguïté sur une référence, la commande est rejettée ou, au mieux, une précision est demandée. Ce n'est pas gênant pour l'utilisateur car celui-ci peut toujours reformuler une commande ou faire disparaître l'ambiguïté en changeant de contexte. Par exemple si l'utilisateur veut désigner une chaise bleue dans un environnement où deux chaises bleues sont visibles, il peut toujours se rapprocher de la chaise souhaitée pour faire disparaître l'autre (bien que cette solution soit contraignante). S'il emploie une expression telle que "déplace la chaise bleue" alors que les deux sont visibles, une interaction stricte refusera la commande et une interaction plus souple fera apparaître les deux chaises bleues avec un rendu particulier, demandant de faire un choix entre les deux (via une désignation gestuelle, par exemple). La prise en compte de la saillance consisterait à appliquer l'action de déplacement sur la chaise bleue saillante. En effet, l'utilisateur peut employer l'expression "déplace la chaise bleue" parce qu'il lui semble évident que la commande ne peut porter que sur la chaise bleue qu'il considère comme saillante.

L'objet de la section 3.3 est de déterminer quels sont les critères qui font qu'un utilisateur trouve évident que sa commande porte sur un objet saillant, et quelles sont les caractéristiques d'un objet qui concourent à sa saillance. Nous verrons ensuite comment traiter chacune des ses caractéristiques.

définitions et hiérarchies :

Le phénomène de saillance est jusqu'à présent relativement peu étudié. Quelques travaux rapportés dans la littérature s'intéressent à la description d'environnements inconnus dans lesquels on doit se repérer. C'est dans ce cadre que [Lynch 1960] identifie le point de vue, la familiarité et les buts de tâche courants comme des facteurs importants, et que [Devlin 1976] 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. Ces aspects de la saillance sont importants et nous y reviendrons, d'autant plus qu'ils peuvent être utilisés de manière générique.

Egalement dans le cas de description de lieux dans une ville (pour un itinéraire routier), [Davis 1989] utilise une classification des repères visuels que l'on peut trouver dans une ville selon leur saillance. Cette classification est sensée représenter ce que le "speaker" pense que le "hearer" 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. L'approche de [Reiter Dale 1992] consiste 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.

Enfin, [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.

La section 3.3 présente une proposition de hiérarchie.

3. Travaux réalisés

3.1. Démarche

La section 3 présente les deux domaines sur lesquels ont portés mes travaux : celui de la désignation et celui de la saillance.

La désignation n'a pas posé de problèmes particuliers au niveau de la démarche : une fois les besoins clairement définis et les concepts classiques étudiés, le travail à réaliser et la manière d'implémenter allaient de soi.

La saillance est un concept peu présent dans les activités de recherche et, lorsque c'est le cas, les approches ne sont pas très approfondies. Les travaux effectués sur la saillance portent le plus souvent sur des situations très simples qui ne présentent que des ambiguïtés faciles à résoudre. Au début de mes recherches, j'ai imaginé un certain nombre de situations présentant de réelles ambiguïtés. Je me suis vite aperçu que les avis sur chaque ambiguïté pouvaient fortement diverger selon les personnes interrogées. En fait, dans certaines situations, les avis étaient partagés avec des scores de l'ordre de 50% – 50%. J'ai alors été amené à définir le concept de saillance et à étudier certaines de ses facettes. J'ai cherché quelle était la part de subjectivité pour chacune de ces facettes, de manière à proposer les situations les moins sensibles. J'ai regroupé une quinzaine de situations (quelques unes d'entre elles sont étudiées dans la section 3.3.3) et j'ai réalisé un questionnaire à choix multiples que j'ai soumis à une trentaine de personnes. L'éventail des personnes interrogées comprenait des informaticiens, des linguistes et des mathématiciens (ni informaticiens ni linguistes).

La méthode que j'ai utilisée pour construire ce questionnaire est la suivante : après avoir défini mon modèle d'utilisation de la saillance par la commande vocale, j'ai décrit en quelques lignes le contexte et les contraintes qui devaient être suivies. Chaque situation a ensuite été présentée de la manière suivante : un schéma montre l'environnement 3D avec les objets susceptibles d'être désignés par la commande vocale ; un texte très court présente l'énoncé de cette commande vocale et le contexte dans lequel elle est effectuée (avec ou sans geste de désignation, généralement en dehors de tout dialogue) ; les réponses possibles correspondent à différents choix d'objets sur lesquels peut s'appliquer l'énoncé. Parmi ces réponses se trouve celle qui me semble le mieux correspondre à mon modèle, c'est-à-dire celle sur laquelle j'espère que les réponses se porteront, afin de prouver que mon modèle est adéquat. Les résultats assez partagés que j'ai récoltés m'ont montré qu'il était difficile de trouver une modélisation satisfaisante de la saillance. J'ai néanmoins tiré de ces résultats quelques commentaires constructifs qui m'ont aidé dans mes recherches.

A propos des algorithmes, en particulier de l'algorithme de regroupement que nous verrons en détail dans la section 3.3.4, de nombreuses situations ont été testées sur le papier avant toute implémentation. Cette méthode a pris quelques jours mais l'implémentation en a été grandement facilitée puisque toutes les étapes de l'algorithme, spécifiées correctement et au niveau le plus détaillé, ont donné des résultats corrects dès la première exécution. Les tests de validité ont été effectués sur une trentaine de situations représentatives non seulement des situations les plus courantes dans un

travail d'aménagement d'intérieur, mais aussi de situations plus particulière permettant de tester tous les aspects de l'algorithme.

3.2. Désignation

3.2.1. Notion de désignation floue

On se place dans cette section dans le cadre d'une interaction utilisant les périphériques 2D classiques (écran, clavier, souris) et basée sur le pointeur.

La sélection d'un objet dans l'environnement virtuel se fait en cliquant sur sa représentation avec le bouton gauche de la souris. La désignation ne devrait pas être plus contraignante si on veut qu'elle soit utilisée. Un clic sur le bouton droit de la souris serait un choix possible (dans COVEN, le choix retenu est un clic sur le bouton gauche de la souris tout en appuyant sur la touche Shift du clavier). Ce mode d'interaction est discret et ne permet donc pas la création dynamique d'un volume entourant l'objet désiré.

Nous nous rabattons ainsi sur la méthode classique du lancé de rayon : comme pour la sélection, un lancé de rayon est effectué dans la direction indiquée par le pointeur. Il peut être représenté dans la scène 3D par une droite. Le lancé de rayon s'arrête sur le premier objet rencontré. A la différence de la sélection, cet objet n'est pas forcément l'objet désigné : il peut s'agir du sol ou d'un mur. C'est le cas par exemple si l'objet est une chaise de profil, c'est-à-dire ne présentant que peu de surface visible au participant, et si le clic a été effectué 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 a priori à quelle distance du participant l'objet se trouve, cette zone doit commencer au niveau du participant et doit s'étendre au delà du point d'arrêt du rayon. En outre, plus l'objet se trouve éloigné du participant et plus la désignation peut être imprécise. C'est pourquoi notre choix de forme de zone consiste en un cône ayant l'œil du participant comme sommet et la direction de désignation comme axe. L'angle au sommet du cône n'est pas déterminé dynamiquement et constitue donc un paramètre empirique.

3.2.2. Implémentation et résultats

L'implémentation de la construction de la zone lorsqu'un geste de désignation est effectué ne pose pas de problème particulier.

Il n'en est pas de même pour la recherche des objets dans la zone. Pour déterminer ces objets, on considère la sphère englobante de chaque objet présent dans l'environnement : on teste l'inclusion ou l'intersection de cette sphère dans le cône. Si ce test est positif, on retient l'objet. Une première approche consistait à considérer les trois cas suivants :

- Désignation d'un seul objet, c'est-à-dire geste de désignation associé à une commande vocale contenant un groupe nominal démonstratif au singulier : parmi les objets retenus, on prend le plus proche de l'axe du cône. Si la zone ne contient aucun objet, un message d'avertissement est indiqué à l'utilisateur.

- Désignation de plusieurs objets en nombre indéterminé, c'est-à-dire geste de désignation associé à une commande vocale contenant un groupe nominal démonstratif au pluriel : on prend tous les objets retenus. Si la zone ne contient aucun objet ou n'en contient qu'un seul, un message d'avertissement est indiqué à l'utilisateur.
- Désignation de plusieurs objets en nombre déterminé, c'est-à-dire geste de désignation associé à une commande vocale contenant un groupe nominal démonstratif au pluriel avec un adjectif numéral *n* : parmi les objets retenus, on prend les *n* plus proches de l'axe du cône. Si la zone ne contient pas suffisamment d'objets, un message d'avertissement est indiqué à l'utilisateur.

Les problèmes liés à cette approche concernent les regroupements : lorsque l'on effectue par exemple un geste de désignation vers un groupe de trois chaises et que l'on émette la commande vocale : "peints ces trois chaises en bleu", il arrive que l'opération s'effectue sur deux des chaises désirées et sur une chaise qui était dans la trajectoire de la désignation mais qui n'était pas sur le même plan que le groupe de trois chaises. Modifier l'angle du cône ne change rien au problème et il est alors nécessaire de reconsidérer la méthode de recherche des objets dans le cas d'un pluriel déterminé.

Une deuxième approche consiste à effectuer des regroupements dans l'ensemble des objets trouvés dans la zone : les objets sont partitionnés en groupes et, dans le cas d'un pluriel déterminé, on cherche prioritairement un groupe contenant le bon nombre d'objets. Un algorithme de regroupements est étudié dans la section 3.3.4. Son implémentation et les résultats obtenus s'appliquent aussi bien à la désignation d'un groupe d'objet qu'à la saillance découlant des groupes.

3.3. Saillance

Comme nous l'avons vu à la section 2.4.3, la notion de saillance appliquée à un objet dans une scène correspond au degré de particularité de cet objet par rapport aux autres objets. La saillance peut être utilisée lorsqu'une commande de l'utilisateur aboutit à une ambiguïté, afin de proposer une solution à cette ambiguïté. L'objet de la section 3.3.1 est d'étudier les différents cas possibles qui peuvent aboutir à une ambiguïté et pour lesquels l'utilisation de la saillance est envisageable. Les caractéristiques qui font qu'un objet devient saillant peuvent être de plusieurs natures. L'objet de la section 3.3.2 est de dresser la liste de ces caractéristiques, de les classer et d'étudier les manières d'effectuer leur traitement. Nous nous intéresserons en particulier à la saillance due à la disposition spatiale des objets dans la scène : la section 3.3.3 explore les techniques classiquement utilisées dans ce domaine, la section 3.3.4 expose le choix et les adaptations que j'ai été amené à faire, et la section 3.3.5 présente les résultats que j'ai ainsi obtenus.

3.3.1. Utilisation en langage naturel

Dans toute la section 3.3, nous nous plaçons dans le cadre d'une application de type aménagement d'intérieur, c'est-à-dire une application dont les tâches consistent à manipuler des objets visualisés sur une scène. Le geste de désignation est autorisé et peut être associé à la commande vocale, dont les énoncés sont du type : action sur un ensemble d'objets, cet ensemble étant le résultat d'une référence.

Le geste de désignation n'est ici intéressant que s'il est utilisé en tant que désignation floue d'un ou de plusieurs objets. Il a ainsi un rôle de spécification d'un sous-ensemble des objets de la scène.

Il n'est pas nécessaire de reprendre tous les types de groupes nominaux étudiés dans la section 2.4.1 afin de déterminer ceux qui peuvent amener à un traitement relatif à la saillance des objets. Il suffit de repérer quelles figures de style peuvent amener à la saillance et de retrouver les types d'énoncés correspondant. Il apparaît en effet clairement que tous les types d'anaphores réfèrent à un ou plusieurs objets précis contenus dans l'historique du dialogue et donc que la saillance ne peut être utilisée que lors d'une désignation directe. Une désignation directe peut se faire :

- dans le cas d'une **référence** simple (cf. section 2.4.2) par un groupe nominal défini : "le X" dans un environnement contenant plusieurs X, "les X" dans un environnement contenant plusieurs X, tout ceci sans geste de désignation et sans que l'historique du dialogue ne contienne de X[9].
- dans le cas d'une **co-référence asynchrone** par un groupe nominal défini : "le X" dans un environnement contenant plusieurs X, sans geste de désignation et l'historique de dialogue contenant plusieurs X. L'emploi de "les X" dans le cas d'une co-référence asynchrone se résout en considérant l'anaphore.
- dans le cas d'une **co-référence synchrone** par un groupe nominal défini, par un groupe nominal démonstratif ou par un pronom démonstratif, bien que tous ces emplois soient maladroits : ils correspondent à une restriction de l'espace visuel et à une désignation dans cet espace restreint. Maintenant que nous avons étudié le geste de désignation, on peut considérer que le traitement doit être effectué dans la zone floue de désignation : "le X" est utilisé pour extraire le X saillant lorsque la zone contient plusieurs X (si elle n'en contient qu'un, l'appel au contexte référentiel relatif à la saillance n'est pas effectué). Le geste de désignation implique l'emploi du démonstratif mais le traitement réalisé dans la zone de désignation est le même que pour le groupe nominal défini. Enfin, l'emploi du démonstratif est autorisé car il utilise l'historique du dialogue pour retrouver X mais il ne réfère pas forcément au(x) même(s) X. L'emploi d'un groupe nominal démonstratif pluriel peut cependant être ambigu : lorsque l'on désigne par le geste une chaise fixe dans un environnement contenant des chaises fixes et des chaises à roulettes, l'expression "ces chaises" peut sous-entendre "les chaises de ce type". Il s'agirait alors d'un glissement au générique et la saillance des chaises ne serait pas utilisée. Cette expression est néanmoins linguistiquement incorrecte et, même si nous l'utilisons parfois en langage parlé, son emploi reste rare. C'est pourquoi nous ne l'évoquerons plus.

Comme on l'a déjà dit, aucun de ces trois types d'emplois n'est vraiment satisfaisant du point de vue de la linguistique : ils sont utilisés en dernier recours. Le cas le plus fréquent faisant appel à la saillance est l'emploi de "le X" alors qu'il y a plusieurs X dans l'environnement. L'ambiguïté porte sur l'identification du X référencé.

Lorsque le système est confronté à une ambiguïté qui reste présente dans chacun des quatre contextes référentiels explicités dans la section 2.4.2, il cherche à la résoudre de la manière suivante :

1. Il suppose une erreur de reconnaissance : comme il n'y a pas d'ambiguïté avec l'expression "les X" et que la différence entre "le X" et "les X" est phonétiquement minime, on peut

[9] X représente une catégorie d'objet avec ou sans groupe qualificatif.

supposer que le module de reconnaissance vocale n'a pas tout simplement fait une confusion.

2. Il suppose une erreur de l'utilisateur. Dans l'exemple donné, le système peut supposer que l'utilisateur n'a cru voir qu'un seul X.

Dans ces deux cas, un message doit être indiqué à l'utilisateur afin que celui-ci recommence sa commande vocale.

La saillance apparaît comme un troisième moyen de lever l'ambiguïté : en supposant que la saillance d'un X a été considérée par l'utilisateur lors de sa commande vocale, on peut retrouver le X en quantifiant et en prenant en compte la saillance de chaque objet.

Les commandes vocales dans COVEN sont énoncées en anglais. Ceci ne pose pas de problème dans la mesure où les mécanismes de langage naturel mis en jeu sont sensiblement équivalents dans les deux langues. En effet, comme le montre le tableau ci-dessous, chaque type de groupe nominal français a un équivalent en anglais. Les groupes nominaux démonstratifs sont employés sans marque déictique dans COVEN. D'autre part, les déictiques purs sont utilisés indifféremment. Notons enfin que les erreurs de reconnaissance sont tout aussi possibles en anglais qu'en français : nous avons vu la confusion possible entre "le X" et "les X" (il n'y a qu'un seul son de différent entre "*la* chaise" et "*les* chaises"). En anglais, la confusion peut porter sur le nom : entre "the *chair*" et "the *chairs*".

Au niveau du calcul de référents, les syntaxes sont très proches comme le montre le tableau suivant :

syntaxe	**exemples en français**	**équivalents en anglais**
groupes nominaux définis	*l'*objet *le* bureau *la* chaise bleue *les* chaises en plastique bleu	*the* item *the* desk *the* blue chair *the* blue plastic chairs
groupes nominaux démonstratifs	*cet* objet *ce* bureau *cette* chaise bleue *ces* chaises en plastique bleu	*this* item *this* desk *this* blue chair *these* blue plastic chairs
groupes nominaux démonstratifs avec marque déictique	*cet* objet-*ci* *ces* objets-*ci* *cet* objet-*là* *ces* objets-*là*	*this* item *these* items *that* item *those* items
pronoms définis	déplace-*le* déplace-*les*	move *it* move *them*
pronoms démonstratifs avec marque déictique ou anaphorique	déplace *ceci* déplace *celui-ci* déplace *ceux-ci* déplace *cela* déplace *celui-là* déplace *ceux-là*	move *this* move *this one* move *these (ones)* move *that* move *that one* move *those (ones)*
déictiques purs	*ici* *là* *là-bas*	*here* *there* *over there*

3.3.2. Définitions

proposition de hiérarchie :

Notre approche consiste à reprendre le principe de hiérarchie de [Reiter Dale 1992] (cf. section 2.4.3) et à le généraliser en tenant compte des problèmes de subjectivité et de dépendance au contexte évoqués par [Lynch 1960] et [Devlin 1976]. L'usage des coefficients de [Edmonds 1993] est surtout intéressante dans une optique de génération. C'est pourquoi nous ne les utiliseront pas.

Nous ne nous intéressons ici qu'à la saillance visuelle, c'est-à-dire au contraste visuel entre un objet et les autres objets de la scène. Il va de soi que d'autres types de perception peuvent entrer en jeu, notamment la perception d'émissions sonores. Le contexte est donc réduit aux seuls éléments graphiques pouvant être perçus à l'écran à un instant précis.

Afin d'étudier la saillance possible d'une caractéristique d'un objet, il est nécessaire de dresser la liste des caractéristiques pouvant distinguer visuellement un objet par rapport à d'autres. Une classification simple et générique pourrait être :

1. **saillance par sa catégorie :**

   Premier exemple : dans une scène contenant plusieurs chaises et une table, la table est saillante.

   Deuxième exemple : dans une scène contenant des chaises dont une chaise à roulettes (et des objets qui ne sont pas des chaises), la chaise à roulette est saillante.

   L'intérêt du deuxième exemple par rapport au premier est qu'il n'existe parfois qu'un seul mot pour désigner deux objets de catégories voisines. C'est en particulier souvent le cas dans les interfaces en langage naturel qui sont limitées en vocabulaire.

   Il est à noter que la catégorie est toujours explicite dans la commande vocale : même dans un environnement ne contenant que des chaises, et donc pour lequel l'expression "l'objet" serait suffisante, on utilise préférentiellement "la chaise".

2. **saillance par ses caractéristiques physiques :**

   Parmi les caractéristiques physiques d'un objet, citons la taille, la géométrie, le matériau, la couleur, la texture.

   Exemples : dans une scène contenant des chaises dont une chaise pour enfant, bien plus petite que les autres, cette dernière est saillante ; dans une scène contenant des chaises dont une a un pied cassé, cette dernière est saillante.

   Il s'avère en outre difficile de donner une priorité à l'une des caractéristiques physiques par rapport à une autre. On peut reprendre un exemple intéressant de [Reiter Dale 1992] : parmi deux chiens dont l'un est grand et blanc et l'autre est petit et noir, on veut désigner ce dernier. Il s'avère que, dans ce cas, on utilisera plus facilement l'expression "le chien noir" que "le petit chien". Ceci pourrait vouloir dire que la caractéristique couleur est légèrement plus saillante que la caractéristique taille. Nous pensons plutôt que la couleur est utilisée ici car le contraste entre noir et blanc est précis, contrairement au contraste entre petit et grand qui reste indéfini.

3. **saillance par ses fonctionnalités :**

   [Devlin 1976] distingue la saillance visuelle de la saillance due aux fonctionnalités de l'objet. Or nous pensons que ces fonctionnalités peuvent être perçues visuellement. 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.

   Premier exemple : dans une scène contenant plusieurs ordinateurs, un ordinateur allumé est saillant si tous les autres ordinateurs sont éteints.

   Deuxième exemple : dans une scène contenant un bureau et plusieurs chaises, une chaise étant accolée au bureau, les autres étant détachées de tout meuble, la chaise accolée au bureau est saillante. Cet exemple peut être repris en considérant que parmi les chaises, il n'y en a qu'une qui fait face au bureau.

4. **saillance spatiale ou saillance par sa localisation dans la scène par rapport aux autres objets et au participant :**

   Le terme localisation d'un objet dans la scène sous-entend la position 3D de l'objet et son orientation.

   Exemples dans une scène contenant plusieurs chaises et éventuellement des objets qui ne sont pas des chaises : une chaise très proche du participant est saillante si les autres chaises sont éloignées ; une chaise isolée est saillante si toutes les autres chaises sont groupées ; une chaise sur une table est saillante si toutes les autres chaises sont sur le sol ; une chaise renversée est saillante si les autres ne le sont pas ; une chaise tournant le dos au participant est saillante si les autres lui font face[10].

5. **saillance par son incongruité :**

   Un objet dans une situation incongrue est en infraction avec une règle implicite, culturelle ou fonctionnelle.

   Exemple : dans une scène contenant plusieurs objets dont un dans une position inhabituelle, cet objet est saillant. Cet exemple est à rapprocher de l'exemple de la chaise renversée dans une scène où les autres chaises ne le sont pas : une chaise renversée n'est pas forcément en situation inhabituelle, par exemple si elle est posée les pieds en l'air sur une table pour permettre à la femme de ménage de faire son travail. On revient dans ce cas à la saillance par la localisation dans la scène. Par contre, une chaise renversée à même le sol est vraisemblablement en situation inhabituelle et relève alors de la saillance par le comportement. Ce type de saillance dépendant étroitement du contexte est difficile à repérer.

6. **saillance par sa dynamique :**

   Premier exemple : dans une scène contenant un objet animé et plusieurs objets inanimés, l'objet animé est saillant.

   Deuxième exemple : dans une scène contenant un objet en mouvement et plusieurs objets statiques, l'objet en mouvement est saillant.

[10] Ces deux derniers exemples peuvent être interprétés différemment, comme on le voit ci-après.

7. **saillance par sa mise en évidence par l'application :**

   Exemple : dans une scène contenant plusieurs chaises dont une chaise est mise en valeur par un rendu visuel propre à l'application, cette dernière chaise est saillante. Parmi les rendus couramment utilisés dans les environnements virtuels, notons : le rendu en fil de fer, les textures transparentes, l'affichage de la boîte englobante de l'objet. Ces rendus ont des significations particulières liées à l'interaction ou à la tâche. Dans COVEN par exemple, les objets sélectionnés apparaissent avec leur boîte englobante visible, la couleur de celle-ci dépendant du participant à l'origine de la sélection.

8. **saillance indirecte :**

   En allant à fond dans l'exploration des possibilités de saillance, on peut également considérer qu'une chaise devant une bureau est saillante par rapport à une série de chaises devant une table, en particulier si le bureau est recouvert de dossiers alors que la table est recouverte d'une nappe et de couverts. Cette saillance peut être expliquée en supposant qu'il existe un certain transfert de caractéristiques du bureau vers la chaise, plus généralement entre deux objets accolés. Dans un même ordre d'idée, la chaise accolée au seul bureau de la pièce est saillante car ce bureau l'est probablement. Autrement dit, il peut également y avoir transfert de saillance entre deux objets accolés.

Une définition générale d'un objet saillant visuellement serait donc un objet qui se distingue des autres objets de la scène par au moins l'un des types de caractéristiques cités ci-dessus. Cette liste est probablement non exhaustive et affinable.

subjectivité et dépendance au contexte :

Nous avons vu que la subjectivité pouvait affecter ces caractéristiques. 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). L'intérêt de cet exemple est d'insister sur le fait que même un objet fait par nature pour être vu et donc pour être saillant peut voir sa saillance remise en cause par la subjectivité. Parmi les types de saillance évoqués ci-dessus, la saillance par les catégories physiques et par la localisation dans la scène peuvent être affectées par la subjectivité (les autres types de saillance font intervenir des caractéristiques trop précises pour qu'il y ait confusion). En effet, 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 familiarise avec certains types d'objets : un objet d'un autre type sera saillant pour nous et peut-être pas pour une autre personne qui l'aura côtoyé pendant des années. La saillance par la localisation dans la scène pose également de nombreux problèmes de subjectivité : la proximité peut être prépondérante pour une personne alors que la distance à l'axe de visée peut l'être pour une autre (nous en reparlerons à la section 3.3.3). La plupart des phénomènes de saillance peuvent donc être affectés par la subjectivité. Comment alors aborder la prise en compte de ces phénomènes ? Clairement, l'approche devra se fonder sur des heuristiques paramétrables, adaptables aux particularités perceptives d'un utilisateur. La seule solution est de trouver la méthode de résolution la plus générique possible et de ne rien conclure dans les cas particuliers. C'est ce que nous essayerons de faire pour la saillance par la localisation dans la section

3.3.3. Cet aspect du problème relève de la psycholinguistique et nous ne nous y attarderons par plus qu'il n'est nécessaire.

Maintenant que toutes les caractéristiques d'un objet ont été énumérées et étudiées, il nous faut spécifier la façon dont elles peuvent être utilisées.

usage de ces caractéristiques :

Il s'agit maintenant de déterminer, dans la liste de ces caractéristiques visuelles, celles que l'on peut raisonnablement retenir et exploiter lors du calcul de référents, c'est-à-dire les caractéristiques à l'origine de points de vue référentiels permettant de résoudre une ambiguïté. Ces caractéristiques doivent être implicites (c'est-à-dire non explicitement mentionnées dans l'énoncé), et suffisamment fortes pour déterminer un contexte référentiel.

La saillance par la catégorie ne semble pas pertinente. Il est en effet peu vraisemblable que dans une commande vocale, l'utilisateur désigne l'objet autrement que par sa catégorie.

La saillance par les caractéristiques physiques pose un peu le même problème : 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" pour désigner celle-ci, même si cette expression est quelque peu coûteuse. (Remarque : Reste le cas où une information sur une caractéristique physique est sous-entendue par l'action : dans le même environnement que précédemment, la commande vocale "peints la chaise en rouge" utilise l'expression minimale "la chaise" mais le fait de vouloir la peindre en bleu sous-entend qu'elle n'est pas bleue. Ce phénomène appelé axiologie permet de résoudre l'ambiguïté sans faire appel à la saillance. La saillance par les caractéristiques physiques ne semble donc pas pertinente).

La saillance par les fonctionalités semble pertinente : dans l'exemple de l'ordinateur allumé parmi plusieurs ordinateurs éteints, il est probable que l'expression "l'ordinateur" référera à celui qui est allumé. Une manière simple et suffisante de traiter ce type de saillance est la suivante : le groupe nominal ne contenant aucune précision sur la fonction du référent, il faut séparer les référents possibles en deux catégories correspondant aux deux fonctions possibles (s'il y a plus de deux fonctions possibles, l'ambiguïté multiple ne peut pas être résolue) ; on calcule alors le nombre de référents possibles appartenant à chacune de ces deux catégories : si une et une seule des catégories contient exactement un référent possible, ce référent est retenu ; dans le cas contraire, l'ambiguïté ne peut pas être résolue.

La saillance par la localisation dans la scène doit également être traitée : 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 "la chaise" pour la désigner. Le traitement est ici complexe puisqu'il fait intervenir de nombreux paramètres : la perception de la scène par le participant, la disposition des objets les uns par rapport aux autres. Le traitement de ce type de saillance a été l'objet principal du développement réalisé pendant ce mémoire et est donc étudié en détail ci-dessous (aux sections 3.3.3, 3.3.4 et 3.3.5).

La saillance par l'incongruité doit être traitée. Le traitement consiste à repérer l'objet qui est en situation inhabituelle, ce qui ne pose pas de problèmes à la condition suivante : chaque caractéristique de l'objet pouvant donner lieu à une incongruité doit pouvoir être comparée à 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 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. Ces trois états de référence peuvent être décrits par le vecteur orientation et par l'altitude de la chaise, qui sont des caractéristiques quantifiables.

La saillance par la dynamique de l'objet doit elle aussi être traitée : un objet en mouvement dans un environnement où tous les autres objets sont statiques est très facilement repérable, à tel point qu'il retient probablement toute l'attention de l'utilisateur. Sa désignation par un groupe nominal défini minimal est donc tout à fait envisageable. Le traitement à effectuer est identique à celui de la saillance par l'incongruité.

La saillance due à la mise en évidence de l'objet par l'application est la première à être traitée car elle est à la base de l'interaction. L'affichage d'un objet avec un rendu particulier est par exemple effectué pour montrer que la suite de l'interaction s'appliquera sur cet objet. Il n'y a dans ce cas aucune ambiguïté et aucun traitement supplémentaire n'est donc à envisager.

La saillance indirecte est difficile à repérer. A mon avis, le transfert de saillance ne doit pas être l'objet de traitements car l'utilisateur exprimera probablement le lien entre les deux objets : il employera par exemple la relation fonctionnelle : "la chaise du bureau".

On déduit de tout cela que dans un environnement comprenant une interaction en langage naturel, il semble intéressant d'explorer les contextes liés à la saillance fonctionnelle, spatiale, dynamique et incongrue en complément des contextes référentiels linguistiques. Dans le cadre de COVEN, la saillance par les fonctionnalités n'est pas traitée car les objets ne possèdent qu'un seul état fonctionnel ; la saillance dynamique n'est pas traitée car tous les objets sont inanimés et immobiles ; la saillance par l'incongruité n'est pas traitée car un objet ne peut être mis en situation inhabituelle que par l'utilisateur, ce qui est contraire au but d'aménagement de la scène (lors de la création d'un objet, cet objet apparaît dans une situation habituelle et, lorsque l'utilisateur le manipule, on suppose que la situation finale de l'objet l'est aussi). Il ne reste ainsi que la saillance spatiale à traiter.

3.3.3. Approches sur la saillance spatiale

Le but de cette section est de déterminer des objets candidats pour la résolution d'une référence en utilisant la saillance spatiale. Le système fait une proposition à l'utilisateur : dans tous les cas, une confirmation est démandée.

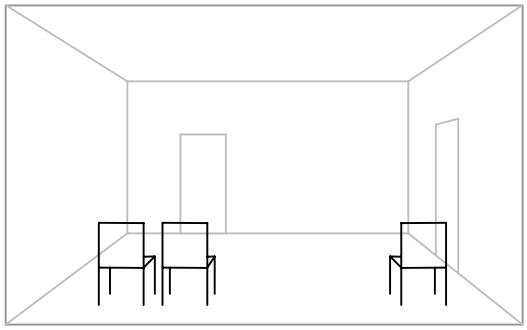
La saillance spatiale dans un environnement en 3D peut être décomposée selon deux axes :

- saillance due au groupement des objets : par exemple, saillance d'un objet isolé.
- saillance due à la distance de l'objet au participant : par exemple, saillance d'un objet proche.

Un autre type de saillance spatiale pourrait dépendre de l'orientation des objets, sans tenir compte de leur positions : dans un environnement composé de plusieurs chaises tournant le dos au participant, une chaise tournée vers lui peut être considérée comme saillante. Comme nous allons le voir; ce type de saillance est difficile à traiter.

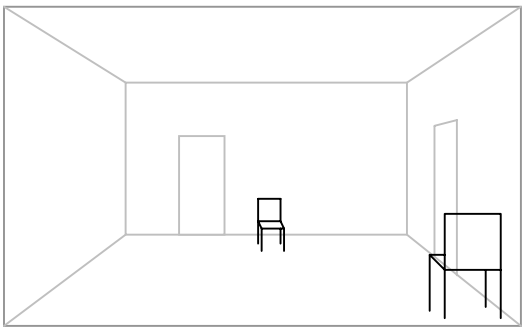
La saillance due au groupement des objets peut donner lieu à plusieurs cas de figure : dans un environnement contenant par exemple au moins trois chaises ainsi que d'autres objets, l'expression "la chaise" pourra être utilisée si une des chaises est isolée et si les autres sont groupées (en un ou plusieurs groupes). De même, l'expression "les chaises" pourrait être utilisée s'il n'existe qu'un seul groupe de chaises, les autres chaises étant isolées et dispersées. Ce dernier emploi est néanmoins

critiquable. En fait, comme on le voit sur le schéma ci-dessous, autant l'expression "la chaise" peut désigner la chaise isolée, autant l'expression "les chaises" désigne plutôt les trois chaises visibles que les deux chaises groupées.



En effet, une expression du type "la chaise de droite" serait préférable à "la chaise" pour désigner la chaise *c* bien que cette dernière expression peut d'une part être employée, d'autre part être comprise. (Il faut en effet considérer que la machine est un interlocuteur et il faut donc lui parler comme on parlerait à un humain, c'est-à-dire que les expressions utilisées doivent être naturelles et compréhensibles[11]). L'expression "les chaises" n'est pas naturelle pour désigner les chaises *a* et *b* (les expressions "les chaises de gauche" voire "les deux chaises" sont bien plus naturelles). D'autre part, un humain à qui on indiquerait "les chaises" dans ce cas comprendrait sûrement les chaises visibles. C'est en tout cas ce qu'a révélé mon questionnaire dans lequel cette situation était présentée.

La saillance due au point de vue du participant peut donner lieu à plusieurs cas de figure : en gardant l'exemple de l'expression "la chaise", une chaise proche du participant alors que les autres sont dans l'arrière plan pourra ainsi être désignée, de même qu'une chaise dans l'axe de vue du participant alors que les autres sont en bordure du champ. Il n'existe pas vraiment de hiérarchie entre ces deux saillances : comme le montre le schéma ci-dessous, il est difficile de dire quelle est la chaise la plus saillante lorsque l'environnement contient deux chaises, l'une étant proche mais en bordure du champ, l'autre étant dans l'axe de vue mais éloigné du participant.



[11] Ceci se rapproche de la notion de "speaker" et de "hearer" vue précédemment : une expression du type "la chaise" pourrait être comprise par le "hearer" mais ne serait pas employée par le "speaker".

En effet, quelle est la chaise perçue en premier ? C'est difficile à dire. Avec une interface immersive, c'est-à-dire avec un casque de visualisation, la réponse serait peut-être plus facile à donner : la chaise en bordure de l'image serait vraiment en bordure du champ de visée de l'utilisateur et serait donc moins bien perçue.

Les groupements et le point de vue du participant doivent être pris en compte simultanément. Les situations à traiter peuvent alors être très diverses comme le montrent les quatre schémas suivants :

schéma 1 :

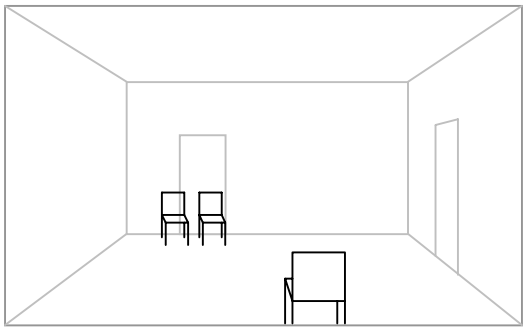


schéma 2 :

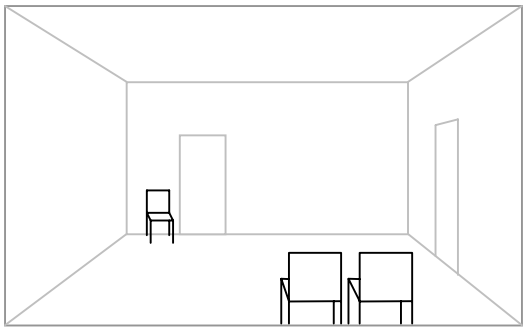


schéma 3 :

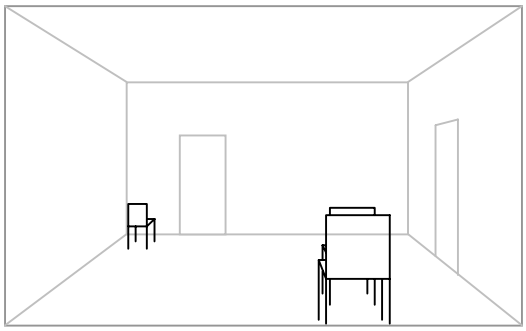
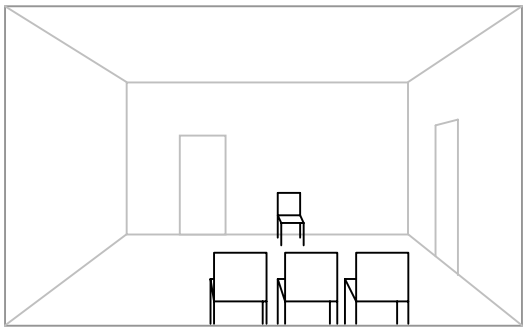


schéma 4 :

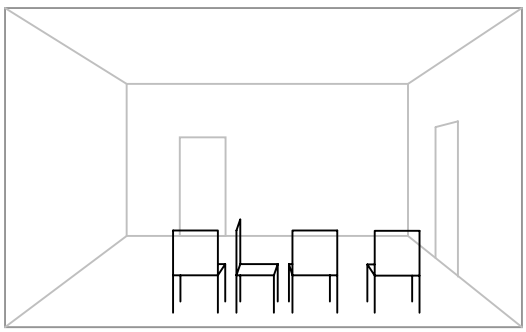


En disposant (schéma 1) le groupe de chaises à l'arrière plan et la chaise isolée au premier plan, on renforce le contraste. La chaise isolée est particulièrement saillante car elle est proche du participant et dans son axe de visée. De même, dans le schéma 2, le contraste entre le groupe et la chaise isolée est renforcé par leur disposition sur des plans différents. Cependant, dans les deux cas, l'expression "les chaises" désignera toujours les trois chaises visibles. Pour le schéma 1 en particulier, une explication possible est la suivante : lorsqu'un objet est particulièrement saillant, il doit appartenir à l'ensemble référencé. La chaise au premier plan et dans l'axe de visée du participant doit ainsi être associée aux deux chaises groupées lorsque l'expression "les chaises" est utilisée.

La situation présentée sur le schéma 3 montre un des problèmes qui apparaissent en 3D : la visibilité des objets. Une chaise peut en cacher une autre ou en cacher au moins une partie. Le schéma 3 montre un environnement composé de trois chaises dont deux groupées. Dans cette situation, l'expression "la chaise" ne désignera peut-être pas la chaise isolée mais plutôt la chaise la plus proche du participant, bien que cette dernière chaise fasse partie d'un groupe. La proximité au participant pourrait donc être un critère plus important que l'isolement. Il est à noter cependant que les personnes soumises à cette situation ont considéré en majorité que l'expression "la chaise" était incompréhensible, ce qui montre d'une part que la chaise à moitié cachée compte autant que les autres, d'autre part qu'à cause de la présence d'une chaise très proche du participant, la chaise isolée au fond n'est plus saillante.

Le schéma 4 est un exemple de situation sur laquelle les avis de personnes interrogées sont surprenants : comme on vient de le voir, l'expression "les chaises" désigne sûrement les quatre chaises visibles. Or il s'est avéré qu'un nombre non négligeable de personnes ont plutôt compris dans ce cas les trois chaises groupées au premier plan. Ceci en remet pas en question la désignation d'un groupe d'objet par un groupe nominal défini au pluriel indéfini (que nous avons choisi de dissocier de la saillance) mais prouve que le premier plan est un emplacement clef et que sa coordination avec le groupement est un critère puissant de saillance.

Nous avons déjà vu que la saillance pouvait dépendre de l'action associée au groupe nominal défini. Les schémas 1 et 2 permettent de mieux s'en rendre compte : la commande "peints les chaises en bleu" s'appliquera probablement aux trois chaises visibles alors que la commande "mets une table près des chaises" s'appliquera plutôt aux deux chaises groupées. La référence est dans le premier cas l'objet de l'action et dans le second cas la désignation d'un secteur de l'espace (donc nécessairement un objet ou plusieurs objets groupés). Dans le cadre de COVEN, la référence ne peut être qu'objet de l'action car la désignation de positions ou de secteurs de l'espace ne peut se faire que par le geste. La saillance peut donc dans ce cas être étudiée indépendamment de l'action. A titre d'exemple, le schéma suivant montre une situation dans laquelle l'expression "la chaise" ne réfère pas au même objet selon l'action (nous avions montré dans la section 3.3.2 que l'orientation était résolue par une application du principe d'axiologie) :



Les quatre chaises sont alignées, la plus à droite etant décalée un peu vers l'extérieur et l'une des trois autres étant retournée. L'expression "retourne la chaise" réfère ainsi à cette dernière alors que l'expression "pousse un peu la chaise à gauche" réfère à celle qui décalée vers la droite.

En résumé, notre approche consistera à :

- Prendre en compte les regroupements des objets, c'est-à-dire séparer les groupes d'objets des objets isolés. Lorsque la référence est au singulier, la résolution se fait s'il n'existe qu'un seul objet isolé. Lorsque la référence est au pluriel, le contexte référentiel correspondant à la saillance n'est pas utilisé dans le traitement, et on passe donc au contexte suivant qui correspond à tous les objets inclus dans le champ de perception du participant.
- Prendre en compte le point de vue du participant, selon deux critères : la proximité et la distance à l'axe de vue. Le critère de proximité est plus important que celui de distance à l'axe de vue.
- Ne pas prendre en compte le recouvrement partiel des objets : comme on l'a vu, un objet à moitié caché peut compter autant qu'un objet entièrement visible. Bien entendu, un objet entièrement caché, que ce soit par un pan de mur ou par un autre objet, n'est pas pris en compte lors du calcul de référents, de même qu'un objet non inclus dans le champ de perception du participant.

Nous allons maintenant voir comment implémenter le traitement de la saillance selon cette approche.

3.3.4. Implémentation

approches :

Une première approche tenant compte simultanément des groupements et du point de vue du participant consiste à traiter l'image 2D perçue par le participant. Chaque objet est décrit par sa taille apparente et par son abscisse et son ordonnée dans le plan. La taille apparente n'est pas proportionnelle à la distance du participant puisque certains objets peuvent être plus gros que d'autres, même entre des objets de la même catégorie. Il est donc nécessaire de tenir compte également de la profondeur.

La profondeur est un paramètre à prendre en compte lors des regroupements : si l'on regroupe les projections 2D des objets, deux objets sur l'axe de visée du participant seront très proches voire partiellement confondus sur l'écran alors que l'un pourra être devant les yeux du participant et l'autre au fond de la pièce. C'est un inconvénient de taille car nous avons vu que le contraste entre premier plan et arrière plan était au moins aussi important que le contraste entre un objet à gauche et un objet à droite. A moins de mettre en place un système de pondération des ordonnées, il est nécessaire d'effectuer les regroupements en 3D.

Nous sommes dans un environnement virtuel sur une station de travail graphique. Les algorithmes de projection et de Z-buffer sont donc cablés en hard. Les informations qu'ils calculent sont donc inaccessibles pour le soft. La profondeur doit donc être recalculée. En outre, si l'on veut appliquer le regroupement sur les projections 2D des objets, il est également nécessaire de recalculer de manière soft ces projections. Ce dernier calcul peut être simplifié en ne tenant compte que des sphères englobantes des objets (pour ne traiter ensuite que des disques dans le plan), mais se pose alors le

problème du recouvrement : on a choisi de tenir compte des objets partiellement cachés or un disque peut en cacher complètement un autre. On en déduit que la seule solution rigoureuse est de reprogrammer entièrement les algorithmes de projection et de Z-buffer. Cette solution très lourde peut néanmoins avoir des avantages : on pourrait par exemple implémenter un object-buffer comme décrit dans la section 2.3.1.

Le principal intérêt de cette solution est la prise en compte immédiate du point de vue du participant : il est très facile de déterminer quels sont les objets proches du participant, de même que ceux proches de son axe de vue. Le principal inconvénient est sa lourdeur et sa complexité. Comme on est amené à grouper les objets en 3D, on peut se demander si une solution consistant à grouper en 3D avec les coordonnées directes des objets n'est pas préférable.

Cette deuxième approche est en effet intéressante puisqu'elle ne nécessite aucune projection : il suffit d'effectuer les regroupements à partir des coordonnées des centres de gravité des objets qui sont directement accessible dans la base de données contenant les caractéristiques des objets de la scène. Cette solution a néanmoins quelques inconvénients : d'une part le point de vue du participant n'est pas pris en compte, d'autre part le centre de gravité ne suffit pas pour décrire un objet. Ce deuxième inconvénient est facile à résoudre en tenant compte des sphères englobantes. Le point de vue du participant peut d'autre part être pris en compte de deux manières : soit par une pondération avant les regroupements, soit par une pondération après les regroupements.

Afin de choisir entre ces deux méthodes, considérons deux objets dans la scène. Ces objets peuvent être :

- accolés : c'est le cas lorsqu'ils se touchent ou lorsqu'ils sont très proches l'un de l'autre. On pourrait dire par exemple que deux objets sont accolés lorsqu'il est impossible de placer un autre objet entre eux deux.
- dans un même voisinage : c'est le cas lorsque les objets ne sont ni très proches ni très éloignés. Les objets font apparemment partie d'un même groupe mais un autre objet peut par exemple est placé entre eux.
- éloignés : c'est le cas lorsque la distance entre les deux objets est telle qu'en aucun cas les objets ne peuvent être associés.

Considérons maintenant deux objets dans un même voisinage. Si le participant se place entre les deux et un peu en retrait de manière à les voir tous les deux dans son champ de vision, il en percevra un à sa gauche et l'autre à sa droite. Autrement dit il les trouvera éloignés l'un de l'autre. Si maintenant le participant se place loin des deux objets, il les percevra à peu près dans la même direction et les trouvera quasiment accolés.

Cet exemple prouve que le point de vue du participant a une influence sur les groupements, et donc qu'il doit être pris en compte par une pondération avant l'algorithme de groupement.

algorithmes :

L'ouvrage "Reconnaissance des formes" [Belaïd Belaïd 1992] fait le point sur les techniques de regroupement. Les différentes approches explicitées sont les suivantes : classification automatique, discrimination fonctionnelle, méthodes statistiques bayésiennes, *k* plus proches voisins, méthodes stochastiques, méthodes connexionnistes et méthodes structurelles. Parmi ces sept approches, seule la classification automatique est envisageable dans notre contexte. Sa théorie est la suivante :

Un objet est représenté par un vecteur de *n* composantes (équivalent à *n* caractéristiques). Classifier des vecteurs consiste à regrouper ces vecteurs en classes qui doivent vérifier les propriétés suivantes :

- **compacité** : les points représentant une classe donnée sont plus proches entre eux que des points de toutes les autres classes.
- **séparabilité** : les classes sont bornées et il n'y a pas de recouvrement entre elles.

La constitution des classes est fondée sur la notion de proximité. Nous allons donner une définition de la proximité en utilisant la notion de distance puis nous développerons deux approches qui les utilisent : le groupement hiérarchique et le groupement non hiérarchique.

distance : il est naturel de considérer qu'un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres. Il est donc nécessaire de définir la distance entre un point et une classe à partir de la distance entre points. Une distance vérifie les quatre propriétés suivantes : séparabilité (la distance entre deux points distincts est strictement positive), réflexivité, symétrie et inégalité triangulaire. Les distances classiques sont les distances de Hamming, euclidienne et du maximum. Nous considérerons 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 : le saut minimal qui consiste à prendre le minimum des distances, et le diamètre maximal qui consiste à prendre le maximum des distances. Afin de rester dans la logique du minimum, nous prendrons le saut minimal. Nous n'avons pas retenu la distance ultramétrique[12] qui s'écarte de notre application sur un espace 3D.

groupement hiérarchique : le groupement hiérarchique est une méthode de classification automatique qui 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 caractérisé par un indice d'agrégation (ou niveau) : l'indice d'agrégation d'un groupe est 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. Dans le cadre de notre application, la recherche de la partition idéale se fait en considérant d'une part des seuils (distance minimale entre deux objets pour qu'ils soient considérés comme éloignés l'un de l'autre, distance maximale entre deux objets pour qu'ils soient considérés comme accolés), d'autre part en tenant compte de la commande vocale : une expression du type "le X" impliquera par exemple la recherche d'une partition contenant plusieurs groupes et un seul objet isolé.

[12] La distance ultramétrique entre deux points se calcule en prenant des intermédiaires et en retenant le minimum des distances calculées. Son intérêt est que, quel que soit le point d'un groupe, il est à égale distance de tous les points d'un autre groupe (tout triangle est isocèle pointu). Se reporter à [Belaïd Belaïd 1992].

groupement 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 correspond à la distance entre les deux objets correspondant aux nœuds. La seconde consiste à établir une partition à partir d'éléments suffisamment représentatifs de chaque classe par un procédé iteratif qui établit les meilleurs représentants et la meilleure partition à chaque étape. Ces deux méthodes sont intéressantes lorsqu'il y a plusieurs milliers d'objets à classer et nous ne les utiliserons donc pas.

En tenant compte des choix que nous avons faits, l'algorithme général de groupement se deroulera de la manière suivante :

- En entrée, les paramètres suivants seront utilisés :
    1. Zone dans laquelle les objets seront traités : soit la pyramide de visée, soit une zone de désignation si un geste de désignation a été effectué.
    2. Nombre d'objets recherchés, par exemple : 1 pour "le X", 3 pour "les trois X".
    3. Catégorie d'objets, par exemple : "chaise" pour "la chaise en plastique bleu", aucune pour "cet objet".
    4. Liste des propriétés des objets recherchés, par exemple : "en plastique" et "de couleur bleue" pour "la chaise en plastique bleu", aucune pour "la chaise".
    5. Position et orientation du participant.
    6. Liste de tous les objets de la scène, avec pour chacun d'eux : la catégorie, la liste des propriétés (matériau et couleur), la position, le rayon de la sphère englobante.
- La première étape de l'algorithme consiste à repérer les objets perceptibles sur lesquels se fera le traitement. Dans le cas d'un geste de désignation, ce sont les objets dont les sphères englobantes sont inclues ou intersectent le cône représentant la zone de désignation. Dans le cas contraire, ce sont les objets dont les sphères englobantes sont inclues ou intersectent la pyramide de visée du participant. Le traitement s'applique à tous ces objets et non uniquement aux objets vérifiant la catégorie et les propriétés demandées. Ce choix a été fait car 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.
- Une deuxième étape consiste à préparer la classification en construisant un demi-tableau contenant les distances des objets deux à deux et en construisant la première partition, c'est-à-dire celle où chaque objet forme un groupe. Les distances sont calculées de la manière suivante : à la distance euclidienne entre les centres de gravité des deux objets, on retire 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 (0.5 dans notre implémentation) est utile lorsque les objets sont de grande taille, comme des tables par exemple : la distance entre les centres de gravité (coefficient nul) de deux tables accolées est déjà assez importante ; la distance corrigée par les rayons des sphères

englobantes (coefficient égal à 1) peut être négative[13] et doit donc être corrigée. 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 donc à cette étape que se fait la prise en compte du point de vue du participant, mais seulement selon l'éloignement : l'écart par rapport à l'axe de visée n'est pas pris en compte.

- L'algorithme proprement dit : à chaque étape, les deux groupes les plus proches sont groupés, la distance entre les deux étant le niveau d'agrégation du groupe et donc celui de la partition ainsi obtenue. La nouvelle partition est retenue en plus des précédentes, le nouveau demi-tableau de distances prend la place de l'ancien. Un exemple montrant les étapes de cet algorithme est donné ci-dessous :

  On considère 5 objets A, B, C, D et E. Une première partition est la suivante :

objet	A	B	C	D	E
groupe	G1	G2	G3	G4	G5
niveau du groupe	–	–	–	–	–

  Les distances deux à deux entre chacun de ces 5 groupes sont données dans le tableau suivant :

d	G1	G2	G3	G4	G5
G1	–	10	8	10	13
G2		–	34	2	41
G3			–	26	**1**
G4				–	29
G5					–

  La première itération regroupe les groupes G3 et G5. Soit G6 le nouveau groupe. La nouvelle partition et le nouveau demi-tableau des distances sont les suivants :

objet	A	B	C	D	E
groupe	G1	G2	G6	G4	G6
niveau du groupe	–	–	1	–	1

d	G1	G2	G4	G6
G1	–	10	10	8
G2		–	**2**	34
G4			–	26
G6				–

[13] Une distance ne doit bien entendue jamais être négative. Si la soustraction des rayons des sphères englobantes amène à un réel négatif, on considère que la distance entre les deux objets est nulle. Cela ne peut être le cas que pour des objets accolés. Ce n'est donc pas gênant.

La troisième itération regroupe les groupes G2 et G4. Soit G7 le nouveau groupe. La nouvelle partition et le nouveau demi-tableau des distances sont les suivants :

objet	A	B	C	D	E
groupe	G1	G7	G6	G7	G6
niveau du goupe	–	2	1	2	1

d	G1	G6	G7
G1	–	**8**	10
G6		–	26
G7			–

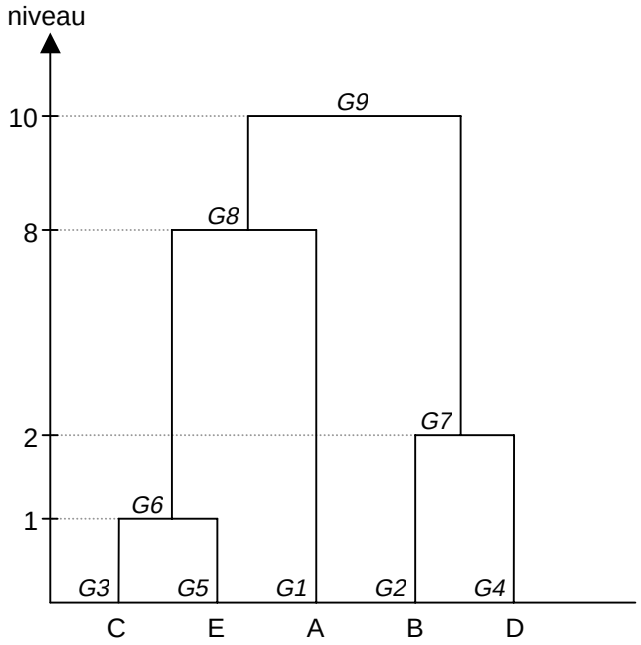
La quatrième itération regroupe les groupes G1 et G6. Soit G8 le nouveau groupe. La nouvelle partition et le nouveau demi-tableau des distances sont les suivants :

objet	A	B	C	D	E
groupe	G8	G7	G8	G7	G8
niveau du groupe	8	2	8	2	8

d	G7	G8
G7	–	**10**
G8		–

La cinquième et dernière itération regroupe les deux groupes restant, c'est-à-dire les groupes G7 et G8. Soit G9 le nouveau groupe. G9 contient les cinq objets et son niveau d'agrégation est 10.

La hiérarchie indicée correspondant est la suivante :



- La dernière étape consiste en l'analyse des partitions obtenues : on enlève les partitions dont le niveau d'agrégation est trop élevé ou trop bas, puis on étudie chaque partition en partant de celle ayant le niveau d'agrégation le plus élevé. L'étude d'une partition se fait de la manière suivante : en ne considérant que les objets ayant la catégorie et les caractéristiques recherchées, on compte le nombre de groupes contenant le nombre recherché d'objets (ce nombre peut être 1 si l'on recherche un objet isolé). Si ce nombre de groupes est nul ou supérieur strictement à 1, on passe à la partition suivante. Si ce nombre est égal à 1, on place le ou les objets correspondants dans une liste qui sera le résultat positif de l'algorithme. Si toutes les partitions donnent des résultats négatifs, l'algorithme n'est pas concluant et renvoie un code de retour indiquant au gestionnaire des contextes référentiels de passer au contexte suivant.

Il faut noter que l'algorithme ne doit être concluant que dans les cas où la saillance est vraiment importante. En effet, le but n'est pas de trouver un résultat à tout prix puisque l'ambiguïté reste tout à fait valable au niveau linguistique ; le but est de détecter les cas où parmi les objets traités, il y en a un certain nombre qui se distinguent des autres et que ce nombre correspond à celui indiqué dans la désignation faite par la commande vocale.

L'algorithme utilise trois paramètres empiriques : deux seuils appliqués aux indices d'agrégation lors de l'analyse des partitions et un coefficient de proportionnalité pour la prise en compte de la distance de l'objet au participant.

Les deux seuils peuvent être justifiés de la manière suivante : leur but est de limiter l'analyse des partitions aux partitions pertinentes. Au-delà d'un certain indice d'agrégation, les groupements ne représentent plus grand chose puisqu'ils consistent à regrouper des objets distants de plusieurs dizaines de mètres. Comme nous sommes dans un environnement représentant un intérieur, on peut considérer que la distance maximale entre deux objets pour qu'ils appartiennent à un même groupe est de deux ou trois mètres. De même, deux objets à vingt centimètres l'un de l'autre sont forcément groupés puisque l'objet le plus petit (une lampe de bureau) a une taille déjà plus importante. Il est ainsi logique de ne pas essayer de trouver un groupe ayant le nombre d'objets voulu dans une partition d'un niveau inférieur à vingt centimètres. En résumé :

- Plus on baisse le seuil minimal, plus on autorise un résultat à être trouvé dans un environnement composé d'objets placés tous dans la même zone de l'espace.
- Plus on augmente le seuil minimal, plus on empêche un résultat d'être trouvé dans un environnement composé d'objets placés tous dans la même zone de l'espace, mais moins les groupes trouvés dans les autres cas seront compacts.
- Plus on baisse le seuil maximal, plus les groupes seront compacts et donc plus le résultat trouvé sera pertinent (mais moins on aura de chances de trouver un résultat).
- Plus on augmente le seuil maximal, plus on aura de chances de trouver les objets dans le cas d'une commande "les trois X" mais plus ces objets risqueront d'être distants les uns des autres.

Le coefficient de proportionnalité peut être déterminé de la manière suivante : le participant met deux objets dans un coin d'une pièce, l'un à sa gauche et l'autre à sa droite, les deux paraissant bien distincts, et se place dans le coin opposé de la pièce de manière à distinguer les deux objets comme appartenant à un même groupe. Dans sa nouvelle position, le participant met deux objets groupés du même type juste devant lui. On part d'un coefficient nul, ce qui contribue à grouper les deux objets

proches bien avant les deux objets à l'autre bout de la pièce. On augmente petit à petit le coefficient jusqu'à ce que les deux groupements se fassent à peu près au même niveau d'agrégation. On garde le coefficient ainsi déterminé. Les effets approximatifs de ce coefficient sont les suivants :

- Plus on le diminue, plus le groupement se fera en 2D dans le plan de la pièce, sans tenir compte du point de vue du participant.
- Plus on l'augmente, plus le point de vue du participant sera pris en compte exagérément, c'est-à-dire que plus des objets éparpillés dans le lointain auront tendance à être groupés avant des objets bien groupés en premier plan.

appel à la saillance :

Revenons maintenant sur les conditions d'appel à la saillance comme nous les avons abordées dans la section 3.3.1. L'implémentation s'effectue dans le cadre d'une interaction gérant déjà un certain nombre de contextes référentiels (se reporter à la section 0 pour la présentation générale de ces contextes). Dans COVEN, les cinq contextes référentiels sont les suivants, en comptant le contexte référentiel relatif à la saillance qui n'était pas géré dans l'état initial du projet :

1. **Le focus** contient les référents calculés lors de la commande précédente. A la tête de la liste de ces objets se trouvent ceux sur lesquels s'est appliquée cette commande et qui forment le focus préférentiel.
2. **L'historique du dialogue** contient la liste des focus précédents. Huit focus sont conservés.
3. **Les objets saillants** constituent le contexte référentiel que nous venons d'étudier.
4. **Les objets perceptibles** correspondent aux objets présents dans le champ de visée du participant.
5. **Tous les objets de l'environnement** consistuent également un contexte référentiel permettant par exemple de répondre à une commande portant sur un objet que l'on ne voit pas, que l'on n'a pas manipulé depuis au moins huit commandes mais qui existe dans l'environnement.

Lorsqu'une commande vocale est émise, le gestionnaire de contextes commence par chercher le référent dans le premier contexte référentiel, c'est-à-dire le focus. Si le focus ne contient pas d'objet de la catégorie et ayant les propriétés souhaitées, le contexte référentiel suivant est testé, et ainsi de suite. Selon le type de groupe nominal utilisé et l'association éventuelle d'un geste de désignation, tous les contextes référentiels ne sont pas testés.

L'algorithme consistant à indiquer si le contexte référentiel des objets saillants doit être exploré ou non se déduit facilement de l'analyse faite à la section 3.3.1 :

```
si geste de désignation
alors (il s'agit d'une co-référence synchrone)
      si l'expression contenant une référence est un groupe nominal défini ou
         un groupe nominal démonstratif ou un pronom démonstratif, au singulier
         ou à un pluriel défini
      alors on fait appel au contexte référentiel des objets saillants
            (dans le cas du pronom démonstratif, on récupère la catégorie et
            les propriétés dans le focus ou dans l'historique du dialogue)
      fin si
sinon si l'expression contenant une référence est un groupe nominal défini
         singulier ou pluriel défini
      alors on fait appel au contexte référentiel des objets saillants
            (il peut s'agir d'une co-référence asynchrone ou d'une référence
            simple selon l'existence ou l'absence d'un historique du dialogue)
      fin si
fin si
```

Le contexte référentiel correspondant aux objets saillants est placé entre l'historique du dialogue et le contexte des objets perceptibles. La priorité de la saillance sur la perception globale est logique : les contextes référentiels relatifs à la perception visuelle sont testés dans un ordre correspondant à un nombre décroissant de contraintes. Le contexte référentiel correspondant aux objets saillants implique une contrainte de perceptibilité et une contrainte de saillante, le contexte référentiel correspondant aux objets perceptibles n'implique plus qu'une seule contrainte qui est celle de la perceptibilité et le contexte référentiel correspondant à tous les objets de l'environnement n'implique plus aucune de ces deux contraintes.

La priorité de l'historique du dialogue sur la saillance découle de la priorité de la mémoire du discours sur la perception visuelle qui est une théorie classique : le fait d'utiliser le discours pour désigner en effectuant une référence implique que le discours est le point de départ pour résoudre cette référence. La perception visuelle n'est prioritaire que lorsque l'attention est mise sur elle, c'est-à-dire que lorsqu'un geste de désignation est effectué [Moulton Roberts 1994].

3.3.5. Résultats

traitement de la saillance :

L'appel à la saillance n'est pas toujours réalisé, même dans des cas qui paraissaient propices à son traitement. Ceci est du au traitement préalable des contextes référentiels anaphoriques, en particulier du contexte relatif à l'historique du dialogue.

Le discours est le point de départ du calcul de référents car la référence est effectuée par le discours [Moulton Roberts 1994]. Cette approche est valable dans le cadre du dialogue homme-homme. En effet, lorsque deux interlocuteurs parlent entre eux, ils ne perçoivent l'environnement qui les entoure que secondairement par rapport aux sujets de leur dialogue.

Le dialogue homme-machine est différent, en particulier dans le cadre d'un environnement virtuel. La perception visuelle est gérée par une machine qu'il est difficile de dissocier de la machine chargée du dialogue. L'utilisateur a en effet l'impression de parler à une seule machine. La perception visuelle

n'est donc plus un environnement du dialogue mais fait partie du dialogue. Il apparaît ainsi possible de considérer que la perception visuelle est une sorte de focus préférentiel du dialogue et est donc prioritaire par rapport à l'historique de ce dialogue.

efficacité des regroupements :

L'algorithme de regroupement a été testé sur une trentaine de situations : des situations comportant un objet isolé et un ou plusieurs groupes d'objets de la même catégorie afin de tester la saillance d'un objet isolé ; des situations comportant un groupe de *n* objets, un ou plusieurs groupes d'objets de cardinalité différente de *n*, et éventuellement quelques objets isolés afin de tester la saillance d'un groupe de *n* objets ; ainsi que des situations ne contenant que des groupes d'objets de cardinalité différente de *n* afin de tester l'absence de résultat quand des expressions telles que "l'objet" et "les *n* objets" sont employées. Quelques situations ont également été testées en association avec un geste de désignation flou.

Les résultats sont en général satisfaisants, même lorsque les objets sont nombreux. Le temps de calcul reste raisonnable dans la mesure où il n'augmente que très peu le temps pris par le calcul de référents. Les seuils, tant qu'ils restent dans des limites sensées, n'ont que peu d'inflluence sur les résultats. Le fait de ne prendre en compte que les objets perceptibles donne lieu à des situations telles que la suivante : on crée deux groupes de chaises, on change de point de vue de manière à voir un groupe dans sa totalité et à ne voir qu'une seule chaise de l'autre groupe. Après avoir vidé l'historique du dialogue, une expression telle que "peints la chaise en bleu" s'appliquera à la seule chaise visible du deuxième groupe. Ce résultat correspond bien à l'algorithme mais peut paraître bizarre lorsque le changement de point de vue est minime et consiste par exemple à se rapprocher du premier groupe tout en gardant la cohésion du second en mémoire. Cet exemple pourrait se résoudre en tenant compte de la mémoire de la perception visuelle dont nous venons de parler.

En résumé, le regroupement paraît être un critère intéressant de saillance, en particulier lorsque l'on cherche un nombre défini d'objets.

4. Conclusion

4.1. Résumé

4.1.1. Désignation

La désignation par le geste offre donc une alternative à la sélection d'un ou de plusieurs objets. Le geste de désignation approximatif est un moyen naturel et simple d'utilisation. L'ambiguïté causée par le manque de précision du geste est levée par une recherche adéquate du ou des objets indiqués. Afin d'être plus efficace, cette recherche doit, dans le cadre d'une interaction multimodale associant geste de désignation et commande vocale, s'appuyer sur les énoncés de l'utilisateur. La désignation multiple repose sur le regroupement des objets.

4.1.2. Saillance

Dans une interaction en environnement virtuel basée sur la commande vocale, le calcul de référents est chargé de faire le lien entre les groupes de mots utilisés et les objets virtuels. Lorsqu'une ambiguïté est rencontrée, la saillance peut permettre de lever cette ambiguïté. Il faut pour cela que parmi les objets sur lesquels se porte l'ambiguïté, un seul possède une caractéristique saillante, et que l'utilisateur ait considéré cette saillance lors de son énoncé. Parmi la liste des caractéristiques que nous avons étudiées, la saillance spatiale apparaît comme le seul type intéressant de saillance non explicite dans l'énoncé et nécessitant ainsi un traitement particulier. Ce traitement consiste en sa prise en compte à l'intérieur d'un contexte référentiel que nous avons placé juste avant le contexte lié à la visibilité des objets. Selon notre approche, la saillance spatiale se base sur les regroupements d'objets et sur la distance des objets au participant. L'algorithme que nous avons implémenté permet de traiter simultanément ces deux aspects. Les résultats que nous avons obtenus montrent que la prise en compte de la saillance permet de simplifier l'interaction. Les cas non couverts sont essentiellement dus à une gestion insuffisante du contexte, en particulier du contexte de tâche.

4.2. Prolongements de l'étude

4.2.1. Désignation

Autant le geste de désignation est irremplaçable dans le cadre du dialogue homme-homme, autant il peut être remplacé dans le cadre du dialogue homme-machine, comme par exemple par la solution plus efficace de l'eye-tracking.

Un système d'eye-tracking, le plus souvent associé à un casque de visualisation, consiste en un rayon infra-rouge qui vient "lire" dans les yeux de l'utilisateur la direction de son regard. Ce système est intéressant dans la mesure où un équivalent de notre zone de désignation pourrait être une zone d'attention correspondant à la zone couverte par le regard. De même que le champ de vision du participant est modélisé par une pyramide de visée, on pourrait modéliser ce que l'on appelerait le champ d'attention par un cône. Ce cône serait maintenu en permanence, son axe serait la direction du regard.

Ainsi, on pourrait effectuer une opération sur un objet rien qu'en le regardant et en émettant une commande vocale. En tenant compte de la méthode de recherche de groupes d'objet que nous avons implémentée, le système pourrait également permettre la désignation d'un groupe d'objets.

D'autre part, afin de pallier au problème de détermination de l'objet visible en un pixel de l'image projetée (autrement que par un lancé de rayon qui est une méthode lente), on pourrait maintenir en temps réel une base de données de type grille ou quadtree recouvrant cette image projetée et dont chaque cellule contiendrait l'identifiant de l'objet visible dans la zone rectangulaire correspondante. La recherche d'objets dans une zone particulière serait ainsi possible en 2D de manière rapide. L'utilisation de cette base de données avec un système d'eye-tracking serait également un moyen intéressant de recherche des objets regardés par l'utilisateur.

4.2.2. Saillance

Un premier prolongement de l'étude serait la gestion d'un contexte de tâche, c'est-à-dire la prise en compte du but du participant lors du calcul de référents et donc lors de la prise en compte de la saillance. Ceci consiste dans un premier temps à utiliser l'action courante (c'est-à-dire le verbe de la phrase) lors du calcul de référents, et dans un deuxième temps à utiliser les actions précédentes. On peut aller plus loin en considérant qu'une tâche d'aménagement d'intérieur implique un but à atteindre et des étapes intermédiaires par lesquelles on doit passer. On pourrait alors considérer que l'utilisateur place d'abord l'objet le plus gros (par exemple le bureau) qui reste le seul de sa catégorie et autour duquel viennent se placer les autres meubles. La saillance de ces meubles serait alors à reconsidérer.

Un autre prolongement possible de l'étude serait de considérer un contexte référentiel lié à la mémoire de la perception. Pour l'instant, les contextes référentiels liés au discours et à la perception sont totalement séparés. En choisissant comme nous l'avons fait la prépondérance de la mémoire du discours sur la perception, nous arrivons à des situations parfois aberrantes. Si par exemple l'utilisateur manipule une chaise bleue et se tourne ensuite vers une chaise rouge, une expression du type "la chaise" désignera la chaise bleue, même si celle-ci n'est plus visible. Une solution consisterait à séparer les contextes non pas selon le critère du mode utilisé (discours ou perception) mais selon un critère indépendant du mode et dépendant de la précision ou de l'atténuation due au temps. Les contextes référentiels correspondant à la mémoire du discours et à la mémoire de la perception seraient ainsi pris en compte simultanément.

Bibliographie

[Belaïd Belaïd 1992] BELAID A. & BELAID Y. — Reconnaissance de formes, Méthodes et applications. — InterEditions, Paris, 1992.

[Bellalem Romary 1993] BELLALEM N. & ROMARY L. — Le dialogue homme-machine multimodal : vers la compréhension du geste de désignation. — Actes L'interface des mondes réels et virtuels, Montpellier, 1993.

[Bellalem Romary 1995] BELLALEM N. & ROMARY L. — Langue et geste pour le dialogue homme-machine finalisé. — Actes 01Design'95 – Aspects communicatifs en conception, Europia & GDR-PRC Communication Homme-Machine, Autrans, 1995.

[Burdea Coiffet 1993] BURDEA G. & COIFFET P. — La réalité virtuelle. — Hermès, Paris, 1993.

[Dale Reiter 1996] DALE R. & REITER E. — The role of Gricean Maxims in the Generation of Referring Expressions. — Proceedings of the 1996 AAAI Spring Symposium on Computational Models of Conversational Implicature, Stanford University, California, USA, 1996.

[Davis 1989] DAVIS J. R. — Back Seat Driver : Voice Assisted Automobile Navigation. — Ph.D. thesis, Massachussetts Institute of Technology, 1989.

[Devlin 1976] DEVLIN A. S. — The "small town" cognitive map : Adjusting to a new environment. — in G. T. Moore and R. G. Golledge, editors, Environmental Knowing : Theories, Research and Methods. Dowden, Hutchinson and Ross, 1976.

[Edmonds 1993] EDMONDS P. G. — A Computational Model of Collaboration on Reference in Direction-Giving Dialogues. — Ph.D. thesis, University of Toronto, Canada, 1993.

[Edmonds 1994] EDMONDS P. G. — Collaboration on reference to objects that are not mutually known. — Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), pages 1118-1122, Kyoto, 1994.

[Foley et al. 1990] FOLEY J. D., VAN DAM A., FEINER S. K. & HUGHES J. F. — Computer Graphics. Principles and Practice — Addison-Wesley Publishing Company, 1990.

[Fraczak et al. 1998] FRACZAK L., LAPALME G. & ZOCK M. — Automatic generation of subway directions : salience gradation as a factor for determining message and form. — Proceedings of the Ninth Workshop on Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada, 1998.

[Fraser Gilbert 1991] FRASER N. M. & GILBERT G. N. — Simulating speech systems. — Computer Speech and Language, 5, pages 81-99, 1991.

[Gaiffe 1992] GAIFFE B. — Référence et dialogue homme-machine : vers un modèle adapté au multimodal. — Thèse de doctorat en informatique de l'université de Nancy I, 1992.

[Gaiffe et al. 1994] GAIFFE B., REBOUL A. & ROMARY L. — Références et gestion du dialogue. — Actes de TALN'94, Marseille, 1994.

[Grice 1975] GRICE P. — Logic and conversation. — in P. Cole and J. Morgan, editors, Syntax and Semantics : vol. 3, Speech Acts, pages 43-58, Academic Press, New York, 1975.

[Heeman Hirst 1995] HEEMAN P. A. & HIRST G. — Collaborating on Referring Expressions. — Computational Linguistics, 21, 1995.

[Heilig 1992] HEILIG M. — The cinema of the future. — Presence : Teleoperators and Virtual Environments. MIT Press, vol. 1, no. 3, 1992.

[Jolivalt 1995] JOLIVALT B. — La réalité virtuelle. — Presses Universitaires de France, collection "que sais-je ?", Paris, 1995.

[Lynch 1960] LYNCH K. — The image of the city. — MIT Press, 1960.

[Mathieu 1997] MATHIEU F.-A. — Prise en compte de contraintes pragmatiques pour guider un système de reconnaissance de la parole : le système COMPPA. — Thèse de doctorat en informatique de l'université Henri Poincaré, Nancy I, 1997.

[Mignot 1995] MIGNOT C. — Usage de la parole et du geste dans les interfaces multimodales - étude expérimentale et modélisation — Thèse de doctorat en informatique de l'université Henri Poincaré, Nancy I, 1995.

[Moulton Roberts 1994] MOULTON J. & ROBERTS L. D. — An AI Module for Reference Based on Perception — Proceedings of the AAAI workshop on Integration of Natural Language and Vision Processing, Ed. P. McKevitt, Seattle, 1994.

[Pouteau et al. 1994] POUTEAU X., ROMARY L. & PIERREL J.-M. — Voix, geste et multimodalité : quand dire c'est faire faire. — Actes Congrès ERGO-IA'94, pages 491-500, Biarritz, 1994.

[Pouteau 1995] POUTEAU X. — Dialogue de commande multimodal en milieu opérationnel : une communication naturelle pour l'utilisateur ? — Thèse de doctorat en informatique de l'université Henri Poincaré, Nancy I, 1995.

[Reiter Dale 1992] REITER E. & DALE R. — A fast algorithm for the generation of referring expressions. — Proceedings of the 14th International Conference on Computational Linguistics (COLING-94), pages 232-238, 1992.

[Romary 1993] ROMARY L. — "Mets ça ici" où quand "ici" dépend de "ça" : l'interprétation de "ici" dans des énoncés de positionnement. — Workshop Le dialogue homme-robot en langage naturel, Caen, 1993.

[Romary 1994] ROMARY L. — Sens et action, ou comment aménager son salon. — Actes de TALN'94, Marseille, 1994.

[Sabah 1990] SABAH G. — L'intelligence artificielle et le langage, représentations des connaissances. — 2ème édition, Hermès, Paris, 1990.

[Schang 1997] SCHANG D. — Représentation et interprétation de connaissances spatiales dans un système de dialogue homme-machine. — Thèse de doctorat en informatique de l'université Henri Poincaré, Nancy I, 1997.

[Slater Usoh 1993] SLATER M. & USOH M. — Representation systems, perceptual position, and presence in Immersive Virtual Environments. — Presence : Teleoperators and Virtual Environments. MIT Press, vol. 2, no. 3, 1993.

[Streit 1997] STREIT M. — Active and Passive Gestures - Problems with the Resolution of Deictic and Elliptic Expressions in a Multimodal System. — Proceedings of the Workshop : Referring Phenomena in a Multimedia Context and Their Computational Treatment, Madrid, 1997.