Interprétation des énoncés langue+geste en situation de dialogue homme-machine.

Laurent Romary, Nadia Bellalem

Interprétation des énoncés langue+geste en situation de dialogue homme-machine

Laurent Romary

Nadia Bellalem

Centre de Recherche en Informatique de Nancy

CRIN-CNRS & INRIA Lorraine

Bâtiment Loria, B.P. 239, F-54506 Vandœuvre Lès Nancy

romary@loria.fr, nbell@loria.fr

1. Présentation générale

L'objectif de cet article est de donner une idée générale des problèmes liés à la définition de systèmes de dialogue homme-machine, et plus particulièrement des facteurs et des mécanismes qui régissent l'interprétation d'énoncés multimodaux combinant langue et geste. L'ensemble des travaux présentés ici se situent dans une problématique résolument pluridisciplinaire puisqu'ils associent des résultats issus de la linguistique, de la psychologie et de l'informatique, dans le cadre de modèles qui - à court ou moyen terme - se veulent opérationnels.

La notion même de communication entre un être humain et un ordinateur peut à la fois enthousiasmer ou faire frémir. Comme tout nouveau secteur scientifique, le domaine du dialogue homme-machine génère des incertitudes propres à éveiller l'imagination et donc à masquer la réalité des choses. De fait, nous sommes loin encore de pouvoir reproduire une qualité de dialogue telle que celle présentée dans le film de Stanley Kubrik *2001 Odyssée de l'espace*. Dans cet œuvre, le système HAL 9000 peut aussi bien jouer aux échec, interagir au niveau du pilotage d'un vaisseau spatial ou discuter de problèmes métaphysiques avec les membres de l'équipage. En dehors même des problèmes

de reconnaissance de la parole - que nous n'aborderons pas ici - les capacités de compréhension des systèmes que nous présenterons ne permettent pas de tels changements thématiques, car dans tous les cas, leur mise en œuvre correspond à des applications concrètes de commande ou de recherche d'information.

Après avoir présenté les concepts fondamentaux associés à la notion d'interface et de dialogue, nous aborderons les principales méthodes d'investigation qui caractérisent le domaine du dialogue homme-machine. Nous verrons en particulier qu'il est possible de s'appuyer sur des données expérimentales pour guider la définition de modèles de communication entre un utilisateur et un système informatique dans un cadre hautement finalisé. Nous essayerons alors de dresser un bilan rapide des connaissances actuelles dans le domaine du dialogue multimodal en montrant le nombre important des facteurs linguistiques et perceptifs qui régissent de telles communications. Enfin, nous essayerons d'esquisser un modèle d'analyse des références multimodales compatible avec les mécanismes linguistiques sous-jacents.

2. Des interfaces au dialogue

D'un certain point de vue, chercher à construire un système de dialogue homme-machine, correspond à redéfinir une médiation possible entre un être humain et un objet qu'il cherche à utiliser pour réaliser une certaine tâche. Cependant, il y a loin de l'idée que l'on se fait d'un marteau par exemple, à l'image qui nous vient quand on parle de dialogue. Peut-être est-il nécessaire d'éclaircir quelque peu les choses, en considérant la nature profonde d'une possible relation entre l'homme et la machine. De fait, si l'on prend un petit peu plus de recul, on constate que l'homme agit sur son environnement de deux façons principales. Soit il manipule directement les objets sur lesquels il souhaite opérer, soit il exprime ce qu'il désire faire à l'un de ses congénères qui va se charger de réaliser l'action correspondante. Ainsi, nous sommes habitués à interagir soit avec des objets relativement inertes (le marteau), soit avec des

individus possédant des capacités comparables aux nôtres à qui nous transmettons des intentions. Si maintenant il s'agit de placer une situation de dialogue homme-machine dans ce schéma, la situation risque d'être un peu plus complexe. En fait, l'hypothèse que nous souhaitons faire est que l'essentiel réside dans la perception que l'utilisateur humain va avoir du système informatique avec lequel il interagit et corrélativement du niveau de compréhension auquel il s'attend de la part de celui-ci. Concevoir une interface consiste alors à définir un comportement d'ensemble du système informatique qui permette à l'utilisateur de réaliser sa tâche dans le cadre d'une certaine modalité d'interaction que celui-ci est prêt à accepter.

Nous proposons ainsi de conceptualiser la relation entre l'homme et la machine sous la forme de deux logiques principales : d'une part, une relation directe entre l'homme et les opérations qu'il souhaite réaliser, que nous qualifierons de logique du « faire », et d'autre part, une logique du « faire faire », où l'action passe par une phase de communication entre l'homme et la machine. La logique du faire, c'est la logique du marteau. L'ordinateur a le rôle d'un outil (ou de plusieurs) que l'utilisateur dirige à sa guise. De la sorte, celui-ci n'a pas à indiquer quelles intentions sous-tendent son activité puisque lui seul doit gérer la séquence d'opérations susceptible d'aboutir au but qu'il s'est fixé. C'est dans cette logique que ce situe la plupart des interfaces actuelles d'ordinateurs (telles le Macintosh ou l'environnement Windows), et l'on voit (cf. figure 1.) que l'utilisateur a la charge de planifier une séquence d'actions plutôt que de transmettre une quelconque consigne à la machine.

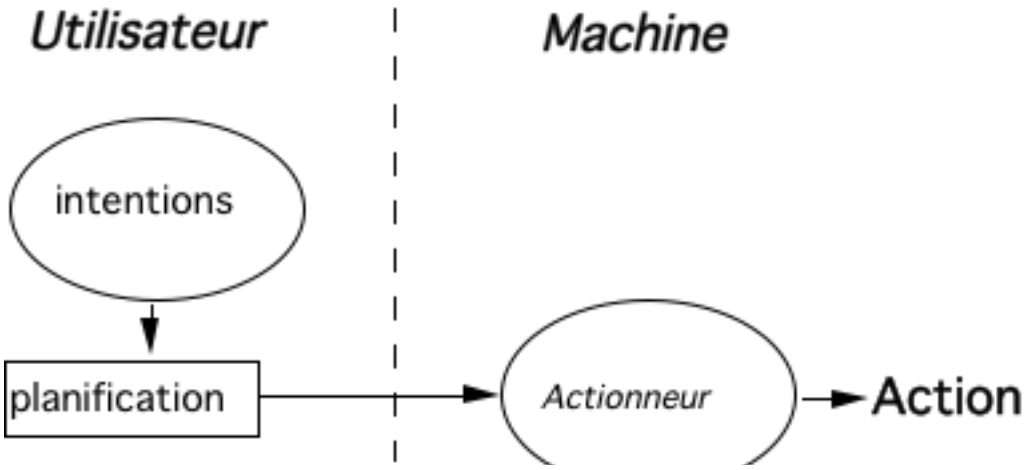


Figure 1 : Interaction homme-machine dans une logique du « faire »

A l'inverse, une logique du faire-faire subordonne toute action à une phase de communication au cours de laquelle l'utilisateur va chercher à transmettre une intention plus ou moins abstraite que la machine devra décoder. La situation est alors proche de celle présentée figure 2, où la machine doit prendre à son compte une partie des opérations de planification qui permettrons de passer d'une intention donnée à la séquence d'actions permettant de la réaliser. La machine n'est plus alors un outil inerte, mais bien plus un *partenaire* qui possède un certain nombre de compétences, tant sur le plan de la communication, que dans le domaine de la tâche à réaliser. Il est très important de réaliser ici que l'usage de la langue comme mode de communication entre l'homme et la machine ne peut se faire que dans le cadre de cette deuxième logique, puisque la langue est par essence un vecteur d'intentions qu'il faut pouvoir interpréter et non pas un instrument conçu directement pour l'action.

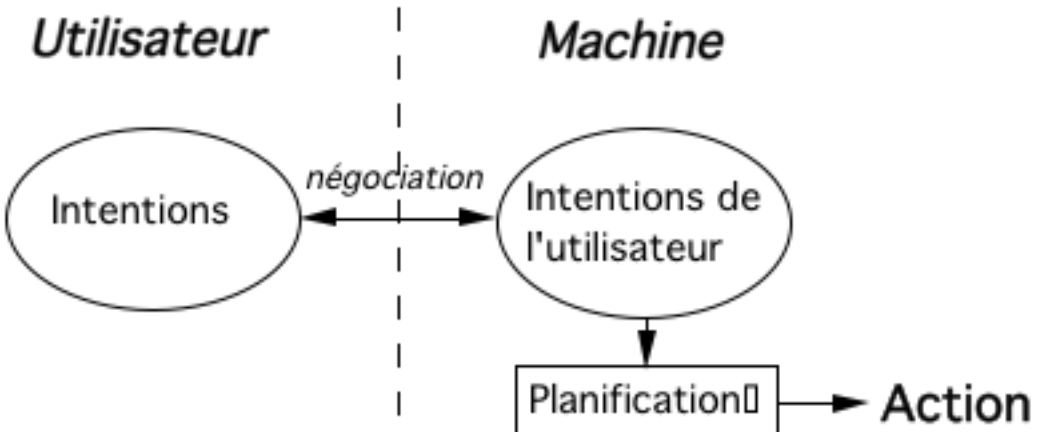


Figure 2 : Interaction homme-machine dans une logique du « faire faire »

La double conception du faire et du faire faire n'est pas innocente dans la mise un œuvre d'une interface homme-machine. En effet, elle peut être vue comme deux métaphores élémentaires dans lesquelles on souhaite plonger un utilisateur humain dans son face à face avec une machine. Une logique de l'action lui fera voir l'ordinateur comme un objet similaire au râteau, violon et autre fouet de cuisine, alors que dans une perspective de communication, il se placera dans la même position que celle qu'il occupe lorsqu'il commande une pièce montée à un pâtissier. Il s'adressera à un partenaire spécialiste du domaine considéré et qui l'encourage à exprimer une intention abstraite éventuellement négociée (« Vous savez, pour moins de trente personnes, je vous conseillerais plutôt une génoise… »). Ce double caractère se retrouve si l'on considère maintenant le rôle du geste dans chacune de ces conceptions. Concevoir une interface relevant d'une logique du faire implique que l'usager soit sûr d'agir effectivement sur l'environnement avec lequel il entre en contact. Ainsi, tout comme les touches d'un téléphone sont des zones prédéfinies pour réagir à une pression, certaines portions de l'espace graphique d'un ordinateur doivent être connues de l'utilisateur comme étant réactives et de ce fait associées à la réalisation d'une action particulière. Inversement, une logique du faire faire sera associée à une structure spatiale négociée entre l'utilisateur et la machine-

partenaire, par exemple par l'association d'une expression démonstrative particulière et d'un geste (cf. *infra*). On peut ainsi prendre l'exemple d'un clique souris dans une fenêtre de texte qui sera interprété différemment suivant l'énoncé qui l'accompagne :

- une lettre : « Remplace cette lettre par un *i* »
- un mot : « Mets ce mot en italique »
- un paragraphe : « Mets ce paragraphe à la fin de la section courante »
- toute la fenêtre : « Déplace le texte en arrière plan »
- etc.

Il est clair que si nous souhaitons introduire plus de souplesse dans nos relations à venir avec les machines électroniques, sans avoir à ingurgiter les quelques centaines de pages des manuels utilisateur qui nous sont livrés actuellement (par exemple celle de Microsoft Office), le seul schéma qui soit viable à terme pour le grand public est celui de la deuxième logique et nous nous placerons dorénavant dans cette perspective. Celle-ci peut être reformulée sous la forme d'un échange d'égal à égal avec un partenaire électronique doué de capacité de compréhension et partageant avec l'utilisateur humain une certaine connaissance de la tâche que tous deux ont à réaliser. Comme on peut le voir sur la figure 3, le partage de cette information peut passer par la présentation visuelle, via une interface graphique par exemple, de l'état courant de la tâche.

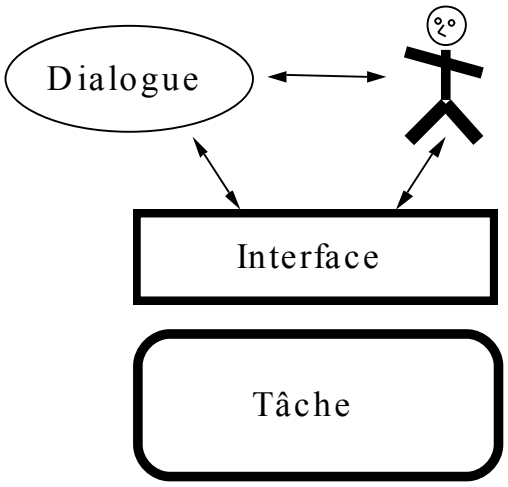


Figure 3 : Placement symétrique de l'utilisateur et du système de dialogue vis à vis de la tâche.

Avant d'aborder plus précisément différents aspects méthodologiques liés aux dialogues homme-machine, il nous reste à faire ici quelques remarques quant aux liens que peut avoir ce domaine avec ce que nous pourrions appeler globalement les sciences du langage. Tout d'abord, remarquons qu'il est impensable de concevoir un système de dialogue sans faire reposer celui-ci sur l'ensemble des connaissances liées au lexique, à la syntaxe, à la sémantique ou à l'interprétation des langues naturelles, puisqu'il s'agit bien de passer d'un message à son interprétation ultime. De fait, à quoi servirait un système reposant sur un analyseur syntaxique à très haute performance mais ne possédant pas de réel module d'interprétation référentielle ? Il est à noter que l'observation de corpus de dialogues homme-machine simulés (par la technique dite du Magicien d'Oz[1]) montre que la complexité du type de dialogues dont nous parlons ici résulte plus des nombreux effets contextuels que ceux-ci contiennent, que de la complexité structurelle des énoncés eux même. Le deuxième point à considérer est que la réflexion sur le dialogue homme-machine finalisé permet de se faire

[1] Ce type d'expérience consiste à mettre un utilisateur devant un système de dialogue, qui se trouve être remplacé par un compère présent dans une autre pièce. Les résultats obtenus permettent (dans une certaine mesure) d'observe les comportements interactionnels d'usagers dans le cadre de tâches finalisées.

une idée peut-être un peu moins abstraite que ce qui peut être dit dans d'autres cadres de la notion de sens. Qu'est-ce donc que comprendre pour un système de dialogue finalisé ? N'est ce pas réaliser une ou plusieurs actions qui satisfassent les intentions exprimées par l'utilisateur ? D'une certaine manière et sans entrer véritablement dans les conséquences d'une telle analyse, il semble possible de s'orienter vers une sémantique de l'action qui peut aussi se voir comme une sémantique du but. Une telle sémantique consisterait, pour chaque énoncé de l'utilisateur, à se demander quel but peut y être associé. Ainsi, dans le cadre d'un système de réservation de billets de chemin de fer, l'analyse d'un énoncé tel que *existe-t-il des trains pour Paris en début de matinée ?* doit pouvoir comprendre le fait que le locuteur souhaite effectivement connaître un horaire précis de train et ne pas simplement répondre *oui*.

3. Méthodes et observations

Comme on le voit, la définition de systèmes de dialogue homme-machine est un processus complexe nécessitant une réflexion à la fois linguistique, puisque l'on s'appuie sur la langue, et informatique pour les aspects de modélisation proprement dits. La difficulté qui se pose alors au chercheur est de se fixer une méthodologie qui lui permette de maitriser l'ensemble de ces facteurs sans pour autant perdre de vue son objectif final qui est d'obtenir un système qui sache effectivement dialoguer.

De fait, deux cadres méthodologiques peuvent être envisagés pour l'étude du dialogue homme-machine : soit réaliser des maquettes de système de dialogue qui sont ensuite évaluées et modifiées suivant un processus itératif, soit mettre en oeuvre des expériences de simulation où le système de dialogue est remplacé - dans le cadre d'un protocole précis - par un interlocuteur humain qui communique avec l'utilisateur et agit sur la tâche.

Intuitivement, la première méthode semble celle qui correspond le mieux aux objectifs visés, puisqu'il s'agit de mettre en œuvre et d'évaluer des systèmes qui

préfigurent ceux que l'on souhaite véritablement rendre opérationnel. Pourtant, elle est relativement difficile à mettre en œuvre et ce pour tout un ensemble de raisons. Ainsi, nous avons vu qu'un système de dialogue repose sur un ensemble de modules correspondant à différents traitements (reconnaissance de la parole, analyse lexicale et syntaxique, interprétation du message, réalisation de l'action, gestion du dialogue et des retours à l'utilisateurs) qui tous doivent fonctionner pour que le système complet soit opérationnel. Tout ceci représente donc un effort important que l'on ne peut envisager que dans le cadre de projets de grande ampleur, et non pas pour tester occasionnellement telle ou telle hypothèse de recherche.

L'autre possibilité consiste donc à remplacer le système de dialogue envisagé par une simulation où un humain va prendre à son compte l'ensemble des fonctionnalités - et éventuellement des contraintes - associé à un système supposé réel. Une expérience de simulation repose en général sur une structure classique impliquant d'une part un ou plusieurs expérimentateurs (des « compères ») et d'autre part une application informatique (ou une base de connaissances) auquel peuvent accéder ces expérimentateurs. Suivant un certain protocole, on présente aux utilisateurs un scénario d'intéraction présentant les caractéristiques de la tâche ainsi que les modes d'actions (oral, geste etc.) qu'il peut utiliser pour arriver à ses fins.

A titre d'illustration, nous pouvons présenter les grandes lignes d'une expérience menée au sein de notre laboratoire en collaboration avec les ergonomes de CERMA au début des années 90 (Dauchy et al., 93 ; Mignot et al., 93). La situation qui était considérée dans cette expérience était une tâche d'aménagement intérieur d'un appartement représenté graphiquement sur un écran tactile branché sur un Macintosh. Chaque utilisateur pouvait s'exprimer ou agir sur l'interface à l'aide de la parole et du geste suivant un ensemble de consignes qui lui étaient présenté, tout en étant surveillé via une caméra vidéo placée dans la salle d'expérimentation. Par exemple, l'un des scénarios était

décrit de la façon suivante :

> *Vous choisissez sur catalogue un frigidaire, un évier, une cuisinière, un meuble de rangement ainsi qu'une table et quatre chaises assorties de façon à les placer dans la cuisine.*
>
> *Disposez ensuite sur la maquette les articles sélectionnés.*

Bien que ce type de tâche puisse paraître particulièrement simple, il faut remarquer qu'il permet de réaliser des observations relativement précises de l'usage de la langue et du geste en contexte multimodal. Sans reprendre l'étude exhaustive faite par Christophe Mignot (95) de cette expérience, nous pouvons remarquer la richesse des phénomènes référentiels observés. Ainsi, dans le dialogue suivant[1], on constate l'usage de descriptions définies (*le réfrigérateur, la cuisinière*), de pronoms anaphoriques (*le*), de pronoms démonstratifs (*celui-ci* + geste) ou de déictiques (*ici* + geste). Chacun de ces usages semble bien correspondre à différents niveaux de représentation de la tâche d'une part et de l'espace d'autre part, qu'un système automatique devrait nécessairement être en mesure de traiter.

U : Le réfrigérateur celui-ci DOWNUP{réfrigérateur moyen} nous allons le poser là DOWNUP{mur est face porte} pour voir.	S : Compris. ACTION : place réfrigérateur moyen mur est face porte S : Et maintenant ?
U : Déplaçons le réfrigérateur DOWNUP{réfrigérateur moyen} ici DOWNUP{coin sud-ouest} à côté de la cuisinière ; hm hm	S : Compris. ACTION : déplace réfrigérateur moyen coin sud-est S : C'est fait.

4. Interprétation d'énoncés multimodaux - quelques facteurs

[1] Dans cet exemple, 'U' indique les énoncés de l'utilisateur et 'S' ceux du système simulé. Les gestes effectués par l'utilisateur ont été retranscrits à l'aide d'indications du type DOWNUP{objet}.

4.1 Cadre général

Comme le montrent assez bien les données expérimentales présentées dans la section précédente, la mise en œuvre de modèles - et *a fortiori* de systèmes - de dialogue homme-machine multimodal est rendu particulièrement complexe de par la conjonction de nombreux facteurs qui interagissent constamment. Ainsi, en plus des phénomènes classiques associés à l'interprétation de toute suite de phrases ou d'énoncés, interviennent différents paramètres relatifs à l'espace de travail *perçu* par le locuteur et qui conditionnent la réalisation de la composante gestuelle de son acte de communication.

Remarquons tout d'abord que le *geste* n'est pas une entité parfaitement identifiable dans sa fonction de communication. Il peut, dans le cas spécifique des langages nationaux ou internationaux de signes pour les sourds-muets (voir par exemple Naughton, 96 dans un contexte « bilingue »), participer de façon quasi-exclusive à la communication. A l'inverse, une communication téléphonique ne permet de transmettre à son interlocuteur aucune information gestuelle alors que celle-ci est toujours présente au niveau de l'émetteur… Dans le même ordre d'idée, l'observation de toute communication orale montre que seul un petit nombre de gestes semblent être intentionnellement produits pour, par exemple, désigner un objet, représenter une forme, ou exprimer une certaine longueur. Ces différents gestes ne sont d'ailleurs pas clairement isolés du reste des mouvements des doigts, des mains ou des bras. On ne peut que constater le fort degré de contextualisation des gestes, qui les rend quasi-ininterprétable quand on ne dispose pas par exemple du message oral qui les accompagne. Enfin, il est parfois nécessaire d'inclure dans la notion de geste l'ensemble des postures corporelles ou faciales qui peuvent, tout aussi bien que l'ensemble main-bras, apporter des informations essentielles au dialogue.

Pour clarifier quelque peu la situation, Claude Cadoz (92), introduit une classification relativement générale des gestes pouvant être observés chez un être

humain. Il distingue ainsi les quatre catégories suivantes :

- le geste *iconique* qui par sa forme directement caractérisable représente un objet ou un concept (on peut simplement penser aux codes de la tête pour confirmer ou infirmer quelque chose) ;
- le geste *épistémique*, qui consiste à toucher directement un objet pour en percevoir la forme, la texture ou tout autre caractéristique. Ce type de geste n'a pas de visée communicative, mais plutôt informationnelle ;
- le geste *ergatif*, qui vise à agir sur un objet. Ce geste n'est de fait lié à aucun contenu informationnel. Il est en quelque sorte l'instrument de la perspective du « faire » que nous avons exposé au début de cet article ;
- enfin, le geste *déictique* vise à identifier un objet ou une propriété dans l'environnement communicationnel de celui qui le réalise.

Dans la perspective stricte d'un dialogue homme-homme ou homme-machine, seuls les gestes iconiques et déictiques doivent en général être pris en compte, puisque ce sont les seuls qui nécessitent une interprétation de la part d'un interlocuteur. Cependant, la plupart des équipes de recherche restreignent encore plus le spectre des gestes à considérer, car il n'existe à l'heure actuelle aucune description unifiées de ces deux catégories somme toute assez hétérogènes. Ainsi, le début des années 90 a vu apparaître différents systèmes (e.g. Braffort, 92) permettant de reconnaître un ensemble de gestes iconiques définis à l'avance. L'avantage d'une telle méthode est qu'elle est rapidement implantable sur machine à l'aide d'algorithmes classique de reconnaissance des formes. A l'inverse, un certain nombre d'études (e.g. Caelen, 91) ont porté leur attention sur la place du geste dans des énoncés multimodaux simples comportant des déictiques tels que *ici* ou *ça*. Ces études ont cependant trop souvent, afin d'aller au plus vite au stade de l'implantation sur machine, réduit à l'extrême la complexité des gestes considérés, ainsi que le finesse des descriptions linguistiques associées aux expressions déictiques. Pourtant, malgré les différentes critiques qu'elles suggèrent, ces expériences ont permis de mettre

en évidence les directions principales de recherches qu'ils faut maintenant suivre pour intégrer pleinement le geste à des systèmes de communication homme-machine.

Si l'on se fixe comme objectif de définir des systèmes qui s'intègrent dans le cadre d'une communication que l'on qualifierait de spontané, il faut bien reconnaître qu'il faut tendre vers des échanges dont la composante principale est plutôt langagière que gestuelle. On a ainsi observé les limitations en terme de bande passante communicationnelle de systèmes tels que Socrates, mis en place par les chemins de fer français pour la commande de billets. Ces systèmes, qui reposent exclusivement sur un écran tactile, engendre très vite une frustration liée à l'impossibilité de communiquer l'information que l'on juge la plus pertinente à un moment donné. Le système est en quelque sorte pré-cablé pour présenter successivement une série de fonctions que l'utilisateur peut désigner. A l'inverse, la langue permet d'accéder directement à une intention abstraite (e.g. *Je voudrais un billet pour Paris*), quitte à ce qu'un sous-dialogue complète certaines informations manquantes et ne peuvant être inférées à partir du contexte.

La conséquence d'une telle option d'un point de vue gestuel est qu'il semble raisonnable de laisser de coté les gestes purement iconiques pour se concentrer sur les gestes strictement coverbaux et plus particulièrement ceux qui participent à des opérations de référence. Bien souvent, et probablement pour des raisons de facilité, on qualifie ces gestes de co-référentiels, alors que, comme nous allons le détailler, il est préférable que considérer qu'il n'existe qu'un seul acte de référence auquel participent conjointement geste et expression langagière.

4.2 Geste, langue et perception

Après avoir présenté le cadre général dans lequel nous concevions une communication spontanée combinant langue et geste, notamment pour des actes référentiels, essayons de préciser les rôles respectifs de chacun de ces deux modes. Selon nous, le rôle essentiel d'un geste de référentiation est de *focaliser*

l'attention de l'interlocuteur sur une sous-partie de l'espace de perception (visuel) partagé entre celui-ci et celui qui s'exprime. Remarquons immédiatement qu'il ne peut y avoir de communication gestuelle sans perception visuelle. Même si cela parait être une évidence, c'est l'occasion de rappeler que d'une part, le geste est perçu par un autre sens que le signal auditif, avec les conséquences que cela peut entraîner en termes de différentiation des traitements, et que d'autre part il marque le rôle prépondérant des critères spatiaux dans l'interprétation de l'acte référentiel global (langue + geste). L'objet (ou référent) désigné par un geste est immédiatement considéré comme un élément de cet espace perçu, éventuellement mis en perspective par rapport à d'autres éléments similaires. Cette constatation nous permet dès lors d'introduire une notion de *contraste spatial* associé à tout acte de désignation gestuel et support du résultat final de son interprétation. Très concrètement, une telle notion permet de prédire qu'un geste va être d'autant plus précis que l'objet désigné est étroitement entouré de « contre-cibles », à savoir d'autres objets susceptibles de jouer un rôle équivalent ou complémentaire au moment de l'énonciation. Ainsi, si l'on souhaite désigner un tableau sur un mur, un simple mouvement de la tête est suffisant si celui-ci est isolé dans la pièce où le locuteur se trouve, alors qu'un geste relativement précis du doigt sera nécessaire s'il est entouré de deux autres tableaux. D'un point de vue algorithmique, une telle constatation permet de régler la finesse d'interprétation des trajectoires gestuelles sur un écran tactile par exemple, pour interpréter des désignations qui pointent plus ou moins précisément sur l'objet visé en fonction du contexte.

La perspective adopté pour le geste permet d'expliciter de façon particulière le rôle de la langue dans un énoncé multimodal en disant qu'il *indique le mode de sélection des référents dans l'espace de visualisation*. Si nous essayons d'éviter ici toute théorie générale de la référence linguistique, nous pouvons en effet rapprocher la place de l'expression référentielle à celle d'un filtre qui va

« calibrer » l'espace visuel, alors que le geste, comme nous l'avons vu, va plutôt jouer un rôle de sélecteur. Le cas du démonstratif illustre relativement bien ce point de vue, puisque l'on va considérer qu'une expression telle que *cet homme*, accompagné d'un geste, va transmettre à l'allocutaire l'instruction de ne percevoir devant lui qu'un ensemble de personnes, dont l'une sera plus particulièrement mis en évidence par le geste qui lui pointe dessus. La langue apporte là des contraintes relativement précise sur les propriétés que doit posséder le *cadre* dans lequel s'effectue l'acte de référence ainsi que sur la nature du contraste que le geste va opérer (dans le cas du démonstratif, il s'agit d'un contraste intra-catégoriel). Au contraste spatial que nous avions introduit pour le geste vient donc s'adjoindre un *contraste linguistique* qui agit de façon complémentaire.

Comme nous l'avons remarqué, la perception joue enfin un rôle important dans l'expression même des trajectoires gestuelles. Si nous essayons d'aller un peu plus loin ici, nous pouvons préciser différents facteurs perceptifs qui vont contraindre de façon importante l'interprétation que peut se faire un locuteur d'une scène donnée et donc toute expression référentielle relative à cette scène.

La perspective dans laquelle nous nous placons ici est de déterminer l'organisation spatiale associée à un champ perceptif donné. L'objectif n'est pas pour nous de détailler les processus réels ou supposé tels qui sont impliqués dans l'acte de perception, mais bien de nous centrer sur les résultats attendus en termes de structures afin de mieux comprendre les paramètres qui influent sur les actes référentiels. Une première observation que nous pouvons faire est que l'appropriation perceptive d'une scène combine à la fois les effets propres au champ perceptif (contraste, couleurs, formes etc.) et ceux liés à l'expérience individuelle de celui qui observe, en particulier sa *connaissance* du domaine perçu et ses *intentions* vis à vis de ce domaine. En fonction de ces différents facteurs, un certain nombre de formes et de structures spécifiques à la situation globale de

perception vont émerger. Ainsi, un champ de fleurs a peu de chance d'être perçu de la même façon par un peintre qui cherche son inspiration et par un botaniste qui recherche une espèce rare.

La psychologie peut ici nous apporter quelques éléments de réponse, notamment dans le cadre de la Gestalttheorie qui suggère quelques principes fondamentaus liés à un acte de perception (Forgus, 66) :

- toute perception tend à structurer le champ perceptif ;
- les formes se distinguent comme des éléments ayant une unité globale se détachant d'un fond non structuré ;
- la perception d'une forme est associée à la perception d'une signification ;
- toute forme possède une certaine prégnace qui la rend plus robuste à la perception ;
- on observe une préservation des caractéristiques de la forme (principe de constance).

Dans une situation de dialogue finalisé où la tâche est présentée graphiquement à l'utilisateur, ces principes prennent une valeur particulière et permettent de cerner les caractéristiques des gestes qui vont servir à désigner les objets perçus. D'une manière générale, les formes présentées sont en générale parfaitement identifiables (forte iconicité) et les effets de Gestalt portent principalement sur la possibilité de repérer des regroupements de formes similaires. Sur cette base, nous avons ainsi pu identifier (Bellalem, 95 ; Bellalem et Romary, 95 a et b) que l'une des dimensions essentielles d'analyse devait porter sur la propension d'un geste à soit désigner un objet ou un groupe d'objets qui se distingue clairement de son entourage (*désignation centrale*), soit marquer les frontière d'un groupe d'objets pour le détacher d'autres objets similaires (*désignation périphérique*).

4.3 Bilan partiel

L'ensemble de cette section nous a permis de mettre en évidence un certain

nombre de paramètres d'origines diverses intervenant dans la production, et donc dans l'interprétation, d'énoncés multimodaux combinant langage et désignation gestuelle. Il est maintenant temps d'observer des données réelles et d'envisager les traitements automatiques qui, au stade actuel de nos connaissances, peuvent être réellement implantées. Nous allons le faire bien sûr au travers d'une tâche particulière qui préfigure le type d'interactions que nous souhaitons voir réalisées dans nos recherches à venir, en portant plus particulièrement notre attention sur l'analyse des trajectoires gestuelles.

5. Interprétation d'énoncés multimodaux - analyse et interprétation de trajectoires gestuelles

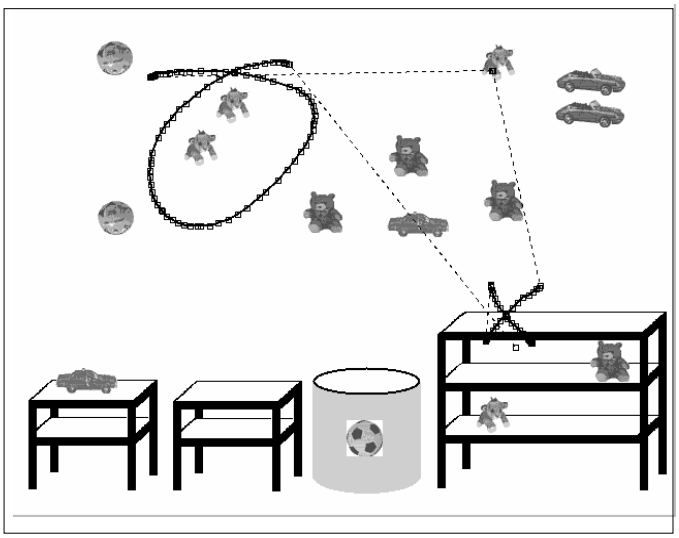
La compréhension automatique d'un geste de désignation nécessite au moins deux étapes de traitement. La première consiste à étudier les caractéristiques de la trajectoire afin d'en extraire les parties les plus significatives du point de vue de l'acte de désignation, la seconde consiste à interpréter la trajectoire relativement à la scène sur laquelle le geste est intervenu et relativement au message langagier accompagnant le geste. La première étape est dite "analyse structurelle", son objectif est la recherche de *singularités* dans la trajectoire. Il s'agit de portions de trajectoire présentant du point de vue de certains paramètres tels que la courbure ou la vitesse d'exécution du geste de fortes variations dénotant une possible intentionnalité de la part de l'utilisateur. Quant à la seconde étape, dite d'interprétation, elle consiste à recherche les objets de la scène, candidat à ou aux désignations contenues dans le geste en tenant compte de la forme de la désignation et de la répartition spatiales des objets. Rappelons qu'un geste désigne l'ensemble du mouvement accompagnant l'énoncé verbal et de ce fait un geste peut contenir plus d'une désignation.

Nous présentons dans ce qui suit les différentes étapes de traitement des trajectoires gestuelles sur un geste effectué à l'aide d'un stylet sur un écran tactile. Ces traitements s'articulent autour des trois composantes suivantes :

- l'analyse du geste et la recherche des parties significatives,
- la représentation de la répartition spatiale des objets dans la scène, tentant de représenter la structuration issue de la perception visuelle,
- l'interprétation du geste par rapport à la scène c'est-à-dire la détermination des objets candidats aux différentes désignations apparaissant dans le geste.

5.1 Présentation de l'exemple traité

L'exemple choisi est issu d'une expérience de type magicien d'Oz menée au CRIN (Wolff et al., 97) dont l'objectif est la constitution d'un corpus multimodal. La tâche choisie est une tâche de rangement. Il s'agit plus précisément de ranger les objets situés dans la partie supérieure de l'écran sur les différentes étagères et réceptacles situés en partie inférieure de l'écran. Voici l'un des gestes effectués dans le cadre de cette expérience.



« Prends ces deux peluches et celle-là et ranges les ici.»

5.2 Capture du geste

Le signal gestuel est une liste d'événements de la forme :

```
(Type-événement, x, y, temps, [n° bouton])
```

où Type-évènement est un code indiquant s'il s'agit d'un déplacement (6), un début de contact (4) ou une fin de contact (5) avec l'écran tactile, (x, y) sont les coordonnées dans le plan de l'écran du point où se produit l'événement, et [n° bouton] indique si le contact est effectué à l'aide du doigt ou du stylet.

Sur la base de ce codage, voici par ce qui a été recueilli pour le geste de pointage de notre exemple :

4	463	62	65893	0	Contact au point (463,62) à l'instant 65876
6	464	62	66004	0	Déplacement au point (464,62) à l'instant 66004
5	464	62	66115	0	Fin de contact au point (464,62) à l'instant 66115

5.3 L'analyse structurelle

La trajectoire observée contient quatre zones de contact avec l'écran qui sont analysées séparément. Il s'agit d'identifier la forme de ces sous-trajectoires sur la base de l'étude du tracé de la courbure, de la vitesse d'exécution du geste et de l'angle de courbure. Ces paramètres sont étudiés lorsque la trajectoire n'a pas été préalablement identifiée comme une droite ou un simple point de contact avec l'écran.

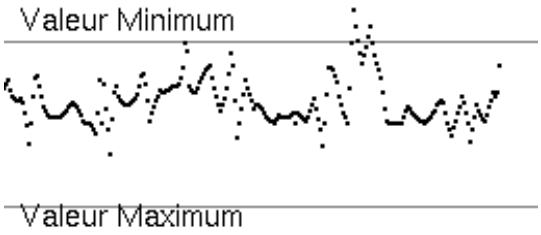
5.3.1 Analyse de la première zone de contact

Nous présentons sur la figure suivante les différentes étapes de l'analyse structurelle aboutissant à la modélisation de la forme du geste. En (1) ne sont représentés que les points résultant de la capture du geste, en (2) s'ajoute le tracé de la B-spline cubique[1] approximant au mieux la trajectoire initiale. Enfin en (4), figurent tous les éléments de la trajectoire (le point d'intersection représente par une croix, les début et fin de segments par une barre) issu de l'étude du tracé de la courbure présentée en (3).

(1) Evénements recueillis

(2) La spline calculée

[1] Il s'agit d'une interpolation des points de la trajectoire par une courbe polynomiale permettant de calculer en particulier des dérivées d'ordre 1 et 2. Ces dérivées participent elles-même au calcul de la courbure et de la vitesse le long de la trajectoire.



(3) Tracé de la courburecorrespondant au geste

(4) Le geste segmenté

La modélisation de la trajectoire est donc la suivante. Il s'agit d'une suite de 6 éléments que nous allons expliciter :

(COURBE 0 25) Une courbe commençant au point 0 et se terminant au 25ème point.

(INTERSECTION 25 207 67 AI) 1er recoupement correspondant au point d'intersection.

(COURBE 26 212)

(INTERSECTION 212 207 67 AI) 2er recoupement correspondant au point d'intersection

(COURBE 213 240)

(VITESSE-NULLE 240 132 68 NIL) Un point où la vitesse d'exécution du geste est nulle.

Le premier élément indique la présence d'une courbe commençant au point 0 de la B-spline et se terminant au 25ème point de celle-ci. Le second élément correspond au 1er recoupement du point d'intersection situé au 25ème point de la spline au point de coordonnées (207, 67) de l'écran. Le dernier élément de la modélisation indique que la vitesse de déplacement au 240ème et dernier point de la spline est nulle 5 ce qui n'est pas étonnant puisque le sujet a ici terminé sa trajectoire.

L'étude de cette modélisation permet donc d'identifier comme partie significative de cette trajectoire l'entourage fermé délimité par les deux points de recoupement du point d'intersection qui s'étend du 25ème au 212ème points de la B-spline(ENTOURAGE 25 212 207 67).

5.3.2 Analyse de la 2ème zone de contact

L'analyse des événements recueillis montre qu'il s'agit d'un point de contact avec l'écran, il s'agit très probablement d'un geste de pointage au point de coordonnées (464, 62). La suite des événements étant la suivante :

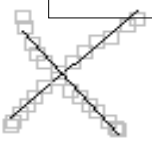
(4 463 62 65893 2)

(6 464 62 66004)

(5 464 62 66115 0).

5.3.3 Analyse des 3ème et 4ème zones de contact

L'analyse montre qu'il s'agit de deux droites dont les meilleures approximations sont calculées. On peut remarquer qu'il s'agit d'un geste multicontact contrairement aux deux précédents et qui est identifié comme une croix.

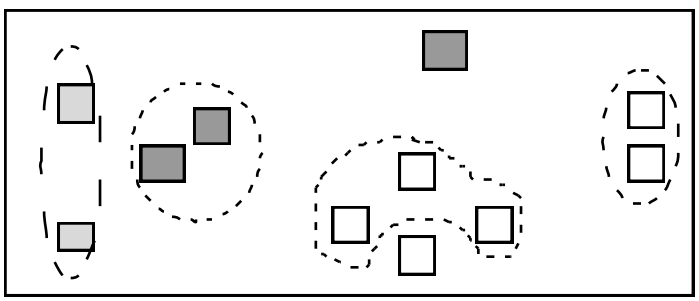


Tracé des événements et de la meilleure approximation pour le geste multicontact en forme de croix.

5.4 Calcul des objets candidats aux désignations

La scène présentée au sujet comporte des objets identiques en nombre variables. On peut remarquer qu'il n'y a pas de superposition d'objets ce qui facilite quelque peu l'interprétation des désignations.

Il est cependant nécessaire de construire une représentation de la répartition spatiale des objets qui fait apparaître des groupes d'objets perceptifs comme indiqués sur la figure suivante (les représentations ont été simplifiées) :



L'interprétation des désignations détectées permet d'identifier comme candidats :

- pour la désignation correspondant à l'entourage : les objets 3 et 4.qui sont contenus dans la surface délimitée par l'entourage.
- pour la désignation correspondant au pointage : l'objet 9. Remarquons que l'interprétation d'un geste de pointage par rapport à la scène sur laquelle il intervient ne consiste pas simplement à trouver l'objet directement situé au point de contact mais l'objet ou le groupe d'objets situés dans la périphérie du point de contact.
- pour la désignation en croix qui ne correspond à aucun objet à déplacer, on détermine une zone correspondant à la surface occupée par la croix.

5.5 Analyse conjointe avec l'énoncé langagier

Il s'agit maintenant de prendre en compte les expressions référentielles détectées dans l'énoncé accompagné du geste et d'établir une correspondance avec les désignations détectées dans le geste sur la base d'une synchronisation des plages temporelles. Dans l'énoncé de notre exemple figurent trois expressions à valeur déictique : " ces deux peluches", " celle-là" et "ici" qui mises en relation avec les trois désignations détectées fournissent les informations suivantes :

- *ces deux peluches* + entourage (objet3, objet4) : l'expression indique que le référent comporte deux objets de type "peluche", conditions vérifiées par le groupe (objet3, objet4). Le référent est donc ici le couple (objet3, objet4).
- *celle-là* + pointage (objet9) : l'expression démonstrative reprend le type du référent précédent "peluche" qui correspond effectivement au type de l'objet9 qui est donc le bon référent.
- *ici* + croix (surface) : le déictique "ici" correspond à lieu déterminé par un geste, ce qui est effectivement le cas. Il faut noter que le point d'intersection de

la croix détermine l'origine de la surface qui peut être étendue aux dimensions des objets concernés par le déplacement ou la création.

6. Conclusion

Nous avons essayé dans ce papier de présenter une vue d'ensemble des difficultés qui se posent pour mettre en œuvre des systèmes de communication homme-machine multimodaux. En particulier, nous avons voulu montrer qu'au delà de réponses technologiques ponctuelles, il s'agit d'un domaine qui reste véritablement à comprendre et à conceptualiser dans le cadre de collaborations qui ne peuvent être que pluridisciplaires. Sur le fond, l'étude de la multimodalité permet de donner un éclairage nouveau aux phénomènes référentiels en les placant au centre d'une interaction entre langue, geste et perception. De la sorte, les modèles linguistiques à venir se doivent d'être compatible avec les contraintes issues de la prise en compte des structures perceptives et s'ancrer encore plus sur les effets contextuels dans le cadre de modèles pragmatiques étendus. Au bout du compte, l'informatique peut elle aussi jouer son role d'intégrateur en s'assurant, via la définition de modèles implantables, que les résultats des autres disciplines ne sont pas contradictoires.

7. Références

Bellalem Nadia 1995, *Etude du mode de désignation dans un dialogue homme-machine finalisé à forte composante langagière : analyse structurelle et interprétation*, Thèse d'informatique, Université Henri Poincaré-Nancy 1.

Bellalem Nadia et Laurent Romary 1995, Langue et geste pour le dialogue homme-machine finalisé, *Communication en conception*, EuropIA, Paris, p. 185-201.

Bellalem Nadia et Laurent Romary 1995, Reference interpretation in a multimodal environment combining speech and gesture, Actes First International Workshop on Intelligence and Multimodality in Multimedia Interfaces, Edinburgh.

Braffort Annelies, Thomas Baudel et Daniel Teil 1992, Utilisation des gestes de la main pour l'interaction homme-machine, Actes *IHM'92*, Paris/1, p. 193-196.

Cadoz Claude 1992, Le geste canal de communication homme/machine - la communication instrumentale, *Technique et Science Informatique, 13*, 1, p. 31-61.

Caelen J. 1991, Interaction multimodale dans ICP-Draw, expérience et perspective, Actes *IHM'91*, GRECO-PRC Communication Homme-Machine.

Dauchy P., C. Mignot et C. Valot 1993, Joint speech and gesture analysis : some experimental results, Actes *Eurospeech 93*, , p. 1315-1318.

Forgus R. H., *Perception*, McGraw-Hill Book Compagny, 1966.

Mignot C. 1995, *Usage de la parole et du geste dans les interfaces multimodales - Etude expèrimentale et modèlisation*, Doctorat d'Université, Université Henri Poincarè, Nancy I.

Mignot C., C. Valot et N. Carbonell 1993, An Experimental Study of Future ``Natural'' Multimodal Human-Computer Interaction, Actes *INTERCHI'93* 1993 Conference on Human Factors in Computing Science INTERACT'93 and CHI'93, Amsterdam (The Netherlands).

Naughton Karen 1996, Spontaneous Gesture and Sign - A Study of ASL Signs Co-occuring with Speech, Actes *WIGLS* (Workshop on the Integration of Gesture in Language and Speech), Newark and Wilmington (Delaware), p. 125-134.

Wolff Frederic, Laurent Romary et Nadia Bellalem 1997, Perception et action dans le cadre d'une interface homme-machine multimodale: Etude experimentale, Actes Colloque JOSC.