# Interaction et multimodalité

Jean Caelen

# INTERACTION ET MULTIMODALITÉ

**Jean Caelen**

Laboratoire CLIPS-IMAG
BP 53, Domaine universitaire
38041 Grenoble Cedex 9
Jean.Caelen@imag.fr

***Résumé*** *: Cet article pose les problèmes généraux, concepts et principes concernant les interfaces homme-machine multimodales :*
- *l'usage et l'adéquation des modes de communication ;*
- *le contexte d'interaction : concurrent, alterné, composé, parallèle ;*
- *la gestion des événements bas niveaux (cohérence, chronologie, redondance, etc.) ;*
- *la fusion des informations ;*
- *leur interprétation (coréférence) ;*
- *les niveaux d'abstraction et de représentation.*

*On situe quelques uns de ces points à travers une taxonomie des interfaces. On aborde le difficile problème de la coréférence entre les modes, placé dans une perspective de fusion d'informations auquel on propose une solution. Une application multimodale de conception de dessin ICPdraw est décrite pour en illustrer le propos. ICPdraw est construit sur une architecture en couches comprenant :*
- *des cartes de reconnaissance et de synthèse dédiées ;*
- *des serveurs d'événements ;*
- *un gestionnaire des modes ;*
- *un contrôleur de dialogue ;*
- *une interface de communication avec le noyau fonctionnel de l'application.*

***Mots-clefs :*** *Interface homme-machine, Interaction multimodale, Communication homme-machine*

## 1. INTRODUCTION : L'INTERACTION HOMME-MACHINE

On distingue plusieurs situations dans lesquelles la machine informatique peut apporter son concours à la communication humaine ou à la communication homme-machine (ce dernier terme est d'ailleurs à utiliser avec beaucoup de prudence). Il s'agit de situations où,

- la machine est *médiatrice*, elle met en communication au-delà des distances, des personnes qui collaborent à un même objectif ou travaillent ensemble (via un collecticiel (Pankoke, 89), une téléconférence (Stefik, 87), etc.). Dans ce cas les documents échangés ou manipulés devraient être multimédias pour être pleinement informatifs ;

- la machine anime une *réalité virtuelle* en étendant les capacités d'expression et de création humaines par immersion de l'individu dans un monde avec lequel il réagit (essentiellement par interaction gestuelle mais bientôt par interaction langagière) (Brooks, 88) ;
- la machine est *partenaire*, elle collabore à une tâche faite par un utilisateur avec qui elle dialogue pour comprendre ses buts, voire ses intentions, afin de rendre la session de travail plus efficace.

À ces situations s'ajoute le fait que l'interaction peut être multi-sensorielle (Coutaz et al., 90). Elle est dite alors multimodale si elle met en jeu simultanément et de manière coopérante, plusieurs modalités sensorielles et motrices de l'être humain ou *modes de communication* - vision, parole (entendue et produite), geste (mouvement, désignation, écriture, dessin), etc. - et qu'il y a nécessité pour la machine de comprendre puis d'interpréter les informations circulant sur les différents canaux d'entrée-sortie ou *médias* (*IHM'91*, *IHM'92*). Une interaction efficace ne s'obtient pas en juxtaposant simplement plusieurs modes indépendamment les uns des autres dans une machine dont l'architecture logicielle n'a pas été pensée pour la composition des modes entre eux ; l'utilisateur doit pouvoir dire, typiquement, « mets ça là » en désignant par le geste un objet (→ ça) et un lieu (→ là) : ce message, d'essence multi-sensorielle, n'est donc interprétable qu'en fusionnant les informations parlées et gestuelles. Ceci fait clairement naître un problème de résolution de référence spatiale et temporelle intermodale qui passe par une interprétation complexe des événements multimodaux.

Pour le concepteur d'interfaces, les situations 1-3 énumérées ci-dessus présentent des points communs qu'il s'agit de réutiliser au maximum. Il est clair qu'il existe un premier niveau de transport des informations : ce niveau est multi-média. Mais ces situations et les systèmes qu'elles engendrent se distinguent ensuite par les couches successives de gestion et d'interprétation des informations véhiculées en entrée et en sortie. Dans la situation 1 (machine médiatrice) le système peut être doté d'une gestion asynchrone (Bastide, 92) (boîte à lettres par ex.) comme d'une gestion synchrone (collecticiel d'édition de documents par ex.) des informations, accompagnée de divers niveaux de dialogue pour finir par une gestion des rôles des participants (qui règle la question des droits et des privilèges attachés aux objets partagés). Dans la situation 2, les aspects comportementaux de l'individu (à travers le temps réel et la simulation) sont mis en valeur au détriment du dialogue proprement dit qui se réduit au schéma élémentaire action-réaction, contrairement à la situation 3 qui privilégie la relation individu-machine (plus exactement la relation opérateur-tâche) où l'individu est seul face à la machine et où celle-ci doit donc avoir des capacités de communication inspirées de la communication humaine pour rendre l'exécution de la tâche efficace et fiable. Le degré de finesse de l'interprétation des informations multimodales est donc variable selon ces systèmes. Malgré tout, au-delà de la plus ou moins grande sophistication de ces processus, on distingue trois types de traitement propres aux interfaces multimodales : la gestion des modes, la fusion des informations en entrée, leur fission en sortie.

Dans l'article ci-après nous n'aborderons que ces aspects de l'interface, principalement sous l'angle de la multimodalité (Caelen, 92b), restreinte à la voix et au geste (par souci de simplification de l'exposé). Il sera essentiel de garder en permanence un double point de vue : celui de l'utilisateur et celui de la machine

(plus exactement de la technologie), puisque l'interface est le point de passage de l'un à l'autre et doit dans sa conception, être un compromis viable entre les exigences ergonomiques et les contraintes technologiques.

## 2. ÉLÉMENTS D'ERGONOMIE

Les interfaces multimodales posent de nouveaux problèmes d'ergonomie cognitive au sujet des modes de communication qui sont offerts à l'utilisateur : parole, geste, vision, etc. Se pose donc le problème de l'adéquation des modes de communication venant s'ajouter aux problèmes plus classiques de l'adéquation des représentations et des traitements qui doivent être compatibles avec les objectifs et le raisonnement humain pour toute machine (Scapin, 86), (Barthet, 88). En se restreignant aux modes seuls, l'adéquation concerne les modalités sensorielles et motrices proprement dites, mais aussi leur utilisation optimale vis-à-vis de la tâche, de la charge cognitive, des représentations mentales, des types d'usagers, etc. On peut résumer ces exigences de manière schématique comme suit :

| | | | | |
|---|---|---|---|---|
| $\mathcal{H}$ | ☞ | $\mathcal{M}$ | + | $m$ |
| *modalités sensorielles* | | | | *modes* |
| *représentations* | | *modèles, objets* | | |
| *raisonnements* | | | | *traitements* |

dans ce schéma $\mathcal{H}$ représente l'utilisateur (ou opérateur), $m$ la machine dont il n'a qu'une représentation abstraite et $\mathcal{M}$ le monde perceptible pour lui (métaphorique, réel ou virtuel) qui donne sens à ses représentations. Les modalités sensorielles et motrices de $\mathcal{H}$ doivent s'accorder aux modes de $m$ ainsi que, respectivement, les raisonnements et les traitements. Le terme multimodal, relevant de l'aspect multisensoriel de l'individu, nécessite la prise en compte :

- de l'usage et de l'adéquation des modes à la tâche ;
- des stratégies d'interaction adaptées aux compétences et performances des usagers ;
- de la gestion des événements bas niveaux (cohérence, chronologie, redondance, etc.) prenant en compte les capacités perceptives et motrices de l'individu ;
- des niveaux communs d'abstraction et de représentation ;
- de la présentation et des points de vue multimodaux ;
- et plus généralement du modèle de l'usager dans sa dimension cognitive.

Ce qui devrait apparaître à l'utilisateur au niveau de l'interface est le reflet de la structure des données et des tâches en machine. Les contraintes technologiques ont imposé jusqu'à maintenant cette structure, tendance qui se renverse petit à petit au profit de l'utilisateur et donc de l'ergonomie. À l'autre extrémité, la question qui se pose maintenant est de savoir jusqu'où structurer l'interface en fonction du

comportement de l'utilisateur. Celui-ci a, en effet, une activité organisée soit par sa formation, son expérience de la tâche, la pratique, son savoir-faire et les objectifs qu'il doit atteindre : dans une certaine mesure son activité est planifiée, il a une intention de départ et un plan d'action qu'il réorganise en fonction des contraintes de la machine. Il a des habitudes et des « préférences » dans les modes de communication, des expressions idiosyncrasiques voire des « manies ». On constate par exemple, à l'heure actuelle, des « habitudes gestuelles » nouvelles chez les utilisateurs de systèmes informatiques possédant une souris. Ces habitudes sont parfois loin d'être optimales. Souvent un langage serait plus adapté s'il pouvait être mis en œuvre efficacement. Examinons ces points, en opérant deux classes parmi les modes, les modes langagiers et les modes non-langagiers.

## 2.1 Les modes langagiers

Vis-à-vis de l'emploi des divers modes de communication peu d'études nous permettent de dire à l'heure actuelle si des interfaces multimodales seront plus fiables, plus souples et plus efficaces, puisque ces interfaces ne sont pas encore réalisées. Cependant, par des expériences partielles on a pu montrer que la parole est souhaitée dans certains cas. Nous donnons quelques résultats ci-après (Brandetti et al., 88), (Falzon, 90) pour des interfaces utilisant la parole en entrée :

- la satisfaction globale des usagers dépend des catégories socioprofessionnelles ;
- l'apprentissage de l'interface est en général plus rapide ;
- l'efficacité dans la réparation des erreurs est améliorée.

mais

- les contraintes d'utilisation sont parfois limitantes (contexte bruyant, confidentialité, etc.) ;
- le niveau de langage compris par la machine nécessite une adaptation de l'usager.

Dans cet état des choses, la conception d'un « dialecte » dérivé de la langue naturelle (plutôt qu'un sous-langage ou qu'un langage formel ou artificiel) paraît être la meilleure solution :

- pour faciliter l'apprentissage des entités et des opérations par l'utilisateur ;
- pour la mise en œuvre en machine car le lexique est bien défini et la syntaxe limitée.

On retrouve ces mêmes spécifications dans les langages opératifs homme-homme où dans des cas extrêmes, il n'y a pratiquement pas de syntaxe, le vocabulaire est limité mais très spécialisé (Falzon, 90). Un tel langage est très lié à la nature de l'application. Par contre, dans des applications grand public liées à des requêtes orales, la richesse de la langue de communication peut être très grande et comprendre de nombreux cas d'ellipse. Devant une machine, les utilisateurs « s'adaptent » en rendant leurs énoncés plus clairs : moins d'ellipses et d'anaphores, syntaxe plus souvent correcte (même si on ne leur demande pas). Du côté de la prosodie (contours intonatifs de la phrase, rythmicité) on a pu se rendre compte d'un phénomène analogue en situation de lecture contrainte par des consignes

d'intelligibilité (Caelen-Haumont, 91) : les locuteurs ont tendance à séparer davantage les mots, voire les syllabes. Inversement, la production verbale se dégrade avec la charge de travail ou la concentration sur l'objectif.

Les limitations dans l'utilisation de la parole ne viennent donc pas uniquement des performances encore insuffisantes des systèmes de reconnaissance de la parole (Siroux et al., 89) puisque les langages reconnus par la machine correspondent à certaines catégories de langages opératifs. Elles tiennent plutôt aux caractéristiques du mode parlé lui-même. Indiquons enfin que par rapport à l'écrit, le mode parlé semble avoir un net avantage, l'écrit étant pénalisé par la présence du clavier qui limite le débit d'entrée des informations et mobilise toutes les ressources sensori-motrices de l'usager. L'introduction de la tablette numérique comme moyen d'entrée peut à cet égard changer les choses (Faure, 93).

## 2.2. Les modes non langagiers

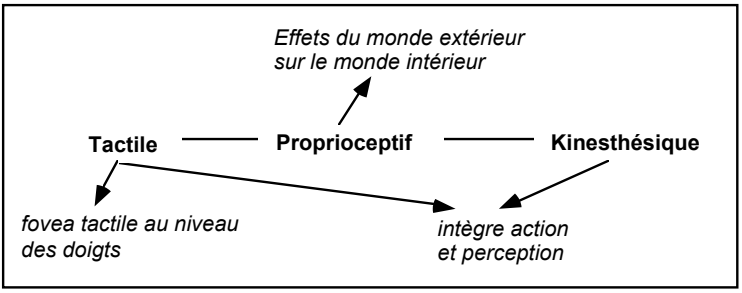
***Le geste***

La communication entre humains sous-tend presque toujours un langage, c'est-à-dire un échange de symboles via un code partagé par les interlocuteurs. L'interaction est une forme de « communication » à un niveau non-symbolique : interagir c'est émettre des commandes ou faire des actions et recevoir des stimulations. Le geste participe clairement de l'interaction (appuyer sur un bouton par ex.) mais ne remplit pas forcément de fonction de communication au sens strict du terme. Le geste n'est apparu que récemment comme mode d'interaction avec la machine car (a) on s'est rendu compte assez tardivement qu'il pouvait être utilisé de manière complémentaire avec les autres modes - chez l'humain le contrôle du geste est transmodal, il passe par la vision, l'audition ou la proprioception (Hécan et al., 75) - et (b) il nécessite une technologie assez lourde pour sa capture (Rubine, 91).

À travers le geste, l'action et la perception sont fortement imbriquées : toute perception est action et réciproquement. Le geste est donc très difficile à modéliser. Pour l'être humain on compte quelque 700 muscles, 110 articulations et une centaine de degrés de liberté. Comme en vision il existe une sorte de « fovea » active à travers laquelle s'effectue l'échange d'énergie avec l'environnement. Cette fovea est surtout localisée au bout des doigts.

On distingue trois fonctions du geste (Cadoz, 92) :

- ergotique (ou ergative), qui est une transformation énergétique, matérielle de (et sur) l'environnement (le geste vu comme une force, une énergie) ;
- épistémique, qui permet par le toucher notamment de prendre conscience et connaissance des objets du monde, d'acquérir des connaissances sur l'environnement. La fonction épistémique peut être définie par le triplet (T, P, K) T=tactile, P=proprioceptif, K=kinesthésique, dans lequel les trois termes sont inséparables. En effet pour prendre conscience d'une forme il faut émettre une action sur cette forme (pression, mouvement, etc.) pour en recevoir des effets ;




***La fonction épistémique du geste***

- sémiotique, qui donne au geste une fonction communicante (langage gestuel des sourds-muets par ex.), ou de manière plus restreinte qui accompagne d'autres langages (parole notamment) dans une fonction complémentaire de désignation (déixis), de rythme, d'iconicité, etc. La fonction sémiotique peut se subdiviser en domaines, comme :
  - l'idéographie, le dessin, l'écriture ;
  - les appuis langagiers ;
  - le geste instrumental ;
  - la communication langagière que l'on retrouve de nouveau ici : on peut par exemple définir des langages sémantiques ou de signes (langage des sourds-muets) ou des langages totalement artificiels et conventionnels pour l'interface homme-machine dans laquelle par exemple la sélection remplit le rôle du complément d'objet direct du verbe de commande, le geste d'excitation le rôle du verbe (double-clic prenant valeur de « ouvrir le fichier ») et le geste modifieur le rôle de l'adjectif ou de l'adverbe (lentement, rapidement, par ex.). Notons que certains gestes remplissent plusieurs fonctions simultanément, celui du chef d'orchestre par exemple qui marque un rythme (appui), une expression musicale (langage) et engage les musiciens à entamer leur partie (commande).

De l'ensemble de ces points de vue, le griffonnage et l'écriture manuscrite ne sont pas des gestes si seul compte le résultat produit et non la manière de le produire. Est-ce le geste qui est sémiotique ou sa « trace » laissée sur le périphérique d'entrée ? Le « geste » sur le clavier ou la souris, qui se réduit à sa fonction d'activation de boutons pour lesquels seule la notion du « où » compte et non le « comment » et le « quand », est-il encore un geste ?

Ainsi donc le geste pour une interface homme-machine présente des aspects très différents voire opposés : on peut passer d'une interface instrumentale - intégrant l'instrument lui-même pour laquelle le geste est action et perception - à une interface où le geste perd tout matérialité : il est vu (au sens propre du terme) et reconnu comme trace d'un autre signifié.

***La vision***

La vision peut permettre de capter des gestes à l'aide de caméras, et parmi ceux-là des expressions faciales (Turk, 91). Ces expressions sont identifiées par des méthodes de reconnaissance de forme classiques (calcul d'indices à partir d'images

pour lesquelles il faut s'affranchir des contraintes d'orientation, puis classification). Des recherches peuvent aussi être envisagées pour synchroniser la reconnaissance de la parole sur le mouvement des lèvres. La vision pour capter le geste n'ajoute qu'un problème technique supplémentaire sans modifier le rôle fondamental de ce dernier. Il en va autrement de la vision en robotique si l'on accepte le robot comme une interface mécanique avec le monde. À ce moment la vision devient un organe de perception du robot pour ses déplacement et son repérage dans l'espace. On retrouve, là encore, la vision comme contrôle du geste (soit de déplacement soit de préhension) (Brooks, 88).

### 2.3. Adéquation des modes

Dans les interfaces graphiques, le paradigme de « manipulation directe » semble avoir atteint ses limites (Buxton, 93) : on ne peut désigner ce que l'on ne voit pas ; la séquence *sélection-opération* (qui renverse l'ordre objet-verbe habituelle dans le langage) est souvent inadéquate. Ainsi, tout plaide actuellement en faveur de l'introduction de la multimodalité dans les interfaces. Bien que peu d'études ergonomiques systématiques aient pu être faites (puisque ces interfaces ne sont pas vraiment disponibles encore), on a pu faire quelques constations générales telles que celles-ci :

***Mode parlé***

usage en entrée : commandes, macro-commandes (mots isolés, parole continue),

usage en sortie : guides, exemples, requêtes, explications, relances, (synthèse, phrases à trous)

***Mode écrit***

usage en entrée : identificateurs, nombres (clavier, tablette graphique)

usage en sortie : explications détaillées (écran)

***Mode gestuel***

usage en entrée : désignation 2D ou 3D (souris, gant numérique, écran tactile), langage de signes (caméra), action ergative (clavier interactif)

usage en sortie : retour d'effort, effet tactilo-kinesthésique

***Mode visuel***

usage en entrée : orientation du sujet, expression du visage

usage en sortie : graphiques, images, animation (synthèse d'images, graphique animé)

## 3. PROBLÉMATIQUE DES INTERFACES MULTIMODALES

Les problèmes qui distinguent les interfaces multimodales des interfaces classiques naissent de la diversité des modes en entrée et en sortie dont il faut analyser, interpréter et générer les informations de manière croisée et dépendante. Ces problèmes concernent :

- la gestion des modes aux niveaux (Bourguet et al., 92)
  - des événements (chronologie, synchronie),
  - des informations (unités, actes),
  - et du contexte interactionnel ;
- la fusion / fission des informations au niveaux
  - morphosyntaxique,
  - sémantique et/ou pragmatique (résolution de la coréférence),
  - actionnel (intégration de la multimodalité au niveau de la couche interaction / dialogue) ;
- l'échange des informations avec les autres modules de l'interface et le noyau fonctionnel de l'application.

À chaque mode, est associé un modèle de représentation des informations qu'il véhicule. Ce modèle dépend de la granularité des événements de bas niveau sur laquelle il est construit. Ainsi pour un « geste » le système délivre des vecteurs de coordonnées de points dans le temps alors que pour la parole ce sont des chaînes de caractères correspondant à des mots ou des phrases reconnues ou même le signal échantillonné brut. Les fréquences d'échantillonnage de ces données sont différentes d'un média à l'autre. En se limitant au problème des entrées (le problème étant symétrique pour les sorties) on peut voir aussi, avec un regard fonctionnaliste, des « couches » de traitement dans les interfaces allant d'un niveau concret, les signaux, à un niveau abstrait, le déclencheur de l'action. Ce sont :

(a) l'acquisition des signaux fournis par l'usager,

(b) leur reconnaissance automatique,

(c) la compréhension des signes qu'ils véhiculent,

(d) leur interprétation coréferentielle,

(e) la construction d'un message actionnel multimodal.

Le cheminement des informations passe par une mise en forme, une représentation abstraite, une fusion et enfin une transmission à la couche « contrôle du dialogue » qui se trouve de fait posé au niveau le plus haut.

## 4. LA GESTION DES MODES

Pour avancer clairement dans la problématique présentée ci-dessus, il est important de bien distinguer les événements qui reflètent l'organisation physique des actes, des informations (ou unités qui les composent).

### 4.1. Événements, informations

***Définition d'un événement*** : un événement est un début, ou une fin d'un signal externe à la machine : il signale un changement perceptible sur un média. Cette définition est centrée sur la machine et non sur l'utilisateur, plus précisément sur les canaux d'entrée-sortie que nous appelons médias.

***Définition d'une information*** : une information est une unité signifiante, mais qui ne prend pas la même signification pour l'usager et pour la machine. C'est :

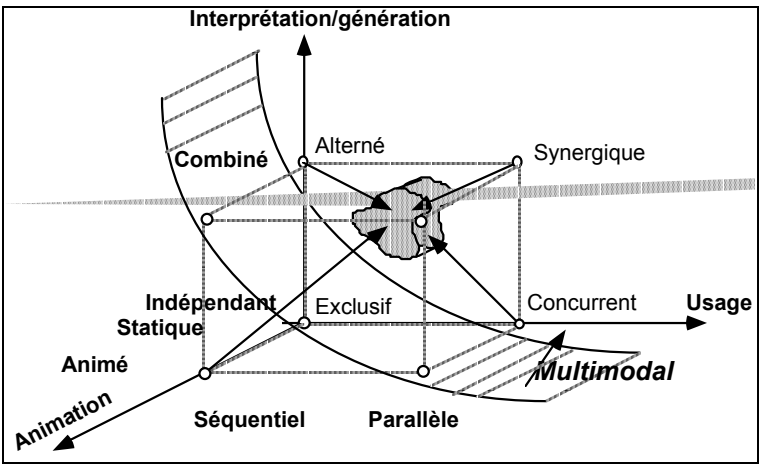
- une unité sémiotique pour l'usager ;
- une unité référentielle pour la machine.

***Définition d'un acte et d'une action :***

- un acte est une suite d'unités sémiotiques émises (ou reçues) par l'usager. Cette suite est portée par un signal (parole, geste) délimité par des marqueurs propres à chaque mode (pauses pour la parole, enfoncement, relâchement de la souris, etc.). Cette suite a une organisation temporelle définie par une syntaxe. Pour la parole l'acte est une notion qui correspond à celle d'*acte de langage* ;
- une action est une opération effectuée par la machine manifestée par un changement d'état rendu perceptible ou non à l'utilisateur.

## 4.2. Le contexte interactionnel

***Définition du contexte interactionnel*** : le contexte interactionnel est le triplet {usage des modes, dépendance des informations, animation}. Le premier attribut dénote l'usage (de facto les capacités du système) séquentiel ou parallèle des modes, le second l'indépendance des informations véhiculées sur les médias et le troisième la dynamique du monde c'est-à-dire les actions à effet continu et les actions à effet instantané. Nous ne nous intéresserons qu'aux deux premiers attributs qui définissent quatre contextes interactionnels : exclusif, concurrent, alterné et synergique (Caelen, 91), (Coutaz, 92). Le contexte « exclusif » est celui de tout système informatique qui possède au moins deux médias distincts en entrée ou en sortie que l'on utilise de manière indépendante. Ce cas se situe en dehors du cadre de cet article car il n'est pas vraiment multimodal. Les trois autres sont décrits ci-après.




***Le domaine des systèmes multimodaux : alterné, concurrent et synergique***

***Le contexte « Concurrent »***

Il se définit par :

- usage parallèle des modes : sans contraintes temporelle
- indépendance des canaux qui entraîne : pas de coréférence intermodale entre les unités, $L\{u_{ij}(k), u_{i',j}(l)\} = \emptyset$ pour $i \neq i'$.

Propriétés : On ne peut traiter les déictiques dans ce contexte et l'anaphore est mal résolue lorsque la référence est portée par un autre mode.

***Le contexte « Alterné »***

Il se définit par :

- usage séquentiel des modes : $Début_i(k) \geq Fin_{i'}(k-1)$ avec $i \neq i'$ ;
- indépendance des unités qui entraîne : pas de contraintes coréférentielles.

Propriétés : L'anaphore est bien résolue lorsque la référence est portée par un autre mode, la déixis peut être résolue. L'usage alterné des modes entraîne une lourdeur programmatique des actions qui pénalise la coordination perceptivo-motrice de l'usager.

***Le contexte « Synergique »***

Il se définit par :

- usage séquentiel des modes : aucune contrainte ;
- indépendance des unités qui entraîne : pas de contraintes coréférentielles.

Propriétés : L'anaphore est bien résolue lorsque la référence est portée par un autre mode, la déixis également. L'usage synergique semble être la meilleure solution si l'on sait résoudre les problèmes coréférentiels intermodaux, c'est également le plus économique au niveau sensori-moteur. Mais nous verrons qu'elle pose problème pour traiter les anticipations ou les retards.

***Le contexte interactionnel (dans un système dynamique)***

Un système est dit dynamique s'il est capable de gérer différents contextes interactionnels. Le contexte interactionnel a été décrit ci-dessus. C'est le triplet $C$ = {usage des modes, dépendance des informations, temporalité}

usage des modes : il est déterminé par la boucle action-perception et les contraintes mécaniques du système

ex. Mettre(Objet, Lieu)

« mets ça ici » < dg(ça) < dg(ici) => alterné
(« mets ça ici » ≈ dg(ça)) < dg(ici) => synergique(p+)
(« mets ça » ≈ dg(ça)) < (« ici » ≈ dg(ici)) => synergique
« mets » < (« ça » ≈ dg(ça)) < (« ici » ≈ dg(ici)) => synergique(g+)

avec

« » = acte de parole
dg = acte de désignation gestuelle
p+ = dominance du mode parole
g+ = dominance du mode gestuel

dans le dernier cas le geste rythme la parole et la détermine temporellement. Les événements sont synchrones et les informations dépendantes ; on en déduit que le contexte interactionnel est synergique à dominance gestuelle.

dépendance des informations : elle est déterminée par les relations sémantiques/pragmatiques entre les unités

ex. dg(triangle) ≈ « déplace le cercle » => concurrent

les deux actes sont synchrones et indépendants car l'objet désigné triangle ne coréfère pas avec l'objet cercle de l'acte de parole. On en déduit le contexte interactionnel « concurrent ».

Ces quelques exemples montrent que le contexte interactionnel se déduit de l'organisation et du contenu même des actes. Cela fait qu'il ne peut être déterminé que de manière indirecte.

### 4.4. Fonctions

En résumé, la gestion des modes est une opération qui consiste à :

- capter les événements en provenance des serveurs de médias (inversement à émettre pour les sorties) ;
- construire les structures événementielles et informationnelles ;
- gérer le contexte interactionnel, en fonction du type d'information et des connaissances transmises par les niveaux adjacents (module de fusion, module de dialogue par exemple) ;
- maintenir un historique pour ce contexte ;
- mettre à profit les connaissances sur l'usager au niveau sensori-moteur (temps de réaction, préférences modales, etc.).

## 5. FUSION-FISSION DES INFORMATIONS

Le problème central dans une interface homme-machine multimodale se situe dans la fusion (en entrée) et la fission (en sortie) des informations intermodales. Placé au-dessus de la gestion des modes, le module qui traite de la fusion (resp. fission) fait le lien avec le module qui traite du dialogue. Ce dernier a des rôles bien définis dans les interfaces monomodales. Ces rôles sont :

- construction d'un univers sémiotique de communication (métaphores, langages, etc.) ;
- structuration et organisation de la communication ;
- gestion et contrôle dynamique de la communication ;
- construction d'une interaction coopérante ;
- réparation des erreurs de communication ;
- aide à l'apprentissage, guidage dans la tâche.

Cerner les fonctions et les limites d'un module de fusion est chose délicate, car sa spécificité peut être contestée (Gaiffe et al., 91) : on pourrait en attribuer tous les rôles au contrôleur de dialogue qui analyserait les informations prélevées au bas niveau et se chargerait de la fusion des informations dans un processus englobant.

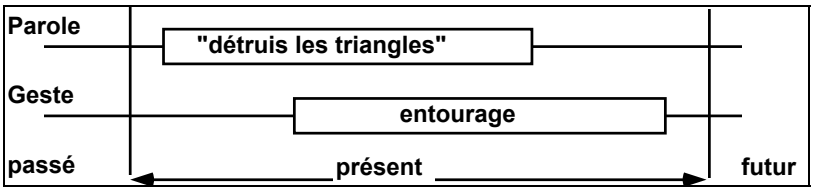
Quelles sont les raisons qui plaident en faveur d'un tel module distinct et spécifique pour les IHM multimodales ?

## 5.1. Exemples de problèmes de résolution de références

Pour introduire la discussion, examinons quelques cas typiques de commandes multimodales simples.

***Référence à une collection d'objets***

Soit l'acte de parole « détruis les triangles » synchrone de l'acte gestuel d'entourage d'objets graphiques illustré ci-dessous :




La notion de *présent* (plus exactement d'épaisseur de présent) étant maintenant bien définie, une interprétation correcte de ces deux actes dépend de l'intention de l'utilisateur et du contexte de production de l'acte. Ce contexte se subdivise à son tour en :

- contexte interactionnel ;
- contexte linguistique ;
- contexte dialogique (ou discursif) ;
- contexte actionnel (ou de la tâche).

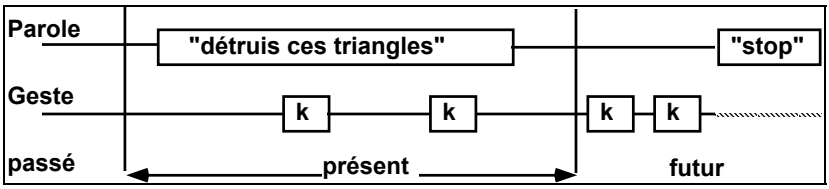
Ces contextes dépendent tous les uns des autres. Illustrons cela sur l'exemple de la manière suivante :

- si contexte interactionnel = synergique, il faut interpréter le message comme un tout signifiant *détruire « tous les » triangles parmi ceux qui sont sélectionnés*. L'article « les » prend alors la valeur d'un déictique et doit être fusionné avec l'information gestuelle. Inversement si l'acte avait été *détruire ces triangles* le contexte linguistique aurait imposé le contexte synergique par le déictique « ces » (au risque d'attendre une information gestuelle dans le futur) ;
- si contexte interactionnel = concurrent (en supposant par exemple une anticipation du geste sur la parole pour des raisons variées), il faut interpréter le message comme un double message signifiant *détruire « les » triangles référencés dans le passé* et *désigner des objets pour l'action future.* L'article « les » prend alors la valeur d'une anaphore et ne doit pas être fusionné avec l'information gestuelle ;
- s'il y a un conflit inter-modal vis-à-vis du contexte actionnel tel qu'un entourage d'un ensemble d'objets ne contenant pas de triangle en contexte synergique, alors il faut peut être remettre en question ce contexte pour tenter d'interpréter l'acte dans un autre contexte.

Dans cet exemple l'ambiguité du contexte linguistique et l'indétermination a priori du contexte interactionnel créent ce problème d'interprétation.

***Référence à une suite d'objets***

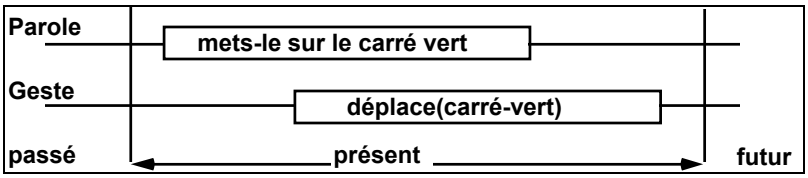
Dans ce deuxième cas l'acte « détruis ces triangles » est sans ambiguïté vis-à-vis de la gestuelle, mais celle-ci est très répétitive puisqu'on désigne tous les objets les uns après les autres par des clics-souris.




Les informations rattachées au présent sont dans un contexte synergique mais qu'en serait-il des autres si un marqueur de fin d'acte (comme « stop ») ne venait en indiquer l'appartenance ? Ici c'est donc le contexte actionnel qui crée l'ambiguïté résolue par le contexte de dialogue.

***Référence à un objet en mouvement***

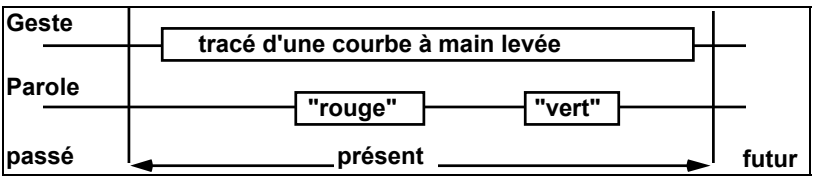
Ici l'acte « mets-le sur le carré vert » porte sur un objet en déplacement et donc virtuellement désigné puisqu'il est saisi à la souris. Est-ce le même qui sert d'objet de référence ?




Un autre cas similaire a été constaté en situation réelle, lorsque voulant optimiser ses actions en profitant au maximum du parallélisme qui lui est offert, l'utilisateur dit « mets-le sur le carré vert » tout en changeant la couleur du carré et détruisant par là la référence à un objet en cours d'action. Contextes actionnels et interactionnels créent ici l'ambiguïté.

***Référence indirecte***

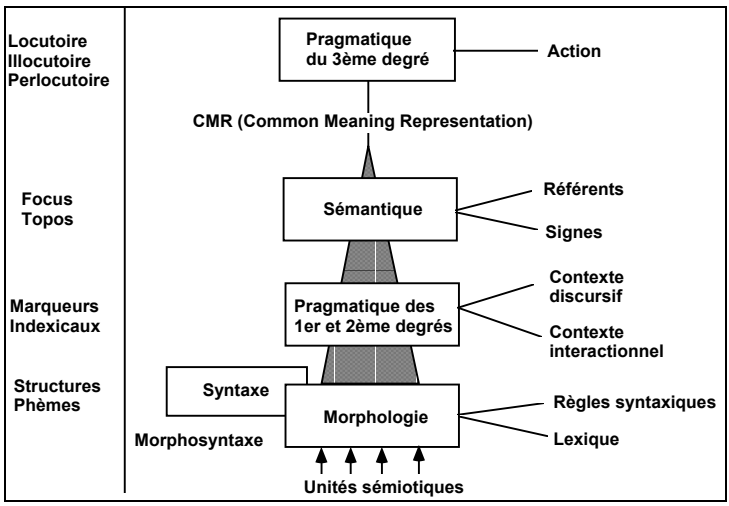
Dans cet exemple l'utilisateur dessine une courbe à main levée et change la couleur des portions de courbe par la voix. La référence « couleur » n'a de valeur que pendant la durée de l'action et se rapporte indirectement à l'objet instancié par l'action.



## 5.2. Les niveaux de fusion

À la suite de ces exemples il est clair que le rôle du module de fusion est de rendre l'interprétation (a) aussi indépendante que possible des contextes dans un premier temps et (b) de permettre une résolution progressive des références pour lever les ambiguïtés dans un deuxième temps. Accessoirement un tel niveau de fusion permet également d'ajouter de nouveaux modes sans avoir à modifier le contrôleur de dialogue en profondeur.

Ces deux contraintes nous conduisent alors à proposer une *fusion progressive* des informations partant des niveaux morpho-syntaxiques pour aboutir à la sémantique selon le schéma suivant :

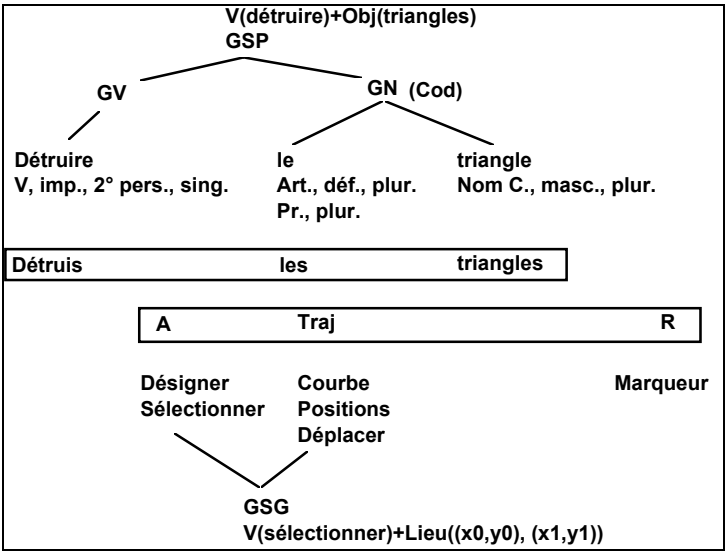



***Les niveaux de fusion***

Dans ce schéma la fusion s'opère à partir des unités collectées dans l'épaisseur du présent et fournit des structures de représentation abstraites (CMR = common meaning representation) débarrassées des composantes modales. Ces structures sont communiquées au contrôleur de dialogue. Détaillons chaque étape de la fusion.

### *Analyse morpho-syntaxique modale*

Une analyse morpho-syntaxique de chaque acte modal est faite sur l'épaisseur du présent. On obtient pour chaque mode une représentation adaptée qui décrit la structure des constituants et la structure fonctionnelle. Par exemple pour le cas (« détruire les triangles »≈ entourage) nous obtenons :




***Analyse pragmatique des 1^er^ et 2^ème^ degrés***

À ce niveau une analyse des indexicaux et des marqueurs pragmatiques par liage intermodal est opérée. Elle permet de relier les éléments référentiels libres d'un mode aux éléments référents des autres modes et de lier les actes entre eux.

Supposons que nous ayions pour l'acte précédent,
Contexte interactionnel = synergique, Contexte du dialogue = directif,
Contexte actionnel = action en cours, (variables transmises des modules connexes)

- analyse des indexicaux : « les »

Le liage de « les », qu'il soit déictique ou anaphorique, aux objets désignés par le mode gestuel se fait en considérant :

$\forall$ (objets) : Contrainte (objet $\in$ Domaine)
objet : $\exists$ objet instancié $\in$ Historique
Domaine = Intérieur(Traj) = Lieu((x0,y0), (x1,y1))
dont la résolution fournit une liste d'objets sans préjuger de leur catégorie sémantique (ce sont des triangles ou non).

- traitement des marqueurs : Prosodie, « R »

Un traitement de la prosodie (que nous n'envisagerons pas dans cet article) permet de fournir des marques sur le type d'acte par exemple la courbe mélodique étant continuellement descendante ici caractériserait l'énonciation d'« assertive »

« R » analysé comme marqueur de fin d'action gestuelle

Ces deux informations permettent de penser que l'acte est complet et donc qu'une hypothèse raisonnable est qu'il forme un tout. On pose l'hypothèse que Contexte interactionnel = synergique.

***Raisonnement sémantique (spatio- temporel)***

Ce raisonnement aboutit à la construction d'une CMR (Common Meaning Representation) par instanciation de schémas (d'action et d'objet). Ces mécanismes ressortissent de mécanismes complexes d'interprétation sémantique du langage naturel. Ils mettent en œuvre des bases de connaissance des actions et des objets ainsi que des règles d'inférence pour instancier ces schémas sur la situation courante. Leur degré de généralité font leur relative indépendance des domaines d'applications. Le formalisme adopté ci-dessous pour la représentation des connaissances utilise une grammaire de cas multimodale.

$ indique un prototype ou un nom de classe

Lien est un attribut qui permet de lier deux informations multimodales par ex. pour le lexème « ici » le lien permet de mettre en correspondance un autre mode désignant un $Lieu

La base de connaissances des actions

Action : Détruire
Activation = double-clic($SObj) | Verbe($Sdétruire)
OBJ = GN($Dominant = $SObj) | clic($SOb)
Temps = GP(prép($CTemps).GN) | Adv($CTemps)

Action : Mettre
Activation = mvt-clic($SObj) | V($SDéplacer)
OBJ = GN($Dominant = $SObj) | clic($SObj)
Lieu = GP(prép($CLieu).GN) | Adv($CLieu) | clic($SLieu)
Temps = GP(prép($CTemps).GN) | Adv($CTemps)

...

La base de connaissance des objets

Triangle : Sorte-de Objet-géométrique
Taille : GA($Dominant = $SObj) | mvt-clic($SObj)
Couleur : CN($Dominant = $SObj) | clic($SPalette
Position : (x,y)
Actions : {Détruire, Mettre}
Monde-référentiel : artificiel
Lien = Synonyme($Pyramide)

...

Le lexique des grammaticaux et indexicaux

çà : démonstratif = simple-clic($Objet) | Démonstratif($Ce)
Lien = Déictique($Objet)

les : pronom personnel, pluriel = P-Pers($Le)
Lien = Anaphore($Objet)

les : article-déf., pluriel = Art-déf($Le)
Lien = Déictique($Objet)

ici : adverbe de lieu = Adv-Lieu($ici)
Lien = Déictique($Lieu)

...

Les règles

Les règles permettent d'instancier les schémas d'action et d'objet à partir des actes multimodaux mis en correspondance au cours des étages de traitement précédents. On distingue :

• les règles d'activation des actions par recherche du verbe (mot pris au sens large, soit un mot du langage soit une unité gestuelle caractéristique). Dans le cas où deux verbes sont trouvés, d'autres règles examinent la redondance ou les conflits. S'il y a redondance un seul schéma est poursuivi, s'il y a conflit plusieurs schémas sont instanciés,

• les règles d'activation des objets par spécialisation à partir de la liste d'objets fournie à l'étape précédente,

• les règles d'affinement des attributs des objets et de repérage dans des mondes possibles,

• les règles d'instanciation des attributs de temps et de lieu des actions, par recherche des informations à travers la structure syntaxique et le lexique.

***Solutions***

Les hypothèses (multiples) sont transmises au contrôleur de dialogue qui assure la liaison avec les niveaux supérieurs. Ces hypothèses sont représentées sous forme de schémas instanciés chaînés entre eux.

## 6. UN EXEMPLE D'INTERFACE MULTIMODALE : ICPDRAW

ICPdraw est une application de dessin (type MacDraw™) multimodale (Caelen, 92a). Elle est de type opérative dirigée par les objets. L'utilisateur dispose d'une palette d'outils graphiques et de menus de fonctions. Il peut les activer par la parole, l'écriture ou le geste (via la souris). Le logiciel prévoit l'ensemble des fonctions habituelles nécessaires au dessin : sélection de plusieurs objets par pointage ou entourage ou désignation verbale, déplacement des objets par la voix ou par « draggage » avec la souris, changement des coloris, etc. Les objets géométriques peuvent être groupés, cachés ou saisis par des « poignées ». Le contexte interactionnel est synergique.

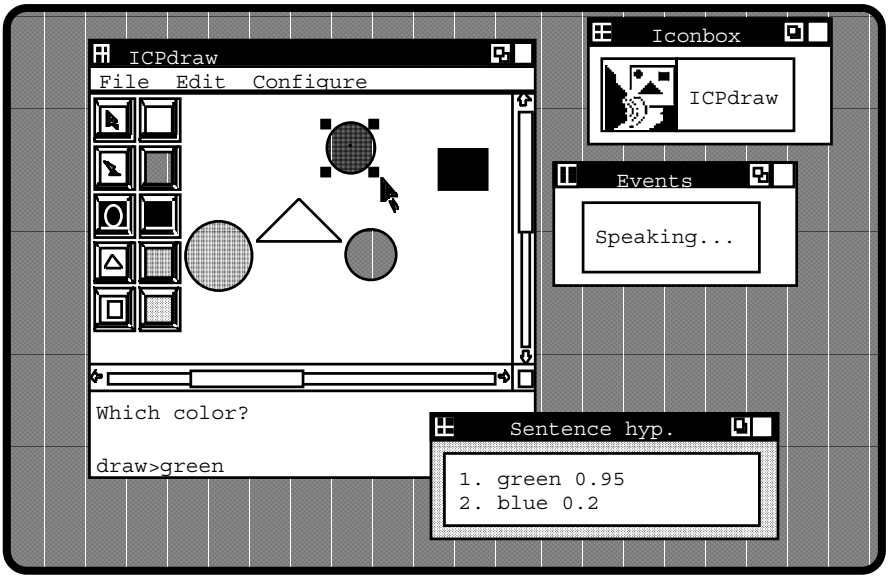
La figure suivante est un exemple d'écran pour l'application ICPdraw. Il est composé de quatre fenêtres, la première (ICPdraw) est découpée en zones et menus définissant l'espace de travail graphique et l'espace de dialogue écriture, la deuxième (Iconbox) contient un logo, la troisième (Events) visualise l'état des canaux de communication et indique, par exemple, à l'utilisateur quand il peut parler, la quatrième (Sentence hyp.) visualise les résultats de la compréhension (les 4 meilleures hypothèses sont retenues et rangées selon un score décroissant).

### 6.1. Les langages de manipulation

Le langage abstrait de manipulation des objets d'ICPdraw est défini par la **forme logique** d'une commande qui est : **Action**(<$\mathbf{arg_1}$><$\mathbf{arg_2}$>...<$\mathbf{arg_n}$>) à laquelle est rattachée une composante parlée ou une composante gestuelle.

**Action** représente une tâche élémentaire. Elle est dénotée par le « verbe » de la phrase ; $\mathbf{arg_i}$ sont des arguments de l'action. Ils sont de type GN (groupe nominal) ou GP (groupe prépositionnel), le nom du GN est en général un objet de l'applica-

tion et les adjectifs du GN des attributs de cet objet (comme couleur, taille, etc.) pour la parole et des gestes de désignation ou des trajectoires pour la souris.




***Exemple d'écran pour l'application ICPdraw***

### *1. Le langage oral*

Dans ce langage tous les éléments peuvent être facultatifs : « dessine cercle vert » ou « non... vert » sont par exemple deux formes admissibles. La grammaire est la suivante :

| | | |
|---|---|---|
| Action | → | V.GN1.Lieu2 |
| Action | → | V.Pr |
| Réit | → | GN1.Lieu2 |
| Rectif | → | non.GN1.Lieu2 |
| Rectif | → | plus.AdjT |
| GN1 | → | Dét2.AdjT.N.AdjC |
| Pr | → | {le, les} |
| V | → | {dessine, déplace, détruis, change, annule, sélectionne, duplique, quitte} |
| Dét | → | {le, les, un, deux, trois, quatre, ce, ces} |
| AdjT | → | {grand, petit} |
| N | → | Obj |
| Obj | → | {carré, cercle, triangle} |
| AdjC | → | {blanc, noir, bleu, jaune, rouge, rouge, vert} |
| Lieu | → | GN2 \| LocP2 \| LocA |
| GN2 | → | LocP1.N.AdjC |
| LocP1 | → | {sous le, sous ce, sur le, sur ce, à côté du, à côté de ce} |
| LocP2 | → | {à droite, à gauche, en haut, à gauche, au centre} |
| LocA | → | {ici, là, là-bas, vers ici, vers là, par ici, par là} |

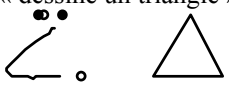
***2. Le langage gestuel***

Un langage gestuel artificiel a été ajouté aux commandes gestuelles habituelles que l'on trouve dans les interfaces à manipulation directe. Il se formalise par :

« dessine un carré »

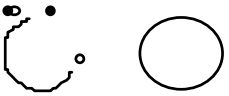
k(Lieu).A(Lieu).Traj(carré).R

« dessine un triangle »



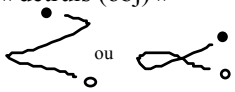
k(Lieu).A(Lieu).Traj(triangle).R

« dessine un cercle »



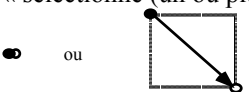
k(Lieu).A(Lieu).Traj(cercle).R

« détruis (obj) »




A(Obj).(Traj(Z) | Traj(å)).R

« sélectionne (un ou plusieurs) objet(s) »




k(Obj) | A(Lieu).Traj.R

« déplace (obj) »

A(Obj). Traj.R

« duplique (obj) »

k(Obj).A(Obj). Traj.R

« désigne (lieu) » k(Lieu)

Les commandes peuvent être tapées au clavier, énoncées dans un microphone ou entrées par la souris. Le module qui traite des entrées en langage naturel (parlées ou écrites) dispose d'analyseurs linguistiques capables de fournir la structure des constituants (c-structure) et la structure fonctionnelle (f-structure) de la commande. Ils fournissent sous forme de chaîne de caractères les quatre meilleures solutions. Ces analyseurs utilisent le formalisme des grammaires lexicales fonctionnelles et sont décrits en détail dans (Reynier, 90) et fonctionnent avec le système de reconnaissance mis au point au laboratoire dans le cadre du projet Multiwoks (projet ESPRIT II n°2105). La reconnaissance proprement dite utilise des modèles de Markov pour des mots enchaînés (algorithme de Viterbi-Ney).

## 6.2. Résultats et discussion

ICPdraw est une application plate-forme qui a été réalisée sur une station de travail (25 MIPS). Le vocabulaire utile est d'une cinquantaine de mots, la syntaxe est limitée. La puissance de la station est suffisante pour obtenir un dialogue en temps réel compte tenu du fait que la reconnaissance tourne aussi sur le même processeur. Cette plate-forme permet de tester et de valider (a) les concepts du dialogue multimodal ainsi que ceux (b) d'une architecture répartie.

Sur ce deuxième point (b) on sait maintenant distribuer le signal et le traiter (notion de client-serveur) sur plusieurs postes de travail qui ne disposent pas du hardware spécialisé pour l'acquisition et l'écoute du signal ou d'une carte de reconnaissance de la parole.

Sur le premier point (a), elle a permis de faire émerger les problèmes de désignation simultanés par la parole et par le geste - surtout dans le cas ou les objets sont animés (problème de la référence perdue lorsque l'on prononce la commande « détruis le cercle du bas » pendant que l'on déplace ce cercle avec la souris) - et d'initialiser les études sur l'usage des modalités. En effet on ne maîtrise pas encore les situations dans lesquelles les usagers utiliseront préférentiellement tel mode ou tel autre et s'ils auront tendance à se contenter d'habitudes appauvrissantes pour le multimodal. Cependant après une étude restreinte (sur une dizaine de chercheurs du laboratoire), il apparaît que cette interface améliore l'efficacité globale de l'usager en lui permettant d'exécuter plusieurs commandes en même temps : c'est le parallélisme des commandes qui lui permet notamment d'anticiper et de préparer les objets du dessin à l'avance. Ainsi on s'aperçoit que les stratégies de planification sont optimisées en fonction de ces nouvelles possibilités : il devient plus courant maintenant d'afficher pêle-mêle une série d'objets par commande vocale (« dessine un cercle noir », « un rouge », « un vert », etc.) puis de les positionner précisément par la souris (éventuellement en continuant à en afficher de nouveaux par la voix « triangle »). L'utilisateur atteint une telle expertise en peu de temps et simplifie ses commandes en les rendant très brèves et souvent très elliptiques. Mais l'usage du parallélisme produit des effets de bord car les commandes peuvent devenir ambiguës en se recouvrant : une certaine désynchronisation du geste et de la parole, des phénomènes de répétition, etc. augmentent avec la charge cognitive ou la rapidité d'exécution imposée. La tendance naturelle à n'utiliser qu'un seul mode n'est pas prouvée : au contraire, l'utilisateur tend certes à spécialiser les modes (la parole pour des commandes répétitives ou qui ne nécessitent pas la mobilisation du regard, ou la précision du geste c'est-à-dire exigeant une planification motrice ou perceptive importantes) et le geste pour des actions immédiates et élémentaires à forte réactivité. À cet égard les deux modalités cohabitent sans se concurrencer.

La voie est donc ouverte pour des validations plus systématiques qui d'ores et déjà semblent encourager l'approfondissement du paradigme « multimodal » (Bisson et al. 92) à travers des interfaces de plus en plus réalistes (Gourdol, 90), (Bourguet, 92), (Condom, 92), (etc.).

## 8. CONCLUSION

Une interface met en relation les niveaux de structuration des connaissances (signes) de mondes référentiels possibles avec les niveaux d'abstraction pour l'architecture de l'interface. Le passage entre ces niveaux (représentations, concepts, symboles) se fait par un double processus : sur l'axe syntagmatique (combinaison des signes, sur l'axe horizontal du temps) par l' « action », sur l'axe paradigmatique (combinaison des signes sur l'axe vertical) par la « fusion ». Notons également qu'une interface met en relation plusieurs milieux, celui de l'homme, celui de la machine et celui dans lequel tous deux sont plongés, leur environnement. De ce fait interface veut tout aussi bien dire capteur, effecteur, transducteur, miroir (multimodal) de la machine que miroir transmodal de l'homme ou relais méta-modal du monde.

Le concepteur d'interfaces doit prendre en compte l'usager dans ses dimensions cognitive mais ici aussi sensorielle et motrice. Dès la gestion des modes et la fusion-fission des informations sont bien des problèmes propres aux interfaces multimodales.

Nous n'avons pas examiné dans ces articles tous les aspects de la multimodalité. Avec la conception, l'évaluation est une étape fondamentale dans l'élaboration d'une application en vraie grandeur (Coutaz, 90), (Scapin, 86). On s'aperçoit à ce niveau de l'importance des erreurs de compréhension et du problème de leur réparation (Siroux et al., 89). Ces erreurs sont non seulement dues aux faiblesses des modules de reconnaissance mais aussi aux phénomènes d'anticipation motrice/ concurrence *vs* retard/hésitation, aux conflits inter-modaux, aux inattendus. Des études doivent aussi être initiées dans cette voie.

## Bibliographie

Barthet Marie-France (1988). *Logiciels interactifs et ergonomie. Modèles et méthodes de conception*, Dunod-Informatique-Bordas, Paris.

Bastide R., Palanque P. (1991). « Modélisation de l'interface d'un logiciel de groupe par Objets Coopératifs », *Document de travail IHM'91*, GDR-PRC CHM éd., p. 1-10.

Bisson P., Nogier J.F. (1992). « Interaction homme-machine multimodale : le système MELODIA », *Actes ERGO.IA'92*, Biarritz, p. 69-90.

Bourguet M.L., Caelen Jean (1992), « Interfaces Homme-Machine Multimodales : Gestion des Événements et Représentation des Informations », *Actes ERGO-IA'92*, Biarritz.

Bourguet M.L. (1992). *Conception et réalisation d'une interface de dialogue personne-machine multimodale*, Thèse INPG, Grenoble.

Buxton B. (1993). « HCI and the inadequacies of direct manipulation systems » *SIGCHI Bulletin*, Vol. 25, n° 1, p. 21-22.

Brandetti M., D'Orta P., Ferretti M., Scarci S. (1988). « Experiments on the usage of a voice activated text editor », *Proc. Speech '88*, p. 1305-1310.

Brooks F.P. (1988). « Grasping reality trough illusion : interactive graphics serving science », *5th Conf. on Comp. and Human Interaction*, CHI'88.

Caelen Jean (1991). « Interaction multimodale dans ICPdraw : expérience et perspectives », *École de printemps PRC « Communication homme-machine »*, École Centrale, Lyon.

Caelen Jean, Garcin P., Wrëto J., Reynier E. (1992a). « Interaction multimodale autour de l'application ICPdraw », *Bulletin de la Communication Parlée* n° 2, p. 141-151.

Caelen Jean, Coutaz Joëlle (1992b). « Interaction homme-machine multimodale : quelques problèmes », *Bulletin de la communication parlée* n° 2, p. 125-140.

Caelen-Haumont G. (1991). *Stratégie des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions entre modèles syntaxiques, sémantiques, pragmatiques et paramètres prosodiques*, Thèse de doctorat d'état, vol. I et II, Aix-en-Provence.

Cadoz Cl. (1992). « Le geste canal de communication homme-machine. La communication instrumentale », *Actes des Entretiens de Lyon*, CNRS.

Condom J.M. (1992). *Un système de dialogue multimodal pour la communication avec un robot manipulateur*, Thèse Université P. Sabatier, Toulouse.

Coutaz Joëlle et Caelen J. (1990). *PRC communication homme-machine : Opération de Recherche Concertée interface homme-machine multimodale*, GDR-PRC CHM, juin 1990.

Coutaz Joëlle (1990). *Interface homme-ordinateur : conception et réalisation*, Dunod, Paris.

Coutaz Joëlle (1992). « Multimedia and Multimodal User Interfaces : A Taxonomy for Software Engineering Research Issues », *St Petersburg HCI Workshop*, August 1992.

Falzon P. (1990). *Ergonomie Cognitive du Dialogue*, PUG, Grenoble.

Faure Claudie (1993). « Communication écrite, concepts et perspectives », *Actes des Journées du GDR-PRC « Communication Homme-Machine*, Montpellier.

Gaiffe B., Pierrel J.M., Romary L. (1991). « Reference in amultimodal dialogue : towards a unified processing », *EUROSPEECH'91*, *2nd European Conference on Speech Communication and Technology*, Genova.

Gourdol A. (1991). *Voice Paint, Rapport de DEA*, Grenoble.

Hécan H., Jeannerod D. M. (1975). « Du contrôle moteur à l'organisation du geste », Masson, Paris, 1975.

Hutchins E.L., Holla J.D., Norman D.A. (1985). « Direct Manipulation Interfaces », *HCI, Lawrence Erlbaum Ass. Publ.*, 1(4), p. 311-339.

*IHM'91* (1991). Groupe de travail interfaces multimodales, Dourdan, déc. 1991.

*IHM'92* (1992). Groupe de travail interfaces multimodales, Paris, déc. 1992

Pankoke-Babatz U. (1989). *Computer based Group Communication, the AMIGO Activity Model*, Ellis Horwood.

Reynier E. (1990). *Analyseurs linguistiques pour la compréhension de la parole*, Thèse INPG, Grenoble.

Rubine D. (1991), *The automatic recognition of gesture*, PhD thesis, School of computer Science, Carnegie Mellon University, CMU-CS-91-202.

Scapin D.L. (1986). « Guide ergonomique de conception des interfaces homme-machine *», Rapport Technique INRIA n° 77*, octobre 1986.

Siroux J., Gilloux M., Guyomard M., Sorin C. (1989). « Le dialogue homme-machine en langue naturelle : un défi ? », *Annales des télécommunications*, 44, n° 1-2.

Stefik M., Bobrow D., Foster S., Tatar D. (1987). « WYSIWIS : Early experiences with multi-user interfaces », *ACM trans. office information system*, Vol.5, n° 2, april 1987, p. 147-167.

Turk M. and Pentland A. (1991), « Eigenfaces for recognition », *Journal of Cognitive Neuroscience*, Vol. 3, n° 1, p. 71-86.

Valot Claude, Amalberti Roger (1991). *Description et analyse de l'activité de l'opérateur*, École IHM-M, GDR-PRC CHM, École Centrale, Lyon, avril 1991