Systèmes interactifs multimodaux

Jean Caelen
Laboratoire CLIPS/IMAG
Domaine universitaire
BP53, 38041 Grenoble Cedex 9
Jean.Caelen@imag.fr

1. Communication homme-machine vs. Interaction homme-machine

La communication humaine est un phénomène *social* dans lequel l'individu *agit* de façon rationnelle : « la communication relève d'une "tentative d'ajustement" où l'on doit ajouter au transport de l'information, le jeu des rôles et des actes par quoi les interlocuteurs se reconnaissent comme tels, agissent comme tels et fondent ainsi des communautés linguistiques dans un monde humain » [Wittgenstein, 59].

Cette définition résulte elle-même de nombreux *ajustements* car beaucoup de disciplines se sont intéressées à la communication humaine. Les cognisciences retiennent de la communication les aspects liés à la perception, à l'action et au raisonnement du point de vue de *l'individu* ; la philosophie s'intéresse à cet *individu* placé en situation de communication, sur un plan *intentionnel*. L'éthnoscience pose par contre la communication dans une perspective *sociale* : les individus agissent dans un cadre normalisé selon des règles et des conventions qui sont socio-culturellement bien définies. Quant aux technosciences, elles visent à intégrer la machine dans un univers de «communication» humaine dans plusieurs directions : soit (a) dans une perspective de machine *médiatrice* — canal entre des interlocuteurs humains — soit (b) dans une perspective *virtualisante* — la machine anime ou simule des mondes virtuels — soit (c) dans une perspective *opérante* — la machine s'insère dans le processus même de la communication pour devenir l'un des partenaires dans la communication et participer à la résolution d'un problème ou d'une tâche. Dans ce dernier cas, elle est assujettie à comprendre pour participer et collaborer au mieux à la tâche de l'utilisateur. Le terme *communication homme-machine* pourtant couramment employé, semble abusif : la machine n'est pas un être *social*, n'a pas d'intention ni de culture. Elle ne peut pas agir sur le

monde réel et on ne peut pas lui dire : « peux-tu fermer la porte s'il te plaît ? ». Elle n'a de prise que sur son propre monde.

N'a-t-on pas plutôt besoin *d'interagir* que de *communiquer* avec elle.

De fait, la machine procure des outils, des moyens d'accès, pour réaliser une tâche ou permet de partager des données, un logiciel, avec d'autres humains pour travailler de manière collaborative dans un même environnement informatique. Elle se présente donc chaque fois comme un *interacteur*. Sa fonction de communication se résume à la manière de présenter des informations ou de comprendre des instructions. Cette fonction se situe dans une relation opérateur-tâche où la machine a un rôle collaboratif [Falzon, 92]. Mais c'est ici que surgit le paradoxe car pour assumer ce rôle, elle doit avoir des capacités qui lui permettent de comprendre les processus actionnels et dialogiques de l'utilisateur qui puissent la rendre *artificiellement* sociale pour être au minimum «conviviale». Pour ce faire, la machine devrait donc posséder :

— la connaissance de l'opérateur,
— la connaissance du domaine de la tâche,
— des représentations d'elle-même (pour s'adapter),
— les règles de l'intervention pédagogique (aides, guides, exemples),
— les règles du dialogue (principes de négociation, de coopération, de réactivité, etc.),
— des règles de comportement social,

et bien sûr tous les processus inférentiels mettant en œuvre ces connaissances, voire même des capacités de compréhension du langage naturel.…

Les différentes étapes par lesquelles la machine, partant des instructions produites par un interlocuteur humain, tente de les comprendre en les replaçant dans un cadre actionnel et dialogique pour générer des réponses sous forme d'actions après avoir planifié ses réponses en fonction des contraintes interactionnelles, sont représentées dans la fig. 1. Ces étapes sont planifiées généralement par un composant logiciel appelé *contrôleur de dialogue*.

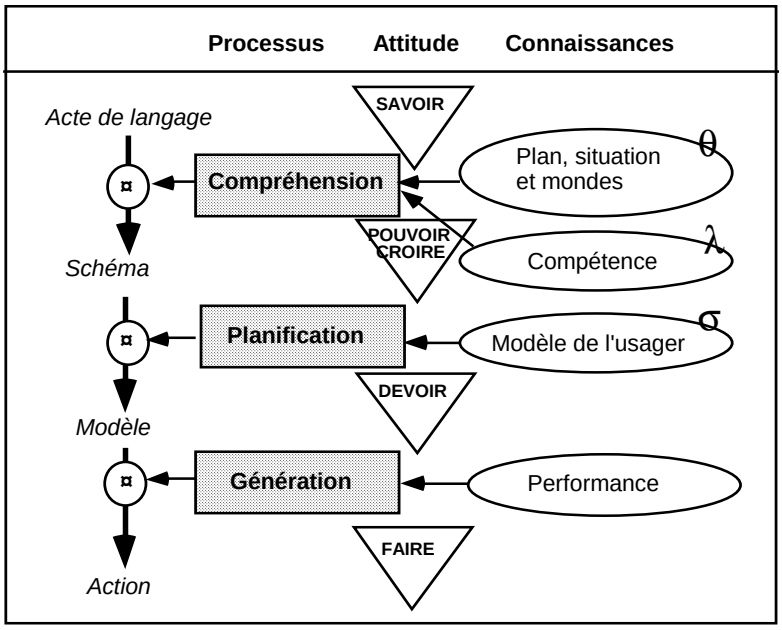

Fig. 1 : les processus inférentiels que doit posséder la machine. Un acte de langage est interprété en regard de la situation, essentiellement le plan d'action, but de la communication avec la machine. Cet acte est ensuite projeté dans un modèle par rapport auquel il est défini en "compréhension" pour finalement provoquer une action selon les performances de la machine manifestées dans une composante de "génération".

Les rôles du contrôleur de dialogue dans une interface homme-machine peuvent être résumés comme suit :

- construction d'un univers commun (mondes)
- réparation des erreurs de communication
- structuration et organisation de la communication
- gestion de la communication
- construction de l'interaction
- mise en contexte
- mise en mémoire, mise à jour, adaptation, apprentissage.

Le contrôle du dialogue

Le contrôle du dialogue par la machine peut s'inspirer des règles observées par les éthnométhodologistes pour la conversation : pour eux, l'interaction fonctionne selon les principes de la *réciprocité des perspectives* et de la *réciprocité des motivations*. Ces principes s'appuient sur la notion d'intercompréhension qui définit le projet d'action de A (ou intention) à travers la réaction qu'il attend de son partenaire B, comme moyen de réaliser son but. Le principe de réciprocité des motivations est l'anticipation par A que son projet, une fois compris, sera accepté par B comme la raison et la motivation "à-cause-de" du projet et de l'action de B [Schütz, 1962]. Si, pour eux, ce principe suffit à régler les niveaux locaux de l'interaction (tours de parole), le deuxième principe, celui de la réciprocité des perspectives est nécessaire pour régler les niveaux supérieurs d'organisation de l'interaction. Ces niveaux sont liés à une conception hiérarchique de l'action dans laquelle ce principe fonde la complémentarité ou la symétrie des rôles des partenaires pour le

guidage des niveaux d'exécution. De lui résultera la *stratégie* utilisée dans l'interaction issue d'un accord entre les partenaires.

En replaçant le dialogue dans cette perspective, on obtient un modèle général qui fait apparaître la nécessité d'un double contrôle : global (similaire à la gestion des perspectives) et local (similaire à la gestion des motivations).

Le contrôle global a pour but de :

- construire les sous-buts communs des interlocuteurs,
- maintenir l'orientation générale du dialogue dans la direction d'une réalisation de ces buts et sous-buts,
- déterminer la stratégie générale,
- évaluer les buts et sous-buts atteints.

Le contrôle local du dialogue consiste à gérer le séquencement des tours de parole c'est-à-dire à comprendre un énoncé vis-à-vis du contexte, puis à générer la réponse et enfin à prédire l'acte suivant. Le travail du contrôleur local peut être schématisé de la manière suivante : il doit,

- tenter d'inférer le plan du locuteur,
- éventuellement réajuster le contexte actionnel précédent,
- émettre des hypothèses sur les niveaux profonds sous-tendus par l'acte en cours, à savoir états mentaux, attitudes cognitives, intentions, etc. de l'usager,
- comprendre le sens de l'acte dans ses trois composantes locutoire, illocutoire et perlocutoire,
- choisir une stratégie de dialogue,
- générer la ou les réponses (actions, aides, questions, réparations, etc.),
- calculer (et prédire) l'ensemble des effets contextuels sur les savoirs, croyances, intentions, actions, etc.

Les stratégies de dialogue

Le dialogue humain est souvent mené avec des stratégies variées qui dépendent du contexte et des interlocuteurs. Ces stratégies peuvent être catégorisées en types de base qui sont :

- *directif*, l'initiative reste toujours du côté de la machine, l'usager doit répondre strictement aux questions qui se réduisent souvent à des choix
- *réactif*, dans ce mode chaque interlocuteur (mais particulièrement la machine) réagit le plus complètement possible au dernier échange. S'il s'agit d'une commande, celle-ci est toujours interprétée et exécutée même au prix d'approximations hasardeuses (prise de décision par défaut)
- *négocié*, ce mode est plutôt celui de deux adversaires : il consiste à minimiser l'espace de concession accordé à son interlocuteur. Chacun défend une position et ne cède que devant les arguments de l'autre. Pour la machine cependant cela revient à adopter une nouvelle position, dans l'espace de concession laissé par l'utilisateur. Cette position reste aussi proche que possible de la position qu'elle vient de céder

• *coopératif*, dans ce mode l'un des partenaire se fait obligation de fournir un maximum d'informations pour aider et orienter son interlocuteur. Mais fournir trop d'informations augmente sa charge cognitive ainsi que celle de son interlocuteur. La règle est donc de fournir l'information la plus pertinente eu égard à la situation et aux interlocuteurs eux-mêmes (principe de la pertinence de Sperber & Wilson, maximes de coopération de Grice)
• *dirigé par les intentions*, est le mode qui consiste à comprendre et interpréter les intentions de son interlocuteur à travers ses actes de langage. Il s'agit pour la machine d'un processus inférentiel pour déduire les objectifs de l'usager à travers la succession de ses actes et de ses dires ainsi que du contexte et de l'évolution de la situation
• *constructif*. est un mode coopératif qui élargit le focus du dialogue au-delà de principes purement rationnels en exploitant les ruptures dialogiques.

2. Le dialogue multimodal

On fera tout d'abord la distinction entre multimédia et multimodalité. Le premier désigne les supports ou les véhicules de l'information le deuxième la substance de l'information :

média : microphone, écran, clavier, souris, caméra, etc.
modalité : parole, vision, écriture, geste, etc.

Le dialogue multimodal est un dialogue qui utilise plusieurs modalités sensorielles : par exemple la parole et le geste. Ces modalités peuvent être utilisées de manière,

- exclusive : seulement une à la fois,
- complémentaire : plusieurs à la fois, chacune complétant sémantiquement l'autre (par exemple le geste de désignation complète l'énoncé « mets-le ici »),
- assignée : exclusive et pour des tâches déterminées
- redondante : plusieurs à la fois, mais de manière redondante (par exemple le geste de désignation et l'énoncé « à droite »),
- équivalente : exclusive mais pour n'importe quelle tâche à priori.

Usage des média

L'interaction homme-machine doit s'appuyer sur une ergonomie d'harmonisation des moyens de communication que sont écran, clavier, souris, voix, image, etc. Par exemple, considérons la commande *"déplacer la fenêtre active vers la gauche"* : deux cas se présentent (a) soit il s'agit de déplacer une fenêtre sur une position précise —un moyen de pointage comme la souris est alors indispensable— (b) soit il s'agit simplement de dégager un espace invisible et le positionnement précis de la fenêtre à déplacer n'est plus nécessaire, auquel cas un ordre oral est plus efficace puisqu'on continue à travailler "mains occupées". Cet exemple montre qu'il n'y a pas équivalence entre une action "souris" et une action "voix" mais qu'elles se complètent en entrant dans des champs d'action et d'utilisation spécifiques. C'est encore plus vrai lorsque l'on dit *"pousse la fenêtre ici"* en désignant la position voulue par la souris.

De manière générale, il vaut mieux entrer des données --nombres, noms (de fichiers par ex.)-- au clavier (pour des raisons de fiabilité et de taille de vocabulaire), les opérations de mouvements fins --réglage de taille de fenêtre, déplacements, pointage, etc.-- à la souris et ne garder pour la communication orale que des commandes de niveau élevé, par exemple "*ouvrir un fichier sur le lecteur interne* " équivalente à une longue séquence de "clics" sur les menus.

Dans le cas de la réponse orale —pour des messages d'aide, de demande de confirmation ou de renseignements complémentaires, etc.— le problème est exactement symétrique : certains messages sont mieux captés par l'oral que par le texte écrit (messages d'alerte notamment, commentaires, aides).

Les niveaux de langages

On peut distinguer deux niveaux de langage liés l'un **(L1)** au système d'exploitation et/ou au gestionnaire graphique (ouverture de fichiers, déplacements de fenêtres, navigation dans les menus, etc.) c'est-à-dire à l'interface homme-machine et l'autre **(L2)** à l'application. Le premier de ces niveaux reste relativement indépendant de l'application puisqu'il concerne surtout l'interface. Par contre pour le second, chaque application ayant son vocabulaire et sa syntaxe propres, une mise en œuvre spécifique devient obligatoire.

On peut dresser une typologie des applications qui donne des cadres d'interaction et de dialogue différents :

A) Jeu : les buts sont connus des deux partenaires mais la machine n'est pas coopérante sauf peut être pour rappeler de temps en temps les règles du jeu au joueur humain. Dans ce cas la machine doit inférer le plan du joueur et ne pas dévoiler le sien. Le dialogue est très réduit tout au plus lorsque le joueur demande à la machine de faire le point sur la situation.

B) Aide à l'apprentissage : l'objectif est connu et la machine est coopérante, elle peut dans le dialogue expliquer et donner des guides pour atteindre le but. Ces guides dépendent du niveau du joueur et de l'opération en cours.

C) Manipulation d'objets d'un univers (logiciel de dessin, CAO, etc.) : ici il n'y a pas de but à long terme mais une série de tâches de détail à exécuter : la machine ne peut percevoir qu'une certaine intention à court terme de l'utilisateur et doit contrôler *a posteriori* les "manœuvres" effectuées. La prédiction est faible dans le dialogue, les guides ne peuvent être que de vagues suggestions.

D) Tâches planifiées (calcul, saisie, visualisation, etc.) : beaucoup de logiciels fonctionnent sur le principe suivant : pour atteindre un résultat les tâches doivent être ordonnancées. Le dialogue est alors **dirigé par la tâche** : faire A1 puis A2, si A2 échoue alors faire A3 puis refaire A2, sinon faire A4, etc. Les intentions de l'utilisateur sont claires, il doit atteindre un résultat à l'aide d'une méthode décomposée en étapes en un minimum de temps. Le dialogue doit viser à clarifier le cheminement de l'utilisateur dans le dédale des possibilités offertes par le logiciel et lui donner les moyens d'y parvenir : saisie des paramètres convenables, choix des méthodes les plus efficaces, planification correcte des étapes, etc.

E) Consultation et renseignement (bases de données, services, etc.) : ici l'utilisateur

ne sait pas trop ce qu'il cherche, ni comment l'obtenir. Il a des difficultés à formaliser sa démarche. La machine doit alors inférer ses buts, le dialogue doit être **dirigé par les intentions**.

Généralement, dans un dialogue de manipulation d'objets, le langage utilisé est opératif : le vocabulaire est limité, la syntaxe peut être négligée (style abrégé) et la compréhension peut être dirigée par des schémas. Le premier mot de la phrase (souvent un verbe) sert de déclencheur à un schéma et les mots suivants permettent d'orienter la particularisation. Certains incidents de communication peuvent être mis en relation avec le caractère inapproprié du premier mot qui peut orienter sur un schéma incorrect. C'est le cas par exemple, quand le mot déclencheur n'est pas en tête du message ou quand ce mot est polysémique. Ces problèmes doivent donc être pris en compte dans le modèle de la tâche.

Dans un dialogue d'interrogation de bases de données les phénomènes linguistiques sont beaucoup plus complexes. A travers la forme de surface de la demande il faut souvent détecter l'intention de l'utilisateur.

Les langages d'interaction en langue naturelle restreinte

La conception d'un dialecte dérivé de la langue naturelle (plutôt qu'un sous-langage ou qu'un langage formel) paraît être une solution acceptable en dialogue homme-machine :

- pour faciliter l'apprentissage des entités et des opérations par l'utilisateur,
- au niveau de la machine car le lexique est bien défini et la syntaxe limitée.

Dans les langages opératifs homme-homme il n'y a pratiquement pas de syntaxe le vocabulaire est limité mais très spécialisé. Ce langage est très lié à la nature de l'application.

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). Pour la prosodie on a pu se rendre compte d'un phénomène analogue [Caelen-Haumont, 79]. La production verbale se dégrade avec la charge de travail ou la concentration sur l'objectif.

Aspects linguistiques

Le langage d'interaction résulte de différents choix opérés sur les facteurs suivants :

Taille du vocabulaire : par exemple, le langage de commande d'un éditeur graphique ne s'élève qu'à 189 mots (Hauptman et Green, 1983). Dans la plupart des applications on peut donc se contenter d'un nombre de mots assez réduit. Cependant il ne faut pas confondre ceci avec le nombre de mots à stocker dans le lexique du modèle de reconnaissance puisqu'il faut ici toutes les formes lexicales utiles (formes conjuguées, formes accordées, expressions, etc.)

Le vocabulaire noyau : la fréquence des mots, leur banalité et leur occurrence dans des expressions différentes. Le vocabulaire comporte toujours des mots rares et spécifiques.

La syntaxe : souvent la forme impérative ou impersonnelle, la syntaxe peut être restreinte (par ex. 14 règles pour Hendler et Michaelis, 1975), en phrases courtes comportant peu de références pronominales ou elliptiques, de métaphores et de métonymies. Par contre les groupes nominaux peuvent être riches (le petit livre rouge sur la table de gauche).

La sémantique : est surtout caractérisée par la monosémie lexicale. La sémantique est orientée par les objectifs.

3. Dialogue et interface : composants

Dans une architecture logicielle, la couche de dialogue multimodal s'intègre dans l'interface utilisateur : le modèle Seeheim définit simplement les différents composants (Seeheim vient du nom de la première réunion de travail de formalisation des interfaces graphiques).

Le modèle Seeheim

Ce modèle est de type série (Fig. 2). Son UIMS (User Interface Management System) se décompose en trois parties : Présentation, Dialogue, Interface avec l'application.

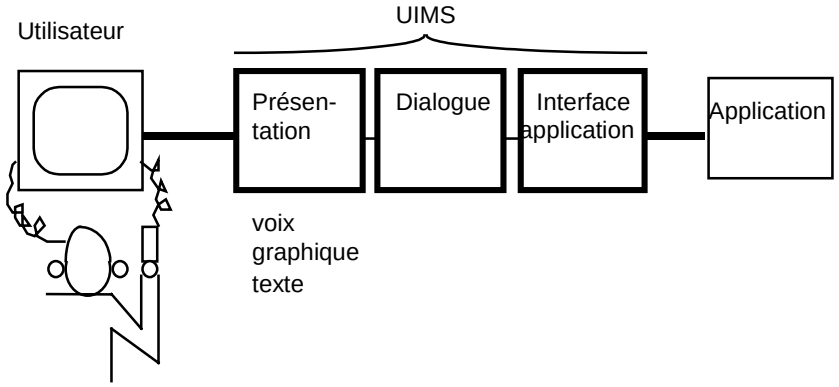

Fig. 2 : Le modèle d'UIMS Seeheim

Le composant Dialogue

Ce composant utilise des connaissances que l'on peut classer en deux groupes :

A. Connaissances statiques

A.1. modèle du langage

- composante reconnaissance: lexique, syntaxe, sémantique,
- composante génération: -idem-

Ces connaissances dépendent de l'application envisagée. Cependant il y a dans le lexique, une partie invariante, ce sont les mots-outils (articles, conjonctions, prépositions, etc.)

A.2. modèle de la tâche

- composante pragmatique: description des objets et de leurs relations relativement à l'application. *On emploie généralement des structures objets.*

- buts et sous-buts: chemins d'accès aux données et aux fonctions et la typologie des tâches. *On emploie ici aussi des structures objets pour définir les tâches et des graphes de dépendance pour décrire l'ordonnancement des tâches de l'application*

A.3. modèle du dialogue

description des diverses situations de dialogue par des scripts ou des scénarios

A.4. connaissances multimodales

dans le cas d'une communication multimodale il faut nécessairement faire le lien entre les événements qui participent au sens du message. Un gestionnaire d'événements spécialisé est alors indispensable.

B. Connaissances dynamiques

B.1. modèle de l'utilisateur

- droits d'accès au système, privilèges, etc.

- connaissances de la machine sur l'utilisateur. *On utilise souvent la logique des croyances dans le cadre de la théorie des intentions.*

B.2. univers de la tâche

- base de faits ou de travail, historique des tâches et des objets de l'univers. *Cette base peut être tenue à jour par l'application elle-même.*

B.3. historique du dialogue

- à court terme
- à long terme.

Le système interactif complet est donc en général un **système basé connaissances** représenté schématiquement sur la (Fig. 3) qui comprend cinq niveaux :
- le niveau des périphériques et des pilotes logiciels,
- le niveau de traitement des entrées-sorties modales : reconnaissance, synthèse, etc.
- le niveau de gestion des événements multimodaux,
- le niveau de dialogue,
- le niveau applicatif (le noyau fonctionnel de l’application).

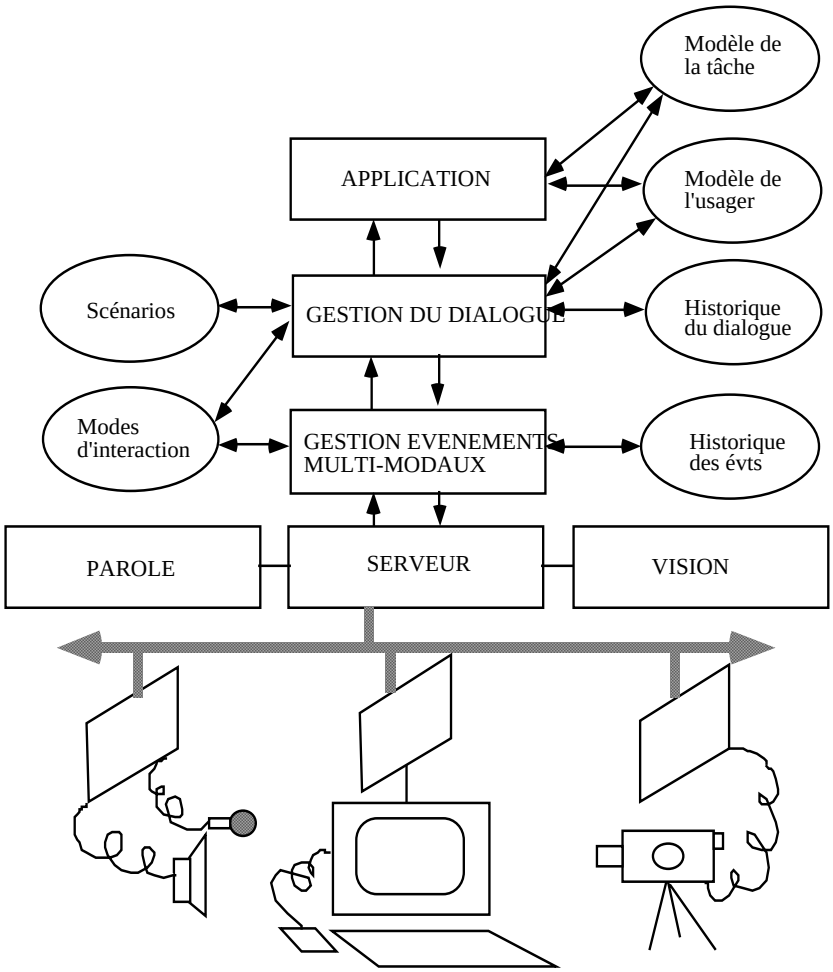

Fig. 3 : Architecture générale d'un système de dialogue multimodal (cas parole+vision).

La composante de dialogue partage certaines connaissances linguistiques avec le module "Parole" et tout particulièrement le lexique. Ce module "Parole" est lui-même un module très complexe (Fig. 4).

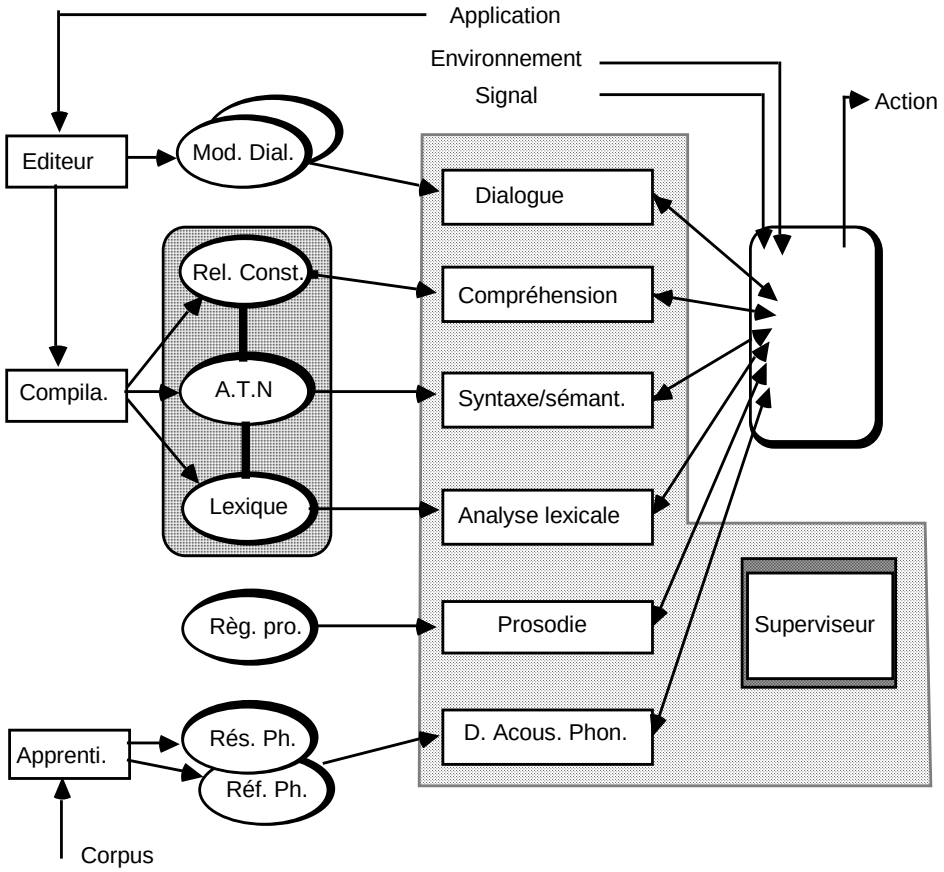

Fig 4 : Architecture du système DIRA [Caelen, 90]. Mod.Dial : modèle de dialogue (et autres bases non explicitées ici), Rel. Cons.: relations entre les constituants, ATN : réseau syntactico-sémantique à transitions augmentées, D.Acous.Phon: décodeur acoustico-phonétique, Rés. Ph.: réseaux acoustico-phonétiques, Réf. Ph : références phonétiques pour la quantification vectorielle.

4. Interaction multimodale

Une interface homme-machine (IHM) multimodale dispose de plusieurs modes d'entrée et de sortie. Ces modes correspondent à certaines des modalités sensorielles et motrices de l'humain. Les problèmes qui distinguent les interfaces multimodales des interfaces classiques naissent de :

- 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.

A 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 bien du son échantillonné. Les fréquences d'échantillonnage de ces données sont différentes d'un média à l'autre. Les problèmes qui se posent dans une interface multimodale sont donc :

(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 «dialogue» [Taylor, 89] qui se trouve de fait posé au niveau le plus haut.

4.1. La gestion des modes

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.).

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).

4.1.1. *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'entrées-

sorties 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.

Il est clair qu'il existe des relations sémantiques entre les unités et des relations temporelles entre les événements.

4.1.2. *Le contexte interactionnel*

Définition du contexte interactionnel : le contexte interactionnel est le triplet {usage des modes, dépendance des informations, dynamicité}. 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 "Concurrent"

Il se définit par :
usage des modes : sans contraintes temporelle (parallélisme possible)
dépendance sémantique : pas de coréférence intermodale entre les unités,

Propriétés : Dans ce contexte, l'anaphore[1] est mal résolue lorsque la référence est portée par un autre mode et la déixis[2] ne peut pas être résolue du tout.

Le contexte "Alterné"

Il se définit par :
usage des modes : deux actes ne peuvent survenir en même temps
dépendance sémantique : pas de contraintes coréférentielles

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

[1] L'anaphore est un phénomène linguistique de renvoi pronominal contextuel. Par exemple dans « Jean aime Marie. Il est jeune », il est une anaphore qui réfère à Jean.
[2] La déixis est un autre phénomène linguistique de renvoi référentiel à un autre mode ou à une situation extérieure. Par exemple dans « ouvre cette fenêtre », cette est un déictique qui nécessite la présence d'un geste de désignation pour être correctement référée.

Le contexte “Synergique”

Il se définit par :
usage des modes : aucune contrainte
dépendance sémantique : pas de contraintes coréférentielles

Propriétés : Dans ce contexte, 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 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 $\mathbf{C_o}$ = {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 sémantique : 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 coréfére 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

d'organisation et du contenu même des actes. Cela fait qu'il ne peut être déterminé que de manière inférentielle.

4.2. Fusion-fission des informations

Le problème central dans une interface homme-machine multimodale se situe dans la fusion (en entrée) et la fission[3] (en sortie) des informations intermodales. Placé au-dessus du module de gestion des événements, le module qui traite de la fusion (resp. fission) fait le lien avec le module qui traite du dialogue.

Cerner les fonctions d'un module de fusion est chose délicate [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 [Wilson et al., 91]. Quelles sont les raisons qui plaident en faveur d'un tel module distinct et spécifique pour les IHM multimodales ?

La discussion générale de cette question est vaste et dépasse le cadre de cet article ; elle devrait porter sur les points suivants :

- Stratégies de fusion
  - Quand ?
    - au plus tôt (précoce)
    - au plus tard (différé)
    - par étapes
  - Comment ?
    - autour d'une structure commune
      - et d'un mode dominant
        - "grammaire" d'unification (langagière bien formée)
      - sans mode dominant
        - "grammaire" multimodale
      - par une théorie de l'action
    - sans structure commune
  - Où ?
    - centralisée dans le contrôleur de dialogue
    - de manière répartie et progressive
  - Avec quelle logique ?

- Critères de fusion
  - de proximité temporelle (règles sensori-motrices)
  - de cohérence structurale et / ou de complétude sémantique
  - d'isotopie sémantique

[3] La fission, non traitée dans cet article, consiste à éclater un message de sortie sur plusieurs modalités, par exemple énoncer « regarde l'icône ici » en faisant clignoter une icône sur l'écran.

fonction du contexte d'interaction
fonction des performances de l'usager [Valot et al., 91]
de logique actionnelle ou intentionnelle [Cohen, 78, 79], [Searle, 83]
etc.

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 au niveau sémantique selon le schéma suivant (fig. 5) :

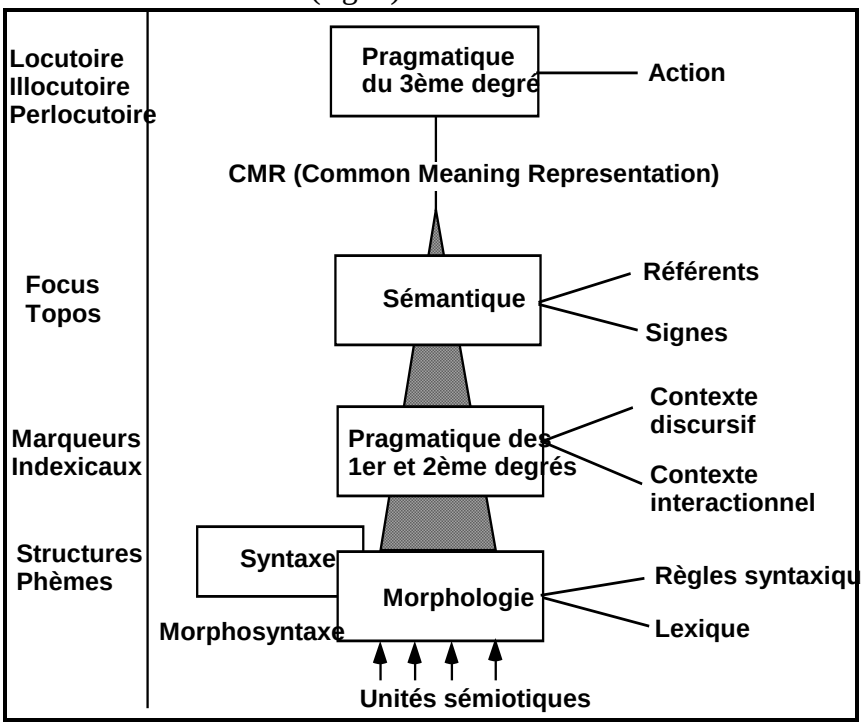

Fig. 5 : Les niveaux de fusion

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

• Analyse morpho-syntaxique modale

Sur une analyse morpho-syntaxique de chaque acte modal, on obtient pour chaque mode une représentation adaptée qui décrit la structure des constituants et la structure fonctionnelle.

• Analyse pragmatique des 1er et 2ème degrés

A 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.

• 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 [Sabah, 88]. 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 fait leur relative indépendance des domaines d'applications.

5. 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 le « dialogue », sur l'axe paradigmatique (combinaison des signes sur l'axe vertical) par le « contrôle ». L'« interaction » se manifeste par une relation plus directe sur le système matériel, c'est-à-dire que la combinatoire syntagmatique est à plus courte portée et la profondeur des « mondes » moins grande que dans le cas du dialogue. 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.

Le concepteur d'interfaces doit prendre en compte l'usager dans ses dimensions cognitive mais ici aussi sensorielle et motrice. Cela donne clairement deux niveaux de traitement : (a) un niveau "bas" pour la gestion des modes et la fusion-fission des informations et (b) un niveau "haut" pour la gestion de l'interaction à travers des couches sophistiquées de dialogue.

Nous n'avons pas examiné dans cet article 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.

Bibliographie

[Austin, 62] AUSTIN J.L., How to do things with words. Oxford U. P., 1962

[Barthet, 88] BARTHET M.F., Logiciels interactifs et ergonomie. Modèles et méthodes de conception. Dunor-Informatique, Bordas, Paris, 1988

[Bastide, 91] BASTIDE R., PALANQUE P., "Modélisation de l'interface d'un logiciel de groupe par Objets Coopératifs", document de travail IHM'91 p 1-10.

[Bisson et al., 92] BISSON P., NOGIER J.F., Interaction homme-machine multimodale : le système MELODIA. Actes ERGO.IA'92, Biarritz, p. 69-90, 1992
[Bourguet et al., 92] BOURGUET M.L. & CAELEN J., "Interfaces Homme-Machine Multimodales: Gestion des Evénements et Représentation des Informations", ERGO-IA'92 proceedings, Biarritz, 1992.

[Bourguet, 92] BOURGUET M.L., Conception et réalisation d'une interface de dialogue personne-machine multimodale. Thèse INPG, Grenoble, 1992

[Buxton, 93] BUXTON B., HCI and the inadequacies of direct manipulation systems. SIGCHI Bulletin, Vol. 25, n°1, p. 21-22, 1993

{Brandetti, 88] BRANDETTI M., D'ORTA P., FERRETTI M., SCARCI S., 1988, "Experiments on the usage of a voice activated text editor", Proc. Speech '88, 1305-1310.

[Brooks, 88] BROOKS F.P., Grasping reality trough illusion : interactive graphics serving science. 5th Conf. on Comp. and Human Interaction, CHI'88, 1988.

{Caelen 91] CAELEN J., Interaction multimodale dans ICPdraw : expérience et perspectives. Ecole de printemps PRC "communication homme-machine", Ecole Centrale de Lyon, 1991.

[Caelen, 92a] CAELEN J., GARCIN P., WRETO J., REYNIER E., Interaction multimodale autour de l'application ICPdraw. Bulletin de la Communication Parlée n°2, p. 141-151.

[Caelen, 92b] CAELEN J., COURAZ J., Interaction homme-machine multimodale : quelques problèmes. Bulletin de la communication parlée n°2, p. 125-140.

[Caelen-Haumont, 91] 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, 1991

[Cadoz, 92] CADOZ Cl., Le geste canal de communication homme-machine. La communication instrumentale. Actes des Entretiens de Lyon, CNRS, 1992.

[Collectif, 91] IHM'91, groupe de travail interfaces multimodales, Dourdan, déc. 1991

[Collectif, 92] IHM'92, groupe de travail interfaces multimodales, Paris, déc. 1992

[Condom, J.M., 92] CONDOM J.M., Un système de dialogue multimodal pour la communication avec un robot manipulateur. Thèse Université P. Sabatier, Toulouse 1992.

[Coutaz et al., 90] COUTAZ J. et CAELEN J., PRC communication homme-machine : Opération de Recherche Concertée interface homme-machine multimodale. Juin 1990.

[Coutaz, 87] COUTAZ J., "PAC: an Implementation Model for Dialog Design", Proceedings of the

Interact'87 conference, Stuttgart, H-J. Bullinger, B. Shackel ed., North Holland, september 1987, pp. 431-436.

[Coutaz, 90] COUTAZ J., Interface homme-ordinateur : conception et réalisation. Dunod éd., Paris, 1990.

[Coutaz, 92] COUTAZ J., “Multimedia and Multimodal User Interfaces: A Taxonomy for Software Engineering Research Issues”, St Petersburg HCI Workshop, August, 1992.

[Cohen, 78] COHEN Ph.R., On knowing what to say : Planning speech acts. Ph.D. Thesis, Technical Report n°118, Department of Computer Science, University of Toronto, January 1978.

[Cohen et al., 79] COHEN Ph.R. et PERRAULT C.R., Elements of a Plan-Based Theory of Speech Acts. Cognitive Science 3, pp. 177-212, 1979.

[Decouchant et al., 88] D. DECOUCHANT, A.DUDA, A.FREYSSINET, M.RIVEILL, X.ROUSSET de PINA, R.SCIOVILLE, G.VANDOME, "GUIDE: an implementation of the Comandos object-oriented architecture on Unix", Proceedings of EUUG Autumn Conference (Lisbon), p 181-193, October 1988.

[Falzon, 90] FALZON P., Ergonomie Cognitive du Dialogue. PUG, Grenoble, 1990

[Faure, 93] FAURE C., Communication écrite, concepts et perspectives. Journée du GDR-PRC “Communication Homme-Machine”, Montpellier, à paraitre, 1993

[Hécan et al. 75] HECAN H., JEANNEROD M., Du contrôle moteur à l’organisation du geste. Masson éd., Paris, 1975.

[Hutchins, 85] HUTCHINS E.L., HOLLA J.D., NORMAN D.A., Direct Manipulation Interfaces. HCI, Lawrence Erlbaum Ass. Publ., 1(4), 1985, p. 311-339.

[Gaiffe et al., 91] GAIFFE B., PIERREL J.M., ROMARY L., Reference in amultimodal dialogue : towards a unified processing. EUROSPEECH’91, 2nd Euopean Conference on Speech Communication and Technology, Genova, Italy, 1991

[Gourdol, 90] GOURDOL A., Voice Paint, rapport de DEA, Grenoble, 1991

[Grice, 75] GRICE H.P., Logic and conversation. in Syntax and Semantic, 3: Speech Acts, P. Cole and J. L. Morgan (Eds), New York Academic Press, pp. 41-58, 1975.

[Fillmore, C.J.] FILLMORE C.J., The Case For Case. Bach E. and Harms R. eds, “Universals in Linguistic Theory”, Holt, Rinehart and Wiston, pp 1-90, New York, 1968.

[Morel, 88] MOREL M.A., Analyse linguistique d’un corpus de dialogues homme-machine. Publications de la Sorbonne Nouvelle, Tomes I et II, Paris , 1988

[Morel, 89] MOREL M.A., Analyse linguistique d'un corpus, Deuxième corpus: Centre d'Information et d'orientation de l'université de Paris V. Paris: Publications de la Sorbonne Nouvelle, 331 p., 1989.

[Pankoke, 89] PANKOKEe-BABATZ U., "Computer based Group Communication, the AMIGO Activity Model", Ellis Horwood, 1989.

[Reynier, 90] REYNIER E., Analyseurs linguistiques pour la compréhension de la parole. Thèse INPG, Grenoble, 1990

[Rubine, 91] RUBINE D., “The automatic recognition of gesture”, PhD thesis, School of computer Science,

Carnegie Mellon University, CMU-CS-91-202, 1991.

[Scapin, 86] SCAPIN D.L., “Guide ergonomique de conception des interfaces homme-machine”, Rapport Technique INRIA no 77, Octobre 1986

[Taylor et al., 89] TAYLOR M.M., NEEL F., BOUHUIS D.G., The Structure of Multimodal Dialogue. Elsevier Science Publishers B.V., North-Holland, 1989

[Sabah, 88] SABAH G., L'intelligence artificielle et le langage. 2 tomes. Hermès ed., 1988 et 1989.

[Searle, 69] SEARLE J.R., Speech Acts. Cambridge U. P., 1969

[Searle, 83] SEARLE J.R., Intentionality. Cambridge U. P., 1983.

[Siroux et al., 89] SIROUX J., GILLOUX M., GUYOMARD M., SORIN C., Le dialogue homme-machine en langue naturelle : un défi ? Annales des télécommunications, 44, n°1-2, 1989.

[Stefik et al.,87] STEFIK M., BOBROW D., FOSTER S., TATAR D., "WYSIWIS: Early experiences with multi-user interfaces" ACM trans. office information system, Vol.5, n°2, April 1987, p 147-167.

[Turk, 91] TURK M. and PENTLAND A., “Eigenfaces for recognition”, Journal of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.

[Valot et al., 91] VALOT C., AMALBERTI R., Description et analyse de l’activité de l’opérateur. Ecole IHM-M, Ecole Centrale, Lyon avril 1991

[Vernant, 92] VERNANT D., Modèles projectifs et structure actionnelle du dialogue. in Recherches sur la philosophie et le langage, Du Dialogue, Vrin éd., 1992.

[Wilson, 91] WILSON M.D., An architecture for multimodal dialogue, Workshop ESCA, Venaco, 1991