HAL
open science

Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales

Meriam Horchani

To cite this version:

Meriam Horchani. Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales. domain_stic.inge. Université Joseph-Fourier - Grenoble I, 2007. Français. NNT : . tel-00258072

HAL Id: tel-00258072

https://theses.hal.science/tel-00258072v1

Submitted on 21 Feb 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse

présentée par

Meriam Horchani

pour obtenir le grade de
Docteur de l'Université Joseph Fourier - Grenoble 1
(arrêtés ministériels du 5 juillet 1984 et du 30 mars 1992)
Spécialité : Informatique

Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales

Soutenance le 17 décembre 2007 devant le jury suivant :

Présidente	Catherine Garbay	Directeur de Recherche CNRS, LIG
Rapporteurs	Philip Gray	Senior Lecturer, University of Glasgow
	Jean-Claude Martin	Maître de Conférence habilité, Paris 8 / LIMSI
Examinateurs	Noëlle Carbonell	Professeur des Universités, UHP / LORIA
	Frédéric Landragin	Chargé de Recherche CNRS, LATTICE
Directeurs	Laurence Nigay	Professeur des Universités, UJF / LIG
	Franck Panaget	France Télécom R&D

Thèse préparée au sein

de l'équipe Ingénierie de l'Interaction Homme-Machine (IIHM)
du Laboratoire d'Informatique de Grenoble (LIG)

et du laboratoire Technologie (TECH) de France Télécom R&D

Remerciements

Merci aux membres du jury d'avoir ... accepté d'être membres du jury ! En particulier, merci à Philip Gray et à Jean-Claude Martin pour leurs lectures attentives, leurs retours, leurs conseils et leur soutien.

Merci à Laurence d'avoir accepté d'être ma directrice, et d'avoir assumé cette charge malgré la distance. Merci pour les démarches administratives, mais surtout, pour le temps (les nuits ...) consacré, les avis éclairés, les longues explications, l'initiation au monde de la recherche.

Merci à Franck de m'avoir proposé un sujet de thèse sur mesure. Merci pour l'encadrement au quotidien, la connaissance partagée, le confort de travail au sein de France Télécom et la confiance accordée.

Merci à ceux qui, autres que Franck et Laurence, ont contribué au contenu de ce mémoire, par leur soutien technique ou scientifique. En particulier, merci à Benjamin, à Dominique, à Thierry et à Vincent.

Merci à ceux qui ont contribué à me donner l'impression d'appartenir à une famille scientifique. En particulier, merci à Frédéric et à Cyril, ainsi qu'aux participants aux rencontres des jeunes chercheurs en IHM, version 2006.

Merci à ceux qui ont partagé mon bureau, et m'ont sans doute inspirée sans que je le sache grâce aux discussions à bâtons rompus : merci à Olivier, pour ses conseils qui m'ont servi trois ans plus tard, pour son enthousiasme et pour son amitié ; merci à Magalie, pour son regard sur la vie et sur la recherche, pour sa détermination communicative et pour sa présence la dernière année ; merci à Carole, pour les deux ans et demi dans le même bureau (dont un à s'apprivoiser ;-)) et merci à Matthieu pour son regard « neuf » et pour son humour.

Merci aux membres du laboratoire TECH/EASY que je n'ai pas encore remerciés et qui ont contribué à rendre agréable le quotidien à France Télécom. En particulier, merci à Florence pour ses conseils avisés en relations humaines, à Kris et à Isabelle pour les relectures, à Farid pour nos grandes discussions sur la science, les chercheurs, les ingénieurs, les ergonomes, etc., à Julien M. pour la pause du soir durant les quatre derniers mois. Merci pour tous les moments passés ensemble autour d'un café, d'un verre ou d'une table, à Pierre, à Étienne, à Karl, à Jean-François, à Estelle, à Marie (et aux autres inspirées, d'ailleurs !), à Morgane, à Laetitia, à Joseph M., à Maryline, à Arnaud, à Aurélie, à Sylvain (et j'en oublie, je suis sûre !). Plus spécialement, merci à Sophie et à Julien, témoin de mes peines, de mes colères, de mes agacements plus que de mes joies (...), et surtout pour les centres d'intérêts partagés, les papotages et les

rigolades.

Merci aux membres de l'équipe IIHM qui m'ont accueillie lors de mes passages à Grenoble ou qui ont répondu présent « au bout du mail ». En particulier, mille mercis à David pour le soutien administratif à distance lors du dépôt du dossier de soutenance. Merci particulier à Benoît pour les dossiers d'inscription et les courriers déposés au secrétariat ou envoyés à Lannion, pour la découverte des petits restos grenoblois, pour les discussions sur le passé, ainsi que pour les conseils et le soutien durant les derniers mois.

Merci à l'Ecole Navale et aux membres de l'IRENAV qui m'ont accueillie avant la fin de ma thèse et m'ont donné tout le temps nécessaire pour mener à bien la rédaction, la dernière conférence et la soutenance.

Un merci tout particulier à Anne-Claire, pour avoir répondu présent à chaque fois que j'ai eu besoin d'une relecture, en anglais ou en français. Pour les fiches et les films qui auraient dû contribuer à améliorer mon anglais ! Pour son accompagnement durant ces trois années, et pour son intérêt constant pour mes travaux depuis cinq ans.

Ce mémoire et sa soutenance constituent pour moi l'aboutissement d'un cheminement dont les prémisses remontent bien au-delà du début de ma thèse. Il est donc impossible de remercier toutes les personnes qui y ont contribué, directement ou indirectement, généralement de façon complètement inconsciente. Il n'empêche que ma pensée va régulièrement à ces personnes, témoignage de ma reconnaissance bien plus important que ne le serait une liste à l'ordre nécessairement trompeur.

Et puis, comme dirait Brel, et puis il y a ... le compagnon du quotidien, celui qui a été là, jour après jour, soir après soir. Qui a accepté la mono-maniaquerie typique des thésards et à laquelle je n'ai pas échappé, ainsi que la concurrence de ma thèse. Qui a contribué, tout autant qu'elle, à celle que je suis aujourd'hui. Qui m'a permis de garder les pieds sur terre. Qui m'a accompagnée dans mes choix de vie et dans mon évolution. Qui a été un soutien indispensable durant les derniers mois, émotionellement et logistiquement. Qui sait plus que quiconque ce que ces trois ans m'ont appris et coûté. Merci à lui d'avoir été témoin (pas que auriculaire ...) de tout ça et de m'avoir accompagnée dans cette expérience qui n'engageait que moi.

Sommaire

Remerciements **i**

Sommaire **iii**

Introduction **1**

I Espace-problème : modalités, combinaison de modalités et communication naturelle **7**

1 Espace terminologique : modalité **9**
1.1 Théorie de la communication de Shannon : un canevas intégrateur . . . 10
1.2 Notion de "modalité" en informatique 12
1.3 Notion de "modalité" dans d'autres domaines 30
1.4 Conclusion : notre terminologie 51

2 Combinaison de modalités **55**
2.1 Combinaison des modalités en informatique 56
2.2 Combinaison des modalités dans d'autres domaines 71
2.3 Conclusion : notre approche de la combinaison de modalités 78

3 Notre approche : vers une communication multimodale naturelle **81**
3.1 L'utilisateur et le contexte au centre de la communication naturelle . . . 81
3.2 Communication humain-machine naturelle 83
3.3 Communication naturelle : les approches existantes 87
3.4 Approche choisie et ses hypothèses 89
3.5 Limitations du cadre d'étude 93

II Espace-solution : vers le choix conjoint des stratégies de dialogue et de présentation **97**

4 Dialogue et interaction multimodale : vers une approche intégrée **99**

4.1 Dialogue humain-machine 100
4.2 Interaction humain-machine 113
4.3 Vers une approche intégrée de la communication humain-machine 127
4.4 Conclusion 145

5 Stratégies de dialogue et de présentation : un choix conjoint 147
5.1 Motivations et existant 148
5.2 Stratégie de dialogue et stratégie de présentation 153
5.3 Composant de choix de stratégies de dialogue et de présentation 160
5.4 Discussion et perspectives 175

6 Spécifier les choix de stratégies de dialogue et de présentation 181
6.1 Motivations et existant 182
6.2 Expérimentation de référence sur le service Santiago 185
6.3 Éditeur de spécification du composant de choix 189
6.4 Discussion et perspectives 201

7 Réalisations logicielles 205
7.1 Composant de choix de stratégies de dialogue et de présentation 205
7.2 Éditeur graphique de spécification du composant de choix 211
7.3 Plate-forme de simulation du composant de choix 222
7.4 Exemples implémentés 225
7.5 Conclusion 243

Conclusion 245

Bibliographie 249

Annexes 263

Annexe 1 - Caractérisation de la convivialité et de la coopération dans le dialogue 263

Annexe 2 - Critères d'utilisabilité 265

Annexe 3 - Extraits du fichier XML pour la simulation d'entrée du composant de choix 271

Annexe 4 - Extraits du composant de choix généré avec l'éditeur graphique dans le cas du système-exemple @mie 275

Table des matières **283**

Table des figures **291**

Table des tables **295**

Glossaire **297**

Sigles et acronymes **301**

Introduction

Sujet

Nos travaux de recherche ont trait à la conception et à la modélisation logicielles des systèmes interactifs. Parmi ces systèmes, nous étudions les systèmes de dialogue multimodal naturel grand-public.

Si les systèmes informatiques sont nés de besoins scientifiques, leur succès depuis vingt-cinq ans revient à leur utilisation grand-public. De plus en plus répandus, les utilisateurs sont de plus en plus familiarisés et exigeants. Nous constatons néanmoins des niveaux de familiarisation variables : les jeunes qui grandissent avec Internet n'ont pas les mêmes appréhensions ni les mêmes utilisations que les "seniors", marché auquel s'intéressent de plus en plus de fournisseurs de services. Cet usage est encouragé par la multiplication et la miniaturisation des terminaux d'une part, et par la convergence tant des services que des supports d'autre part. La diversité des terminaux et leur taille font des systèmes informatiques des outils indispensables omniprésents et la convergence des services et des supports assure un accès quasi-permanent aux systèmes d'information en ligne.

Ces trois constats - l'utilisation grand-public, la multiplication et la miniaturisation des terminaux et la convergence des services et des supports - modifient l'utilisation des systèmes informatiques, en particulier des systèmes d'information. Les machines sont démystifiées, l'utilisation se fait ubiquitaire et l'utilisateur ne veut plus être tributaire des choix des concepteurs. Alors que les systèmes informatiques ont longtemps été conçus de façon à être adapté à un utilisateur et une utilisation types, et le sont encore, la recherche doit œuvrer pour dépasser cette approche de la conception afin que les systèmes ne semblent pas figés et donnent l'impression de communiquer naturellement avec les utilisateurs. Cela passe par la prise en compte des particularités de l'utilisateur et de l'utilisation. Nos objectifs se situent dans ce contexte.

Objectifs et démarche

La communication naturelle intègre plusieurs principes qui ont fait l'objet d'études depuis le début de l'informatique, que ce soit en Intelligence Artificielle (IA), ou, par la suite, en Interaction Homme-Machine (IHM). Il s'agit notamment d'assurer l'utilisabilité, i.e. la souplesse et la robustesse, non seulement des interfaces mais plus lar-

gement des systèmes, le choix du paradigme de communication adéquat en fonction de la situation d'utilisation (système-outil, système-partenaire ou système-médiateur [Beaudoin-Lafon, 2004], le comportement convivial, coopératif du système [Siroux *et al.*, 1989, Sadek, 1999]. Si Beaudoin-Lafon [Beaudoin-Lafon, 2004] prône la conception de l'interaction et non des interfaces, nous pensons qu'il faut aller plus loin et chercher à concevoir la communication. La principale différence réside dans le fait que l'interaction se concentre sur l'action de l'utilisateur sur la machine alors que la communication comprend aussi l'action de la machine sur l'utilisateur - et l'utilisation à venir - par un travail sur le comportement du système, que ce soit sur sa réaction ou sur la présentation de cette réaction.

Cette appréhension de la communication humain-machine nous conduit à travailler sur la détermination de la sortie des systèmes. Nous nous focalisons sur les systèmes d'information grand-public. Considérant que, au même titre que la multiplication des capacités d'action de l'utilisateur sur la machine contribuent à une spontanéité et à un naturel de la communication, la multiplication des moyens d'expression de la machine a un rôle à jouer sur l'appréhension des systèmes par l'utilisateur. Nous nous intéressons donc aux systèmes d'information multimodaux en sortie grand-public. En sortie, de tels systèmes sont en mesure de répondre à l'utilisateur en combinant plusieurs modalités informatiques de sortie et/ou plusieurs modalités sensorielles, mais aussi de n'utiliser qu'une seule modalité informatique ou sensorielle si l'information présentée ou le contexte d'utilisation s'y prête, ou, plus simplement, si l'utilisateur le demande. Autrement dit, le comportement des systèmes multimodaux que nous considérons tient compte des contraintes de présentation existantes.

De façon à garantir une communication naturelle et les principes sous-tendus, les contraintes de présentation doivent avoir un impact non seulement sur la présentation de la réponse du système, mais aussi sur sa réaction, de façon à garantir non seulement l'accessibilité sensorielle et actionnelle de l'utilisateur aux informations et aux capacités d'action, mais aussi l'accessibilité cognitive qui limite sa charge mentale et l'accessibilité rhétorique qui assure l'accès aux informations et aux capacités d'action les plus pertinentes. Les contraintes de présentation doivent donc influencer à la fois la stratégie de présentation (i.e., la forme) et la stratégie de dialogue (i.e., le fond) adoptée par le système. La prise en compte de ces contraintes à ces deux niveaux et la garantie des trois types d'accessibilité identifiées nécessite de s'appuyer sur les connaissances actuelles sur l'humain en tant qu'utilisateur et en tant que sujet.

Si l'impact de la communication dans le sens utilisateur vers système est depuis longtemps étudié, la communication dans le sens système vers utilisateur reste encore le parent pauvre de la communication humain-machine. Aussi, la prise en compte des contraintes de présentation sur le comportement du système nécessite, d'une part, de travailler sur la conception de tels systèmes et, d'autre part, de proposer des outils pour pouvoir étudier l'impact des choix de stratégies et de dialogue sur l'utilisateur et sur l'utilisation. Ce constat est à l'origine de nos deux principales contributions.

Nous avons étendu un système d'information au cas de la multimodalité en sortie qui nous sert d'exemple tout au long de l'exposé de nos travaux. On trouvera ci-dessous une description rapide du système, qui nous permet aussi d'immédiatement illustrer le

type de systèmes visés dans nos travaux.

@mie, un exemple de référence

@mie (Annuaire Multimodal Intelligent d'Entreprise) est un système qui permet aux salariés d'une entreprise d'avoir accès, entre autres, à des informations sur leurs collègues (en particulier, leurs prénoms, noms, photos, adresses courriels, numéros de téléphones fixes et portables, numéros de bureau, etc.). L'utilisateur peut saisir une requête dans un formulaire ou énoncer oralement sa requête en langue naturelle. Par exemple : " Je cherche le numéro de Carole " ou " Dites-moi le numéro de Carole ". Nous considérons deux modalités de sortie et leurs combinaisons : les informations sont présentées en combinant - ou non - la modalité auditive <haut-parleurs, langage naturel oral> et la modalité visuelle <écran, hypertexte (incluant des photos)> d'un terminal mobile. La figure 1 présente des présentation proposées par le système @mie.

FIG. 1 – Exemples de sorties multimodales du système @mie

Organisation du mémoire

Le mémoire comprend deux parties, schématisées à la figure 2. Une première partie nous permet de définir l'espace-problème. Pour cela, nous étudions les notions de "modalité" (chapitre 1) et de "combinaison de modalités" (chapitre 2) de façon à expliciter notre appréhension de la communication humain-machine en général, multimodale en particulier. Parce que les systèmes d'information impliquent une machine, mais aussi un humain (sic !), nous extrayons ces deux notions de travaux en informatique ainsi que de travaux en sciences humaines. Nous terminons notre positionnement en soulignant la nécessité de travailler sur une communication humain-machine naturelle et la présentation de notre démarche pour y contribuer (chapitre 3).

Une deuxième partie nous permet de décrire notre espace-solution. Nous commençons par une étude des travaux qui visent à une approche intégrée de la communication humain-machine en rapprochant les paradigmes dialogique et actionnel (chapitre

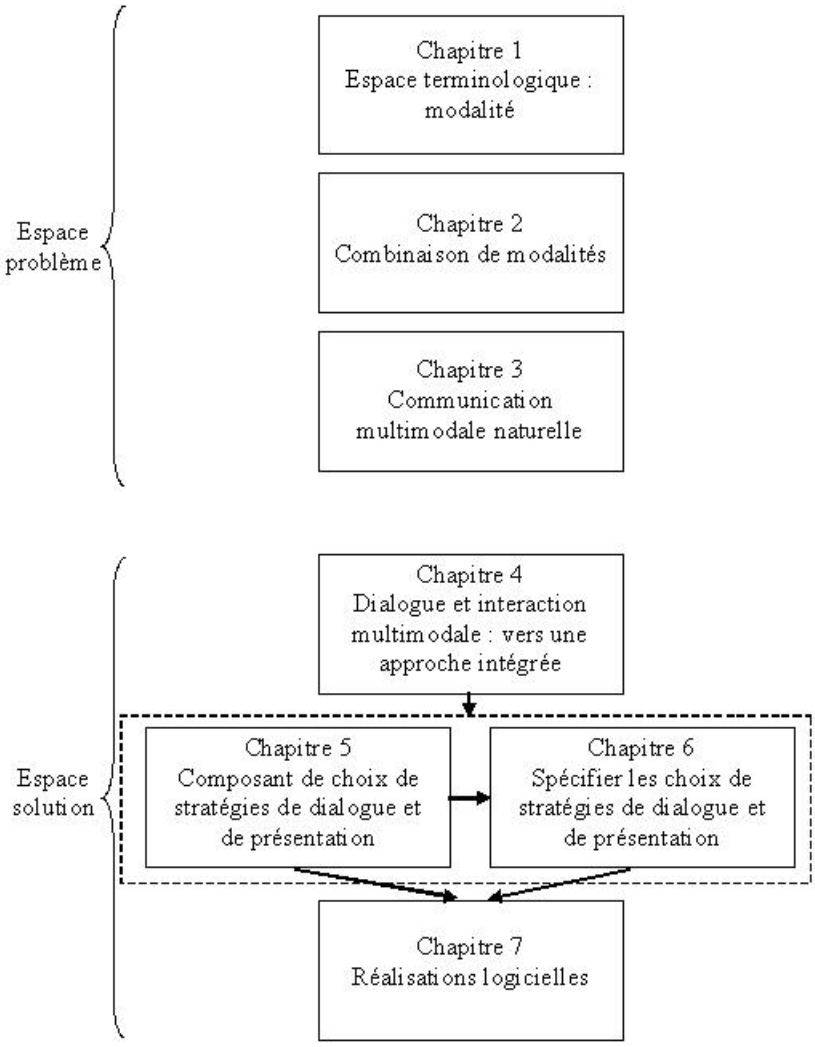

FIG. 2 – Schématisation de l'organisation du mémoire

4). Puis, nous décrivons la solution conceptuelle que nous proposons pour la prise en compte des contraintes de présentation dans le choix des stratégies de dialogue et de présentation des systèmes d'information multimodaux (chapitre 5). Ensuite, nous présentons un outil dédié à l'étude du choix de stratégies de dialogue et de présentation sur l'utilisateur et l'utilisation, exploitable dans la conception de tels systèmes (chapitre 6). Nous terminons par les réalisations logicielles de la solution conceptuelle et de l'outil d'étude proposés. En conclusion, nous soulignons les points contributifs de nos travaux et nous développons les perspectives de travaux futurs.

Première partie

Espace-problème : modalités, combinaison de modalités et communication naturelle

Chapitre 1

Espace terminologique : modalité

Dans le cadre de la communication humain-machine et humain-humain, le terme "modalité" a diverses acceptions qui ne permettent pas d'isoler aisément une définition de la notion de "modalité". De façon plus large, les termes "modalité", "mode" ou encore "média" sont couramment utilisés les uns pour les autres et parfois définis de façon identique par des auteurs différents. Les raisons invoquées pour expliquer cette imprécision tiennent essentiellement :

- aux nuances dues à la langue d'écriture des auteurs [Martin, 1995]. Par exemple, la notion de "médium" en français s'est abstraite du latin alors que celle de *medium* en anglais en est restée assez proche ;
- les spécificités dues aux champs d'étude des auteurs, aux problèmes et aux applications considérés [Martin, 1995, Nigay et Coutaz, 1996]. Des variations peuvent d'ailleurs être faites dans une même référence en fonction du point de vue adopté par l'auteur [Martin, 1995].

Nous considérons que les spécificités en question influencent non seulement la terminologie adoptée mais encore, et surtout, l'appréhension de la communication humain-machine. C'est cette appréhension qui détermine la façon dont est conçue la notion de "modalité" et, par extension, la façon dont plusieurs modalités peuvent être combinées. Par conséquent, il est impossible de cerner la notion de "modalité" - et par extension celle de "combinaison de modalités" - en listant les auteurs qui emploient le terme "modalité" et en répertoriant les définitions qu'ils utilisent. Afin d'avoir une vue aussi juste que possible de la notion de "modalité", il convient de dégager cette notion des études de la communication humain-machine et de la communication humain-humain, même si le terme "modalité" n'est pas employé. Objet de ce chapitre, cette étude terminologique est importante pour nos travaux et nous permet d'expliciter notre appréhension de la communication humain-machine multimodale.

Parce que nous pensons que la conception de systèmes de communication humain-machine doit être centrée sur l'humain, il nous a paru primordial de tenir compte des connaissances et interprétations actuelles de la perception humaine au sens large - incluant l'assimilation des informations perçues. C'est pourquoi nous avons également étudié les appréhensions de la communication humain-humain dans des sciences plus

"humaines", en essayant d'y cerner les notions de "modalité" (et de "combinaison des modalités" dans le chapitre 2) même si le terme "modalité" n'y apparaît pas toujours. Nous avons exclu les travaux sur la communication humain-machine-humain, dite "médiatisée", car le pas nous semble trop important pour rapprocher (1) une communication entre deux agents communiquants dont l'un au moins est humain et (2) une communication incluant un relais artificiel susceptible d'enrichir de façon variable le message transmis. De plus, notre revue se concentre sur les situations de communication qui n'impliquent que deux agents : ne sont donc pas étudiés les travaux en sociologie de la communication et ceux sur les systèmes informatiques collaboratifs. Ce choix est motivé par le fait que les systèmes informatiques sur lesquels nous travaillons sont, originellement, des systèmes de dialogue, n'impliquant que deux intervenants. Enfin, pour ne pas ajouter au flou terminologique existant, nous préciserons le terme utilisé par les auteurs anglophones cités en cas de traduction controversable et nous aurons recours au terme anglais pour les cas les plus litigieux.

Dans ce chapitre, nous étudions la notion de "modalité" dans le domaine de l'informatique puis dans d'autres domaines avant de conclure par la terminologie que nous adoptons. Nous débutons par un rappel de la théorie de la communication de Shannon.

1.1 Théorie de la communication de Shannon : un canevas intégrateur

La théorie de la communication - dite aussi "théorie de l'information" - de Shannon [Shannon, 1948] sert souvent de référence tant dans les travaux sur la communication humain-machine que ceux sur la communication humain-humain [Corraze, 1980, Escarpit, 1991, Battail, 1997].

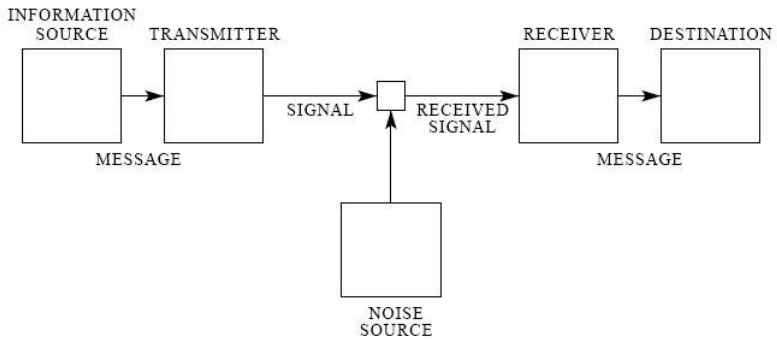

FIG. 1.1 – Diagramme schématique d'un système de communication selon la théorie de la communication de Shannon (extrait de [Shannon, 1948])

La théorie de la communication de Shannon est essentiellement mathématique. Elle porte sur la communication à son niveau le plus élémentaire, sans tenir compte des supports physiques qui la permettent. Comme le rappelle la figure 1.1, Shannon consi-

dère que cinq éléments principaux interviennent dans un système de communication, auxquels s'ajoute un élément perturbateur :

- une source (*an information source*) qui est à l'origine du ou des messages à transmettre. Les messages possibles cités par Shannon sont tous formalisables mathématiquement ;
- un émetteur (*a transmitter*) qui traduit le message en un signal transmissible par le canal considéré ;
- un canal (*the channel*) qui permet de transmettre le message de l'émetteur au récepteur. Shannon parle aussi de *medium*[1] de transmission ;
- un récepteur (*the receiver*) qui traduit le signal en un message compréhensible par le destinataire ;
- un destinataire (*the destination*) auquel est destiné le message ;
- du bruit (*noise source*), nécessairement présent, et qui peut perturber la transmission du message de l'émetteur au récepteur.

La source et le destinataire sont deux entités séparées. Par conséquent, le message doit être transmis de l'une à l'autre. Même si Shannon ne parle pas explicitement de code, l'émetteur et le récepteur sont en quelque sorte un codeur et un décodeur qui permettent de transmettre le message de la source vers le destinataire via un canal donné : Shannon met en évidence que le message est transformé pour pouvoir être transmis et perçu. ***Nous considérons donc qu'émetteur et récepteur renvoient respectivement aux capacités d'expression et aux sens de perception de l'humain, ainsi qu'aux dispositifs des terminaux informatiques dédiés à la production et à la récupération de messages.*** De plus, la notion de "modalité" peut se rapprocher indifféremment :

1. du format du message tel qu'il est produit par la source ;
2. du format du message tel qu'il est perçu par le destinataire si ce format n'est pas identique au précédent ;
3. du format du message tel qu'il est transmis par le canal si on considère comme un tout la source et l'émetteur d'une part et la destination et le récepteur d'autre part ;
4. du code utilisé par l'émetteur ;
5. du code utilisé par le récepteur si ce code n'est pas le même que le précédent.

Battail [Battail, 1997] souligne que, si l'on considère la source, le canal et le destinataire quelconques, rien n'assure qu'il y ait compatibilité entre la source et le canal d'une part, et entre le canal et le destinataire d'autre part. Pour assurer une certaine compatibilité, il peut être nécessaire de normaliser source, canal et destinataire. Ce constat nous conduit à deux remarques. Premièrement, ***cerner le phénomène de communication nécessite de caractériser les éléments impliqués dans l'appréhension choisie*** : c'est l'objectif du chapitre en cours, en considérant la théorie de la communication comme cadre unificateur général de cette étude. Deuxièmement, ***il serait fastidieux de vouloir faire une liste exhaustive et définitive de toutes les sources,***

[1] Terme utilisé par l'auteur.

tous les canaux et tous les destinataires possibles. En effet, ils sont dépendants des émetteurs et des récepteurs existants qui sont en évolution constante : c'est pourquoi, à une exception près, nous ne nous attardons pas dans les paragraphes qui suivent sur les taxonomies existantes ou possibles des "modalités"/"modes"/"médias"/"canaux", etc.

1.2 Notion de "modalité" en informatique

Pour étudier la notion de "modalité" en informatique, nous décrivons les principaux espaces, taxonomies et modèles qui lui sont dédiés par ordre chronologique. Nous concluons par une synthèse sous la forme d'un tableau comparatif. Pour des raisons de lisibilité, les principales conclusions que nous tirons dans chacune des sections sont écrites en gras et en italique.

1.2.1 Espace des interfaces selon Frohlich

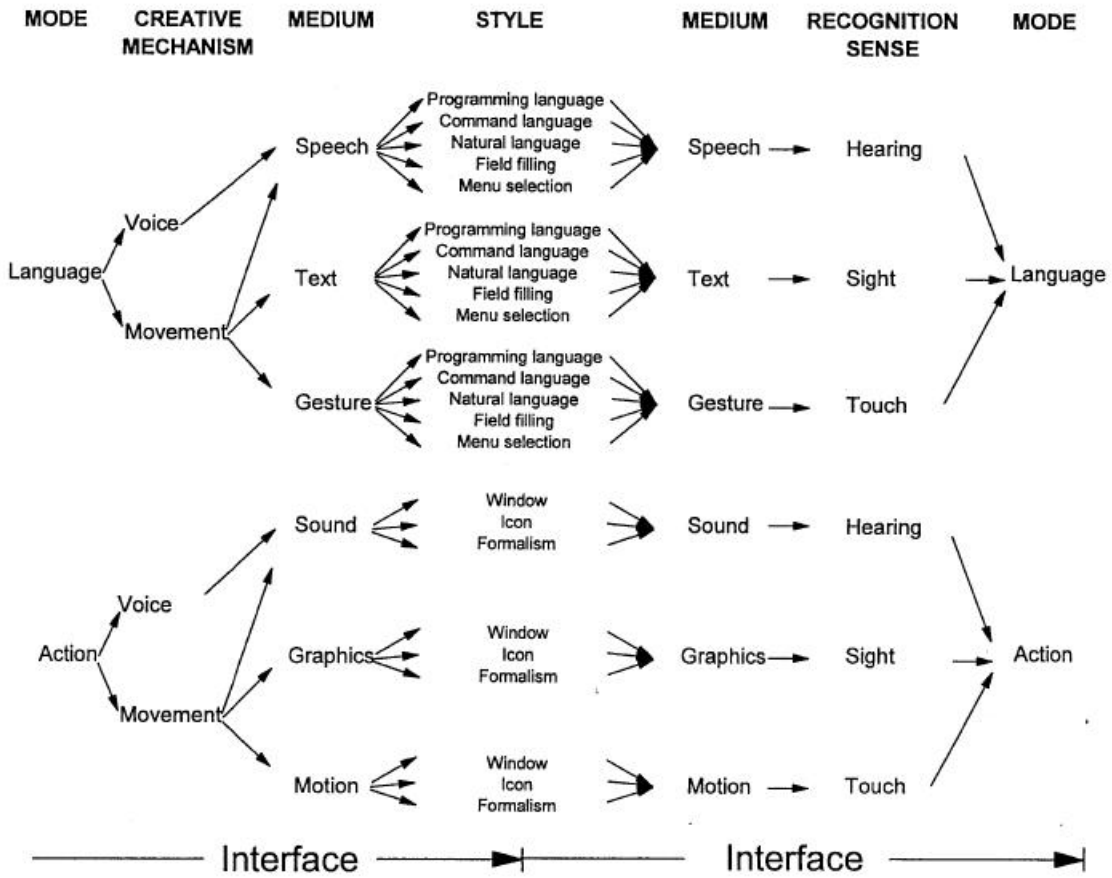

FIG. 1.2 – Espace des interfaces selon Frohlich (extrait de [Frohlich, 1996])

Frohlich propose l'un des plus anciens espaces de classification des interfaces dédiés aux interfaces multimédias. Notre étude s'appuie sur l'analyse de [Frohlich, 1991] par [Martin, 1995, Nigay et Coutaz, 1996] et sur [Frohlich, 1996].

Comme le montre la figure 1.2, Frohlich propose un espace qui distingue :

- les modes (*mode*) : ils correspondent aux "états dans lesquels différentes actions de l'utilisateur peuvent avoir le même effet" [Nigay et Coutaz, 1996]. Les deux modes identifiés par Frohlich sont le langage et l'action : le langage se rapporte à une métaphore conversationnelle alors que l'action se rapporte à une métaphore du monde réel ;
- les *media*[2] (*medium*) : ce sont des "systèmes représentationnels qui permettent l'échange d'informations" [Nigay et Coutaz, 1996]. Frohlich identifie six *media*, qui dépendent directement du mode qu'ils concrétisent : le langage peut se décliner en parole (vocalisation à fonction communicationnelle), en texte (trace visible d'un mouvement à fonction communicationnelle) ou en geste (mouvement communicationnel sans trace visible) et l'action peut se décliner en son, (toute forme audible autre que la parole), en graphique (toute trace visible autre que le texte) ou en mouvement (tout mouvement autre que le geste) ;
- les styles (*style*) : ils correspondent "aux classes reconnues de méthodes qui rendent possible l'interaction" [Nigay et Coutaz, 1996]. Si les styles du langage produisent des expressions, ceux de l'action produisent des événements.

Frohlich considère que la communication humain-machine se fait à travers deux interfaces distinctes, l'une d'entrée quand l'utilisateur communique "vers" la machine, l'autre de sortie quand la machine communique "vers" l'utilisateur. Modes et *media* ne sont pas articulés de la même façon en entrée et en sortie. Dans le sens humain-machine, le mode est exprimé grâce à des mécanismes de production (*creative mechanism*) qui le déclinent en un ou plusieurs *media*. Dans le sens machine-humain, les sens de perception (*recognition sense*) permettent d'interpréter le message transmis par le *medium* et d'extraire le mode. Frohlich met en évidence qu'un même mécanisme de production en entrée peut entraîner la transmission d'un message sur plusieurs *media* alors qu'en sortie le *medium* conditionne le sens de perception.

Étude comparative et prise de position ***La notion de "mode" est à un niveau différent des autres notions de l'espace de Frohlich et définit avant tout le paradigme d'interaction observé***, en l'occurrence avec ou sans une dimension langagière forte. La notion de "*medium*" n'a aucune dimension physique et ce dernier est déterminé par le mode d'une part, et par le sens de perception mobilisé d'autre part. Le style désigne un format de présentation. ***La notion de "modalité" peut correspondre aussi bien à chacun de ces trois éléments, qu'à tous à la fois.*** Les pendants des mécanismes de production et des sens de perception n'étant pas détaillés par Frohlich, il est difficile d'extraire une notion de "modalité" applicable aux agents tant naturels qu'artificiels. ***Cela aurait pourtant été dans la continuité de sa***

[2] Dans le sens latin originel ou selon l'acception anglophone, d'où un pluriel sans "s" et la typographie en italique.

distinction entre entrée et sortie qui permet la prise en compte de la liberté d'interaction différente de l'utilisateur.

1.2.2 Caractérisation des modalités représentationnelles selon Bernsen

À l'opposé de l'espace de Frohlich, l'espace de caractérisation de Bernsen [Bernsen, 1994, Bernsen, 1997] est dédié uniquement aux interfaces en sortie. Cet espace est issu d'une étude de la représentation possible des modalités motivée par la question suivante : quelle est la "meilleure" modalité pour représenter les informations à échanger entre le système et l'utilisateur ?[3]. Nous nous concentrons sur la représentation choisie par Bernsen, sans nous attarder sur la taxonomie qu'il en extrait : du propre constat de Bernsen, cette taxonomie est susceptible d'évoluer en fonction des technologies inventées et des systèmes développés.

Bernsen distingue les notions de "modalité sensorielle" et de "modalité représentationnelle". Une modalité sensorielle renvoie à l'un des sens humains. Une modalité représentationnelle est une façon de représenter l'information dans un format physique particulier et correspond à une sortie possible des systèmes de communication humain-machine. Bernsen la définit comme la composition d'un couple < *medium* d'expression, profil>.

Un *medium* d'expression (*medium of expression*) est une réalisation physique d'une information. Bernsen identifie trois *media*[4] d'expression courants en sortie des systèmes : le graphique, l'acoustique et l'haptique. Les *media* d'expression ont des propriétés propres qui les rendent perceptibles uniquement grâce à certaines modalités sensorielles. Ces propriétés sont appelées "canaux d'information" (*information channel*) par Bernsen. Elles contribuent à caractériser finement les modalités représentationnelles.

Le profil est identifié via quatre propriétés binaires de base. Chaque modalité représentationnelle est :

- linguistique ou non-linguistique (*linguistic/non-linguistic*) : cette propriété indique l'utilisation - ou non - d'un système syntaxico-sémantico-pragmatique. Ce système n'est pas forcément un langage naturel ou artificiel de façon formelle ;
- analogue ou non-analogue (*analogue/non-analogue*) : cette propriété fait référence à la similarité - ou non - entre la représentation et ce qu'elle représente. Par définition, une modalité représentationnelle ne peut être linguistique et analogue à la fois ;
- arbitraire ou non-arbitraire (*arbitrary/non-arbitrary*) : cette propriété distingue les modalités représentationnelles organisées selon un système considéré comme conventionnel de celles qui ne sont pas organisées selon un système considéré comme conventionnel. Par définition, une modalité représentationnelle ne peut

[3]Traduction libre de : "*given any particular set of information which needs to be exchanged between user and system during task performance in context, identify the input/output modalities which constitute an optimal solution to the representation and exchange of that information*".

[4]Reprenant le terme "*medium*" utilisé par l'auteur dans son acception latine et anglophone, le pluriel est sans "s", sans accent et en italique.

être à la fois linguistique et arbitraire, ni analogue et arbitraire ;
- statique ou dynamique (*static/dynamic*) : si cette propriété faisait initialement référence à l'absence ou à la présence d'une composante temporelle dans la représentation [Bernsen, 1994], Bernsen l'a étendue par la suite à la notion de "liberté d'examen perceptuel" (*freedom of perceptual inspection*) [Bernsen, 1997] : une modalité représentationnelle statique laisse l'utilisateur libre de son appréhension de l'information, alors qu'une modalité représentationnelle dynamique le contraint dans cette appréhension. Ainsi, une modalité représentationnelle dépendante du temps peut être considérée comme statique et *vice et versa*. L'exemple donné par Bersnen est une alarme qui se déclenche à intervalle régulier jusqu'à ce que l'utilisateur l'arrête : bien qu'incluant une caractéristique temporelle, cette modalité représentationnelle est maîtrisée par l'utilisateur et est donc considérée comme statique.

Bernsen s'appuie entre autres sur ces propriétés pour définir une taxonomie des modalités représentationnelles à quatre niveaux [Bernsen, 1997] présentée dans la figure 1.3. Les deux premiers niveaux sont obtenus en fusionnant les profils de modalités représentationnelles qui peuvent l'être (en l'occurrence, les formes statiques et dynamiques des modalités représentationnelles haptiques et acoustiques) et les modalités représentationnelles dont les propriétés sont incompatibles. Au niveau le plus haut (*super level*), Bernsen fait abstraction de la propriété statique/dynamique et du *medium* d'expression et obtient quatre groupes de modalités :
- les modalités linguistiques (*linguistic modalities*) qui sont linguistiques, non-analogues et non-arbitraires ;
- les modalités analogues (*analogue modalities*) qui sont analogues, non-linguistiques et non-arbitraires ;
- les modalités arbitraires (*arbitrary modalities*) qui sont arbitraires, non-linguistiques et non-analogues ;
- les structures de modalité explicites (*explicit modality structures*) qui sont ni linguistiques, ni analogues, ni arbitraires.

Le niveau générique (*generic level*), qui a été le premier à être établi, correspond à la déclinaison du niveau le plus haut en tenant compte du *medium* d'expression, ainsi que de la propriété statique/dynamique dans les cas opportuns.

Les deux derniers niveaux résultent de l'introduction de nouvelles propriétés. Les propriétés utilisées pour le niveau atomique (*atomic level*) sont spécifiques au niveau le plus haut et s'appliquent à certaines modalités représentationnelles identifiées au niveau générique. Les propriétés utilisées au niveau sous-atomique (*sub-atomic level*) permettent de raffiner encore certaines modalités définies au niveau atomique.

Étude comparative et prise de position L'étude de Bernsen caractérise le format de présentation des informations par un système informatique. Il se focalise sur la sortie des machines, considérant, dans la continuité de Frohlich, qu'entrée et sortie ne doivent pas être traitées de la même façon. Les propriétés spécifiques du niveau sous-atomique caractérisent la notion de "style" proposée par Frohlich dans son espace des interfaces (*cf.* la section 1.2.1).

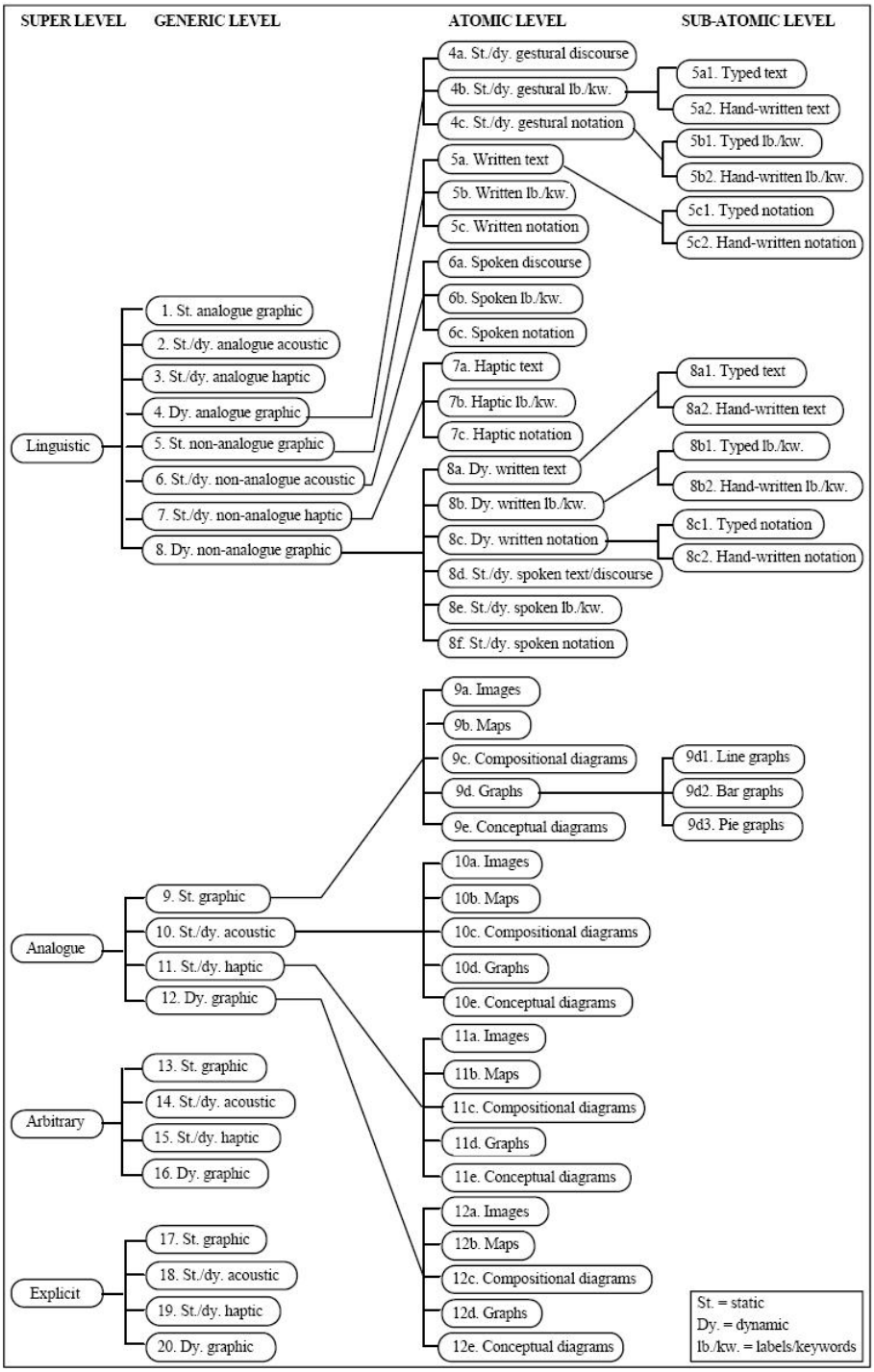

FIG. 1.3 – Taxonomie des modalités représentationnelles de sortie de Bernsen (extrait de [Bernsen, 1997])

Plusieurs travaux prolongent la caractérisation des modalités représentationnelles proposées par Bernsen. Entre autres, [Ratzka, 2006] affine la caractérisation de Bernsen en tenant compte des spécificités perceptuelles et manipulatoires des *media* d'expressions. ***Cette démarche peut être utile dans le choix d'un medium - i.e. d'un sens de perception à exploiter - ou son exclusion*** pour la production d'un message par les systèmes. Mais nous considérons qu'***intégrer une telle caractérisation dans un système pour déterminer le choix de sa sortie multimodale est irréalisable***, en raison de la complexité de cette caractérisation (comme le montrent les affinements proposés par Bernsen et par Ratzka).

1.2.3 Définitions pour les systèmes multi-sensori-moteurs selon Nigay et Coutaz

De façon à proposer un espace de référence pour les systèmes "multi-sensori-moteurs", Nigay et Coutaz définissent les notions de canal, de dispositif physique, de niveau d'abstraction, de langage d'interaction et de modalité [Nigay, 1994, Nigay et Coutaz, 1996].

L'appréhension d'un système multi-sensori-moteur (*i.e.* multimédia ou multimodal) passe par la détermination des canaux de communication disponibles. À la suite de Frohlich, entrée et sortie sont distingués. Un canal de communication d'entrée, respectivement de sortie, regroupe les dispositifs physiques d'entrée, respectivement de sortie. Un dispositif d'entrée est un récepteur de système - artificiel ou humain - et un dispositif de sortie un effecteur de ce système. Chaque canal peut disposer d'un ou plusieurs dispositifs physiques.

Les informations reçues ou transmises par un système grâce à ses dispositifs physiques peuvent avoir différents niveaux d'abstraction. Un niveau d'abstraction permet d'exprimer le degré de transformation pour passer d'une information perçue ou émise à l'information comprise par le système. Cette transformation est possible grâce à un langage d'interaction. Un langage d'interaction se définit par une grammaire dont les symboles terminaux sont produits ou perçus par les dispositifs physiques. Plus le langage d'interaction est réduit, plus le niveau d'abstraction est bas ; plus le langage d'interaction est complexe, plus le niveau d'abstraction est haut.

Les notions de "dispositif physique" et de "langage d'interaction" permettent de caractériser la communication humain-machine. Pour la notion de "modalité", les auteurs choisissent de laisser ouvert le niveau d'abstraction, de façon à ce qu'il soit restreint en fonction des besoins de concepteurs : une modalité est identifiée comme un dispositif d'entrée/sortie, un langage d'interaction, ou un couplage des deux. Cette définition de la modalité comme un couple <dispositif physique, langage d'interaction> est notamment exploitée dans [Vernier, 2001]et [Mansoux, 2005]). Nigay et Coutaz se positionnent d'ailleurs par rapport à la distinction entre média et modalité de la façon suivante. Elles considèrent qu'un média relève d'un bas niveau d'abstraction - matériel pour les systèmes artificiels ou perceptuel pour les humains - alors qu'une modalité inclut un plus haut niveau d'abstraction, avec un aspect représentationnel et/ou interprétationnel.

Étude comparative et prise de position L'approche de Nigay et Coutaz est clairement centrée machine même si elle peut s'appliquer à l'humain. Leurs définitions, et leur appréhension de la communication humain-machine de façon générale, ont le mérite d'être claires et organisées. ***Leur démarche ne va pas à l'encontre de celle de Bernsen (cf. la section 1.2.2) dont la notion de "medium" est à rapprocher de celle de "dispositif physique" avec une dimension plus matérielle que perceptuelle et le profil permet clairement de caractériser le langage d'interaction.*** Notons que la notion de "langage d'interaction" n'est pas nécessairement linguistique au sens de Bernsen. Un langage d'interaction permet juste - et c'est déjà un pas important dans la formalisation de la notion de "modalité" - de rappeler et de statuer que la communication est nécessairement codifiée et organisée. ***Cette définition se rapproche de la notion implicite de "code" suggérée par Shannon*** (*cf.* la section 1.1).

1.2.4 Communication humain-machine et notion de "modalité" selon Martin

Martin [Martin, 1995] définit plusieurs termes qui explicitent sa conception de la communication humain-machine. La communication est un échange d'énoncés. Chaque énoncé peut être composé d'une ou plusieurs informations. Une information est caractérisée par différents attributs, qui peuvent conditionner sa nature : par exemple, un événement est une information possédant un attribut "date". Les informations sont exprimées et perceptibles via des modalités. Une modalité, ou un mode, est un processus (dans le sens de "programme en cours d'exécution") d'analyse ou de synthèse, qui n'est pas nécessairement élémentaire : une modalité peut être le résultat de la combinaison de modalités (*cf.* le chapitre 2). Les modalités incluent les logiciels qui utilisent un dispositif physique. Un dispositif physique est ce qui permet à un système informatique d'acquérir ou de restituer des données : c'est donc le pendant des sens humains, appelés aussi modalités sensorielles.

De façon plus formelle, Martin distingue deux définitions de la notion de "modalité", l'une - statique - basée sur un ensemble de données possibles et l'autre - dynamique - basée sur un jeu particulier de ces données. Nous nous contentons ici de donner la définition dynamique qui suffit à comprendre le fonctionnement d'une modalité. Une modalité telle que définie dynamiquement par Martin est caractérisée par :

- les données analysées à un instant i parmi un ensemble de données possibles. Chaque donnée est une information qui peut être caractérisée par des attributs (par exemple une date de "perception", une position spatiale, une ou plusieurs modalités d'origine et une étiquette pour les informations communiquées via une souris ou un écran) auxquels il est possible d'accéder ;
- un processus spécifique à cette modalité ;
- un ensemble de résultats parmi un ensemble de résultats possibles ;
- un ensemble de paramètres ayant effectué un contrôle entrant parmi un ensemble de paramètres susceptibles d'effectuer un contrôle entrant ;
- un ensemble de paramètres constituant le contrôle sortant parmi un ensemble de

paramètres susceptibles d'effectuer un contrôle sortant.

La distinction entre les notions de "média" et de "modalité" est claire : les deux ne sont pas au même niveau, une modalité étant un processus alors qu'un média est un support sans rétroaction possible. Ce dernier n'intervient donc pas directement dans la communication humain-machine .

Étude comparative et prise de position Nous identifions des points de contact entre les définitions de Martin et celles de Frohlich (*cf.* la section 1.2.1). En effet, les notions de "*recognition sense*" et de "*productive mechanism*" de Frohlich ont la même dimension de processus que la "modalité" de Martin. Par contre, l'approche de Martin place au même niveau entrée et sortie mais aussi humain et système : entrée et sortie ne sont pas distinguées et chaque terme utilisé par l'humain a son pendant pour les machines. Cette schématisation de la communication humain-machine confère à cette dernière un aspect dynamique qui ne ressort par dans l'approche de Nigay et Coutaz (*cf.* la section 1.2.3). Toutefois, les deux démarches sont proches (1) par la volonté de formaliser la notion de "modalité" et la production ou l'acquisition des informations sur une modalité donnée et (2) par le niveau d'interprétation ou de génération de ces informations. En effet, la définition formelle de la notion de "modalité" en tant que processus et ensemble de données, de résultats et de paramètres peut être rapprochée de celle de "langage d'interaction". Une modalité selon Martin peut être aussi bien un langage d'interaction ou un dispositif physique tels que définis par Nigay et Coutaz. On rejoint alors la définition la plus ouverte de Nigay et Coutaz selon laquelle une modalité peut être un langage d'interaction, un dispositif physique ou un couplage des deux. La principale différence est que ***Martin choisit de ne pas dissocier les dispositifs physiques de leurs processus de traitement interne***, ce que font Nigay et Coutaz. Notons que plusieurs dispositifs physiques peuvent nécessiter le même processus de traitement interne.

1.2.5 Interfaces multimodales selon Bellik

Dans son appréhension des interfaces multimodales, Bellik [Bellik, 1995] se positionne par rapport aux notions de "média," de "mode" et de "modalité".

Pour Bellik, un média est un dispositif physique qui permet l'acquisition ou la diffusion d'information. Il s'appuie sur le fait qu'au sens commun le média désigne un support physique d'information.

La notion de "mode" se réfère aux organes humains et aux médias, plus précisément à leur mobilisation pour percevoir ou produire un message. À la suite de Frohlich, sont distingués les modes d'entrée et de sortie. Chez les humains, les modes de communication d'entrée correspondent aux cinq sens de perception (le visuel, l'auditif, l'olfactif, le gustatif et le tactilo-proprio-kinesthésique qui correspond au toucher étendu à l'haptique et à l'équilibre) et ceux de sortie aux moyens d'expression (le mode gestuel et le mode oral). Pour les machines, aucune liste des modes d'entrée et de sortie n'est proposée : elle dépend des médias existants, qui sont en constante évolution. Le mode, en particulier de sortie, contraint la nature des informations : si un agent - humain ou artificiel - produit

des informations sur un mode de sortie donné, ces informations seront perceptibles par l'agent-interlocuteur uniquement via certains modes d'entrée. Il est donc possible d'établir des correspondances entre médias d'entrée et de sortie et modes humains de perception et d'expression.

Chaque mode répond – ou non – à plusieurs critères, inspirés de la caractérisation des modalités par Bernsen (*cf.* la section 1.2.2). Les critères sélectionnés par Bellik [Bellik, 1995] sont les suivants :

- le critère temporel : il indique l'existence d'une "animation" pour les modes visuel et tactilo-proprio-kinesthésique et distingue les gestes statiques des gestes dynamiques du mode gestuel (entraînant la prise en compte du temps et de sa pertinence sémantique). Il correspond à la propriété spatial/dynamique de Bernsen dans sa définition initiale ;
- le critère spatial : c'est le fait que l'information soit transmise par le biais d'une ou plusieurs dimensions ;
- le critère langagier : c'est un critère à valeurs continues valable pour tous les modes. C'est une extension de la propriété langagier/non-langagier de Bernsen avec des valeurs continues au lieu d'être binaire ;
- le critère d'analogie : c'est la ressemblance entre le réel et sa représentation (image, son synthétisé, etc.). Ce critère correspond à la propriété analogue/non-analogue de Bernsen ;
- le critère de prosodie : il met l'accent sur l'importance dans la prosodie dans le mode considéré.

Comme le synthétise le tableau 1.1, Bellik identifie les critères pertinents pour chaque mode, plutôt que d'établir une taxonomie des combinaisons possibles à l'image de celle de Bernsen. Dans sa démarche, le critère temporel n'est pas pertinent pour le mode auditif, étant donné qu'un son inclut forcément une dimension temporelle.

	Critère temporel	Critère spatial	Critère langagier	Critère d'analogie	Critère de prosodie
Mode oral			X		X
Mode gestuel	X	X	X		
Mode visuel	X	X	X	X	
Mode auditif		X	X	X	X
Mode TPK	X		X	X	

TAB. 1.1 – Modes humains de sortie (haut du tableau) et d'entrée (bas du tableau) et critères pertinents selon Bellik (adapté de [Bellik, 1995])

La notion de "modalité" correspond, quant à elle, à une forme concrète particulière d'un mode de communication. Plus précisément, elle fait référence à la structure des informations telle qu'elle est perçue par l'humain. Chaque mode peut donc comporter plusieurs modalités. Les critères pertinents pour un mode sont valables pour les modalités qui en découlent. De plus, Bellik note qu'il peut y avoir plusieurs médias correspondant à une seule modalité.

Étude comparative et prise de position ***Bellik propose une vision unifiée des références citées.*** En effet, à la suite de Frohlich (*cf.* la section 1.2.1), il distingue entrée et sortie, qu'il caractérise grâce à des critères inspirés de Bernsen (*cf.* la section 1.2.2). ***Il fait ainsi le lien entre certains modes d'entrée et de sortie, à travers des critères communs. Ce lien pourrait être exploité pour le choix d'une modalité de sortie en fonction de la modalité d'entrée mise en œuvre par l'utilisateur.*** De plus, la notion de "média" de Bellik correspond à celle de dispositif physique et celle de "modalité" à celle de langage d'interaction de Nigay et Coutaz (*cf.* la section 1.2.3). Bien que la notion de "processus" ne soit pas explicite chez Bellik, on peut aussi rapprocher sa notion de "modalité" de celle de Martin (*cf.* la section 1.2.4), du moins du point de vue de ce qui est perceptible par l'utilisateur et si l'on admet qu'un dispositif physique délivre un certain format d'information. La notion de "mode" de Bellik est du même niveau que les notions de "*creative mechanism*" et "*recognition sense*" de Frohlich (*cf.* la section 1.2.1).

1.2.6 Notions de base pour le modèle de référence des systèmes intelligents de présentation multimédia

Définissant un modèle de référence pour les systèmes intelligents de présentation multimédia - plus connus sous l'acronyme d'IMMPS pour "*Intelligent MultiMedia Presentation Systems*" -, [Bordegoni *et al.*, 1997] introduit une vision de la communication qui permet de passer des contenus à transmettre aux informations perceptibles. La figure 1.4 synthétise les différentes notions impliquées.

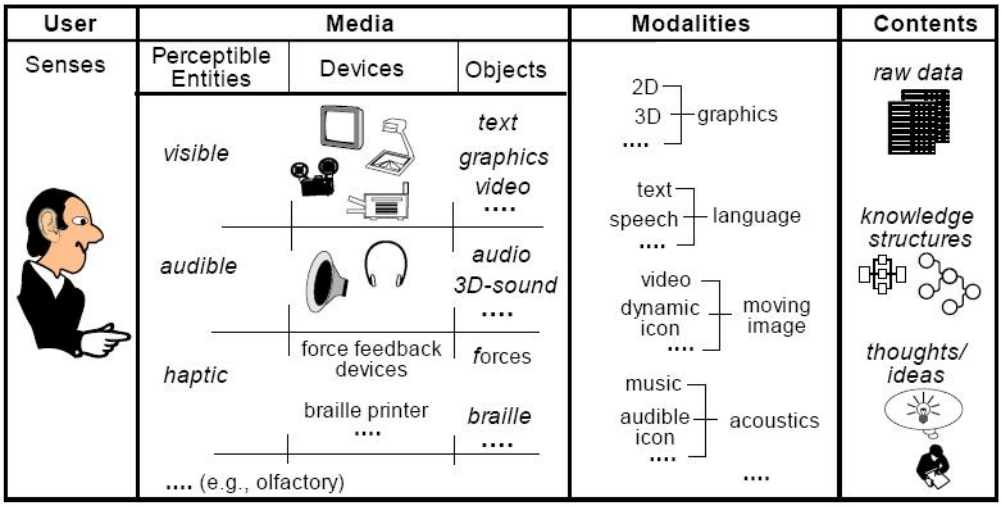

FIG. 1.4 – Des contenus à transmettre aux informations perceptibles dans les systèmes intelligents de présentation multimédia (extrait de [Bordegoni *et al.*, 1997])

Tout comme dans la théorie de la communication Shannon (*cf.* la section 1.1), il y a, à une extrémité de la chaîne de communication, une source et, à l'autre extrémité, un destinataire. La source, à droite de la figure cherche à transmettre un contenu particulier

de façon à ce qu'il soit perceptible par le destinataire, à gauche de la figure. Si la source peut aussi bien être humaine qu'artificielle, le destinataire est toujours considéré comme humain. La condition *sine qua non* pour que cette transmission soit possible est que le contenu soit présenté de façon perceptible par le destinataire. Cette approche amène les auteurs à donner plusieurs sens au terme "*medium*" en fonction du point de vue adopté. Ce terme peut renvoyer à :

- un espace physique pour matérialiser un contenu de façon à ce qu'il soit perceptible. Cette définition se rapproche de la notion de "support" ou de "dispositif physique", avec la distinction du sens de perception mobilisé (*i.e.* visible, audible, haptique, etc.). Un "*medium*" ainsi défini peut être caractérisé par les différentes dimensions physiques nécessaires pour produire un contenu perceptible dans l'environnement considéré ;
- un type d'information et/ou un format de représentation. Des objets respectant un format de représentation donné (*media objects*) sont rendus perceptibles grâce à un ou plusieurs dispositifs physiques.

C'est la notion de "modalité" (*modality*) qui permet de passer du contenu tel que le conçoit la source d'une part aux objets définis selon un format de représentation donné d'autre part. Plus précisément, une modalité fait référence à un format ou à un mécanisme d'encodage des informations qui permettant de présenter ces informations au destinataire dans une forme physique concrète. Le choix de la modalité conditionne, de par les caractéristiques de cette dernière, les *media* possibles.

Étude comparative et prise de position Pour poursuivre la comparaison avec la théorie de la communication de Shannon (*cf.* la section 1.1), les modalités correspondent au code ou au codage choisi et utilisé pour émettre le message. Les *media* peuvent être, en fonction du point de vue adopté, le canal de transmission ou l'ensemble du message codé. Soulignons que, suivant l'approche adoptée pour poser le modèle de référence des IMMPS, ***une modalité peut être aussi bien un processus (i.e. un codage) que ce qui permet la réalisation de ce processus (i.e. un code) et un medium peut être aussi bien un format, que le résultat de l'application de ce format ou encore ce qui permet de rendre ce résultat perceptible***. Si l'on se réfère aux approches déjà citées, (1) la modalité de façon générale correspond à la combinaison du *creative mechanism*, du *medium* et du *style* considérés par Frohlich (*cf.* la section 1.2.1) car celui-ci ne parle pas explicitement de mécanisme de codage mais la notion de "*creative mechanism*" semble indiquer un processus dynamique ; (2) la modalité en tant que processus peut être rapprochée de la modalité telle que définie par Martin (*cf.* la section 1.2.4) ou du *creative mechanism* de Frohlich mais défini plus spécifiquement ; (3) la modalité en tant que format d'encodage renvoie à la notion de "langage d'interaction" de Nigay et Coutaz (*cf.* la section 1.2.3), à celle de "modalité" telle que conçue par Bellik (*cf.* la section 1.2.5) ou encore à ce qui définit le profil de modalité chez Bernsen (*cf.* la section 1.2.2) ; (4) le *medium* en tant que format semble correspondre à une généralisation du style de Frohlich ; (5) le *medium* en tant que résultat d'application d'un format de représentation peut être rapproché du *medium* tel que défini par Bernsen ou du *recognition sense* de Frohlich ; (6) enfin, le *medium* en

tant que dispositif permettant de rendre le résultat d'application du format perceptible correspond au dispositif physique de Nigay et Coutaz et au média de Bellik.

Cette approche permet donc de rassembler les approches existantes et de donner une vue globale de l'appréhension des notions de "*medium*" et de "*modalité*" et de leurs places dans la communication humain-machine. Toutefois, ceci entraîne un positionnement flou : la distinction entre la "modalité en tant que code" et "le *medium* en tant que format" ne nous semble pas très claire. ***Le flou entre ces deux définitions nous insite à considérer que les travaux sur les IMMPS ne sont pas spécifiques aux systèmes multimédias et peuvent parfaitement être intégrés et assimilés aux travaux sur la présentation multimodale d'informations.***

1.2.7 Caractérisation des modalités de sortie selon Vernier

La thèse de Vernier [Vernier, 2001] constitue l'une des premières études françaises à focaliser exclusivement sur la multimodalité en sortie. Adoptant la définition de "modalité" de Nigay et Coutaz, il propose une caractérisation des modalités qui distingue les dispositifs physiques d'une part et sur les langages d'interactions d'autre part.

Vernier considère deux niveaux pour caractériser un dispositif physique. À un premier niveau, un dispositif physique est caractérisé par le sens humain qui permet la perception du message produit, à savoir la vue, l'ouïe, l'odorat, le goût ou le toucher. À un deuxième niveau, un dispositif physique est caractérisé par les propriétés propres du sens impliqué. Par exemple, un écran est caractérisé à un premier niveau par la vue et à un deuxième niveau par la couleur, la forme, etc. Notons que Vernier ne détaille pas les valeurs possibles de la caractérisation des dispositifs physiques à un deuxième niveau et ne fait *a fortiori* pas de taxonomie des dispositifs physiques possibles ou actuels.

En ce qui concerne les langages d'interaction, Vernier considère qu'ils permettent de définir des classes de modalités, un même langage d'interaction pouvant être concrétisé grâce à des dispositifs physiques différents. Aussi, il s'appuie sur les quatre propriétés identifiées par Bernsen (statique/dynamique, linguistique/non-linguistique, arbitraire/non-arbitraire, analogique/non-analogique[5]) pour caractériser les langages d'interaction. Toutefois, il y apporte une nuance : de façon à prendre en compte la dimension temporelle de la communication humain-machine qui conduit à une disponibilité variable des modalités au cours de l'interaction, Vernier considère que chaque modalité peut combiner les deux valeurs binaires d'une même propriété. De plus, il introduit deux nouvelles propriétés précisant si une modalité est :

- pure/composée : une modalité pure est une modalité atomique (c'est-à-dire dont le langage d'interaction est atomique) ; une modalité composée est une modalité non-atomique (c'est-à-dire dont le langage d'interaction n'est pas atomique). L'atomicité d'une modalité dépend directement de la granularité choisie par le concepteur et permet une récursivité dans la composition des modalités ;
- passive/active/passive-active : cette propriété permet de faire le lien entre les modalité de sortie et les modalités d'entrée. Une modalité purement informative est dite passive ; une modalité de sortie qui peut être utilisée en entrée (par exemple,

[5]Propriété correspondant à celle que nous avons traduite par "analogue/non-analogue".

un bouton) est dite active ; une modalité qui peut être tour à tour passive ou active durant l'interaction est dite passive-active (par exemple, les items d'un menu déroulant qui sont actifs de façon contextuelle). Vernier précise qu'une modalité composée d'au-moins une modalité active, respectivement passive-active, est aussi considérée comme active, respectivement passive-active.

Selon l'approche de Vernier, une modalité est donc caractérisée par :
- sa nature statique, dynamique ou statique/dynamique ;
- sa nature linguistique, non-linguistique ou linguistique/non-linguistique ;
- sa nature arbitraire, non-arbitraire ou arbitraire/non-arbitraire ;
- sa nature analogique, non-analogique ou analogique/non-analogique ;
- sa nature pure ou composée ;
- sa nature passive, active ou passive-active ;
- le sens humain de perception ;
- les propriétés de ce sens humain.

Étude comparative et prise de position Le positionnement de Vernier par rapport à la notion de "modalité" et par rapport à l'appréhension de la communication humain-machine est le même que celui de Nigay et Coutaz (*cf.* la section 1.2.3). Sa contribution est double. D'une part, il établit comme propriété de caractérisation la composition ou non d'une modalité donnée, là où Bernsen s'était contenté de signaler une composition possible. D'autre part, il introduit la dimension dynamique de la communication, *i.e.* la disponibilité des modalités et la variation possible des propriétés au cours du temps. C'est cette deuxième contribution qui nous semble la plus importante car elle introduit la possibilité qu'une modalité n'est pas disponible dans une situation donnée. ***Cette non-disponibilité peut alors être considérée par le système comme une contrainte de présentation.***

1.2.8 Dialogue humain-machine selon Landragin

Les travaux de Landragin [Landragin, 2004a] focalisent sur le dialogue et cherchent à donner les moyens aux machines d'être aussi performantes que l'humain dans ce type de communication. À cette fin, Landragin se concentre sur la problématique de la référence aux objets dans les situations dialogiques. Il identifie trois principales modalités en jeu : la parole et le geste en tant que "modalités d'expression" et la perception visuelle en tant que "modalité de support" [6]. Il souligne que les modalités d'expression qu'utilisent les interlocuteurs font référence à la modalité de support. Les messages produits n'auront donc aucune signification sans cette modalité particulière. Landragin va même plus loin en mettant en avant le fait que les objets de la modalité de support peuvent être plus ou moins saillants, voire constituer des groupes perceptifs, impactant ainsi l'utilisation des modalités d'expression. De façon plus précise, Landragin définit la modalité d'expression comme étant la composition d'une dimension physiologique (comprenant un sens de perception et une intensité) et d'un langage d'interaction. Pour pouvoir interpréter

[6] Les termes "modalité de support" et "modalités d'expression" employés par l'auteur.

une modalité, un système doit avoir accès au langage d'interaction sous-tendu et à au moins un dispositif physique pour capturer les informations transmises par la modalité en question. La relation entre modalité et dispositif physique n'est pas bijective : plusieurs modalités peuvent être concrétisées par un seul dispositif physique et plusieurs dispositifs physiques peuvent traiter une seule modalité. Landragin emploie également le terme latin "*medium*" pour désigner un dispositif physique et le terme de "mode" pour désigner une modalité. Landragin ne traite pas explicitement de modalités de sortie, mais son approche s'y applique aussi. La prise en compte de la dimension physiologique de la modalité pour le choix et/ou l'utilisation de la modalité est moins triviale dans le cas des sorties.

Étude comparative et prise de position Landragin identifie deux canaux mobilisés pour la transmission des informations dans la communication humaine. Il s'agit du canal visuo-gestuel et du canal audio-oral. Landragin précise que chaque canal peut transmettre une ou plusieurs modalités : par exemple, le geste peut être spécifié en fonction de la partie du corps qui est utilisée pour le produire, définissant ainsi un canal plus spécifique. Notons que le nom donné aux canaux par Landragin spécifie à la fois le sens de perception visé et le type de modalité d'expression utilisé (si ce qui se transmet par la parole est considéré comme étant oral). Par extension, ***nous pourrions définir un canal de communication humain-machine dans le sens machine vers humain par le couple <sens perceptif visé, dispositif physique utilisé>.***

Notons que la notion de "modalité du support" n'apparaît pas, à notre connaissance, dans d'autres approches de la communication humain-humain et humain-machine. Lorsqu'elle est présente, elle est intégrée au contexte de communication, notion vague s'il en est, et n'a donc pas un rôle aussi central que celui que lui donne Landragin. En tant que source d'informations, elle a pourtant un rôle à jouer, y compris dans la communication dans le sens machine vers utilisateur : dans ce cas, elle peut permettre une redondance des informations principales ou la présentation d'informations additionnelles. Lorsqu'elle est numérique, ***le recours à la modalité de support peut constituer une plus-value pour le système qui peut la transformer de source passive en source active*** : par exemple, il peut agir sur une carte (point clignotant, zoom) pour attirer l'attention de l'utilisateur sur un élément particulier et de façon plus précise qu'un geste déictique généralement plus vague. ***Les modalités de support doivent, autant que faire se peut, toujours être disponibles pour l'utilisateur*** : ainsi les informations hypertextuelles doivent-elles, si le support le permet, être présentées, même si l'utilisateur communique oralement.

1.2.9 Modèle de la communication selon Clémente

Clémente propose un modèle de la communication [Clémente, 2004] valable tant pour les agents naturels (*i.e.* les humains) qu'artificiels (*i.e.* les systèmes informatiques). Les paragraphes qui suivent précisent les différences et distinctions pour chaque type d'agent. Nous n'exposons pas ici l'ensemble du modèle de Clémente et nous ne détaillons pas les taxonomies proposées, cherchant avant tout à appréhender la vision d'ensemble

de Clémente sur la communication afin d'isoler la notion de "modalité" telle qu'il la considère en sortie des systèmes multimodaux de communication humain-machine.

Pour Clémente, communiquer, c'est avant tout agir et percevoir, même si des processus de raisonnement sont nécessaires pour la production et la compréhension de messages. Action et perception sont des capacités duales. Alors que les capacités perceptives d'un agent lui permettent de capter les informations, ses capacités d'action lui permettent d'en extérioriser. Ces capacités sont réalisées grâce à des dispositifs, respectivement sensoriels et moteurs, mais l'auteur préfère conserver le terme "capacité" à celui de "dispositif" considérant qu'une capacité est potentiellement composée de plusieurs dispositifs. Clémente appelle "motricités" les capacités humaines[7] d'action, "sens" les capacités humaines de perception, "effecteurs" les capacités artificielles[8] d'action et "capteurs" les capacités artificielles de perception. Clémente précise que la combinaison d'un objet et d'une capacité sensorielle ou motrice (qui correspond à une prothèse) produit une nouvelle capacité sensorielle ou motrice.

Clémente identifie deux grands canaux utilisés pour communiquer : l'environnement - quand il n'y a pas de contact physique entre les interlocuteurs - et le destinataire (respectivement l'émetteur, selon que le système soit en "action" ou en "perception") - lorsqu'il y a contact physique avec l'émetteur (respectivement le destinataire). Clémente souligne la nécessité de connaître la nature des informations pour déterminer les capacités d'action à mobiliser lors de la production d'un message, en particulier dans le cas où le canal est l'environnement ambiant. Ces canaux comprennent un certain nombre de modes.

Pour Clémente, un mode de communication est une forme physique, concrète de transmission d'énergie ou de matière présente initialement, de façon continue ou discrète, dans l'environnement des interlocuteurs et avec laquelle les dits interlocuteurs sont potentiellement en contact. Un mode est produit - respectivement perçu - par une capacité d'action - respectivement de perception. C'est la nature du message, et non la "nature" de l'émetteur ou celle du destinataire, qui détermine le mode de communication. Autrement dit, un mode est un support de transmission de messages et est à rapprocher de la notion de "canal" de Shannon. Clémente propose une taxonomie des modes de communication utilisés par des agents artificiels ou naturels. Cette taxonomie renvoie généralement, à un premier niveau, au codage et au support de transmission utilisés (rayonnement électromagnétique, vibrations acoustiques, électrique, magnétique, réseau et téléphonique) et à des niveaux inférieurs à la capacité sensorielle mobilisée (pour le mode "contact") ou au codage de transmission à un grain plus fin (par exemple, longueurs d'onde, protocoles réseaux, etc.).

Clémente distingue les média[9] d'acquisition, de production et d'interaction, qui interviennent directement dans la communication. Un média d'acquisition - ou média sensoriel - regroupe l'ensemble des dispositifs permettant de transformer un message reçu en un message interprétable par l'agent. Recevant des informations transmises sur un mode unique, il n'opére aucun traitement sémantique sur ces informations. Un média

[7]Dans le sens où l'agent considéré est humain.
[8]Dans le sens où l'agent considéré est artificiel.
[9]Singulier et pluriel de "média" sont invariables pour Clémente.

de production regroupe l'ensemble des dispositifs permettant à partir d'un contenu sémantique de produire des informations transmissibles par un ou plusieurs modes. Média d'acquisition et média de production font le pont entre les agents et l'environnement, autrement dit entre capacités sensorimotrices et modes. Ils sont regroupés sous le terme générique de "média d'interaction".

Étude comparative et prise de position La notion de "média d'interaction" correspond à celle de "média" chez Bellik (*cf.* la section 1.2.5) et à celle de "dispositif physique" chez Nigay et Coutaz (*cf.* la section 1.2.3). Les capacités sensorimotrices sont à rapprocher des notions de "*creative mechanism*" et de "*recognition sense*" évoquées par Frohlich (*cf.* la section 1.2.1). Le terme "motricité" a la même dimension dynamique que le terme "modalité d'expression" utilisé par Landragin (*cf.* la section 1.2.8), avec toutefois une dimension plus actionnelle que communicationnelle. Les termes "capteurs" et "effecteurs" renvoient directement aux termes "*transmitter*" et "*receiver*" employés par Shannon, plaçant la machine dans une forte position passive. ***Nous leur préférons les termes de "capacités d'action" et de "capacités de perception" qui dénotent de la liberté d'interaction laissée tant à l'utilisateur qu'au système.***

La définition de la notion de "modalité de communication" proposée par Clémente est double. D'une part, à l'exemple de Bellik, une modalité renvoie à une structuration des informations de façon concrète, perceptible. Cette structuration se fait en suivant certaines règles rarement arbitraires. D'autre part, une modalité renvoie à une structuration abstraite des informations : elle se rapproche alors plus de la notion de "langage d'interaction" de Nigay et Coutaz. Pour Clémente, une modalité est en fait une structuration possible de l'information, que celle-ci soit matérialisée ou non. ***Sa démarche rejoint celle adoptée pour la notion de "modalité" dans le cadre de la proposition d'un modèle de référence pour les systèmes intelligents de présentation multimédia.***

Reprenant les travaux de Bernsen (*cf.* la section 1.2.2) et ceux de Bellik, Clémente spécifie l'ensemble des critères qui permettent de caractériser une modalité donnée. Ces critères sont au nombre de 17 (auxquels s'ajoutent quatre autres pour caractériser la coopération possible d'une modalité avec les autres modalités). Nous ne le détaillons pas car cette large caractérisation, bien qu'utile pour l'étude des systèmes multimodaux, nous semble difficile à exploiter : trop de critères tue les critères. De plus, ***il nous semble délicat de prétendre établir une liste exhaustive de tous les critères à prendre en compte, ces derniers dépendant directement de l'état des connaissances sur l'assimilation des informations par l'humain***. Ce sont principalement ces deux raisons qui nous ont poussés à ne pas faire une revue des caractérisations ni des taxonomies des modalités possibles.

1.2.10 Composants d'interaction selon Rousseau

Rousseau [Rousseau, 2006] appelle "composants d'interaction" les éléments permettant la communication entre humains et machines. Trois types de composants sont identifiés : le mode, la modalité et le média. Les définitions adoptées sont celles de

Bellik avec une focalisation sur la sortie des systèmes informatiques.

Un mode de sortie renvoie aux systèmes sensoriels de l'utilisateur. Si les modes visuel, auditif et tactilo-proprio-kinesthésique sont couramment employés en sortie des systèmes informatiques, Rousseau précise qu'à ce jour (pour reprendre son expression), les moyens techniques actuels limitent le recours aux modes gustatifs et olfactifs. Nous pouvons ajouter que l'impact de ces modes sur l'humain et leur interprétation ne sont pas maîtrisés.

Une modalité de sortie correspond à la structure de l'information perçue par l'humain, quelque soit la structure identifiée ou utilisée par la machine. Rousseau précise que modalité et attribut de modalité ne doivent pas être confondus. En effet, une modalité se caractérise par un ensemble d'attributs, dont une partie au moins lui est spécifique. Si seules les valeurs des attributs varient entre deux structures d'information, alors il ne s'agit pas deux modalités différentes. Ainsi, deux textes écrits en deux langues utilisant le même alphabet relèvent d'une même modalité avec un attribut - la langue - qui les spécifie alors que deux textes écrits dans deux alphabets différents sont deux modalités à part entière, chaque langue ayant ses propres attributs spécifiques.

Un média de sortie n'est rien d'autre qu'un dispositif physique ou un groupe de dispositifs physiques de sortie permettant la concrétisation d'une modalité.

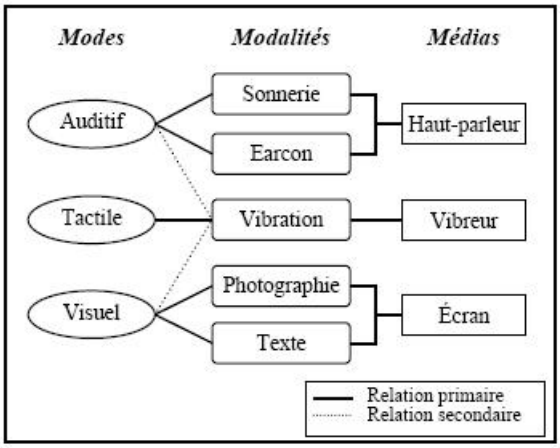

Fig. 1.5 – Exemple de diagramme des composants d'interaction dans le cas d'un téléphone portable en réception d'appel (extrait de [Rousseau, 2006])

Rousseau souligne que mode, modalité et média peuvent être organisés dans un diagramme des composants d'interaction. Un média peut réaliser une ou plusieurs modalités, plusieurs modalités peuvent être générées en utilisant un même média et, logiquement, plusieurs médias peuvent exprimer une même modalité. Une modalité est perçue via un mode et plusieurs modalités peuvent être perçues grâce à un seul mode. Rousseau n'indique pas s'il considère qu'une modalité peut être perçue grâce à la mobilisation de plusieurs modes : la question se pose dans le cas typique des avatars, suivant qu'ils sont considérés comme une composition de modalités ou comme une unique modalité. Dans les diagrammes des composants d'interaction tels que celui présenté en

1.5, modalités et modes entretiennent deux types de relations : une relation principale lorsque le mode considéré est celui qui est communément utilisé pour percevoir la modalité et une modalité secondaire dans le cas contraire, *i.e.* la perception de cette modalité par ce mode est un effet de bord de la relation principale. La nature des relations entre modes et modalités est complètement dépendante de la perception humaine individuelle et donc du public considéré.

Étude comparative et prise de position La notion d'"attribut" peut être rapprochée de celle de "propriété" selon Bernsen (*cf.* la section 1.2.2), avec un grain plus fin, ainsi que de celle de "critère" utilisée par Bellik (*cf.* la section 1.2.2) et par Clémente (*cf.* la section 1.2.9). Par ailleurs, les deux types de relations entre modalité et modalité mis en évidence dans les diagrammes des composants d'interaction est particulièrement intéressante : ***elles permettent de prendre en compte les effets de bord éventuels d'une modalité donnée***. La notion de "relation secondaire" explicite ce que Clémente a cherché à mettre en évidence avec sa taxonomie des modes de communication.

1.2.11 Synthèse : modalité en informatique

Pour une vue d'ensemble sur les différentes approches, la tableau 1.2 synthétise et propose une comparaison entre les termes principaux utilisées dans les références citées : les termes les plus proches sont placés dans une même colonne.

Ce tableau fait ressortir que la notion de "modalité" peut s'apparenter à :

- la notion de "dispositif physique". Celui-ci est utilisé par les machines pour percevoir ou produire des messages. Sont également utilisés les termes de "média (d'interaction)", de "*medium*" et de "modalité ou mode de support" ;
- les notions de "capacité de perception" mais aussi de "capacité d'action" pour les humains, pendants des dispositifs physiques des machines. Les auteurs parlent de "mode", "*medium*", "motricités" et "sens", "*creative mechanism*" et "*recognition sense*" et plus couramment "modalité (sensorielle)" ;
- la notion d'"organisation abstraite ou concrète des informations émises ou perçues". Le terme "langage d'interaction" pour l'organisation abstraite est aussi utilisé et renvoie au codage et à l'organisation des informations ;
- une combinaison de deux de ces notions.

Comme nous l'avons déjà souligné, nous avons fait le choix de ne pas détailler les propriétés/critères/attributs identifiés ou possibles pour caractériser les modalités. En effet, comme le soulignent Bernsen et Rousseau, ces propriétés sont en partie spécifiques à une modalité. De plus, elles résultent de l'analyse de la communication et des connaissances actuelles sur la perception humaine au sens large. Par conséquent, cette liste est loin d'être exhaustive et elle est pourtant déjà longue. Elle nous semble donc difficile, voire impossible, à exploiter pour la conception de systèmes multimodaux. Toutefois, le lecteur intéressé pourra se reporter entre autres à [Bellik, 1995, Bernsen, 1994, Bernsen, 1997, Clémente, 2004, Ratzka, 2006, Vernier, 2001].

Pour compléter notre espace terminologique en prenant en compte la dimension humaine, nous étudions, dans la section suivante, l'appréhension de la notion de "modalité"

[Bellik, 1995]	"média"	"modalité"	"mode" (humain)
[Bernsen, 1994] /[Bernsen, 1997]		caractérisation par un profil	"*medium*" (humain)
		"modalité"	
[Bordegoni *et al.*, 1997]	"*medium*"	"modalité"	
			"modalité" (informatique) ou "*medium*" (humain)
		"modalité"	
[Clémente, 2004]	"média d'interaction"	"modalité"	"capacités de perception/d'action"
[Frohlich, 1991]		"*medium*"	"*creative mechanism / recognition sense*"
[Landragin, 2004a]	"modalité ou mode de support" (étendu au contexte perceptible)	"modalité d'expression" (incluant une intensité)	
[Martin, 1995]	"modalité" ou "mode"		"modalité sensorielle" - "sens" (entrée humaine)
[Nigay et Coutaz, 1996]	"dispositif physique"	"langage d'interaction"	/
	"modalité"		
[Rousseau, 2006]	"média"	"modalité"	"mode" (humain)
[Vernier, 2001]	"dispositif physique"	"langage d'interaction"	/
	"modalité"		
[Wahlster, 2006] [Wahlster, 2006]	/	/	"modalité" (humaines) ou "mode" (informatiques)

TAB. 1.2 – Synthèse et comparatif des termes utilisés pour définir la notion de "modalité"

dans d'autres domaines que l'informatique.

1.3 Notion de "modalité" dans d'autres domaines

La multimodalité dans les systèmes informatiques n'auraient sans doute pas vu le jour si l'humain ne communiquait pas naturellement de façon multimodale. Non pas uniquement plurimodale, avec la cohabitation de plusieurs capacités d'action et de perception, mais bien de façon multimodale, avec la combinaison de ces capacités de façon simultanée. Même en prenant l'exemple des langues de signes, où a priori aucun son n'est émis, l'information n'est pas seulement transmise par les gestes, elle l'est

aussi par les expressions faciales. Les informations transmises par ces deux procédés ne sont pas les mêmes, car les supports (*i.e.* l'équivalent des dispositifs physiques) et la construction des messages (*i.e.* les langages d'interaction sous-tendus) sont différents. Il s'agit de modalités différentes.

Néanmoins, les travaux des sciences expérimentales et des sciences humaines utilisent rarement le terme de "modalité" lorsqu'ils traitent de la communication humaine. Extraire cette notion peut donc sembler encore plus délicat qu'en informatique. Toutefois, parce que la communication humaine est par essence multimodale, parce qu'elle a été étudiée dans des domaines aussi variés que l'histoire, les sciences de l'information et de la communication, l'éthologie, la base biologique des neurosciences, la psychologie ou encore les sciences cognitives, et ce bien avant que l'informatique s'y intéresse, il est important de s'étudier ces travaux.

Nous avons choisi de regrouper les travaux étudiés par domaine. Pour chacun de ces domaines, les principales conclusions que nous tirons sont mises en gras et en italique. Nous commençons par une approche de la communication humaine du point de vue de l'histoire.

1.3.1 Point de vue issu de l'histoire

Comme cela est souligné dans [Calvet, 1996], les anthropologues et les historiens considèrent généralement que l'écriture est rattachée à la langue et que, même si la plupart des écritures semblent avoir été d'abord pictographiques[10] ou idéographiques[11], le summum de leur développement est de se doter d'un alphabet. Cette idée est motivée par le fait qu'un alphabet permet de noter aussi bien les consonnes que les voyelles et rend ainsi possible une trace écrite parfaitement fidèle à ce qui est dit. De plus, cette trace est durable - dans la mesure, bien sûr, où l'alphabet est conservé.

Cependant, certains chercheurs, tels que Leroi-Gourhan, Clottes ou encore Courtin, considèrent que les premières "écritures"[12] ont transcrit des gestes et non des sons. Bien que controversée, Calvet croit en cette hypothèse. Pour lui, ce n'est pas l'oralité qui est à l'origine de l'écriture, mais la rencontre entre gestualité et picturalité. Plus précisément, "l'écriture est la picturalité asservie à la gestualité", et non à l'oralité. L'écriture, bien que trace de l'oral, est le résultat de la picturalisation du geste. Frohlich semble le rejoindre, étant donné qu'il considère que l'écriture est la trace visible du mouvement.

Calvet identifie donc trois grands groupes de moyens d'expression, qui peuvent s'apparenter à la notion de "modalité" mais qui se rapprochent surtout des *media* de Frohlich : l'oralité, la gestualité et la picturalité (qui intègre l'écriture). Il ne tient pas compte du fait que le geste est à l'origine des écritures. Il ne tient pas non plus compte de la particularité constitutive du caractère durable des écritures, à savoir l'utilisation d'un *medium* (au sens latin du terme), d'un support "de conservation" indépendant du support

[10] Un pictogramme est un dessin représentant un message sans référence à sa forme linguistique.

[11] Un idéogramme est un signe graphique qui représente une idée. Pour Calvet, les idéogrammes ont ceci de plus que les pictogrammes qu'ils constituent un système et sont donc une écriture à part entière.

[12] Il s'agit, en particulier, des "mains négatives", pourtours de mains peintes (les couleurs sont donc inversées comme pour des négatifs photographiques) avec des doigts "sectionnés".

de production. Se focalisant sur les caractéristiques des messages obtenus, il constate que les messages gestuels sont, par définition, fugaces, au même titre que les messages oraux. Les messages picturaux, inscrits sur un support comme le sable, les parois ou le papier, résistent - encore que ce soit de façon plus ou moins efficace en fonction du support d'inscription utilisé - au temps, et ce même si l'interprétation est susceptible de se perdre.

Étude comparative et prise de position La rémanence de l'écrit est connue en informatique. Elle transparaît dans la distinction entre écrit et oral. Cette distinction pourrait nous amener à penser que ***la rémanence est propre aux informations perceptibles visuellement. Ce n'est pourtant pas le cas : un geste, sans contact physique et qui n'est pas filmé, est perçu visuellement, il n'est pourtant pas rémanent.*** Même sans support d'inscription physique (par opposition à numérique), l'écriture peut être rémanente : c'est le cas d'un texte affiché à l'écran, dans la mesure où l'utilisateur le contrôle[13].

Nous en concluons que deux propriétés sont souvent omises dans les travaux en informatique. La première est ***la rémanence***, propriété aussi importante que la dimension dynamique d'une modalité telle que définie par Bellik (*cf.* la section 1.2.5). La deuxième est ***le contrôle de l'utilisateur sur la rémanence*** de façon à décrire la mesure dans laquelle l'utilisateur subit ou non l'information présentée avec la modalité considérée : elle se rapproche de la propriété statique/dynamique de Bernsen (*cf.* la section 1.2.2). De façon plus générale, l'appropriation de la machine passe par la garantie du contrôle de l'utilisateur : dans le cas de nos travaux sur les sorties des systèmes, ***nous considérons que ce contrôle doit aussi se manifester par la possibilité continue de l'utilisateur d'avoir recours à la capacité d'action qu'il souhaite***. Par exemple, même si le système présente une liste de solutions oralement via un téléphone portable, il doit exploiter l'écran de façon à permettre à l'utilisateur d'avoir recours au clavier et à un éventuel pointeur lors de sa prochaine interaction.

1.3.2 Points de vue issus des sciences de l'information et de la communication

Comme le montre la figure 1.6, la schématisation de la théorie de la communication de Shannon est adaptée dans les sciences de l'information et de la communication afin d'intégrer plus explicitement la notion de "code" à travers les éléments "codeur" et "décodeur".

Comme le rappelle Escarpit [Escarpit, 1991], la théorie de la communication de Shannon est avant tout une théorie du rendement informationnel qui place l'énergie en

[13]Sans rentrer dans les détails, notons que la notion de "document" est interrogée par ce que Bachimont appelle "la raison computationnelle". Alors qu'autrefois un document était une trace d'informations palpable, manipulable, appréhendable sous le contrôle de celui qui le tient en main, un document désigne aussi aujourd'hui des objets numériques, virtuels non palpables et qui peuvent ne pas être contrôlés par celui qui les appréhende. Notamment, la "trace" peut disparaître aussitôt produite (*e.g.* les animations possibles dans les présentations informatisées), remettant en question la rémanence de l'écriture. Le lecteur intéressé peut se reporter aux travaux de Bachimont [Bachimont, 2004].

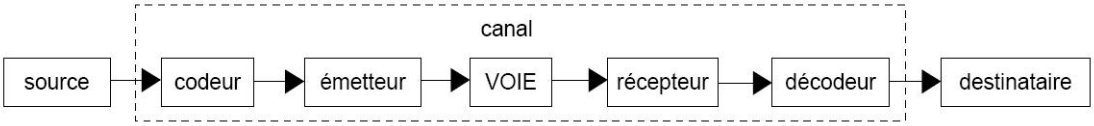

FIG. 1.6 – Diagramme schématique d'un système de communication en sciences de l'information et de la communication (extrait de [Escarpit, 1991])

tant que matérialisation de l'information au centre du processus de communication. S'il y a transfert d'information entre la source et le destinataire, il y a transfert d'énergie entre l'émetteur et le récepteur. La fonction du codeur est de transformer l'information en énergie en fonction du code utilisé. Le décodeur est chargé de l'opération inverse. La voie qui transporte l'énergie et le canal qui regroupe l'ensemble de la chaîne entre la source et le destinataire sont distingués. Dans le cas de la communication orale, source et destinataire sont des parties des centres nerveux ; le codeur est la zone de ces centres nerveux en mesure de transformer le message pour qu'il puisse être transmis par l'émetteur, en l'occurrence le système musculaire, l'appareil respiratoire et l'appareil de phonation ; la voie est l'air ambiant ; le décodeur est constitué de l'oreille externe, du tympan des osselets et de l'oreille interne ; le décodage se fait au niveau des centres nerveux.

Le principal reproche fait à Shannon est de se concentrer sur le canal, minimisant la source et ignorant les aspects psychologiques et sociologiques de l'avant et de l'après-canal. En particulier, Escarpit ne nie pas que le canal est une problématique importante qui a un impact sur le message transmis, mais il considère que la source ne peut être ignorée car elle conditionne aussi le message. De plus, les aspects psychologiques et sociologiques de l'avant-canal et de l'après-canal ont un impact sur "l'efficacité" de transmission des informations de la source au destinataire.

De plus, lathéorie de la communication n'est pas applicable à la communication humaine spontanée. En effet, pour cela, il faudrait que les moyens de communication répondent à des codes. Un code peut être défini comme une convention préalablement établie. Afin d'assurer le lien avec l'information, le code doit être connu tant du codeur et que du décodeur. Escarpit rappelle que, d'un point de vue pratique, un code est une liste de signes. L'approche saussurienne définit un signe comme étant composé de deux éléments, qui sont :

- le signifié : c'est la partie perceptible du signe ;
- le signifiant : c'est la partie imperceptible du signe, sa dimension abstraite.

Dans une approche mathématique de la communication telle que la théorie de la communication de Shannon, chaque signe du code doit répondre aux conditions suivantes :

- à un signifiant doit correspondre un seul signifié et réciproquement : il n'y a donc ni polysémie, ni homonymie ;
- les signifiants doivent être équiprobables.

Or les moyens de communication impliqués dans la communication humaine - et qui

pourraient correspondre à la notion de "modalité" - répondent rarement à ces conditions. Quatre moyens de communication encore utilisés aujourd'hui sont identifiés par Breton et Proulx [Breton et Proulx, 2002]. Il s'agit, dans l'ordre chronologique d'apparition :

- du geste : il peut être associé à des sons vocaux limités, comme c'était le cas durant la Préhistoire, ou accompagner l'oral comme c'est encore le cas aujourd'hui. Dans ce dernier cas, les gestes ne peuvent prétendre être un moyen de communication à eux-seuls ;
- de l'oral : en présentiel, les auteurs considèrent l'oralité comme l'accomplissement de la communication par le partage d'un même espace physique et perceptuel ;
- de l'image : elle inclut le documentaire et le reportage filmé mais non les pictogrammes qui relèvent de l'écrit. Rien ne prouve que l'image était une forme de communication à ses débuts avant l'invention des écritures ;
- de l'écriture : Breton et Proulx considèrent qu'elle vient de l'image et qu'elle s'est rapprochée du son avec l'alphabet.

Breton et Proulx excluent les communications chimiques et les communications auditives autres qu'orales (*e.g.* la musique). Concernant ces dernières, ils considèrent qu'elles ne sont plus, aujourd'hui, des moyens de communication à proprement parler.

Aucun des moyens de communication répertoriés, y compris ceux qui sont langagiers, ne respectent les deux critères formellement constitutifs d'un code. Preuve en est la difficulté à concevoir des grammaires en traitement automatique de la langue susceptible de couvrir au mieux la communication langagière humaine.

La notion de "canal" telle que définie par Shannon n'est valable que pour les communications non-médiatisées. Une communication non-médiatisée est une communication qui ne fait intervenir aucun *medium*, *i.e.* aucune technologie constituant une extension du corps et des sens humains et susceptible de diffuser la communication humaine [Escarpit, 1991]. Dans ce type de communication, Escarpit identifie deux émetteurs, le "système moteur" et le "système phonique" et trois canaux, le "canal auditif", le "canal visuel" et le "canal tactile"[14]. Il précise que pour chaque canal de communication, un langage spécifique, une forme de code[15], est nécessaire. Il reconnaît donc l'existence d'un langage "tactile", d'un langage "visuel" et d'un langage "auditif" ou "phonique". Chacun de ces langages a des caractéristiques qui lui sont propres et que nous considérons comme des propriétés/attributs/critères de modalités.

Dans le cas des communications médiatisées, une notion supplémentaire intervient, celle de "support". Breton et Proulx distinguent trois supports de communication en fonction de la catégorie des moyens de communication qu'ils expriment. Il s'agit des supports de l'oralité, des supports de l'image et des supports de l'écriture. Ils n'évoquent pas les supports du geste qu'ils semblent assimiler aux supports de l'image. Par ailleurs, la médiatisation rend possible le production d'un document [Escarpit, 1991]. Un document est "un objet informationnel visible ou touchable et doué d'une double indépendance par rapport au temps :

[14] Les termes entre guillemets sont employés par l'auteur

[15] Nous parlons de "une forme de code" car ce sont bien des langages qui ne remplissent pas les deux conditions qualifiant les codes, tel qu'explicité plus haut.

- la synchronie : indépendance interne du message qui n'est plus séquence linéaire d'événements, mais une juxtaposition multidirectionnelle de traces ;
- la stabilité : indépendance globale de l'objet informationnel qui n'est plus un événement inscrit dans l'écoulement du temps mais un support matériel de la trace qui peut être conservé, transporté, reproduit."

Si Escarpit considère que la synchronie est une convention établie pour être en accord avec le fonctionnement de l'esprit humain, la stabilité rend nécessaire la réintroduction du temps à la "lecture" du document par adjonction du mouvement. Dans le cas de l'écrit, ce mouvement est le balayage visuel nécessaire à la lecture. Le document est caractérisé par sa disponibilité permanente pour un balayage volontaire qui lui permet de tenir lieu de mémoire externe. Il se distingue du semi-document dont le balayage n'est pas actif et qui intègre une succession temporelle interne : c'est le cas des enregistrements vidéo et audio.

La mise en forme des messages est appelée "genre de communication" par Breton et Proulx. Ils en distinguent trois principaux, généralement combinés :

- le genre argumentatif : c'est la rhétorique, l'art de convaincre. Les arguments utilisés sont divers :
  - ceux qui s'appuient sur une autorité ;
  - ceux dits "de communauté", faisant appel à des présupposés communs ;
  - ceux dits "de cadrage", présentant le réel sous un certain point de vue ;
  - ceux qui fonctionnent par analogie ;
- le genre expressif : c'est celui qui permet l'extériorisation du ressenti, selon une vision généralement subjective ;
- le genre informatif : c'est le genre qui se veut le plus objectif.

Étude comparative et prise de position ***Nous rejoignons Escarpit sur sa critique de la théorie de la communication de Shannon selon laquelle la source, tout comme le canal, conditionne le message.*** De plus, à partir du moment où l'on admet que source et destinataire ne sont pas identiques, comme c'est le cas dans la communication humain-humain et humain-machine, nous pensons qu'il ne faut pas omettre l'impact du destinataire sur le contenu et l'expression du message. En effet, nous considérons que ***la source et le destinataire permettent une première sélection des canaux*** en éliminant ceux qui ne sont pas utilisables par recoupement des canaux que la source est en mesure d'utiliser et ceux auxquels le destinataire est en mesure d'accéder.

Comme nous l'avons précisé, ni Escarpit ni Breton et Proulx ne font référence à un canal ou à un support de communication chimique. La raison à cela est sans doute que l'humain ne contrôle pas les informations chimiques qu'il est susceptible de produire. Des travaux portent pourtant sur le Web odorant [Exhalia, 2004] et plus largement sur la diffusion d'odeurs dans des magasins ou dans des lieux publics. Notons, par ailleurs, que les canaux évoqués par Escarpit correspondent aux modes d'entrée de Bellik et que ses systèmes émetteurs correspondent aux modes de sortie de Bellik (*cf.* la section 1.2.5). De plus, les supports de communication identifiés par Breton et Proulx sont différents des trois moyens d'expression de Calvet (*cf.* la section 1.3.1) : les premiers négligent la sin-

gularité du geste et le dernier celle de l'écriture par rapport à l'image. ***La combinaison de ces deux approches nous amène à distinguer quatre moyens d'expression couramment utilisés par l'humain, à savoir l'oral, le geste, l'écrit et l'image.***

La notion de "code" définie par Escarpit représente parfaitement le besoin de la machine pour interpréter ou produire des messages de ou pour l'humain. Cette définition peut se substituer à celle de Nigay et Coutaz (*cf.* la section 1.2.3) pour ce qu'elles appellent un "langage d'interaction", confirmant la démarche centré machine de ces auteurs.

La caractérisation faite par Escarpit de la notion de "document" nous semble importante pour mieux décrire l'objet de nos travaux. Ces derniers ne portent pas sur les documents au sens où l'entend Escarpit : les informations présentées par les systèmes de communication humain-machine considérés ici ne sont pas conservées d'une intervention de l'utilisateur à la suivante. ***Nous ne tenons pas compte de la possibilité qu'a ou que pourrait avoir l'utilisateur de créer une trace (i.e. l'enregistrement ou l'impression) des informations fournies.*** Dans le cas de l'écrit, la mémoire externe considérée est une mémoire à court terme, comme peut l'être l'écrit sur un tableau.

Par ailleurs, ***la plupart des systèmes informatiques, et en particulier ceux sur lesquels nous travaillons, adoptent un genre informatif.*** Soulignons toutefois que d'autres travaux exploitent d'autres genres. Notamment, le genre expressif est de plus en plus d'actualité avec entre autres le recours aux avatars, le genre argumentatif est utilisé pour les systèmes éducatifs et pour les applications commerciales cherchant à encourager certains comportements (par exemple, alimentaires) chez les utilisateurs. ***Les systèmes informatiques qui permettent l'édition et la conception de documents ne relèvent donc pas d'une communication à proprement parler.***

1.3.3 Point de vue issu de l'éthologie

Corraze est l'auteur d'un ouvrage de vulgarisation portant sur les communications non verbales [Corraze, 1980] qui a été précurseur en France. Plus précisément, les travaux présentés portent sur la communication directe, *i.e.* d'individu à individu sans aucune médiatisation, et exclut les communications faisant intervenir un langage, au sens strict du terme. Ne sont pas exclues les communications entre animaux en général, mais nous ne les aborderons pas dans ce document.

Corraze considère comme communication tout comportement qui permet à un groupe de se constituer, de se définir, de s'exprimer et d'établir des relations en son sein. Un dialogue est donc une communication où le groupe est réduit à deux individus. De plus, l'auteur considère que la communication sous-entend une intention de modification des autres intervenants et correspond à l'expression de cette intention. Cette intention n'étant pas toujours consciente, Corraze recherche des caractères objectifs permettant d'appréhender une intention dans une communication donnée.

Pour Corraze, une communication non verbale est une communication entre deux individus particulière dans la mesure où elle ne fait appel à aucun modèle linguistique

humain, qu'il soit oralisé ou non. Il considère trois types de supports[16] chez l'humain utilisés pour exprimer de telles communications. Il s'agit :

- du corps et de ses mouvements : sont incluses les caractéristiques physiques – comme les postures – et physiologiques – dont les sécrétions chimiques du corps ;
- des artefacts liés au corps, comme les tatouages, les scarifications, les vêtements et autres accessoires, les coiffures, etc. ;
- du positionnement des individus dans l'espace.

Corraze aborde quatre types de communication dont le support est corporel, qui sont :

- les communications non verbales posturales et gestuelles : centrales dans la communication, elles se manifestent par des postures et par des gestes. Nous ne reprenons pas la revue de Corraze des travaux sur ce type de communications non verbales, primordial pour la conception d'avatars en particulier pour la dimension co-verbal d'accompagnement du langage, mais hors-contexte pour notre étude. Le lecteur est invité à se référer directement à l'ouvrage ;
- les communications non verbales faciales : elles transparaissent à travers le visage. Les études montrent que chez les humains, ce sont souvent les plus riches. Nonobstant la richesse et le nombre important des communications non verbales faciales, il nous semble qu'elles pourraient aussi être considérées comme des communications non verbales posturales et gestuelles particulières ;
- les communications non verbales chimiques : elles résultent de la production de phéromones. Corraze précise qu'on parle de "phéromone" quand trois conditions sont réunies : l'existence d'une substance chimique, sa diffusion dans l'environnement ambiant et le déclenchement d'un comportement particulier chez les individus en contact. Nous ne rentrons pas dans le détail de la typologie des phéromones, qui restent mal-connues chez l'humain ;
- les communications non verbales cutanées : elles passent par un contact cutané. Sont prises en compte les communications utilisant des pressions, des différences thermiques ainsi que les stimulations algiques.

La communication n'est possible que parce que les informations peuvent circuler dans un milieu physique perceptible. Corraze appelle ce milieu "canal". Il en cite quatre pour les communications non verbales, qui sont :

- les canaux visio-facial et visio-postural, sous-canaux du canal visuel : c'est le support des communications non verbales posturales, gestuelles et faciales, lorsque les communications non verbales sont perçues grâce à la vue ;
- le canal sonore ou auditif : le message est sonore et il est perçue grâce à l'ouïe. Corraze en parle peu, car, de son point de vue, il est constitué essentiellement voire exclusivement du sous-canal audio-vocal, propre aux communications verbales orales ;
- le canal chimique, regroupant ce qu'on pourrait appeler les sous-canaux gustativo-chimique et olfactivo-chimique : c'est, d'après Corraze, l'un des rares canaux permettant une persistance du message grâce à l'utilisation de phéromones. Nous

[16] Terme utilisé par l'auteur.

émettons une réserve sur ce point : en effet, Cozzare cite comme autre canal ayant cette capacité le canal visuel et ce grâce à l'écriture. Or, il compare ainsi deux choses différentes : un canal supportant une communication non-médiatisé, et la perception d'une communication médiatisée. De plus, nous avons déjà dit que la persistance de l'écriture pouvait être relative ;

- le canal cutané ou tactile : il permet les communications non verbales cutanées et comprend donc les sous-canaux prenant en compte des différences de pressions (ou tacts), des différences de température et des stimulations algiques. Le canal tactilo-postural ou canal haptique est un sous-canal du canal tactile : il correspond à un rapprochement physique entre interlocuteurs, jusqu'au contact corporel, prenant le relais du canal visio-postural. L'émission du message se fait alors par la posture et/ou le geste et la réception par perception tactile, voire musculaire. Le canal haptique est donc possible par la présence de pressions dues au geste ou à la posture.

Même s'il admet que l'existence d'un canal sous-entend nécessairement un codage, tel que le propose la théorie de la communication de Shannon, Corraze n'approuve pas cette conception de la communication. En effet, il considère qu'il est artificiel de distinguer source et destinataire : chaque intervenant est à la fois émetteur et récepteur ; chaque agent joue les deux rôles et s'adapte parallèlement à l'autre. Corraze fait une autre critique à lathéorie de la communication de Shannon : il rappelle qu'il y a réaction des interlocuteurs avant même qu'il y ait émission de messages.

Le codage, et bien-sûr le décodage, ne sont possibles que grâce à certaines conditions. En particulier, il faut coder le message de façon à ce qu'il puisse être décodé par le receveur. De plus, il faut qu'il y ait ... un code. Corraze souligne que, lors du codage, se pose le problème de ce qu'il faut coder : par analogie avec la traduction, il ne faut pas considérer un codage littéral, mais un codage du message dans son ensemble, donc contextuel. L'ordre de prise en compte des éléments est donc important. Corraze souligne d'ailleurs que la perte de l'ordre peut conduire à la perte du sens et/ou à l'apparition d'une nouvelle signification. Les mêmes problèmes se posent lors du décodage : quel est le code utilisé ? Qu'est-ce qu'il faut décoder ? Selon quel ordre et quelle granularité ?

Étude comparative et prise de position Ces problèmes liés à la notion de "code" se retrouvent au niveau du langage d'interaction dans la communication humain-machine (*cf.* la section 1.2.3). ***L'approche choisie peut influencer la notion de "modalité", en particulier sa granularité et par conséquent sa combinaison avec d'autres modalités.***

Les différents types de communications non verbales identifiés par Corraze peuvent constituer des modalités. Cependant, il ne s'agit pas du tout de celles qui nous intéressent dans notre étude de la communication humain-machine pour des systèmes d'information n'incluant pas des avatars. Notons que le terme "communication non verbale" admet d'autres définitions que celle de Corraze. Par exemple, Jakobson [Breton et Proulx, 2002] considère que les communications non verbales sont des communications co-verbales, *i.e.* une gestuelle d'accompagnement garantissant qu'il y a une communication en cours et donnant des informations supplémentaires ou complémentaires aux

messages oralisés. Ces dernières peuvent être utilisées dans des systèmes d'information sans avatar grâce à des déictiques visuels (surlignement, points clignotants, etc.). ***Nous excluons les communications non verbales telles que définies par Corraze car les systèmes qui nous intéressent s'appuient, en tant que systèmes d'information, sur la langue naturelle et leurs réponses ne peuvent pas n'inclure aucune information langagière.***

Même s'il ne le fait pas systématiquement, Corraze a tendance à définir les canaux de communication en fonction de ce que Clémente (*cf.* la section 1.2.9) appelle des "capacités d'action et de perception" : ainsi parle-t-il de canal visio-facial, de canal visio-postural, de canal audio-vocal et de canal tactilo-postural. Landragin (*cf.* la section 1.2.8) adopte la même démarche d'appellation que nous trouvons intéressante car elle évite de s'interroger sur le support de transmission des messages considérés, support appelé "mode" par Clémente, "canal" par Shannon (*cf.* la section 1.1) et "forme d'énergie" par Martin (*cf.* la section 1.2.4).

Par ailleurs, si nous comprenons les critiques de Corraze à l'encontre de la théorie de la communication de Shannon et que nous en partageons certaines, nous ne les approuvons pas toutes. ***Nous pensons effectivement important de prendre en compte l'intention et l'expression d'intention de l'interlocuteur, qui sont complètement ignorées dans la théorie de la communication de Shannon.*** Ces deux notions sont d'ailleurs présentes dans les approches informatiques qui mettent l'accent sur le comportement coopératif ou honnête du système (*cf.* chapitre 3). Nous partageons également la critique de Corraze faite à la théorie de la communication de Shannon concernant l'existence d'une réaction des interlocuteurs avant même qu'il y ait émission des messages. D'ailleurs, un message est généralement interprété en fonction d'autres éléments, d'autres événements qui l'accompagnent ou le précèdent, de façon interne à un agent ou externe : c'est ce qui peut être appelé "contexte". ***Dans un système informatique, il est nécessaire d'utiliser des éléments tels que l'historique du dialogue, ou les précédents dialogues pour pouvoir déduire le contexte de communication. Cette prise en compte doit ressortir lors de la production d'une réponse par le système.*** Par contre, nous considérons que ***la distinction entre source et destinataire a lieu d'être***, y compris dans les communications humaines. ***Elle permet notamment à un agent de se positionner dans la communication et, éventuellement d'évaluer l'existence d'un bruit de communication introduit au niveau du canal et de le distinguer de l'interprétation de son interlocuteur.*** De surcroît, rien n'empêche de considérer que deux processus source-destinataire co-existent en parallèle, chaque destinataire étant en même temps source et émettant des informations tout en en intégrant d'autres.

Nous étudions maintenant des approches plus expérimentales que sont les neurosciences et la neurobiologie, la psychologie, ainsi que les sciences cognitives. Ces approches ne s'intéressent pas directement à la communication, mais plutôt à la perception, base nécessaire à toute communication. En effet, c'est par la perception qu'un individu capte les informations venant d'autrui et de l'environnement.

1.3.4 Points de vue issus de la neurobiologie et des neurosciences

Alors que la neurobiologie se concentre sur l'étude du fonctionnement du système nerveux, les neurosciences englobent toutes les recherches qui intègrent le système nerveux, de la composition moléculaire des cellules nerveuses aux processus intelligents réalisés par les centres nerveux en passant par les pathologies d'ordre nerveux et l'interaction du système nerveux avec les différents organes. Nous avons choisi de rassembler ces deux domaines dans ce document car nous nous concentrons sur la perception des informations par le système nerveux.

Dans le langage courant, les notions de "perception" et de "sensation" sont couramment confondues [Purves *et al.*, 2003, Imbert, 2006]. D'un point de vue scientifique, la notion de "perception" correspond à une perception bas-niveau limitée à la détection de stimulations au niveau des récepteurs perceptifs. La notion de "sensibilité" ou "sens" regroupe, quant à elle, l'ensemble du processus perceptif, incluant la perception, la transmission des stimulations perçues - ou pas - vers les centres nerveux et l'interprétation à l'origine de l'identification ou de l'ignorance de la perception à bas niveau. Nous adoptons cette terminologie dans la suite de cette section.

La perception se fait via les récepteurs perceptifs, qui sont des terminaisons nerveuses. En fonction de leur sensibilité à certains types d'énergie, les récepteurs perçoivent des stimulations différentes. Ces stimulations sont exclusivement des variations d'énergie et sont couramment appelées "stimulus". Les informations invariantes et les variables trop faibles - en fonction du degré de sensibilité des récepteurs - (frôlement dans le cas du toucher, bruit trop faible, odeur ou goût trop subtil, etc.) sont ignorées, de même que les stimulations persistantes finissent par ne plus être relayées vers les centres nerveux. Il semble qu'un même récepteur peut être sensible à plus d'un type de stimulation, mais il est plus efficace pour un seul d'entre eux [Imbert, 2006] : on ne peut donc caractériser une modalité exclusivement par les récepteurs perceptifs qu'elle active. Trois types de stimulations sont possibles, les stimulations mécaniques (incluant les mouvements de l'air des ondes sonores), les stimulations photoniques et les stimulations chimiques. Ces trois types de stimulations sont perçues grâce à plusieurs sens. La classification aristotélicienne classique en cinq sens est rarement utilisée par les neuroscientifiques car elle confond la sensation, *i.e.* une expérience subjective généralement consciente, et la sensibilité, *i.e.* une réaction objective, mesurable expérimentalement, consciente ou inconsciente, à l'origine d'une sensation [Imbert, 2006]. Plusieurs classifications sont possibles, en fonction de ou des éléments caractéristiques choisis et du degré d'affinement admis pour les circuits nerveux spécifiques. En rassemblant et en mettant au même niveau les sensibilités identifiées dans [Purves *et al.*, 2003] et dans [Imbert, 2006], nous obtenons les dix sensibilités suivantes :

- la sensibilité visuelle [Purves *et al.*, 2003, Imbert, 2006] ;
- la sensibilité auditive [Purves *et al.*, 2003, Imbert, 2006] ;
- la sensibilité thermique ou algique [Purves *et al.*, 2003] plus largement extéroceptive (super-)cutanée [Imbert, 2006] ;
- la sensibilité olfactive [Purves *et al.*, 2003, Imbert, 2006] ;
- la sensibilité gustative [Purves *et al.*, 2003, Imbert, 2006] ;

- la sensibilité chimio-sensorielle trigéminale [Purves *et al.*, 2003] ;
- la sensibilité chémosensorielle ou voméronasale [Imbert, 2006] ;
- la sensibilité proprioceptive [Imbert, 2006], partie de la sensibilité mécanique dans [Purves *et al.*, 2003] ;
- la sensibilité intéroceptive [Imbert, 2006], partie de la sensibilité mécanique dans [Purves *et al.*, 2003] ;
- la sensibilité vestibulaire [Purves *et al.*, 2003, Imbert, 2006]

La sensibilité visuelle est basée sur les photons perçus par les récepteurs rétiniens. Elle permet d'identifier la position, la taille, la forme, les couleurs, la texture, le déplacement, la direction et la vitesse éventuels des objets. Le message nerveux correspondant au message photonique reçu au niveau de la rétine est déjà le résultat d'un traitement élevé, accentuant les informations les plus riches de la scène visuelle.

La sensibilité auditive est basée sur les ondes sonores dont les vibrations sont transformées et amplifiées de façon à mettre en mouvement le liquide qui remplit l'appareil auditif interne. Ce liquide va déplacer les cils situés à l'extrémité des cellules réceptrices auriculaires déclenchant une série de signaux transmis aux centres nerveux. Les ondes sonores ne sont donc pas transformées en un signal nerveux mais en une configuration particulière d'activité nerveuse. Tous les traitements liés à la reconnaissance et à la compréhension sont faits en aval.

La sensibilité thermique et algique, plus largement extéroceptive (super-)cutanée, dite "tactile", permet de percevoir les informations tactiles, thermiques ou algiques à la surface de la peau. Ce sont les mêmes récepteurs qui permettent la détection d'informations thermiques et algiques d'où leur regroupement dans une même sensibilité. La différence entre ces deux sensibilités se fait au niveau des cellules nerveuses transmettant la stimulation aux centres nerveux. En ce qui concerne la douleur, ont été identifiés des récepteurs cutanés qui n'émettent pas de signal vers les centres nerveux tant que l'intensité de stimulation ne représente pas une menace pour le tissu environnant. Ces récepteurs peuvent être stimulés par différents types de messages, à savoir la chaleur, la pression mécanique et les produits acides (d'où, sans doute, la sensibilité chimio-sensorielle trigéminale particulière prise en compte dans [Purves *et al.*, 2003]). L'intégration de la douleur est encore mal connue et deux théories coexistent. L'une défend une spécificité du système noniceptif et l'autre prône un traitement cérébral actif capable de corriger, amplifier ou supprimer la douleur à partir des messages perceptifs. En ce qui concerne les perceptions cutanées de façon générale, les études semblent montrer que les messages cutanés sont décomposés en constituants élémentaires par des récepteurs spécialisés et qu'il y a ensuite re-composition en un ensemble pertinent permettant d'identifier ce qui est "touché". Des récepteurs libres, *i.e.* non spécialisés, existeraient aussi.

La sensibilité olfactive est basée sur les molécules chimiques perçues par la muqueuse olfactive au niveau du nez. Les études semblent montrer qu'un seul odorant peut activer plusieurs types de récepteurs, que tous les récepteurs étudiés sont sensibles à plusieurs odorants différents mais qu'un odorant donné n'active qu'une unique combinaison de récepteurs. Le message transmis vers les centres nerveux correspond à l'identité moléculaire des produits chimiques perçus olfactivement, ainsi qu'à leur concentration.

L'interprétation et la réaction se feraient au niveau du cerveau, mais on ne sait pas encore dans quelles circonstances (on ne sait même pas si le codage des informations d'origine olfactive se fait spatialement ou temporellement).

La sensibilité gustative est basée sur les molécules chimiques perçues par la muqueuse gustative au niveau de la langue. Les substances sapides contenues dans les produits ingérés sont détectées. Leur nature et leur concentration sont transmises vers les centres nerveux. Il semblerait que les mêmes processus soient impliqués dans les sensibilités olfactive et gustative.

La sensibilité chimio-sensorielle trigéminale est basée sur les molécules chimiques perçues au niveau des récepteurs noniceptifs du visage, du cuir chevelu, de la cornée et des muqueuses des cavités buccales et nasales. Elle sert à la détection des substances chimiques considérées comme irritantes. La nature des substances et leurs concentrations sont transmises vers les centres nerveux.

La sensibilité chémosensorielle ou voméronasale est basée sur les molécules chimiques perçues par des récepteurs situés au niveau du nez mais distincts des récepteurs olfactifs. Elle permet la perception des phéromones et reste très peu connue chez l'humain qui n'a même pas conscience de son existence.

La sensibilité proprioceptive permet de percevoir les mouvements de son propre corps. Les récepteurs, situés au niveau des muscles et des articulations, mesurent la longueur et l'étirement des muscles ainsi que la tension exercée par les attaches tendineuses lors des contractions musculaires. Ceci permet de déterminer la position relative des membres dans l'espace de même que leurs vitesses et leurs directions en cas de mouvements réflexes ou volontaires. La sensibilité proprioceptive est aussi appelée "toucher actif", "perception kinesthésique","perception haptique", etc. Elle nécessite une exploration active, *i.e.* un mouvement. Il semblerait que le décodage des informations spatio-temporelles nécessaire à ce type d'information implique des interactions dynamiques entre signaux moteurs et signaux sensoriels, *i.e.* entre sortie et entrée du corps humain.

La sensibilité intéroceptive se rapproche de la sensibilité proprioceptive mais est spécifique aux viscères. Elle permet de détecter leurs états et les éventuelles douleurs.

La sensibilité vestibulaire s'appuie sur les récepteurs de la sensibilité auditive. Elle détecte l'accélération linéaire ou circulaire de la tête en se basant sur la gravité. Ceci lui permet de déterminer les mouvements absolus de la tête et du corps à l'origine de la sensation, ou non, d'équilibre. Cette sensibilité se place dans un cadre spatial absolu, alors que la sensibilité proprioceptive se situe dans un cadre spatial relatif [Imbert, 2006].

Comme le met en avant cette revue des différentes sensibilités, les récepteurs perceptifs sont des capteurs différents, plus ou moins spécialisés et plus ou moins complexes. De façon générale, lorsqu'une stimulation est perçue, sa nature, son intensité, sa durée et sa position spatiale sont transmises vers les centres nerveux. La transmission des messages perçus au niveau de récepteurs jusqu'aux centres nerveux est du ressort des cellules nerveuses, couramment appelées "neurones". À cette fin, elles utilisent des signaux électriques générés par échanges ioniques. Le passage entre deux cellules nerveuses est appelé "synapse". Ces synapses jouent un rôle primordial dans les sensibilités car ce sont elles qui sont à l'origine de transformations importantes sur les messages

nerveux qu'elles transmettent : certaines vont inhiber et d'autres amplifier les messages nerveux transmis. Les messages arrivent donc considérablement modifiés au niveau des centres nerveux.

Étude comparative et prise de position Si les récepteurs mobilisés renvoient à la notion de "dispositif physique" mise en avant dans notre synthèse sur la notion de "modalité" en informatique (*cf.* la section 1.2.11), celle de "type de stimulation" n'a pas été évoqué dans cette synthèse. Cette notion correspond au canal de Shannon (*cf.* la section 1.1), à la forme d'énergie de Martin (*cf.* la section 1.2.4) et au mode de Clémente (*cf.* la section 1.2.9). ***Plutôt que de rajouter la notion de "type de stimulation"***, notion purement matérielle s'il en est, aux notions auxquelles la "modalité" peut s'apparenter, ***nous préférons l'omettre, considérant, à la suite de Corraze (cf. la section 1.3.3) et Landragin (cf. la section 1.2.8), qu'elle est déterminée par un couplage de la capacité d'action de l'émetteur et de la capacité de perception du récepteur.*** Dans le cas de la communication machine vers humain, ***cela revient à caractériser la modalité*** non pas par rapport au type de transmission du message et au récepteur sensoriel humain impliqué, mais par rapport au dispositif physique impliqué et à la capacité humaine de perception mobilisée ou visée. ***Par rapport à la démarche neuroscientifique, une différence de point de vue est opérée, où ce n'est pas l'humain qui est au centre de l'étude mais le couple <machine ; humain>.***

Les sensibilités identifiées par les neuroscientifiques recouvrent un spectre plus large que les moyens de communication reconnus en sciences de l'information et de la communication. Ceci est dû à la motivation-même de la neurobiologie qui traite de la perception au sens large et ne se concentre pas sur la perception impliquée dans la communication. S'il est important, voire nécessaire, de tenir compte de toutes les sensibilités humaines pour certains travaux en informatique, telles que la réalité augmentée ou la réalité virtuelle, ***les systèmes d'information non immersifs peuvent - nous ne pouvons affirmer qu'ils doivent - se limiter aux sensibilités permettant l'interprétation consciente des informations par l'utilisateur***. En effet, il nous semble que l'appréhension sémantique des informations est déjà suffisamment délicate pour chercher à traiter en plus l'appréhension "instinctive" permise par les sensibilités dont la conscience est plus floue et qui sont encore plus mal connues. ***Les sensibilités ayant, à l'heure actuelle, une portée sémantique avérée sont les sensibilités visuelle, auditive et tactile (en particulier, cas du braille).***

1.3.5 Points de vue issus de la psychologie

Avant d'être la base de la communication, la perception est la base de la construction d'une appréhension stable de l'environnement par les individus. Cette appréhension est primordiale car elle constitue un cadre de référence pour leurs actions [Reuchlin, 1977]. Plusieurs approches du fonctionnement perceptif cohabitent [Delorme, 1994]. Ainsi la perception est-elle considérée tour à tour comme une prise de conscience sensorielle consciente, une pré-conscience du monde environnant permettant d'anticiper l'identifi-

cation d'un changement, impliquant ou non un effet sur l'individu et/ou intégrant par définition un traitement d'information. Bien sûr, les appréhensions contraires ont existé ou existent encore. La tendance actuelle générale place la physiologie au centre des études en psychologie et en particulier en psychologie de la perception. De plus, comme dans les neurosciences et en neurobiologie, la pensée la plus répandue aujourd'hui veut que la perception ne se limite pas au prélèvement d'informations et intègre bel et bien les traitements de ces informations. La distinction entre perception et traitement semble plus floue qu'elle n'a jamais été.

Le cheminement des informations des récepteurs sensoriels aux centres nerveux sont dans l'ensemble bien connus [Delorme, 1994, Reuchlin, 1977]. Outre des voies spécifiques, les mêmes mécanismes de sélection existeraient pour tous les sens. Les études physiologiques ont permis de montrer que la première sélection des informations se fait au niveau des récepteurs sensoriels. En effet, il existe des seuils en deça desquels les variations qui tiennent lieu de stimulations ne sont pas transmises aux centres nerveux. [Reuchlin, 1977] apporte quelques précisions supplémentaires sur la sélection qui se continue lors de la propagation des informations jusqu'aux centres nerveux. Comme cela a déjà été dit dans la section précédente, la transmission des variations se fait par succession de neurones. Or l'information n'est pas transmise telle quelle d'un neurone à l'autre. Entre deux neurones, au niveau des synapses, l'information transmise par un neurone donné est combinée aux informations provenant des autres neurones afférents et re-codées en conséquence. Chaque synapse fonctionne donc comme un pallier de sélection avec, semble-t-il, une position hiérarchique donnée. Par conséquent, des informations peuvent disparaître alors que d'autres sont amplifiées. Arrivées aux centres nerveux, les informations ont donc déjà été remaniées, *i.e.* leur traitement a déjà commencé.

Cette sélection des informations à plusieurs niveaux amène Reuchlin [Reuchlin, 1977] à considérer que le rejet ou non d'un signal perçu est "décidé", au moins au niveau nerveux, par le sujet. Ceci est d'autant plus vrai lorsqu'une stimulation est invariante dans le temps. Le sujet est donc loin d'être inactif même si la sélection - que ce soit des signaux ou des hypothèses perceptives possibles permettant la construction perceptive - est en grande partie inconsciente. Par conséquent, la limite entre perception et action est de plus en plus floue.

Le traitement à proprement parler des informations est moins maîtrisé. Ainsi l'impact de l'inhibition et de l'amplification des messages par les synapses n'est-il pas complètement connu. Il existe toutefois des différences en fonction des sensibilités. Sans remettre en question la classification aristotélicienne des cinq sens, les psychologues se sont surtout intéressés à la vue - au point de s'en servir comme base d'appréhension de la perception en général - puis à l'audition et plus récemment à la proprioception. Les mécanismes de traitement spécifique sont donc plus connus pour la vision que pour les autres sens. Certains mécanismes seraient toutefois valables pour tous les sens. Notamment, il semble qu'une même stimulation est transmise via plusieurs voies jusqu'aux centres nerveux [Delorme, 1994]. En effet, pour chaque sens, il y aurait plusieurs "cerveaux perceptifs" spécifiques, composés d'une variété de neurones spécialisés. Ces neurones spécialisés réagiraient sélectivement à certaines dimensions de l'environnement

perçu. En résulteraient, pour chaque sens, plusieurs centres de traitement. Chacun de ces centres de traitement aurait une fonction différente par rapport à la perception, *i.e.* ne serait pas influencés par les mêmes paramètres. Une même stimulation emprunterait plusieurs voies pour atteindre ces différents centres nerveux de traitement. Une "modalité" ne peut donc être exclusivement définie par la voie qu'elle emprunte, tout comme elle ne peut être exclusivement définie par les récepteurs qu'elle active (*cf.* la section précédente).

Parmi les mécanismes perceptifs, certains semblent innés [Reuchlin, 1977]. Il s'agit notamment des lois d'organisation perceptives étudiées par les gestaltistes pour la perception visuelle. L'hypothèse sous-tendue est que la perception visuelle se fait selon une organisation saillante d'éléments perçus et non pas par association d'éléments. Ceci est à l'origine de différentes définitions pour la notion de "stimulation" : couramment considérée comme un signal physique particulier doté d'une nature, d'une intensité, d'une durée, etc., elle est parfois définie comme un patron de signaux. Les mécanismes perceptifs non-innés semblent dépendre, entre autres, de l'expérience antérieure du sujet, de ses besoins, de sa situation, de son contexte individuel et social, etc. mais aussi des décalages observés, consciemment ou inconsciemment, entre effet attendu et effet produit [Reuchlin, 1977]. Ces dispositions de perception semblent intervenir dès les récepteurs [Imbert, 2006].

Certains travaux en psychologie sont centrés sur l'attention. La psychologie identifie deux types d'attention [Coquery, 1994] : d'une part, l'attention spontanée, involontaire, passive résultant d'un apprentissage ou d'une correspondance stable en stimulus et réponse ; d'autre part, l'attention volontaire et active. Dans les deux cas, l'attention joue un rôle sélectif sur la perception ainsi que sur la réaction du sujet.

La psychologie ne se contente pas d'étudier la perception isolément. Elle s'interroge aussi sur l'impact de la perception sur le comportement. Deux principales approches existent [Paillard, 1994] : celle défendant une boucle stimulus-réponse, et celle défendant une boucle perception-action. Dans le cas de la boucle stimulus-réponse, les entrées sensorielles provoquent des sorties motrices qui correspondent à des réactions (sans traitement interne de plus haut niveau, ou du moins, sans sa prise en compte). C'est une approche issue du béhaviorisme. Dans le cas d'une boucle perception-action, il y a des opérations mentales entre la perception de la stimulation et le déclenchement de l'action. C'est l'approche cognitive, laissant une large place à la notion de "représentation" dans l'intégration de la perception. En effet, les représentations que se fait le sujet du monde environnant, des entités existantes et de lui-même sont à la base même de ses capacités d'action : capable d'anticiper l'effet de ses mouvements, de ses actions, un individu adopte une démarche prédictive et non réactive. Certains, tel que Paillard [Paillard, 1994], défendent l'idée selon laquelle ces deux boucles co-existent. Pour eux, la première correspond à des réflexes innés ou acquis où les informations sont traitées à un niveau sensorimoteur, la deuxième permet de traiter de nouvelles situations, d'y réagir et de prévoir le résultat des comportements adoptés. Dans une telle conception, l'acte moteur n'est pas nécessairement déduit de façon passive. Il est le résultat d'un échange entre sujet et environnement et permet de confronter la représentation interne de l'environnement aux réponses sensorielles obtenues. S'en suit un enrichissement des

connaissances tant sur l'environnement que sur les réactions à appliquer. Ces boucles d'intégration sensorielle à deux niveaux pourraient parfaitement avoir leur place dans les systèmes informatiques : pourraient se cotôyer une boucle entrée-sortie sans intervention de traitements avancés (comme ça peut être le cas pour les clics sur des liens hypertextes ou l'énonciation d'un mot-clé indiqué par le système informatique) et une boucle interprétation-génération dans le cas de demandes plus complexes de l'utilisateur.

Étude comparative et prise de position Depuis la conception centrée utilisateur, ***les études en psychologie sont prises en compte dans les travaux en informatique sur la communication humain-machine par le biais de l'ergonomie***. Les travaux dans ce domaine reprennent donc la plupart du temps la terminologie utilisée par les psychologues. C'est pourquoi nous ne les évoquons pas explicitement dans ce chapitre, ni dans le suivant.

Les points de vue sur la perception en psychologie ne nous fournissent pas directement une définition de la notion de "modalité". Néanmoins, ces études mettent en exergue plusieurs aspects à approfondir dans le cadre de la communication humain-machine. Tout d'abord, la limite entre perception et action, soulignée par l'emploi des termes "mécanismes perceptifs" [Reuchlin, 1977] et "processus perceptifs" [Delorme, 1994], doit être prise en compte dans la production du comportement des systèmes. L'utilisateur est trop souvent considéré comme passif aux informations présentées par les machines, comme en témoigne l'expression de "*recognition senses*" utilisée Frohlich (*cf.* la section 1.2.1). ***La détermination du comportement d'un système doit tenir compte de l'impact des choix de présentation sur l'utilisateur, mis en évidence par un nombre croissant de travaux en ergonomie [Karsenty, 2006, Le Bigot* et al.*, 2006, Fréard* et al.*, 2007], et sur l'interprétation que ce dernier concernant ses capacités et las capacités de la machine [Karsenty, 2000].***

Ensuite, une partie importante des mécanismes perceptifs, plus précisément les mécanismes perceptifs non-innés, seraient très dépendants des caractéristiques des sujets et du contexte. ***Ceci confirme, si besoin est, la place à accorder au contexte - au sens large - dans le choix des réactions des systèmes et de leur présentation à l'utilisateur.***

Par ailleurs, la notion d'"attention" est importante dans la communication humain-machine car l'objectif d'un système présentant des informations est que ces informations soient perçues par l'utilisateur. Plus précisément, ces informations doivent être identifiées, voire mémorisées. ***Il faut donc réussir à attirer ou à mobiliser l'attention de l'utilisateur sur les informations jugées les plus importantes ou les plus pertinentes.*** De plus, étant donné que l'attention prépare l'action, ***elle peut être utilisée pour inciter l'utilisateur à réagir d'une certaine façon***, par exemple à répondre à la machine en utilisant une capacité d'action plutôt qu'une autre. Par ailleurs, les travaux sur l'attention pourraient peut-être permettre de dégager des caractéristiques de "saillance physique", pour reprendre le terme de Landragin [Landragin, 2004a]. Le choix d'une modalité de présentation plutôt qu'une autre pourrait ainsi être fait en fonction de l'information sur laquelle le système veut attirer l'attention de l'utilisateur.

Malheureusement, ***les processus perceptifs tels qu'ils sont expliqués/exprimés à l'heure actuelle en psychologie et les connaissances sur le fonctionnement de l'attention ne nous permettent pas de comprendre la façon dont une information perçue au niveau des récepteurs sensoriels est identifiée, mémorisée et utilisée.***

Pour terminer, les études présentées soulignent l'existence de ***deux boucles d'intégration sensorielle que nous retrouvons en informatique*** dans le modèle de l'interaction instrumentale [Beaudoin-Lafon, 2004]. Y sont distinguées les boucles d'action-réaction (*action-feedback*) et de commande-réponse (*command-response*), qui se situent à deux niveaux de traitement différent. ***Le niveau du traitement sémantique est un autre niveau qui n'est généralement pas considéré ni en psychologie, ni dans les systèmes-outils (system-as-tool [Beaudoin-Lafon, 2004]). Il est, par contre, central dans les sciences cognitives et pour les systèmes-partenaires (system-as-partner [Beaudoin-Lafon, 2004]).***

1.3.6 Points de vue issus des sciences cognitives

Par définition, les sciences cognitives s'intéressent au cerveau. La perception et son intégration par les centres nerveux a une place importante dans ces travaux. Si une partie d'entre eux s'appuient sur des études et des résultats expérimentaux, une autre partie relève de la philosophie et de l'appréhension du rôle des centres nerveux. Pour certains, la perception est subjective : c'est une forme particulière d'expérience où sont donc distingués réel et imaginaire, corps et cerveau, perception et abstraction [Clementz, 2003]. Pour d'autres, ces distinctions sont fausses et "le cerveau est incarné dans le corps" [Berthoz, 1997]. Cette deuxième approche remet en question la distinction entre perception et action, la rendant moins tranchée [Reuchlin, 1977]. Pour mieux comprendre cette approche, nous nous penchons sur la philosophie de la perception conçue par Berthoz [Berthoz, 1997]. Notons que, pour Berthoz, le terme "perception" ne se limite pas à un bas niveau et intègre des mécanismes cognitifs : il parle de "perception" quand d'autres parlent de "sensibilité".

Berthoz défend l'idée que le système nerveux n'est pas uniquement réactif, mais qu'il est proactif. C'est pour lui un simulateur, un "émulateur de la réalité" dont l'objectif est de prédire et d'anticiper les conséquences des actions produites et des actions des sujets environnants. La perception ne se résume donc pas au relevé des stimulations sensorielles et à leur interprétation. Elle est conditionnée par l'action au même titre qu'elle la provoque et elle anticipe ses impacts. En effet, comme nous l'avons déjà évoqués, les récepteurs perceptifs ne mesurent pas des données physiques observées, mais des variations de ces données physiques. Berthoz considère qu'en mesurant les variations d'une grandeur, leurs valeurs futures peuvent être prédites. La perception est - au delà de l'interprétation - une prise de décision par la sélection d'hypothèses sur les mondes interne et externe. Nous retrouvons la conception de la perception de Reuchlin présentée précédemment, ainsi que la boucle perception-action citée par Paillard.

Berthoz lie perception et action, *i.e.* entrée et sortie des agents humains, en affirmant que ce lien est à double sens : si la perception est à l'origine de l'action, l'action permet

aussi la perception. Dans cette boucle, Berthoz considère que le point central est l'action, *i.e.* la sortie, alors que l'étude de cette boucle a longtemps été centrée sur la perception, *i.e.* l'entrée, notamment en psychologie et en neurobiologie. On peut parler de théorie motrice de la perception.

Cette théorie motrice de la perception admet deux types d'actes, les actes réflexes - résultant de la boucle stimulus-réponse évoquée par Paillard - et les actes perceptifs - résultant de la boucle perception-action. Les actes perceptifs sont considérés comme étant adaptatifs (*i.e.* ils s'adaptant aux stimulations initiales) et prédictifs (*i.e.* ils s'adaptent aussi aux stimulations ultérieures supposées). Ils sont aussi anticipatifs. Preuve en est les mouvements semblant prendre en compte une estimation basée sur le passé immédiat, comme c'est le cas des mouvements de déplacement vers un objet ou un individu lui-même en mouvement. Cette anticipation laisse entrevoir l'existence de mémoires neuronales qui maintiennent temporairement les perceptions. Berthoz rapproche les anticipations et les prédictions qui sous-tendent la perception des concepts husserliens de "protentions" et de "rétentions" : de façon succinte, les protentions correspondent aux attentes par rapport aux perceptions à venir ; elles sont générées à partir des rétentions qui correspondent à la mémorisation des perceptions passées.

La place de l'action dans la perception amène Berthoz à ajouter le sens du mouvement - ou kinesthésie - aux cinq sens aristotéliciens classiques. Ce sens correspond à la proprioception dont il est question en neurobiologie et en neurosciences. D'après Berthoz, la particularité de ce sens, et la principale raison pour laquelle il a été si longtemps ignoré, est de mobiliser plusieurs capteurs répartis dans le corps, qui plus est dissimulés et n'émergeant pas à sa surface. À ces capteurs s'ajoutent indirectement la vision (qui mesure le glissement de l'image du monde sur la rétine) et le toucher (qui mesure la pression exercée, les variations de pression, la durée des pressions, le contact, la température et la douleur).

Étude comparative et prise de position La théorie motrice de la perception est applicable à la communication humain-machine, en prenant compte l'importance déterminante des sorties des systèmes pour les entrées à venir, *i.e.* pour les réactions de l'utilisateur. Comme nous l'avons souligné dans la section précédente, les études ergonomiques confirment cet impact [Karsenty, 2006, Le Bigot *et al.*, 2006, Fréard *et al.*, 2007]. ***Il ne s'agit plus alors de concevoir des systèmes puissants en termes d'interprétation des entrées venant des utilisateurs, mais d'orienter ces entrées par un travail centré sur les sorties vers les utilisateurs.*** La machine doit être pro-active, tout comme l'est l'utilisateur, en adaptant par exemple ses réactions en fonction des réactions de l'utilisateur à ses retours antérieurs. ***Ceci implique un fonctionnement en anticipation, ou en protention***, basé sur une mémorisation des échanges passés, par le maintien d'un historique de dialogue par exemple.

Berthoz ajoute le sens du mouvement aux cinq sens aristotéliciens classiques. Sa particularité est de mobiliser des capteurs répartis dans tout le corps, mais aussi la sensibilité visuelle et la sensibilité tactile. En se référant aux travaux en neurobiologie et en neurosciences, il convient d'inclure également la sensibilité auditive utile au positionnement des éléments dans l'espace environnant et qui prend pleinement le relais

de la sensibilité visuelle dans les situations où la vue fait défaut (en particulier, de nuit et pour les aveugles). Influencés sans doute par la distinction encore nécessaire pour les systèmes informatiques entre sorties et entrées, ***le sens du mouvement est plus, pour nous, une capacité d'action qu'une capacité de perception***. Pour autant, nous ne nions pas - au contraire ! - les ***liens forts entre perception et action, entrée et sortie. C'est la revendication de la force de ces liens qui motive nos travaux sur la sortie des systèmes informatiques***.

Berthoz souligne que ***l'action a longtemps été négligée dans l'étude du fonctionnement humain au profit de la perception***. Ceci ressort également dans l'étude des autres travaux en sciences expérimentales (*cf.* les sections précédentes). ***De la même façon, la sortie reste le parent pauvre de la communication humain-machine*** et ce pour plusieurs raisons. L'une d'elle est que les machines doivent être en mesure d'extraire les demandes des utilisateurs pour pouvoir y répondre. Or les utilisateurs dits "novices" ne sont pas toujours à même de s'adapter aux capacités de perception des machines. À l'opposé, il est couramment admis que les utilisateurs peuvent très bien "faire avec" les capacités d'expression des systèmes informatiques. C'est pourtant oublier que les capacités d'expression d'une machine conditionnent les capacités d'expression de son utilisateur et, par conséquent, les entrées suivantes de la machine. Une sortie "bien construire" permet d'anticiper les entrées, et ceci est valable tant pour l'humain que par la machine. Berthoz cite McKay pour expliquer la séparation de l'étude de la perception et de l'action chez l'humain. Pour cet auteur, cette séparation est due à l'existence d'une "subordination philosophique de l'action à la perception, subordination fonctionnelle car on considère que c'est pas la seule perception que la connaissance est acquise (thèse empiriste), subordination dans le temps parce que la perception est considérée comme un précurseur nécessaire à l'action (paléobéhaviorisme), et subordination dans l'ordre des valeurs parce que la vie contemplative est conçue comme supérieure à l'action (Platon)." Un parallèle peut être fait avec le traitement des entrées et des sorties des systèmes : ***il est généralement considéré que les entrées sont plus riches en information et qu'elles conditionnent la réaction du système, alors qu'on peut inverser le problème en considérant que la sortie du système va impacter l'entrée suivante de l'utilisateur et que la construction de la sortie est la base de comparaison entre les réactions de l'utilisateur et les réactions attendues par le système***. Sans compter qu'on a longtemps considéré qu'il est plus facile de générer une sortie que de comprendre l'entrée du système et que l'adaptation se manifeste au niveau de l'ouverture de l'action de l'utilisateur plutôt qu'au niveau de l'adaptation de la sortie en termes de fond et de forme. Enfin, si à la suite de Berthoz, on considère que la dichotomie classique entre perception et abstraction est fausse, on ne peut admettre qu'il faille complètement séparer présentation ou perception (*i.e.* forme) et abstraction (*i.e.* contenu) dans les systèmes informatiques. ***Ceci nous amène, dans la suite de ce document, à défendre l'idée de prise en compte de la forme pour choisir le fond dans la production des réactions des machines.***

1.3.7 Synthèse : modalité dans les autres domaines

Si les sciences humaines se sont intéressées à la communication, les sciences plus expérimentales ont surtout étudié la perception et l'impact de cette dernière sur le comportement en général et non sur la communication en particulier. Dans les deux cas, le terme "modalité" est employé dans son sens littéral, à savoir une forme particulière (de perception, communication, transmission, etc.). Cette étude des travaux portant exclusivement sur l'humain, *i.e.* hors d'un contexte incluant des machines, nous conduit au positionnement suivant.

La notion de "modalité" fait intervenir les capacités d'action et les capacités de perception utilisées pour produire et percevoir l'information présentée. Cette caractérisation évite de devoir définir le ou les supports de transmission utilisés qui sont sous-entendus par les capacités d'action et de perception mobilisées. De plus, cette caractérisation nous permet de laisser de côté la notion de "document". Cette notion renvoie à la possibilité que pourrait avoir l'utilisateur de conserver les informations présentées via des supports physiques ou numériques. Nous considérons que les informations données sont disponibles uniquement jusqu'à la prochaine réaction de la machine (provoquée par une intervention de l'utilisateur). Ces informations sont présentées de façon à être perçues grâce à deux sensibilités, la sensibilité visuelle et la sensibilité auditive : même si elle pourrait trouver tout son sens dans l'utilisation de l'écriture braille, la sensibilité tactile n'est pas courante dans les équipements grand-public sur lesquels nous nous concentrons. L'emploi du terme "sensibilité", ici et dans le reste du manuscrit, est voulu : à la suite des approches en neurobiologie et en neurosciences, nous distinguons sensibilité et perception. Lorsque nous parlons de "capacité de perception", nous ne sous-entendons pas les traitements d'interprétation qui ne sont que partiellement connus. Toutefois, même partielles, ces connaissances doivent être prises en compte dans la détermination de la réaction des systèmes car elles sont à l'origine des comportements des utilisateurs. Elles pourraient faire partie des propriétés/attributs/critères de définition des modalités. Toutefois, parce que leurs tenants et leurs aboutissants ne sont pas parfaitement maîtrisés, il nous semble délicat de chercher à les formaliser. Nous les considérerons donc comme étant implicites dans le choix d'une modalité mobilisant une capacité de perception particulière.

Par ailleurs, profitant du recul des sciences humaines et expérimentales sur l'intégration sensorielle, nous considérons que (1) le support doit être considéré au même niveau que le message ; (2) entrée et sortie d'un système ne peuvent être complètement dissociées et, travaillant sur la sortie, nous devons prendre en compte l'impact de la sortie sur l'entrée suivante, *i.e.* sur l'utilisateur, sur son intégration du message présenté et sur son comportement en conséquence ; (3) le contexte, non seulement physique mais aussi lié à l'historique de la communication, doit avoir une place centrale dans la détermination du comportement des systèmes.

1.4 Conclusion : notre terminologie

Notre exploration de la notion de "modalité" tant en informatique qu'en sciences humaines et expérimentales guide notre appréhension de la communication humain-machine. En particulier, partant du constat que l'informatique aborde la communication comme l'ont fait par le passé les sciences expérimentales, privilégiant l'entrée aux dépens de la sortie, nous pensons qu'il est nécessaire aujourd'hui de se pencher plus spécifiquement sur la sortie des systèmes informatiques. Toutefois, parce que la communication est avant tout une interaction, il est nécessaire de ne pas dissocier sortie et en entrée. Par conséquent, nous appuyant sur des travaux en histoire (*cf.* la section 1.3.1), en sciences de l'information et de la communication (*cf.* la section 1.3.2), en psychologie et en ergonomie (*cf.* la section 1.3.5) et en sciences cognitives (*cf.* la section 1.3.6), nous considérons que :

1. la sortie doit préparer et anticiper l'entrée suivante produite par l'utilisateur - notamment en garantissant ses capacités d'action ;
2. les disponibilités des modalités doivent non seulement être prises en compte (*cf.* la section 1.2.7) mais doivent constituer des contraintes de présentation qui influencent le comportement du système.

Ces différents aspects sont à l'origine de nos choix terminologiques qui sont les suivants.

Notre exploration de l'espace terminologique de la notion de "modalité" en informatique nous conduit à identifier trois notions auxquelles elle peut s'apparenter : celle de "dispositif physique", celle de "capacité de perception" et celle de "langage d'interaction" (*cf.* la section 1.2.11). Ces trois notions sont utilisées isolément ou par couples pour définir la notion de "modalité" en informatique. Notre exploration de l'espace terminologique de la notion de "modalité" en sciences humaines et en sciences expérimentales nous amène à ne pas négliger, comme c'est trop souvent le cas, les notions de "capacités de perception" et de "capacités d'action" de l'utilisateur. Ces deux notions sont directement impliquées dans l'intégration d'une modalité informatique donnée et peuvent contribuer à la caractérisation d'une modalité. De ces deux explorations, nous extrayons deux dimensions qui doivent intervenir de façon équivalente dans la notion de "modalité" et qui sont :

- les capacités de perception et les capacités d'action humaines ;
- les couples <dispositif physique ; langage d'interaction> où le langage d'interaction spécifie la structuration des informations et le dispositif physique présente ou recueille ces informations dans les systèmes informatiques. Ce dernier est le pendant des capacités humaines de perception et d'action pour les machines.

À la suite de Frohlich (*cf.* la section 1.2.1), nous distinguons entrée et sortie. Les capacités de perception humaines sont intégrées dans les modalités sensorielles. Elles interviennent en sortie des systèmes informatiques. Les capacités d'action humaines sont intégrées dans les modalités motrices. Elles interviennent en entrée des systèmes informatiques. Nos travaux se focalisant sur la sortie des systèmes, nous privilégions l'emploi du terme "modalités sensorielles" même si notre raisonnement est valable pour

les modalités motrices. Les couples <dispositif physique ; langage d'interaction> correspondent à la notion de "modalité" du point de vue système. Nous sommes tentés de reprendre le terme de Bernsen : "modalité représentationnelle". Là encore, travaillant sur la sortie des systèmes, nous privilégions, dans la suite de nos propos, les modalités représentationnelles en sortie même si notre raisonnement s'applique aux modalités représentationnelles en entrée.

De façon à lier explicitement ces deux définitions de la notion de "modalité", nous faisons intervenir une troisième définition. Une modalité est aussi un couple <modalité représentationnelle, modalité sensorielle> où les deux éléments sont liés par une relation primaire au sens de Rousseau. De cette façon, la modalité sensorielle visée par la modalité représentationnelle considérée est explicitée et le support de communication/transport/transmission est sous-entendu. Cette dernière définition permet de combiner les approches en informatique d'une part et en sciences humaines et expérimentales d'autre part, que nous avons décrites dans ce premier chapitre. C'est donc une définition unificatrice issue directement de notre étude dont la particularité est de définir la notion de modalité non par rapport aux systèmes informatiques comme c'est généralement le cas en informatique, ni par rapport à l'humain comme c'est généralement le cas dans les sciences humaines et expérimentales, mais par rapport au couple <machine ; humain>. Nous nous inscrivons ainsi dans la lignée de Landragin, des travaux en sciences de l'information et de la communication et des travaux en neurobiologie et en neurosciences (*cf.* les sections 1.2.8, 1.3.2 et 1.3.4).

Dans la suite du manuscrit, nous utiliserons le terme "modalité" indifféremment pour ces trois définitions. Nous focalisant sur la sortie et nous appuyant sur le constat que la capacité de perception impliquée conditionne les dispositifs physiques utilisables, nous considérerons que la référence à un sens de perception indiquera implicitement les dispositifs physiques mobilisés : par exemple, une présentation visuelle sous-entend une présentation perceptible visuellement qui mobilise donc des dispositifs physiques spécifiques, tel que l'écran, susceptibles de produire une présentation visuelle. En cela, nous rejoignons notamment Martin (*cf.* la section 1.2.4) et Rousseau (*cf.* la section 1.2.10). Les modalités sensorielles que nous retenons comme susceptibles de transmettre des informations à portée sémantique sont les modalités visuelle, auditive et tactile. Toutefois, nous privilégions les deux premières dans la suite du mémoire car nos travaux portent sur les systèmes d'information grand-public qui ne sont que rarement équipés, à l'heure actuelle, pour permettre une présentation tactile autre que le braille.

Pour terminer, soulignons que notre exploration de la notion de "modalité" en sciences humaines et expérimentales participe également à notre appréhension de la communication humain-machine en ce qui concerne :

1. la prise en compte du contexte pour la succession des échanges (*cf.* la section 1.3.3) ;
2. la limitation de nos propos aux systèmes d'information sans tenir compte d'un éventuel enregistrement des informations pour un accès ultérieur (*cf.* la section 1.3.2) ;
3. l'importance de l'intention du système pour la détermination de son comportement

(*cf.* la section 1.3.3), *i.e.* de sa réaction et de sa présentation.

L'impact de ces différents aspects sur la communication humain-machine telle que nous l'appréhendons sur les systèmes que nous cherchons à concevoir sera détaillé dans le chapitre 3 et dans le chapitre 5.

Chapitre 2

Combinaison de modalités

L'utilisation de plusieurs modalités dans le cadre de la communication humain-machine et dans le sens machine vers humain implique de se poser la question suivante : comment combiner ces modalités ? Cette question est à la base-même de l'identification des différents types de multimodalités.

Par exemple, Bernsen [Bernsen, 1994, Bernsen, 1997] considère l'utilisation de plusieurs modalités représentationnelles pour présenter des informations en sortie. Sans détailler les combinaisons possibles, il souligne leur utilisation séquentielle ou simultanée. Il considère que de la combinaison de modalités représentationnelles résulte une modalité représentationnelle et distingue les modalités pures, qui ne sont pas le résultat d'une combinaison, des modalités combinées. Néanmoins tous les auteurs cités dans notre exploration de l'espace termonologique (chapitre 1), ne traitent pas de la combinaison de modalités. Ainsi Shannon ne l'envisage pas dans sa théorie de la communication. Certes, Battail [Battail, 1997] souligne que, dans cette théorie, le choix du canal se fait en fonction des propriétés souhaitables propres aux messages transmis. En cas d'invariance des canaux pour des informations à transmettre, autrement dit en cas d'équivalence entre ces canaux, il s'agit de choisir parmi l'ensemble des canaux permettant de transmettre de façon équivalente ces informations. Il est donc admis que plusieurs canaux de communication sont disponibles. Mais rien, ni dans [Shannon, 1948] ni dans [Battail, 1997] n'évoque la combinaison possible de ces canaux. Ceci est sans doute dû au fait que, comme les notions d'"émetteur" et de "récepteur", la notion de "canal" peut être définie de façon à ce que la combinaison de canaux constitue un canal à part entière, laissant de côté la question de la répartition des informations sur les sous-canaux.

Pourtant, la combinaison des modalités pose des problèmes de codage des informations. Comme le souligne Corraze (*cf.* la section 1.3.3), le codage, et bien-sûr le décodage, ne sont possibles que grâce à certaines conditions. En particulier, il faut qu'il y ait ... un code, ou ce que Nigay et Coutaz appellent un langage d'interaction (*cf.* la section 1.2.3). Il faut aussi qu'il y ait un code plus global spécifique à la combinaison des modalités, et non aux modalités-même. Dans le cas où plusieurs modalités sont utilisées, quel code de combinaison des modalités adopter ? Corraze souligne que la perte de l'ordre peut

conduire à la perte du sens et/ou à l'apparition d'une nouvelle signification. Quel ordre de transmission choisir ? C'est toute la problématique de la répartition des informations à présenter sur un ensemble de modalités disponibles.

Dans ce chapitre, nous adoptons la même approche d'exploration que dans le chapitre précédent, en commençant par l'étude des travaux en informatique puis celle des travaux issus d'autres domaines centrés sur le fonctionnement humain.

2.1 Combinaison des modalités en informatique

Nous considérons un ensemble d'études sur la combinaison des modalités en mettant l'accent sur la multimodalité en sortie, objet de nos travaux. Les premiers résultats dans le domaine de l'informatique datent des années 90 : nous présentons trois études de référence, celles de Bellik, Martin et Nigay, les premières thèses consacrées à l'interaction multimodale avec celles de Bourguet. Les autres résultats présentés constituent des extensions à ces travaux de base. Au lieu de présenter les travaux de façon chronologique, nous présentons les trois espaces de référence avec leurs extensions plus récentes : l'espace de composition des modalités en sortie de Vernier repose sur les travaux de Nigay et ceux de Rousseau sur les résultats de Bellik. Avant de présenter ces travaux, nous étudions la combinaison de mode selon Frohlich.

2.1.1 Combinaison des paradigmes d'action et de conversation selon Frohlich

Bien qu'il distingue deux modes, le langage et l'action, Frohlich [Frohlich, 1996] ne donne pas de processus permettant de les combiner dans la présentation des réactions du système. Toutefois, il défend l'idée selon laquelle ces deux modes sont complémentaires. Contrairement à la pensée commune dans les systèmes d'interaction humain-machine de l'époque, il considère que le paradigme d'interaction actionnel (*i.e.* s'appuyant sur la manipulation directe) n'est pas toujours meilleur que le paradigme d'interaction conversationnel. Cette démarche amène Frohlich à proposer une nouvelle philosophie de manipulation directe. Elle s'appuie sur un guide de sélection du mode le plus adéquat, présenté dans le tableau 2.1. D'après ce guide, le choix d'un mode dépend de l'activité en cours de façon générale (*activity metaphors*), des actions que l'utilisateur doit réaliser (*interaction tasks*), des *media* utilisés (*interaction media*), des dispositifs physiques disponibles (*interaction technology*) et des capacités d'action et de perception de l'utilisateur (*interaction context*).

Étude comparative et prise de position Deux points nous paraissent particulièrement intéressants dans ce guide de sélection du mode le plus approprié. Premièrement, Frohlich met en avant le fait qu'***un mode peut être plus approprié que l'autre pour une tâche que l'utilisateur doit accomplir***, que ce soit en entrée ou en sortie des machines. Lorsqu'il présente sa réaction, un système informatique a donc tout intérêt, s'il est en mesure d'anticiper la tâche de l'utilisateur, de ne pas limiter le mode jugé

Application features	**Manual mode**	**Conversational mode**
Activity metaphors	O Looking O Browsing O Exploring O Navigating O Controlling O Monitoring O Constructing O Creating	O Informing O Requesting O Asking O Advice seeking O Understanding O Negociating O Delegating O Problem solving
Interaction tasks	O Selecting seen objects O Executing actions with immediate consequences O Responding immediately to feedback O Identifying relationships between objects	O Selecting unseen objects O Identifying sets of objects O Referring back O Scheduling forward O Repeating actions O Combining actions O Specifying exact values
Interaction media	O Sound O Graphics O Motion	O Speech O Text O Gesture
Interaction technology	O Large displays O Sound effects O Motion sensing	O Small displays O Keyboard only O Voice activation
Interaction context	O Hands and eyes free O Ears busy	O Hands or eyes busy O Ears free

TAB. 2.1 – Paramètres de sélection du mode selon Frohlich (extrait de [Frohlich, 1996]

adéquat pour la réaliser, voire d'encourager l'utilisateur à utiliser ce mode. Deuxièmement, le contexte pris en compte par Frohlich pour la sélection d'un mode plutôt que l'autre est assez réduit : il se résume aux capacités d'action et de perception de l'utilisateur qui sont occupées. Mais il souligne ainsi une idée qui nous semble centrale suite à l'exploration terminologique de la notion de "modalité" présentée dans le chapitre précédent (*cf.* la section 1.4), à savoir que ***les capacités de perception (et d'action) de l'utilisateur doivent être prises en compte pour le choix de la forme des réponses du système et des moyens d'action proposés à l'utilisateur***.

Enfin, si Frohlich note qu'il peut être intéressant d'alterner paradigmes interactionnel et conversationnel en fonction de celui qui est le plus approprié, il n'envisage pas de combiner les deux. Par conséquent, comme le soulignent Martin [Martin, 1995] et Nigay et Coutaz [Nigay et Coutaz, 1996], il ne tient pas compte des problèmes engendrés par la coexistence et la co-utilisation de plusieurs *media* ni de plusieurs modes[1]. Il ne rentre pas non plus dans le détail des risques de surcharge cognitive de l'utilisa-

[1] *Medium* et mode sont utilisés ici dans les sens donnés par Frohlich

teur si le système passe continuellement d'un mode de présentation à l'autre. ***Comme nous l'expliciterons dans le chapitre 4, nos travaux visent à contribuer à ces problématiques.***

2.1.2 Coopération entre modalités selon Martin

Pour Martin [Martin, 1995], un système multimodal est un système capable de faire coopérer plusieurs modalités, que ce soit en entrée ou en sortie des systèmes informatiques. Le cadre d'étude de la multimodalité TYCOON qui en résulte prend en compte deux dimensions : les TYpes et les buts de la COOpératioN entre modalités. Par buts, l'auteur entend les problèmes qui peuvent être – au moins en partie – résolus grâce à la multimodalité. Les buts répondent à la question : "pourquoi la multimodalité ?". Par types, il désigne les différentes façons de coopérer entre modalités. Les types répondent à la question : "qu'est-ce que la multimodalité ?". Pour nos travaux, nous nous concentrons sur les types de coopération.

La multimodalité est la coopération entre plusieurs modalités, *i.e.* entre plusieurs processus. Le résultat d'une coopération est une nouvelle modalité qui peut à son tour se combiner avec d'autres. Une modalité est donc soit un processus élémentaire, c'est-à-dire ne pouvant être considéré comme une coopération entre processus, soit un processus résultant de la coopération entre deux ou plusieurs processus. La fusion correspond à l'établissement de liens entre plusieurs informations et la fission ne survient qu'en cas d'erreur de fusion, lorsqu'il faut séparer des informations qui ont été malencontreusement fusionnées. En sortie, le système n'opère pas une fission mais construit la fusion que l'humain devra réaliser.

S'inspirant des différentes coopérations de modalités sensorielles humaines, Martin distingue cinq types de coopération, valables tant de l'humain vers la machine que de la machine vers l'humain. Il s'agit de la coopération par transfert, de la coopération par spécialisation, de la coopération par équivalence, de la coopération par redondance et de la coopération par complémentarité. Les paragraphes qui suivent détaillent chacun de ces types de coopération.

La coopération par transfert peut porter soit sur les données, soit sur le contrôle. Deux modalités a et b coopèrent par transfert de données quand un résultat obtenu par a peut être utilisé par b. Il y a transfert total si n'importe quel résultat obtenu par a peut être utilisé par b, et transfert partiel sinon. Deux modalités a et b coopèrent par transfert de contrôle quand un résultat obtenu par a peut être utilisé pour contrôler b.

La coopération par spécialisation peut être relative aux données, relative à la modalité ou absolue. Une modalité coopère avec les autres modalités par spécialisation relativement aux données si elle est la seule à analyser un ensemble de données mais qu'elle peut analyser d'autres données. Elle coopère avec les autres modalités par spécialisation relativement à la modalité si elle ne peut analyser qu'un ensemble de données, pouvant elles-même être analysées par d'autres modalités. Enfin, elle coopère avec les autres modalités par spécialisation absolue si elle est la seule à analyser un ensemble de données et qu'elle ne peut analyser que cet ensemble.

La coopération par équivalence peut être vue "comme la possibilité de transmettre

une information sur une modalité qui a été choisie parmi plusieurs", chacune permettant d'obtenir le même résultat pour l'information considérée. C'est donc une équivalence de résultat. D'un point de vue formel, la coopération par équivalence est différente en entrée et en sortie des systèmes informatiques. Dans le sens machine vers humain, deux modalités a et b coopèrent par équivalence pour une modalité c quand, pour tout résultat rc de c, il est possible de former un couple (ra, rb) de résultats de a et b tel que le résultat de a sur rc soit ra et le résultat de b sur rc soit rb. Martin précise qu'il n'y a pas d'intégration dans le cas de la coopération par équivalence, puisqu'il s'agit de déterminer si plusieurs modalités sont équivalentes pour transmettre une information mais de n'en choisir qu'une pour ladite transmission. L'équivalence relève donc plus de la comparaison que d'une réelle coopération. De plus, Martin souligne qu'il s'agit généralement d'une équivalence partielle et que l'information transmissible n'est pas toujours exactement la même.

Comme pour la coopération par équivalence, la coopération par redondance implique une comparaison de modalités pour une information donnée. Mais dans le cas de la redondance, il y a intégration des modalités comparées. Formellement, dans le sens machine vers humain, deux modalités a et b coopèrent par redondance pour une modalité c si tout résultat rc de c permet d'obtenir un couple de résultats (ra, rb) de a et b, tels que ra et rb transmettent une même information et que rc résulte d'une intégration de (ra, rb) selon un certain critère.

La coopération par complémentarité consiste à transmettre des informations différentes sur différentes modalités pour atteindre un but donné. Formellement, dans le sens machine vers humain, deux modalités a et b coopèrent par complémentarité pour une modalité c si tout résultat rc de c permet d'obtenir un couple de résultats (ra, rb) de a et b selon un certain critère, ra et rb transmettant des informations différentes.

En plus des types de coopération, TYCOON repose sur la notion de "critère d'intégration". Un critère d'intégration permet, en présence de deux informations, de faire le choix de les fusionner ou non. Ce critère intervient lors d'un processus qui détermine le produit cartésien des deux ensembles d'informations à fusionner vers un ensemble final d'informations. Les critères d'intégration peuvent être combinés pour former des critères d'intégration complexes. De façon non exhaustive, trois types de critères d'intégration sont identifiés : les critères d'intégration temporelle, d'intégration spatiale ou d'intégration modalitaire, correspondent respectivement à des relations temporelles, spatiales ou de modalité entre événements. Deux événements vérifient un critère d'intégration temporelle, spatiale ou modalitaire si la modalité résultant de la coopération entre leurs modalités d'origine effectue une intégration faisant intervenir un critère temporel, spatial ou modalitaire. Se concentrant sur les relations temporelles entre événements, quatre critères d'intégration temporelle sont définis :

- le critère de coïncidence temporelle stricte, lorsque la différence temporelle entre deux événements est inférieure à un paramètre "durée" ;
- le critère de séquence temporelle stricte, quand la différence temporelle entre deux événements vaut un paramètre "retard" ;
- la relation "avant" entre un événement a et un événement b si l'événement a survient avant l'événement b ;

– la relation "après" entre un événement a et un événement b si l'événement a survient après l'événement b.

De même, des relations spatiales peuvent être définies. En résulte la notion de "critère d'intégration spatiale" si la modalité résultante de la coopération entre les modalités d'origine des deux événements considérés effectue une intégration faisant intervenir un critère spatial. Des relations entre modalités d'origine des informations à fusionner peuvent aussi être prises en compte. Typiquement, il peut y avoir relation modalitaire à partir du moment où deux informations sont issues de la même modalité. D'autres cas peuvent être envisagés, tenant compte des coopérations entre modalité. Il y aura alors critère d'intégration modalitaire entre les modalités d'origine des deux informations à fusionner.

Étude comparative et prise de position Considérant que la coopération entre modalités est une nouvelle modalité, Martin rejoint Bernsen qui distingue modalités pures et modalités composées (*cf.* l'introduction du présent chapitre), ce qui permet de cacher la complexité possible d'une modalité donnée sans la nier. ***Cette complexité doit demeurer transparente pour l'utilisateur bien que devant être traitable par le système*** : par exemple, même si un avatar est considéré par le système comme la combinaison d'une modalité visuelle impliquant l'avatar et ses mouvements et d'une modalité auditive synchronisée correspondant aux messages oraux énoncés, cet avatar doit être perçu par l'utilisateur comme un tout, ou plutôt comme un élément unique mobilisant deux modalités sensorielles dans lequel il n'identifie pas les deux modalités du point de vue de la machine. Chez Martin, les compositions possibles sont précisées par les différents types de coopération entre modalités. La multimodalité résultante est clairement centrée utilisateur, la prise en compte de la fission en sortie de la machine étant assimilée à une fusion à réaliser par l'utilisateur.

Dans TYCOON, les coopérations en entrée et en sortie d'un système sont distinguées. Même si les mêmes coopérations sont utilisées en entrée et en sortie des systèmes, elles ne le sont pas de la même façon. Par conséquent, cette distinction permet une caractérisation plus fine de la coopération entre modalités, d'autant plus avec la définition des critères d'intégration.

Toutefois, la notion de critère d'intégration, ainsi que celle de fusion - qui renvoient à l'intégration et à la fusion à réaliser par l'utilisateur - peuvent être trompeuses. C'est pourquoi ***les termes de "critères de décomposition" et de "fission" nous semblent plus appropriés dans notre approche*** qui se concentre sur les choix multimodaux de la machine. Soulignons que, ***s'ils doivent être explicites pour les concepteurs des systèmes multimodaux, critères d'intégration et de décomposition n'ont pas nécessairement à l'être pour les systèmes*** informatiques.

De nombreux concepts de TYCOON sont partagés avec les espaces CASE et CARE et ses extensions présentés dans le paragraphe suivant.

2.1.3 Classification des systèmes multimodaux : CASE, CARE et composition des modalités en sortie

2.1.3.1 Classification CASE

Une première classification de systèmes multimodaux centrée sur la combinaison des modalités a été proposée dans [Coutaz *et al.*, 1993]. Cette classification considère différents niveaux d'abstraction de l'information véhiculée par un canal de communication. La définition d'un canal de communication a été introduite dans la section 1.2.3. Pour rappel, un canal d'entrée, respectivement de sortie, regroupe l'ensemble des dispositifs physiques d'entrée, respectivement de sortie. Cette classification s'appuie sur deux dimensions.

La première dimension de la classification distingue fusion et fission d'informations. Elle spécifie le lien entre unités informationnelles et canaux de communication. Il y a fusion quand plusieurs unités informationnelles issues d'un ou plusieurs canaux sont combinées pour former une unité d'information globale. Respectivement, il y a fission quand une unité d'information globale est décomposée en plusieurs unités informationnelles réparties sur un ou plusieurs canaux. Les auteurs précisent que la fusion n'est pas spécifique à l'entrée des systèmes ni la fission à leur sortie. Un exemple de fission en sortie est la référence intermodale utilisée par le système (*e.g.* le message oral "l'accueil est ici" avec un point clignotant sur un écran pour indiquer l'accueil sur le plan du site). Un exemple de fusion en sortie est la superposition d'informations en utilisant le même canal de communication (*e.g.* un point clignotant de suivi de parcours superposé sur un plan : il s'agit bien de deux types d'informations distincts présentés via le même canal). Nigay [Nigay, 1994] évoque aussi l'utilisation de la fusion en sortie pour l'adaptation des informations du noyau fonctionnel aux besoins conceptuels de l'interface.

La deuxième dimension de la classification porte sur le parallélisme d'utilisation. Les termes "simultanéité" et "synchronie" sont également employés dans la littérature. Dans la classification présentées ici, deux parallélismes sont pris en compte. D'une part, deux canaux peuvent être utilisés en parallèle pour une unité informationnelle donnée : c'est souvent le cas pour les références intermodales. D'autre part, un même canal peut être utilisé en parallèle pour deux tâches données : c'est notamment le cas quand l'utilisateur fait plusieurs choses indépendantes en même temps, comme travailler sur plusieurs fichiers. Ces deux types de parallélisme ne sont pas distingués dans la classification. Le fait que le parallélisme ne soit pas admis dans un système multimodal renvoie à deux cas possibles : (1) deux canaux ne peuvent pas être utilisés en même temps ou (2) un même canal ne peut pas être utilisé pour deux tâches différentes en même temps.

Ces deux dimensions permettent aux auteurs d'identifier quatre types de multimodalité, à l'origine du nom de cette classification (CASE pour Concurrent-Alterné-Synergique-Exclusif) :

- Concurrent : un seul canal est utilisé et le parallélisme est possible. Par déduction, le parallélisme n'est donc possible que sur la tâche ;
- Alterné : plusieurs canaux peuvent être utilisés et le parallélisme n'est pas admis. L'interdiction de parallélisme porte donc sur les tâches ;
- Synergique : plusieurs canaux peuvent être utilisés et le parallélisme est possible ;

– Exclusif : un seul canal est utilisé et le parallélisme n'est pas admis.

Notons que la fusion, respectivement la fission, n'est pas une phénomène spécifique aux entrées, respectivement aux sorties. Or, on parle généralement de fusion multimodale pour la multimodalité en entrée et de fission multimodale pour la multimodalité en sortie. Cette approche particulière est due au fait que les auteurs considèrent qu'entre les informations concrètes, (multi)modalement allouées et physiquement perceptibles et les informations abstraites, amodales et données internes aux systèmes, il y a plusieurs niveaux d'abstraction. Le passage d'un niveau d'abstraction à un autre nécessite des fusions et/ou des fissions. Par conséquent, il y a plusieurs types de fusion et de fission. Plus précisément, [Nigay, 1994] distingue trois niveaux de fusion et de fission valables tant en entrée qu'en sortie des systèmes. Ces différents niveaux dépendent de la nature des informations manipulées. Pour la sortie, ces trois niveaux sont :

- le niveau sémantique : il porte sur les états ou les intentions du système, *i.e.* en amont des niveaux d'abstraction. Une fission sémantique est la décomposition d'un état ou d'une intention du système en plusieurs états ou intentions qui font sens. La fusion sémantique est le regroupement de plusieurs états ou intentions du système pour définir un unique état plus global du système ;
- le niveau syntaxique : il porte sur les unités informationnelles, *i.e.* au niveau des langages d'interaction qui sont à un haut niveau d'abstraction . Une unité informationnelle est une unité conceptuelle - qui a donc un haut niveau d'abstraction - regroupant un ensemble d'informations indivisible et renvoyant à un concept du domaine pour le système. Une fission syntaxique est la décomposition d'un état ou d'une intention du système en plusieurs unités informationnelles. Une fusion syntaxique est le regroupement de plusieurs états ou intentions du système en une unité informationnelle unique ;
- le niveau lexical : il porte sur les actions physiques du système, *i.e.* au niveau des dispositifs physiques qui sont à un bas niveau d'abstraction. Une fission lexicale est la concrétisation d'une même unité informationnelle avec différents actions, *i.e.* via différentes techniques ou outils. Une fusion lexicale est la concrétisation de plusieurs unités informationnelles distinctes avec une seule action.

Étude comparative et prise de position Nous constatons que ces différents niveaux d'abstraction ne sont pas considérés de façon explicite dans la dimension de parallélisme de CASE. Cette classification pourrait donc être affiné. Nous verrons que ***l'espace de Bellik (cf. la section 2.1.4.1) vise cette distinction explicite entre les différents niveaux d'abstraction du parallélisme.***

2.1.3.2 Propriétés CARE

Tandis que CASE focalise sur l'usage des modalités pour réaliser une ou plusieurs tâches, les propriétés CARE (pour Complémentarité, Assignation, Redondance et Equivalence) décrivent les relations entre modalités pour la réalisation d'une seule tâche [Nigay, 1994]. Ces propriétés s'appliquent aux deux éléments constitutifs d'une modalité que sont les dispositifs physiques et les langages d'interaction (*cf.* section1.2.3). Elles

portent sur l'usage parallèle/synchrone/simultané des modalités pour une unité informationnelle ou une tâche données. Nigay distingue les entrées des sorties des systèmes. En se focalisant sur la sortie d'un système informatique, voici comment sont définies ces quatre propriétés :

- l'équivalence : il y a équivalence entre plusieurs modalités lorsqu'il est possible de choisir entre elles pour présenter une unité informationnelle. L'équivalence est définie comme suit pour les deux éléments constitutifs d'une modalité :
  - l'équivalence entre deux langages d'interaction est totale si toutes les unités informationnelles peuvent être exprimées grâce à ces deux langages ; elle est partielle si seule une partie des unités informationnelles peuvent être exprimées grâce à ces deux langages ;
  - l'équivalence entre deux dispositifs physiques est totale si toutes les unités informationnelles exprimables avec un langage d'interaction considéré sont spécifiables avec ces deux dispositifs physiques ; elle est partielle si seule une partie des unités informationnelles exprimables avec le langage d'interaction considéré sont réalisables avec ces deux dispositifs physiques ;
- l'assignation : il y a assignation lorsqu'une seule modalité permet de présenter une unité informationnelle donnée. C'est le contraire de l'équivalence. L'assignation est définie comme suit pour les deux éléments constitutifs d'une modalité :
  - un langage d'interaction est assigné à une unité informationnelle lorsque cette unité informationnelle ne peut être exprimée que grâce à ce langage ;
  - un dispositif physique est assigné à une unité informationnelle exprimable grâce à un langage d'interaction lorsqu'il est le seul à la pouvoir la concrétiser ;
- la redondance : il y a redondance lorsque la même unité informationnelle est présentée plusieurs fois, de différentes façons, *i.e.* via différentes modalités. La redondance est définie comme suit pour les deux éléments constitutifs d'une modalité :
  - il y a redondance totale entre deux langages d'interaction lorsque ces deux langages d'interaction sont totalement équivalents ; il y a redondance partielle entre deux langages d'interaction lorsque ces deux langages d'interaction sont partiellement équivalents ;
  - il y a redondance totale entre deux dispositifs physiques lorsque ces deux dispositifs physiques sont totalement équivalents ; il y a redondance partielle entre deux dispositifs physiques lorsque ces deux dispositifs physiques sont partiellement équivalents ;
- la complémentarité : il y a complémentarité lorsque "chaque modalité véhicule une partie de l'information qui, prise isolément, est insuffisante pour faire sens" [Nigay et Coutaz, 1996]. La complémentarité est définie comme suit pour les deux éléments constitutifs d'une modalité :
  - il y a complémentarité totale entre deux langages d'interaction lorsque toutes les unités informationnelles exprimables grâce à ces langages d'interaction ne sont pas équivalentes ; il y a complémentarité partielle entre deux langages d'interaction lorsque seule une partie des unités informationnelles exprimables grâce à ces langages d'interaction n'est pas équivalente ;

- il y a complémentarité totale entre deux dispositifs physiques lorsque toutes les unités informationnelles exprimables grâce à un langage d'interaction donné ne sont pas équivalentes par rapport à ces deux dispositifs physiques ; il y a complémentarité partielle entre deux dispositifs physiques lorsque seule une partie des unités informationnelles exprimables grâce à un langage d'interaction donné n'est pas équivalente par rapport à ces deux dispositifs physiques ;

Nigay souligne qu'équivalence et assignation ont trait à l'existence de choix alors que complémentarité et redondance ont trait à la combinaison de choix.

Notons que les travaux de Clémente reprennent les quatre propriétés CARE comme critère de caractérisation des modalités [Clémente, 2004]. Afin d'automatiser la génération multimodale qui combinent plusieurs modalités dans un même énoncé, Clémente définit quatre critères, un par propriété CARE, pour caractériser la validité de la coopération possible - ou non - entre la modalité considérée et les autres modalités pour un type donné d'information.

Étude comparative et prise de position ***Si la redondance résulte d'un constat d'équivalence, la complémentarité peut aussi bien résulter d'un constat d'équivalence que d'un constat d'assignation.*** En effet, deux unités informationnelles qui sont assignées à une modalité de par leur nature, ne peuvent être présentées que de façon complémentaire si elles sont combinées pour former une information plus globale. De plus, ***l'équivalence en sortie doit nécessairement faire l'objet d'un choix, soit de la part du système, soit de la part de l'utilisateur***. Ce constat confère un statut particulier à l'équivalence. En effet, si elle est admise en sortie du système, *i.e.* si le système ne prend aucune décision lorsque deux présentations multimodales sont équivalentes, le déroulement de la communication avec l'utilisateur en est modifié car le système doit alors demander à l'utilisateur de choisir.

À la notion de "transfert" près et l'assignation se rapprochant de la spécialisation, CARE s'accorde avec TYCOON sur les multimodalités possibles. ***Il y a toutefois une nuance entre assignation et spécialisation.*** L'assignation porte sur une donnée particulière en prenant en compte les deux niveaux d'abstraction que sont le dispositif physique et le langage d'interaction. ***C'est donc un cas particulier de la spécialisation qui s'applique à un type de données.***

2.1.3.3 Composition des modalités en sortie

L'espace de composition des modalités en sortie selon Vernier [Vernier, 2001] affine les propriétés CARE, en particulier la complémentarité et la redondance qui impliquent toutes deux une fusion ou une fission. Cet espace s'appuie sur les trois niveaux d'abstraction de fusion décrits ci-dessus d'une part et sur les aspects spatio-temporels de composition d'autre part. En résulte les cinq aspects de composition suivants :

- l'aspect temporel : il décrit les relations temporelles entre les modalités ;
- l'aspect spatial : il correspond à la répartition spatiale des modalités dans le champ de perception de l'utilisateur ;

- l'aspect articulatoire : il spécifique que la composition est à un bas niveau d'abstraction, plus précisément au niveau des dispositifs physiques ;
- l'aspect syntaxique : il renvoie à une composition au niveau des langages d'interaction ;
- l'aspect sémantique : il fait référence à une composition au niveau des informations choisies pour être présentées.

Chacun de ces aspects est précisé en définissant une proximité entre les modalités. S'appuyant sur les relations temporelles identifiées par Allen, cinq schémas de proximité sont définis :

- la composition de modalités éloignées (*cf.* colonne 2 du tableau 2.1) ;
- la composition de modalités avec un point de contact (*cf.* colonne 3 du tableau 2.1) ;
- la composition de modalités avec une intersection non vide (*cf.* colonne 4 du tableau 2.1) ;
- la composition de modalités dont l'une englobe l'autre (*cf.* colonne 5 du tableau 2.1) ;
- la composition de modalités de même étendue (*cf.* colonne 6 du tableau 2.1).

Le recoupement des aspects et des schémas de composition permet de dégager 25 compositions possibles résumées dans le tableau 2.1.

Aspects de composition \ Schémas de composition					
Composition					
Temporelle	Anachronique	Séquentielle	Concomitante	Coïncidente	Parallèle / Simultanée
Spatiale	Disjointe	Adjacente	Intersectée	Imbriquée	Recouvrance
Articulatoire	Indépendance	Fissionnée	Fissionnée + Dupliquée	Partiellement Dupliquée	Dupliquée
Syntaxique	Différente	Complétion	Divergence	Extension	Jumelage
Sémantique	Concurrente	Complé-mentaire	Complémentaire + Redondante	Partiellement Redondante	Totalement Redondante

FIG. 2.1 – Composition des modalités selon Vernier (extrait de [Vernier, 2001])

Étude comparative et prise de position Les aspects de composition temporelle et de composition spatiale renvoient clairement aux critères d'intégration spatiale et d'intégration temporelle de TYCOON (*cf.* la section 2.1.2). L'aspect articulatoire, respectivement syntaxique, renvoie au niveau de fusion/fission, respectivement syntaxique, identifiés par Nigay (*cf.* la section 2.1.3.2). Il intègre les propriétés CARE appliquées aux dispositifs physiques, respectivement aux langages d'interaction. Enfin, l'aspect sémantique renvoie au niveau de fusion/fission sémantique de Nigay. Le critère d'intégration

modalitaire de TYCOON, qui porte sur les modalités, *i.e.* sur le couple <dispositif physique ; langage d'interaction>, n'apparaît pas dans l'espace de composition. Il permettrait pourtant d'intégrer les propriétés CARE au niveau des couples <dispositif physique ; langage d'interaction> considérés comme un tout. Ce niveau de prise en compte de la composition des modalités est plus proche de l'appréhension d'une sortie multimodale par l'utilisateur. En effet, comme le met en avant Landragin [Landragin, 2004a], la communication humaine est multimodale parce que multicanale, mobilisant principalement le canal visuo-gestuel pour le co-verbal (transmis par la voix mais pas par la langue) et le non-verbal (transmis ni par la voix ni par la langue [2]) et le canal audio-oral pour le verbal. Or, dans le cas de la référence linguistique et/ou gestuelle étudiée par Landragin, la multimodalité humaine est à des niveaux temporel et spatial mais aussi modalitaire : les unités informationnelles sont spatialement positionnées et synchronisées et elles sont allouées à une ou plusieurs modalités (et non à un ou plusieurs langages d'interaction et/ou à un ou plusieurs dispositifs physiques). Notons que les schémas de composition considérés dans ces travaux sont la complémentarité, la complémentarité+redondance et la redondance partielle : dans l'expression "mets ça là", le "là" est redondant avec le geste de désignation.

Les schémas de composition affinent la complémentarité et la redondance des modalités. ***Nous considérons que la complémentarité correspond à un recouvrement plus ou moins important (composition avec un point de contact et composition avec une intersection non vide) et la redondance à un recouvrement total (composition de même étendue) : selon cette approche, la redondance partielle est une forme de complémentarité.***

L'espace n'intègre pas l'assignation/spécialisation, l'équivalence et le transfert. Ceci se justifie par le fait que l'espace se focalise sur la composition des modalités et ces trois coopérations ne combinent pas les modalités dans un seul et même message.

Dans nos travaux, nous nous focalisons sur l'aspect sémantique, l'aspect modalitaire (i.e. qui considère le couple <dispositif physique ; langage d'interaction> comme un tout) et, dans une moindre mesure, l'aspect temporel. Nous excluons toutefois la composition concurrente. Ces choix sont motivés par les systèmes sur lesquels nous travaillons et par le niveau de détermination de la sortie du système que nous considérons. En effet, dans le cadre des systèmes d'information, nous nous penchons plus spécifiquement sur le choix de la stratégie de dialogue et de la stratégie de présentation qui constituent le comportement adopté par le système (*cf.* le chapitre 5), sans se préoccuper particulièrement de la concrétisation de ce comportement. Par conséquent, nous sommes contraints de restreindre la problématique traitée et notamment la portée du choix de la stratégie de présentation. En nous limitant au choix des modalités, de leur allocation aux unités informationnelles et de leur répartition temporelle (qu'il est nécessaire de prendre en compte en raison de la dimension séquentielle de la modalité auditive), nous nous concentrons sur le comportement du système à un haut niveau d'abstraction, sans empiéter et sans fermer la porte à un

[2] Notons au passage que la définition de "non-verbal" est différente de celle de Corraze [Corraze, 1980]

travail plus poussé sur la coordination spatio-temporelle et la présentation à un niveau plus fin du comportement du système (par exemple, rythme d'énonciation, typographiqe d'affichage, etc.). L'exclusion de la composition concurrente est dûe au constat que, à un niveau sémantique, il est difficile, voire impossible, pour l'utilisateur de mener de front deux interactions concurrentes avec un système d'information sur une ou sur deux modalités différentes.

2.1.4 Typologie des multimodalités selon Bellik et application à la multimodalité en sortie selon Rousseau

2.1.4.1 Types de multimodalités selon Bellik

Pour Bellik [Bellik, 1995], un système multimodal en sortie est un système capable de déterminer la présentation la plus adéquate pour communiquer avec l'utilisateur en utilisant différentes modalités. De façon à identifier les différents types de multimodalité, Bellik distingue trois dimensions :

- la production d'énoncés : cette dimension permet de préciser si des énoncés indépendants doivent être produits séquentiellement ou parallèlement ;
- l'usage des médias : cette dimension permet de préciser si chaque média doit être utilisé exclusivement ou si plusieurs médias peuvent être utilisés simultanément ;
- le nombre de médias par énoncé : cette dimension permet de préciser s'il y a un seul ou plusieurs médias possibles par énoncé. Dans le deuxième cas, Bellik précise qu'il y a fusion en entrée.

Ces dimensions conduisent à l'identification de huit types de multimodalités, qui définissent un cube présenté dans la figure 2.2. Ces huit multimodalités possibles sont les suivantes :

- impossible : la production des énoncés est séquentielle, il n'y a qu'un média par énoncé et l'usage des médias est simultané. Ce cas est impossible car les valeurs prises par les deux derniers paramètres sont incompatibles ;
- multimodalité exclusive : la production des énoncés est séquentielle, il n'y a qu'un média par énoncé et chaque média est utilisé exclusivement ;
- multimodalité alternée : la production des énoncés est séquentielle, il y a plusieurs médias par énoncé mais chaque média est utilisé exclusivement (pour une partie d'énoncé, donc ; par exemple, dire "je veux aller là" puis cliquer sur une carte) ;
- multimodalité synergique : la production des énoncés est séquentielle, il y a plusieurs médias par énoncé et plusieurs médias peuvent être utilisés parallèlement (dire "je veux aller là" tout en cliquant sur une carte) ;
- multimodalité parallèle exclusive : la production des énoncés est parallèle (on peut adresser plusieurs requêtes en même temps, parallèlement), mais il n'y a qu'un seul média par énoncé et chaque média est employé exclusivement à un instant donné ;
- multimodalité parallèle simultanée : la production des énoncés est parallèle, un seul média est autorisé par énoncé mais plusieurs médias peuvent être utilisés simultanément (on a deux requêtes parfaitement parallèles sur deux médias diffé-

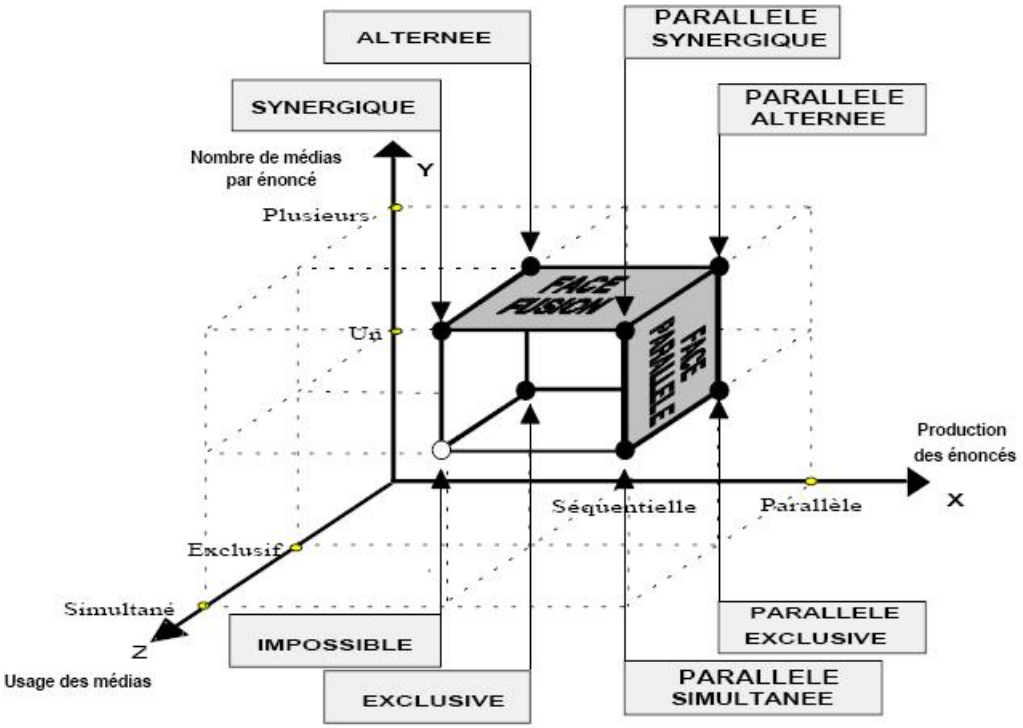

FIG. 2.2 – Types de multimodalité selon Bellik (extrait de [Bellik, 1995])

rents : on peut donc demander quelque chose à l'oral et autre chose via la souris ou le clavier) ;
- multimodalité parallèle alternée : la production des énoncés est parallèle, plusieurs médias peuvent être utilisés pour un même énoncé mais chaque média est utilisé exclusivement à un instant donné ;
- multimodalité parallèle synergique : la production des énoncés est parallèle, il peut y avoir plusieurs médias par énoncé, chaque média pouvant être utilisé parallèlement avec d'autres. C'est le type de multimodalité le plus ouvert.

Comme les autres classifications citées, cette typologie des multimodalités permet l'étude des systèmes multimodaux en sortie mais n'aide pas à leur conception. En particulier, la relation entre énoncés considérés par Bellik est temporelle : il ne rentre pas dans le détail des relations sémantiques entre les énoncés. Sa dimension de "production d'énoncés pourrait, à notre avis, être raffinées en fonction de la relation sémantique entre les énoncés parallèles. Nous laissons ce travail à d'autres, considérant qu'il ne permet pas de faciliter la conception de systèmes multimodaux en sortie.

Étude comparative et prise de position La première dimension prise en compte par Bellik renvoie à la dimension parallèle au niveau des tâches de la classification CASE (*cf.* la section 2.1.3.1). Les deux autres dimensions détaillent la dimension parallèle au niveau des canaux de communication de la classification CASE. ***La typologie de Bellik***

précise donc la dimension parallèle de la classification CASE, répondant à notre critique de manque de précision de cette classification. Étant donné le point de vue qu'il adopte, il n'a pas besoin de considérer la dimension fusion/fission qui est implicite. Son positionnement par rapport à cette dimension est différent des auteurs cités : il considère qu'il y a fusion, respectivement fission, en entrée, respectivement en sortie, lorsque plusieurs médias sont utilisés pour un énoncé donné.

Parmi les types de multimodalité identifiés qui résultent de ces trois dimensions et par rapport à la classification CASE (*cf.* la section 2.1.3.1), la multimodalité exclusive renvoie à l'usage exclusif, la multimodalité alternée à l'usage alterné, la multimodalité synergique à l'usage synergique avec limitation du parallélisme des tâches, la multimodalité parallèle exclusive - dans le cas où il n'y a qu'un seul média pour tous les énoncés - à l'usage concurrent au niveau des tâches (en parallèle sur un même canal), la multimodalité parallèle simultanée à l'usage concurrent au niveau à la fois des tâches et des canaux, la multimodalité parallèle alternée à l'usage synergique restreint pour le parallélisme des canaux, et la multimodalité parallèle synergique à l'usage synergique sans restriction.

Si cette typologie est valable tant en entrée qu'en sortie - la face fusion du cube devenant alors une face fission - ces types de multimodalité en sortie ne sont pas toujours utilisables, selon les médias considérés. En effet, Bellik souligne que les multimodalités parallèles alternées et parallèles exclusives correspondent à des dialogues à plusieurs fils et les multimodalités parallèles simultanées et parallèles synergiques ont trait à des dialogues concurrents[3]. Or, comme nous l'évoquions dans la section précédente, nous pensons que de tels dialogues sont difficiles à gérer pour l'utilisateur mais aussi pour le système, en raison des difficultés à identifier quels énoncés ont trait à un (fil de) dialogue particulier. ***Dans une approche de communication centrée sur l'utilisateur, nous réduisons la complexité en ne considérant que les cas où la multimodalité intervient dans une communication entre un seul humain et une seule machine qui ne peut produire en parallèle que des énoncés ayant trait à un même fil de dialogue.***

2.1.4.2 Systèmes multimodaux et combinaison des modalités en sortie selon Rousseau

S'appuyant sur les définitions de Bellik, Rousseau [Rousseau, 2006] considère qu'un système multimodal en sortie est un système capable de générer une présentation multimodale "intelligente". Une présentation multimodale est une présentation d'informations composée d'un ensemble de couples <modalité, média> liés par des propriétés de redondance et/ou de complémentarité. Une présentation intelligente est une présentation adaptée au contexte. Par contexte, Rousseau désigne toute information susceptible de caractériser l'état ou la situation d'une entité considérée comme une référence. Si cette entité est l'interaction, alors le contexte (d'interaction) est "toute information relative à une personne, un lieu, un intervalle temporel ou un objet considéré comme pertinent pour l'interaction entre l'utilisateur et l'application". Rousseau précise qu'il privilégie

[3]Des dialogues concurrents sont des dialogues à plusieurs fils dont les énoncés sont en parallèle.

la composition des modalités par complémentarité et redondance car il considère que l'assignation et l'équivalence sont implicites en sortie et peu manipulées.

Étude comparative et prise de position ***Nous élargissons la définition de contexte d'interaction adoptée par Rousseau à toute information susceptible de modifier le déroulement de l'interaction entre l'utilisateur et la machine***, même si elle n'est pas considérée comme pertinente pour l'interaction. ***Le contexte s'apparente alors à la notion de "bruit" mise en évidence dans la théorie de la communication*** [Shannon, 1948], dans la mesure où elle renvoie à tout ce qui perturbe non seulement le canal mais aussi les différents éléments de la chaîne de transmission de l'information entre la source et le destinataire. La notion de "contexte" admet donc plusieurs niveaux. Elle nécessite de se positionner par rapport aux éléments admis comme bruits. ***Ne cherchant pas à définir la notion de "contexte" dans le cadre de nos travaux, nous nous focalisons sur les contraintes de présentation comme éléments perturbateurs*** de la communication humain-machine (*cf.* le chapitre 5).

2.1.5 Synthèse : combinaison de modalités en informatique

Nous avons présenté et mis en relation les différents espaces de composition de modalités définis en informatique. Nous constatons que ces espaces ne sont pas en contradiction et qu'ils sont complémentaires, abordant des points de vue différents sur la combinaison des modalités. Les multimodalités possibles résultantes constituent un vaste espace de conception qu'il convient de restreindre pour nos travaux.

Comme nous l'avons déjà dit, nos travaux se focalisent sur l'aspect sémantique et l'aspect modalitaire de la combinaison des modalités (*cf.* la section 2.1.3.3). Nous référant aux schémas de composition communs à CARE (*cf.* la section 2.1.3.2) et à TYCOON (*cf.* la section 2.1.2), nous considérons que la complémentarité inclut les compositions de complémentarité+redondance et de redondance partielle (*cf.* la section 2.1.3.3). Par conséquent, la redondance telle que nous la considérons est de la redondance stricte et la complémentarité inclut de la complémentarité partielle. Si ce choix peut sembler réducteur, il ne l'est pas car nous restons à un haut niveau de détermination du comportement du système. La composition considérée des modalités est au même niveau macroscopique que CARE et TYCOON. Ce positionnement ne contraint pas une détermination plus fine du comportement multimodal du système qui distinguerait les différents cas entre la redondance et la complémentarité strictes et qui constitue une extension possible de notre approche et de nos travaux. Par exemple, une réponse du système déterminée aux niveaux sémantique et modalitaire comme étant constituée d'une liste de solutions présentée de façon complémentaire en combinant oral et hypertexte et d'une invitation orale à une nouvelle requête peut être affinée en aval en spécifiant que la propriété demandée par l'utilisateur est présentée oralement alors que les autres informations qui lui sont liées sont présentées visuellement : la composition modalitaire et sémantique est donc affinée et peut/doit s'y ajouter une composition temporelle, voire une composition spatiale s'il y a lieu, entre les différentes informations.

Comme nous le détaillerons dans le chapitre 5, la précision de la composition sémantique et de la composition modalitaire, et donc la finesse des compositions réalisables en aval, relèvent du choix des concepteurs.

Nous préférons le terme d'"assignation" (*cf.* la section 2.1.3.2) à celui de "spécialisation" (*cf.* la section 2.1.2) car nous considérons qu'une modalité peut ne pas être spécialisée pour une unité informationnelle donnée, mais assignée par le système pour une présentation donnée. D'une telle assignation peut découler une complémentarité, stricte ou partielle : en particulier, une complémentarité à un niveau modalitaire impliquera des assignations aux niveaux syntaxique et/ou lexical/articulatoire. Nous excluons l'équivalence à un niveau modalitaire car elle ne peut être stricte si l'on se réfère à l'effet de modalité identifiée en psychologie et en ergonomie [Lieury, 2005] : une même information exprimée par le même langage d'interaction (*e.g.* le langage naturel) mais présentée par deux dispositifs physiques différents (*e.g.* écran versus haut-parleurs) n'est pas intégrée de la même façon par l'utilisateur et influence la suite de la communication [Karsenty, 2006, Le Bigot *et al.*, 2006]. Nous excluons également le transfert (*cf.* la section 2.1.2) car nous considérons qu'il constitue une utilisation de la présentation durant une communication plus qu'une combinaison possible au sein d'un échange. L'exclusion de la concurrence, qu'elle soit sémantique ou modalitaire, est principalement motivée par une complexité que nous jugeons trop importante, que ce soit pour l'utilisateur ou pour le système (*cf.* la section 2.1.3.3).

2.2 Combinaison des modalités dans d'autres domaines

Pour identifier les combinaisons de modalités prises en compte dans les sciences humaines et expérimentales, nous reprenons les références utilisées dans le chapitre 1, à l'exception de celles qui n'abordent pas ce concept. En particulier, le point de vue issu de l'histoire [Calvet, 1996] n'évoque pas la combinaison des moyens d'expression qu'il a défini. De même, Escarpit [Escarpit, 1991], dont les travaux s'inscrivent dans les sciences de l'information et de la communication, n'aborde pas la combinaison des canaux visuels, auditifs et tactiles. Enfin, le point de vue issu de l'éthologie humaine [Corraze, 1980] identifie un parallélisme des communications non verbales mais ne détaille pas les relations entre ces communications.

2.2.1 Point de vue issu des sciences de l'information et de la communication

Comme nous l'avons expliqué dans la section 1.3.2 du chapitre 1, Breton et Proulx [Breton et Proulx, 2002] identifient quatre moyens de communication utilisés aujourd'hui. Parmi eux, les auteurs notent que le geste joue encore "un rôle important d'appui et de soutien". Ils le considèrent comme un "complément actif à l'oral". Ils citent Jakobson qui propose la notion de "message phatique". Un message phatique est produit par les gestes et les mouvements du corps. Il permet de s'assurer qu'une communication est bien en cours. Cette gestuelle d'accompagnement ou d'appui ne peut prétendre être

un moyen de communication à elle-seule. De plus, [Breton et Proulx, 2002] rattache le documentaire et le reportage filmé à l'image.

Étude comparative et prise de position Oral et gestuelle d'accompagnement et d'appui sont complémentaires pour une même unité informationnelle. Par rapport aux niveaux de composition identifiés, la complémentarité est d'ordre modalitaire, *i.e.* syntaxique et lexical/articulatoire.

Par ailleurs, le rattachement du documentaire et du reportage filmé à l'image nous incite à aller plus loin quant à l'étude des combinaisons entre modalités. En effet, il peut sembler difficile d'admettre que ces deux types de messages relèvent de l'image alors qu'ils transmettent du mouvement et du son. Ce constat nous amène à considérer que certains messages combinent intrinsèquement plusieurs modalités. Dans le cas du documentaire ou du reportage filmé, le contenu inclut du mouvement et du son, qui peut être de l'oral. Mais, dans ce cas particulier, dissocier le contenu en fonction de moyens de communication qui ne sont pas indépendants n'a pas de sens. C'est ce constat qui nous conduit à considérer que ***les avatars relèvent d'une multimodalité particulière*** que nous choisissons de ne pas traiter. ***Le message oral produit par un avatar et les mouvements de lèvres, voire la gestuelle, correspondants ne sont pas indépendants. Les deux sont indissociables*** : il ne s'agit ni de complémentarité, ni de redondance, mais d'***une assignation d'un message à la modalité "avatar" dans son ensemble*** (*cf.* la multimodalité au niveau de l'avatar dans le cadre du projet SmartKom [Poller et Tschernomas, 2006]). Pourtant, chaque sous-unité informationnelle orale peut être redondante avec une unité informationnelle textuelle, sans pour autant qu'il y ait redondance entre l'avatar et le texte. ***La complexité de traitement des avatars résulte de ces relations imbriquées, où une relation au niveau sémantique n'est pas toujours transmise intégralement au niveau modalitaire.***

Pour réduire cette complexité, nous laissons de côté les systèmes incluant des avatars et nous considérons la génération multimodale comme une succession de fissions et d'allocations mono ou multimodales.

2.2.2 Points de vue issus de la neurobiologie et des neurosciences

Même si les connaissances actuelles sur le cerveau ne permettent pas de cerner exactement la façon dont sont intégrées des informations provenant de plusieurs sensibilités, les études s'accordent à dire que les sensibilités ne sont pas complètement indépendantes. Pour commencer, ce constat s'appuie sur la diversité des informations perçues par certains récepteurs [Imbert, 2006]. Cette diversité peut porter sur plusieurs aspects d'une même sensibilité (comme la couleur et la texture dans le cas des cônes du système visuel) mais aussi sur plusieurs sensibilités (par exemple, les récepteurs noniceptifs servent aussi la sensibilité chimio-sensorielle trigéminale). Par ailleurs, plusieurs modalités semblent coopérer pour obtenir une information globale, même si une modalité est traitée de façon individuelle le long d'un trajet sensoriel allant des capteurs aux centres nerveux [Imbert, 2006]. Par exemple, le traitement de la gustation au niveau cérébral

prend sans doute également en compte les informations thermique et tactile fournies par la sensibilité extéroceptive (super-)cutanée. C'est la combinaison de ces différentes informations qui déterminerait la texture et la température des produits ingérés, facteurs entrant en compte dans l'appréciation du produit goûté [Purves *et al.*, 2003, Imbert, 2006]. La coopération entre sensibilités est la plus flagrante pour le système vestibulaire. En effet, il semble qu'une partie des informations perçues par les récepteurs est transmise sans modification, alors qu'une autre partie est intégrée avec les messages issus d'autres sensibilités [Imbert, 2006]. Plus précisément, les informations vestibulaires sont combinées avec des informations visuelles (en ce qui concerne la représentation de l'espace issue de la représentation de la direction et de la vitesse du flux optique) et avec des informations proprioceptives (pour la position des membres). C'est cette combinaison qui permet de déterminer l'équilibre du corps avec plus de précision [Purves *et al.*, 2003, Imbert, 2006]. Imbert [Imbert, 2006] souligne d'ailleurs que, contrairement aux sensibilités visuelle, auditive et extéroceptive (super-)cutanée, il semblerait qu'aucune zone dédiée au système vestibulaire n'existe dans les centres nerveux. Les chercheurs pensent que si une telle zone existe, elle serait "multimodalitaire"[4]. La coopération entre sensibilités pour extraire une information globale sur le monde externe et interne porterait non seulement sur la nature des informations perçues, mais aussi sur le traitement spécifique à chaque sensibilité. Par exemple, [Purves *et al.*, 2003] souligne que le résultat auditif est beaucoup plus rapide que celui des autres sensibilités et que cette rapidité est primordial pour guider les comportements et les actions. Aussi la sensibilité auditive intervient notamment pour faciliter l'orientation du corps et celle de la tête, permettant de prévenir de nouvelles stimulations, en particulier visuelles.

Étude comparative et prise de position ***Les compositions ou coopérations entre modalités identifiées en informatique sont donc en accord avec les études en neurobiologie et en neurosciences.*** Nous constatons aussi que ***les différents niveaux d'abstraction mis en évidence par Nigay*** (*cf.* la section 2.1.3.2) ***peuvent se rapporter aux différentes étapes de l'intégration sensorielle, dans laquelle interviennent vraisemblablement les critères d'intégration de Martin*** (*cf.* la section 2.1.3.2).

2.2.3 Point de vue issu de la psychologie

Un état des connaissances sur la coopération des modalités sensorielles, selon le terme utilisé par Martin [Martin, 1995], est présenté dans [Hatwell, 1994]. Non seulement les modalités sensorielles fonctionnent en parallèle, mais qui plus est les informations récupérées par une modalité donnée sont, au moins en partie, transférables aux autres modalités. L'auteur parle d'"intégration intermodale" pour désigner l'intégration des informations exploitables par plusieurs modalités. Cette intégration se fait sur la base de transferts intermodaux. Un transfert intermodal renvoie à l'utilisation correcte par une modalité d'une information perçue par une autre modalité. Ces transferts

[4] Terme employé par l'auteur.

permettent une économie d'apprentissage et une connaissance cohérente et unifiée du monde. Deux types d'intégration intermodale sont identifiés. D'une part, une modalité activée permet d'obtenir des informations sur une propriété particulière. Cette modalité activée est parfois assignée à la perception de la propriété en question : elle est alors complémentaire des autres modalités sensorielles. D'autre part, plusieurs modalités activées fournissent des informations sur une même propriété. Si la valeur attribuée par ces différentes modalités est identique pour la propriété en question, ces modalités sont redondantes. Ces deux types d'intégration conduisent à deux situations de transfert intermodal. D'une part, l'information acquise par une modalité est utilisée par une autre : la mobilisation des modalités est séquentielle. D'autre part, des informations sur une propriété donnée sont perçues par plusieurs modalités. On peut parler de mobilisation simultanée, synchrone ou parallèle. L'étude de l'intégration intermodale met donc en évidence d'une part la dimension temporelle prise en compte et d'autre part les relations existantes entre les modalités sensorielles. La dimension temporelle distingue intégration séquentielle et intégration parallèle d'une propriété. Les relations identifiées sont la complémentarité, qui découle de la spécialisation, et la redondance, qui découle de l'équivalence.

Plusieurs situations de transfert intermodal sont étudiées en psychologie, certaines relevant d'une mobilisation séquentielle et d'autres d'une mobilisation simultanée. Tout d'abord, même si toutes les modalités mettent en œuvre les mêmes processus, ou du moins des processus proches, pour la récupération et le traitement des informations, la théorie d'une perception centralisée ne semble pas tout à fait exacte. En effet, une telle théorie prône une représentation symbolique interne indépendante des modalités des informations perçues, comme un langage. Or les études ont formellement exclus l'existence d'une telle représentation symbolique. Ensuite, il semble que chaque modalité est plus compétente pour une dimension donnée de perception, en l'occurrence la dimension spatiale pour la vision, la dimension temporelle et successive pour l'audition et, de façon moins nette, la dimension de substance pour le toucher. Cette spécialisation se répercuterait sur le codage en interne des informations, une information étant mémorisée dans la modalité qui est la plus compétente. Il y aurait donc plusieurs stratégies de codage d'une information donnée. Cette théorie ne nie pas l'existence d'un codage amodal, *i.e.* indépendants des modalités, dans la mesure où ce codage est une stratégie parmi d'autres. Cette diversité de stratégies semble être confirmée par l'existence de variations interindividuelles dans la modalité de codage d'une propriété donnée.

L'ergonomie s'appuie sur les démarches scientifiques utilisées en psychologie. Ces démarches sont utilisées pour étudier l'impact des modalités et de leurs combinaisons sur l'utilisateur, sur son comportement et, dans le cas de systèmes d'information, sur le déroulement de la communication avec ces systèmes [Le Bigot *et al.*, 2006, Fréard *et al.*, 2007]. Ces travaux mettent en avant que, pour des informations dont seul le format de présentation varie, *i.e.* le langage d'interaction est identique mais la sensibilité perceptive impliquée est différente (*e.g.* le langage naturel oral et le langage naturel écrit), la mémorisation des informations par les sujets diffère, le comportement de ces derniers varient (par exemple, nombre et longueur des mots, hésitations) et, par conséquent, le cours de communication en est modifiée. Ceci a été observée pour des présentations

monomodales [Le Bigot *et al.*, 2006] mais aussi pour des présentations bimodales [Fréard *et al.*, 2007]. Les psychologues et les ergonomes parlent d'effet de modalité [Lieury, 2005].

Étude comparative et prise de position Nous rapprochons les propriétés des unités informationnelles à présenter dans les systèmes mutlimodaux en sortie de la coopération possible, redondante ou complémentaire, entre modalités sensorielles. Si la redondance résulte de l'équivalence de deux modalités sensorielles pour une propriété donnée, il y a complémentarité quand une modalité sensorielle particulière est spécialisée - ou assignée - pour une propriété. ***Notre constat selon lequel la complémentarité sous-entend au moins une assignation semble donc être justifié dans l'intégration intermodale chez l'humain (et plus largement chez les êtres vivants).*** Notons que les types de coopération de TYCOON [Martin, 1995] s'appuient explicitement sur les travaux de Hatwell [Hatwell, 1994].

De plus, si l'on compare les conclusions en psychologie sur le codage des informations sensorielles avec la multimodalité en sortie des systèmes informatiques, cette diversité des stratégies, certaines (multi)modalement allouées et d'autres amodales, devrait se retrouver dans le traitement des entrées et dans la production des sorties. ***La gestion de la communication humain-machine se fait encore majoritairement de façon amodale, tout comme les premières théories psychologiques ont prôné pendant longtemps une intégration amodale des informations issues des sens chez l'humain et les animaux en général. Le recul des sciences humaines sur les limites de cette amodalité nous incite, comme nous l'expliquerons dans les chapitres suivants, à remettre en question l'amodalité de la détermination du comportement des machines.***

Enfin, ***les études en ergonomie sur l'effet de modalité confirment la difficulté à déterminer l'équivalence entre deux modalités ou deux combinaison de modalités***. En effet, si elles peuvent sembler l'être pour une unité informationnelle donnée, elles ne le sont pas en ce qui concerne l'intégration par l'utilisateur, allant jusqu'à modifier le déroulement de la communication. ***C'est pourquoi nous ne considérons pas, dans nos travaux, l'équivalence à un niveau modalitaire.***

2.2.4 Point de vue issu des sciences cognitives

Berthoz [Berthoz, 1997] ne parle pas de relations entre modalités à proprement parler. Mais le sens du mouvement qu'il propose mobilise des informations issues de plusieurs autres sens. Ceci l'amène à étudier l'intégration intermodale, pour reprendre le terme d'Hatwell [Hatwell, 1994].

Berthoz [Berthoz, 1997] rappelle qu'il est à présent établi que les messages issus des récepteurs sensoriels convergent à différents niveaux du système nerveux. Cette convergence est aisée à constater. Un exemple connu en est le mal des transports, où le mal-être provient d'un décalage entre système vestibulaire et système visuel [5]. De façon générale, il est admis que messages visuels et vestibulaires sont centralisés au niveau de

[5] Il s'agit d'une hypothèse et non d'une certitude, mais cette hypothèse n'a pas encore été contredite.

neurones bien spécifiques. Reste à déterminer dans quelle mesure il y a centralisation. Les sciences cognitives, à la suite de la psychologie et des neurosciences, admettent donc une intégration intermodale des informations perçues. Ceci sous-entend que si une présentation multimodale produite par un système informatique est perçue par l'utilisateur via différentes sensibilités, il y aura vraisemblablement intégration, centralisation, voire fusion au sens où l'entend Martin (*cf.* la section 2.1.2), des informations "sensiblement" allouées.

Berthoz souligne qu'on perçoit plus facilement la position d'un objet qu'on entend et qu'on voit, autrement dit où sensibilités visuelle et auditive sont toutes deux mobilisées. Il note qu'il existe un renforcement réciproque entre les entrées visuelles et auditives au niveau de chaque neurone où ces modalités[6] convergent. Ce renforcement est constaté même s'il y a un décalage entre les temps d'arrivée du message aux différents récepteurs sensoriels : les fenêtres temporelles seraient donc prises en compte. Des conclusions équivalentes portent sur les modalités visuelles et tactiles.

Berthoz signale que ce n'est pas la fusion des informations perçues sur différentes modalités qui est délicate, mais une fusion cohérente. En effet, les informations perçues présentent un certain nombre de différences : elles sont ambiguës, elles utilisent des systèmes de coordonnées différents, elles sont décalées dans le temps, elles ne couvrent pas les mêmes plages de vitesse de mouvement et elles sont bruitées (au sens de Shannon - *cf.* la section 1.1). Berthoz constate que les nombreux modèles mathématiques proposés pour décrire la combinaison des modalités n'ont pas assez pris en compte l'importance de la cohérence de cette combinaison. À ses yeux, ceci est une erreur car la constitution d'une cohérence du monde est nécessaire pour élaborer une représentation mentale de ce monde et surtout, de soi et des autres. Sans cette représentation mentale, il est impossible de communiquer car "il n'est certainement pas possible de construire une hypothèse interne de l'intention de l'autre si l'on a pas réussi à rendre cohérente la perception des relations de son propre corps avec l'environnement et avec toutes les informations qu'il contient".

Étude comparative et prise de position ***Les critères d'intégration*** (*cf.* la section 2.1.2) ***et les niveaux de composition*** (*cf.* la section 2.1.3.3) ***temporels et spatiaux sont en accord avec les fenêtres temporelles et spatiales identifiées pour l'intégration d'un même message arrivant aux récepteurs visuels et auditifs d'une part et visuels et tactiles d'autre part.*** Par ailleurs, appliquant la mise en avant par Berthoz de l'importance de la cohérence dans la fusion d'informations perçues via différentes modalités aux systèmes informatiques, ***nous concluons que ces derniers doivent, dans leurs échanges avec l'utilisateur, donner, produire, générer une représentation cohérente de leur état, de leurs "connaissances" et de leur fonctionnement***. Ceci implique que les concepteurs s'interrogent sur la façon dont les messages, *a fortiori* multimodaux, produits par la machine vont être transmis, perçus, intégrés. Par exemple, si on dé-corrèle auditif et visuel alors que l'auditif fait référence au visuel, la cohérence va poser problème, que ce soit au niveau

[6] Terme employé par l'auteur.

perceptif ou au niveau interprétationnel. ***Ce souci de cohérence nous mène donc à penser que certaines caractéristiques des modalités doivent être prise en compte lors de la détermination du comportement de la machine en tant que contraintes de présentation (cf. le chapitre 5).***

2.2.5 Synthèse : combinaison de modalités dans d'autres domaines

La combinaison des modalités dans les autres domaines que l'informatique a surtout été étudiée du point de vue de la perception. Les quatre utilisations de base des modalités couramment admises dans la communication humain-machine - *i.e.* la complémentarité, la redondance, l'assignation et l'équivalence - y sont identifiées (*cf.* les sections 2.2.2 et 2.2.3). On retrouve également plusieurs niveaux d'intégration, impliquant la prise en compte de critères temporels, spatiaux et modalitaires (*cf.* les sections 2.2.2, 2.2.3 et 2.2.4). Ce dernier niveau peut s'appliquer à des récepteurs perceptifs spécifiques, à rapprocher des dispositifs physiques, mais peut aussi s'opérer à l'échelle d'une modalité sensorielle.

Il nous paraît important de souligner que, même si l'étude de la communication humain-machine ne cherche pas à imiter la communication humaine, elle passe par des approches similaires. En particulier, l'intégration intermodale a longtemps été considérée comme étant amodale à partir d'un certain niveau de traitement (*cf.* la section 2.2.3). Comme cela sera exposé dans le chapitre 4, cette amodalité est aussi défendue dans certaines architectures de systèmes informatiques. Or les travaux en psychologie, en neurosciences et plus largement en sciences cognitives remettent de plus en plus en question cette amodalité, non en la niant totalement, mais en défendant l'existence de plusieurs stratégies d'intégration intermodale. Nous nous appuierons sur ce constat lorsque nous expliciterons, dans le chapitre 3, notre approche de la communication humain-machine en général et de la conception de systèmes d'information multimodaux en particulier.

Notre analyse de la combinaison de modalités dans les sciences humaines nous a amenés à positionner nos travaux par rapport à d'autres aspects. D'une part, l'intégration des informations sensorielles semble confirmer notre point de vue selon lequel l'efficacité de la complémentarité résulte d'une spécialisation sensorielle par rapport à une propriété donnée (*cf.* la section 2.2.3). D'autre part, l'appréhension du documentaire et du reportage filmé dans les sciences de l'information et de la communication (*cf.* la section 2.2.1) nous encourage à considérer que les avatars relèvent d'une multimodalité particulière où les modalités impliquées ne sont pas indépendantes. La complexité de leur intégration dans un système multimodal en sortie, notamment par rapport à leurs coopérations avec les autres modalités, nous conduit, au haut niveau de traitement sur lequel nous nous focalisons, à opter pour une multimodalité résultant d'une succession de fissions et d'allocations d'une ou plusieurs modalités à chaque unité informationnelle considérée. Nous ne nions pas l'existence de la fusion en sortie mais elle se fait en aval, plus précisément après l'affinement de l'utilisation des modalités, au moment de la concrétisation de la présentation à réaliser sur un dispositif physique donné - *i.e.* au niveau articulatoire/lexical.

2.3 Conclusion : notre approche de la combinaison de modalités

De cette étude de la multimodalité, nous concluons que la définition d'un système multimodal en sortie dépend de la définition donnée au terme "modalité". Notre exploration de la notion de "modalité" en informatique et en sciences humaines et expérimentales (*cf.* le chapitre 1) nous a conduits à identifier deux dimensions qui interviennent dans cette notion, qui, en sortie des systèmes, sont les capacités de perception des utilisateurs et les couples <dispositif physique, langage d'interaction> mobilisés par le système. Cherchant à traiter de la multimodalité à un haut niveau d'abstraction en se focalisant sur les aspects modalitaires et sémantiques (*cf.* la section 2.1.5), nous distinguons les deux éléments des couples <dispositif physique, langage d'interaction> : même si les dispositifs physiques permettent la mise en œuvre de la multimodalité en sortie à un bas niveau, ils doivent être pris en compte à un haut niveau d'abstraction car leurs capacités et leurs caractéristiques conditionnent les choix de la sortie multimodale, que ce soit de la réaction du système ou de sa présentation. Il en est de même pour les capacités de perception de l'utilisateur, voire ses capacités d'action (*cf.* les sections 1.3.6 et 1.4). L'identification de chacune de ces dimensions (capacités de perception, dispositifs physiques et langages d'interaction) ainsi que le poids équivalent que nous leur accordons dans notre approche centrée utilisateur, *i.e.* qui prend en compte l'impression de multimodalité qu'a l'utilisateur, nous conduit à considérer qu'un système multimodal en sortie répond à l'une des conditions suivantes :

- il est multi-langage (d'interaction), *i.e.* il mobilise plusieurs langages d'interaction ;
- il est multi-dispositif, *i.e.* il mobilise plusieurs dispositifs physiques ;
- il est multi-sensoriel, *i.e.* il mobilise plusieurs sensibilités chez l'utilisateur.

Pour notre étude, nous considérons la fusion et la fission selon plusieurs niveaux d'abstraction (*cf.* les sections 2.1.3.2 et 2.1.3.3). Toutefois, nous réduisons leur prise en compte dans la production d'une présentation multimodale à la succession de fissions à des hauts niveaux d'abstraction, ce qui n'empêche aucunement une fusion des unités informationnelles mono ou multimodalement allouées avant leur concrétisation, *i.e.* à un bas niveau d'abstraction (*cf.* la section 2.2.5). Nous nous focalisons sur les deux niveaux sémantique (permettant de passer d'une information à un ensemble d'unités informationnelles) et modalitaire (par allocation d'une ou plusieurs modalités à une unité informationnelle) *cf.* la section 2.1.3.3). Ce choix évite des enchaînements de fusions et de fissions à chaque niveau d'abstraction, enchaînements complexes à gérer et qui empêcheraient la restriction de la problématique de la multimodalité en sortie à un haut niveau d'abstraction. Au haut niveau d'abstraction qui nous intéresse, les unités informationnelles amodales ou mono/multimodalement allouées ne sont pas forcément élémentaires : cela dépend directement du choix des concepteurs. Nous considérons toutefois que, au moment de la concrétisation de la sortie, une éventuelle fusion ne peut s'opérer que sur des unités informationnelles élémentaires. Ceci revient simplement à considérer que, en sortie des systèmes, les fissions sont prioritaires sur les fusions et que ces dernières, quels que soient leur niveau, sont traitées le plus tard possible dans la chaîne de traitement. Cette approche ne tient pas compte de l'intégration d'avatars

dans les systèmes informatiques, qui impliquent nécessairement l'imbrication de fusions et de fissions à des niveaux amonts et une complexité accrue des coopérations entre modalités (*cf.* la section 2.2.1).

Que ce soit en informatique ou dans les autres domaines, l'exploitation envisageable des modalités semble être consensuelle. Sont donc admises l'assignation ou spécialisation d'une modalité ainsi que la redondance, la complémentarité et l'équivalence de deux modalités. Toutefois, travaillant sur la détermination de la sortie multimodale des systèmes à un haut niveau d'abstraction (*i.e.* sur les aspects de composition modalitaire et sémantique) et nous appuyant sur l'appréhension de la combinaison de modalités en informatique et en sciences humaines et expérimentales, nous jugeons adéquats les choix suivants : (1) nous préférons utiliser le terme d'assignation à celui de spécialisation, pour y inclure explicitement les cas de choix d'assignation ne découlant pas d'une spécialisation liée au type de l'information (comme c'est le cas de la photographie avec une modalité visuelle) (*cf.* la section 2.1.5) ou à une de ses propriétés (*cf.* la section pouvant conduire à une présentation complémentaire 2.2.3) ; (2) nous excluons l'équivalence modalitaire de façon à prendre en compte l'effet de modalité éventuel pour une même unité informationnelle présentée [Lieury, 2005, Karsenty, 2006, Le Bigot *et al.*, 2006] (*cf.* la section 2.1.5) ; (3) la complémentarité inclut à nos yeux toute redondance partielle et peut impliquer une assignation non explicite. Si ce dernier choix peut paraître réducteur, il ne l'est pas car il ne contraint pas un affinement de la coopération entre modalités à un niveau d'abstraction aval à celui que nous considérons, lors de la construction de la présentation ou de sa concrétisation (*cf.* la section 2.1.5).

Pour terminer, le constat de la similitude du traitement de la multimodalité en informatique avec l'évolution de l'appréhension de l'intégration sensorielle dans les sciences expérimentales, en particulier en psychologie (*cf.* la section 2.2.3), nous pousse à vouloir profiter du recul de cette dernière. C'est en partie l'identification de différentes stratégies d'intégration sensorielle dont certaines ne sont pas amodales qui nous pousse, dans les chapitres suivants (*cf.* le chapitre 4 et le chapitre 5) à défendre une modulation de sélection du contenu indépendante des modalités lors de la détermination du comportement des systèmes. Cette volonté, ainsi que celle portant sur l'anticipation de la communication à venir dans la conception de la sortie des machines (*cf.* la section 1.4), sont à la base de notre approche de la communication humain-machine que nous explicitons dans le prochain chapitre.

Chapitre 3

Notre approche : vers une communication multimodale naturelle

Les deux premiers chapitres ont cerné l'espace terminologique concernant la multimodalité et introduit la terminologie adoptée dans nos travaux. Focalisant sur la multimodalité en sortie pour des systèmes d'information grand-public, il convient maintenant d'étudier l'aspect naturel de la communication que nous défendons. C'est l'objet de ce chapitre, organisé comme suit : après avoir justifié notre volonté de travailler à une communication humain-machine naturelle, nous définissons la communication naturelle. Nous exposons ensuite des travaux qui œuvrent aussi pour une telle communication, avant d'expliciter nos objectifs et notre démarche sur lesquels reposent nos contributions présentées en deuxième partie de ce mémoire.

3.1 L'utilisateur et le contexte au centre de la communication naturelle

Quel que soit le paradigme adopté, toute communication humain-machine implique au moins deux agents dont l'un est artificiel (la machine) et l'autre naturel (l'utilisateur humain). Si des utilisateurs avertis, experts, formés, peuvent s'adapter aux machines, l'adaptation doit être du ressort des machines lorsqu'elles sont grand-public : en effet, rien ne garantit que ces derniers soient assez familiers des systèmes informatiques pour les utiliser sans surcharge cognitive. Si, pour des applications professionnelles ou spécialisées, il est possible de former, voire de formater, le comportement de l'utilisateur, de façon à l'adapter au fonctionnement de ces applications, ceci est impossible pour des systèmes d'information grand-public destinés à des utilisateurs dits "lambda". Contrairement à une pensée commune, un utilisateur lambda est loin d'être un utilisateur "standard" : c'est, par définition, un utilisateur que l'on ne peut caractériser, ne serait-ce parce que les différences interindividuelles de communication avec la machine

sont importantes [Karsenty, 2006]. Sans nier la nécessité de la prise en main nécessaire à toute utilisation d'un système informatique, nous pensons que le défi de la communication humain-machine est de réussir à permettre à *tout* utilisateur d'interagir avec les machines. Pour cela, nous considérons que ce ne sont pas les caractéristiques de la machine, mais celles de l'humain qui doivent principalement guider la conception. Nous focalisant sur la sortie des systèmes informatiques, nous pensons que la conception du comportement du système, *i.e.* sa réaction et la présentation de cette réaction, doit s'appuyer sur les dernières avancées en matière de fonctionnement humain, que ce soit au niveau de la perception sensorielle (à un bas niveau), de la sensibilité ou de l'intégration sensorielle (à un plus haut niveau) ou encore de l'attention et de la mémorisation. Or le fonctionnement humain est loin d'être complètement maîtrisé (*cf.* le chapitre 1 et le chapitre 2). C'est pourquoi nous pensons que l'un des rôles des systèmes informatiques est de contribuer à l'étude de ce fonctionnement, que ce soit par des expérimentations en laboratoire ou par des études d'usage. En effet, nous estimons que plus le fonctionnement cognitif humain sera connu, plus l'informatique pourra en tirer parti pour concevoir des systèmes adaptés.

Si l'utilisateur lambda est difficile à caractériser, les situations d'utilisation ou de communication lambda sont impossibles à cerner. La multiplication des terminaux et leur miniaturisation conduit à une multiplication des services, des usages et des attentes des utilisateurs. Les systèmes ne doivent plus seulement être adaptés aux utilisateurs pour accomplir une tâche, ils doivent aussi être adaptés aux situations. Une situation de communication regroupe non seulement les caractéristiques de l'utilisateur mais aussi les caractéristiques de l'environnement d'utilisation au sens large, qui inclut aussi bien les caractéristiques des dispositifs physiques ou de contexte ambiant que des régularités d'usage dans des environnements similaires [Horchani *et al.*, 2005]. Chacune des caractéristiques d'une situation de communication constitue une contrainte à prendre en compte pour le système. L'adaptation des systèmes aux situations de communication signifie que ces systèmes doivent prendre en compte ces contraintes et doivent, en particulier lorsqu'il s'agit de systèmes d'information, être accessibles à l'utilisateur quelque soit la situation de communication.

La notion d'"accessibilité" est généralement réduite à l'accessibilité sensorielle. Celle-ci est indissociable de l'accessibilité actionnelle - Karsenty parle de transparence [Karsenty, 2000] - qui doit assurer à l'utilisateur de pouvoir avoir accès aux fonctions du système et agir sur ce dernier. Nous parlons donc d'accessibilité sensoriactionnelle (en référence aux capacités de perception et d'action définies dans les chapitres précédents). Nous identifions en plus deux autres types d'accessibilité qui doivent aussi être assurées [Horchani *et al.*, 2005]. La première, l'accessibilité cognitive, est couramment appelée utilisabilité. La deuxième, l'accessibilité rhétorique, peut être rapprochée de la saillance ou de la pertinence [Landragin, 2004a, Landragin, 2004b] en fonction du niveau considéré d'intégration. Si l'accessibilité sensoriactionnelle est la condition *sine qua non* pour avoir accès à l'information présentée, l'accessibilité cognitive dépend de la charge mentale induite par l'utilisation du système et l'accessibilité rhétorique dépend de l'organisation des informations et de leur mise en avant différente dans la présentation de la réaction du système et des capacités d'action dont dispose l'utilisateur.

Un système doit réussir à combiner ces trois accessibilités tout en facilitant son adoption par l'utilisateur. Or, même si cette adoption - ou appropriation - et ces accessibilités sont liées, elles ne vont pas toujours de pair. En particulier, les utilisateurs peuvent ne apprécier certains systèmes, ce qui traduit une mauvaise appropriation, avec lesquels la performance est pourtant bonne [Fernández *et al.*, 2007]. Un équilibre entre les trois accessibilités identifiées et l'appropriation des systèmes doit donc être trouvé.

Les systèmes dits à initiative mixte s'appuient sur l'hypothèse que l'appropriation est facilitée quand l'initiative de communication est partagée entre humain et machine. Nous partageons cet avis et nous pensons que l'appropriation est d'autant plus grande si le système admet l'initiative mixte mais peut, à la suite de la proposition de Frohlich [Frohlich, 1996], choisir le paradigme de communication le plus adéquat, voire laisser ce choix à l'utilisateur. Le système peut alors être utilisé en tant qu'outil (paradigme actionnel) ou se comporter en partenaire de l'utilisateur (paradigme dialogique) (*computer-as-tool* / *computer-as-partner* [Beaudoin-Lafon, 2004]). Il ne s'agit pas de désorienter l'utilisateur en changeant le comportement de la machine en fonction de raisons qui peuvent lui sembler obscures, mais de le laisser contrôler la communication, y compris dans le rôle attribué à la machine (outil, partenaire ou médiateur, si l'on se réfère à [Beaudoin-Lafon, 2004]). Ceci revient à promouvoir ce que nous appelons une "communication naturelle", notion que nous explicitons dans la section suivante.

3.2 Communication humain-machine naturelle

L'aspect naturel de la communication humain-machine est souvent associé à l'entrée du système, autrement dit au comportement de l'utilisateur que le système peut traiter. Ainsi, plus l'utilisateur peut interagir de façon spontanée, pour reprendre le terme utilisé par Landragin [Landragin, 2004a], plus la communication sera considérée comme naturelle. Dans les systèmes de dialogue [Allen *et al.*, 2001, Landragin, 2004a], cela signifie que l'utilisateur peut se permettre des hésitations, des corrections et des interruptions du système à des fins confirmatives, complétives, correctives ou explicatives.

Néanmoins, l'aspect naturel s'applique parfois à la réaction du système, qui est alors qualifié de coopératif ou de convivial, voire de réaliste [Allen *et al.*, 2001]. L'éventuelle dimension naturelle de la présentation n'y est que rarement prise en compte, les systèmes de dialogue se concentrant sur la réaction du système plus que sur la concrétisation de cette réaction. Sadek [Sadek, 1999] considère que la convivialité d'un système transparaît dans sa réaction. Il identifie plusieurs caractéristiques qui concourent à la convivialité. Pour le cas des sorties, les caractéristiques les plus importantes sont les suivantes :

- la flexibilité de l'interaction, qui passe par une structure non figée de l'interaction, autrement dit de l'enchaînement prévu des sous-tâches. Dans les faits, il s'agit de laisser à l'utilisateur la possibilité de s'écarter du cours prévu du dialogue dès qu'il en ressent le besoin, par exemple pour compléter ou corriger sa requête dans le cas des systèmes d'information. Cette caractéristique s'avère primordiale, voire nécessaire, en cas de difficulté de communication. Selon [Allen *et al.*, 2001], cette flexibilité devrait aussi s'étendre au système, qui devrait être en mesure d'ignorer

une intervention de l'utilisateur si elle ne sert pas la tâche que l'utilisateur cherche à accomplir avec le système. Cette flexibilité sous-entend une indépendance entre modèle de dialogue et modèle de(s) tâche(s). Les systèmes résultants sont dit à initiative mixte ;
- la capacité de négociation qui, par exemple, dans le cas des systèmes d'information, laisse la possibilité au système d'entamer un sous-dialogue pour préciser ou clarifier une requête de l'utilisateur. Les demandes de clarification sont souvent considérées comme des réponses coopératives (*cf.* item suivant) ;
- la production de réponses coopératives, *i.e.* de réponses qui s'étendent de façon pertinente au-delà de la stricte question posée. À partir des types de réponses coopératives identifiées dans [Siroux *et al.*, 1989], Sadek identifie les suivantes comme étant les plus importantes :
  - o les réponses complétives ou sur-informatives, qui donnent des informations supplémentaires par rapport à la requête stricte de l'utilisateur ;
  - o les réponses correctives, qui informent l'interlocuteur que certains de ses présupposés sont caducs ;
  - o les réponses suggestives, qui proposent des solutions approchées mais qui ne répondent pas exactement à la requête stricte de l'utilisateur ;
  - o les réponses conditionnelles, qui tiennent compte de conditions non exprimées par l'utilisateur ;
  - o les réponses intensionnelles (dans le sens d'appartenance à une même classe, par opposition à "extension"), qui factorisent un ensemble de réponses par rapport à un élément sémantique particulier ;
- l'adéquation des formes ou styles de réponses, qui concerne la présentation adoptée et sa cohérence avec les demandes de l'utilisateur, le contexte d'interaction, les formes et styles utilisés en entrée, etc.

Le caractère coopératif, qui contribue à la convivialité, a également été caractérisé par Grice [Grice, 1975] en utilisant les quatre maximes suivantes :
- maxime de quantité : la contribution doit contenir autant d'information que nécessaire, mais pas plus. Elle pose la question du choix des réponses sur-informatives ;
- maxime de qualité : la contribution ne doit pas inclure d'information considérée comme fausse, ou pour lesquelles les preuves manquent. C'est le condition *sine qua non* à un comportement coopératif ;
- maxime de pertinence : la contribution doit être pertinente et à propos. La détermination de la pertinence d'une information ou pas se pose dans les systèmes d'information à partir du moment où ceux-ci ne se contentent pas de répondre uniquement à la requête stricte de l'utilisateur ;
- maxime de manière : les ambiguïtés doivent être évitées et le contribution doit être brève et ordonnée. Cette maxime peut sembler relever plus de la forme que du fond. Nous la relions à l'accessibilité rhétorique.

Ces caractéristiques concourent à la convivialité du dialogue, la convivialité participant au naturel de la communication. Elles concernent le contenu de la réaction du dialogue (sorties du système) et ne dépeignent pas la forme de ces réactions. Afin de caractériser aussi le caractère naturel de la présentation, en particulier de la présentation

visuelle, nous nous tournons vers les critères d'ergonomie qui relèvent plus de travaux en interaction homme-machine concernant l'utilisabilité d'un système interactif. L'utilisabilité, nous appelons aussi accessibilité cognitive, combine souplesse d'interaction, qui renvoie aux choix laissés tant à l'utilisateur qu'au système, et robustesse d'interaction, qui vise la prévention des erreurs et l'augmentation des chances de réussite de l'interaction. Parmi les critères liés à la souplesse de l'interaction, nous considérons que les suivants, au moins, ont un impact sur l'aspect naturel d'une communication :

- l'atteignabilité : elle correspond à la capacité du système à permettre à l'utilisateur de passer d'un état observable du système à un autre. Ce critère n'est pas spécifique à ce que le système admet en entrée ni à son comportement en sortie (réaction ou présentation) : il caractérise le tout. Ainsi peut-il transparaître à travers la proposition en sortie de moyens ou capacités d'action pour l'intervention suivante de l'utilisateur (par l'affichage d'un champ de saisie ou l'énonciation d'une invitation pour formuler une nouvelle requête "visuellement" - *i.e.* via le clavier - ou auditivement - *i.e.* oralement). C'est l'un des aspects inclus dans la notion de "transparence" [Karsenty, 2000] ;
- la multiplicité du rendu ou représentation multiple d'un même concept : c'est la capacité du système à fournir plusieurs représentations pour un même concept, que ce soit en termes de granularité (informations plus ou moins détaillées) ou de concrétisation (présentation différente). Si la multiplicité du rendu n'est pas spécifique à la sortie, c'est pour cette dernière qu'elle est la plus importante pour le système et la plus délicate car il doit faire le choix du rendu à présenter, que ce soit en termes de granularité ou de présentation ;
- la réutilisabilité des données d'entrée et de sortie : elle indique que les sorties du système peuvent être utilisées comme des données d'entrée. Elle permet donc de faire le lien entre la sortie du système et l'entrée suivante. Si elle est généralement définie par rapport aux données, elle peut aussi s'appliquer aux capacités d'action, proposées en sortie pour être utilisée en entrée, et se rapprocher alors de l'atteignabilité ;
- l'adaptabilité : ce critère correspond à la personnalisation du système sur intervention explicite de l'utilisateur. C'est un cas particulier de l'adaptation où l'utilisateur dirige celle-ci explicitement. Elle s'applique tant aux données présentées qu'au rendu de la présentation, ainsi qu'au choix des modalités ;
- l'adaptativité : ce critère décrit la capacité du système à s'adapter à l'utilisateur sans intervention explicite de sa part. C'est un cas particulier de l'adaptation centrée sur l'utilisateur mais dirigée par le système. Parce qu'elle peut, par le choix du contenu ou de sa présentation, engendrer une désorientation de l'utilisateur, elle est souvent considérée comme devant être prévisible, cette prévisibilité contribuant au naturel. Pour cela, il convient de viser une cohérence de la réaction et de sa présentation entre les situations similaires de communication[1] ;
- la migrabilité de tâche : elle fait référence à la capacité de délégation dynamique

[1]Notons que, pour certains auteurs, la distinction entre adaptabilité et adaptativité renvoie au moment de l'adaptation, à savoir à la conception ou à l'exécution.

de tâches entre le système et l'utilisateur. Autrement dit, c'est un changement dynamique de l'acteur responsable de l'accomplissement de la tâche, couramment appelée "initiative mixte" dans les systèmes de dialogue.

La souplesse d'interaction peut être rapprochée de la flexibilité de l'interaction : elle la détaille en focalisant notamment sur les capacités d'adaptation de la machine. Elle contribue autant à l'accessibilité sensoriactionnelle qu'à l'accessibilité cognitive, dans la mesure où elle intègre aussi une souplesse d'action et de perception.

La communication naturelle humain-machine est aussi caractérisée par des critères de robustesse de l'interaction, l'autre aspect de l'utilisabilité. Les critères suivants sont les plus importants pour une communication naturelle :

- l'observabilité : elle correspond à la capacité du système à rendre perceptible son état. Dans le cas des systèmes d'information, elle peut être rapprochée de la capacité du système de fournir les réponses coopératives, en particulier à la maxime de qualité de Grice. C'est un autre aspect inclus dans la notion de "transparence" [Karsenty, 2000] ;
- l'insistance : c'est la capacité du système à favoriser la perception de son état par l'utilisateur. Ce critère est donc lié à la pertinence et à la saillance de la présentation de la réaction d'un système d'information. Il rejoint l'accessibilité rhétorique que nous avons définie en la distinguant de l'accessibilité cognitive qu'est l'utilisabilité, ainsi que la maxime de manière de Grice ;
- l'honnêteté : ce critère fait référence à la capacité du système à rendre observable son état de façon à engendrer une interprétation correcte de cet état par l'utilisateur. L'honnêteté doit donc mener le système à informer l'utilisateur de son état, *i.e.* des informations qu'il connaît ou ne connaît pas et de son état mental, de façon interprétable, autrement dit accessible cognitivement, quelque soit l'état de l'interaction. C'est un troisième aspect inclus dans la notion de "transparence" [Karsenty, 2000]. L'honnêteté implique la prise en compte des quatre maximes de Grice. Notons qu'un système peut être honnête tout en refusant de divulguer certaines informations (sur l'emploi du temps personnel de collègues de travail, par exemple) ;
- la curabilité : ce critère renvoie à la capacité de l'utilisateur à corriger une situation non désirée. Aucune action de l'utilisateur ne doit donc être irréversible et une "sous-communication" (au sens de sous-dialogue) peut être entamée par l'utilisateur pour rectifier une requête dans le cas des systèmes d'information ;
- la prévisibilité : elle indique la capacité de l'utilisateur à prévoir, pour un état donné, l'effet d'une action. Cette caractéristique est nécessaire dans le mesure où il y a adaptation de l'interaction à la situation de communication. Comme nous l'avons expliqué précédemment, la prévisibilité doit plus relever de la cohérence par rapport à la métaphore de communication adoptée et de la continuité du comportement (réaction et/ou présentation) du système dans des situations de communication similaires que de la même réaction à une intervention de l'utilisateur. Ceci signifie qu'un système ne doit pas toujours se comporter de façon identique, mais que son comportement doit être conforme à ce à quoi l'utilisateur s'attend.

L'observabilité, l'honnêteté et la curabilité peuvent être rapprochées des réponses coopératives des systèmes d'information, telles que définies par Sadek pour caractériser la convivialité du dialogue.

Outre la convivialité et l'utilisabilité (souplesse + robustesse), un système qui vise à une communication naturelle doit être facile à appréhender, facile à s'approprier par l'utilisateur. La facilité d'appropriation et d'apprentissage font l'objet d'études en ergonomie et psychologie. Si la convivialité et l'utilisabilité sont difficiles à caractériser, l'appropriation l'est encore plus et des études d'usage semblent nécessaires. Toutefois, cette facilité d'appréhension est étudiée dans des travaux qui cherchent à reproduire des éléments typiques d'une communication inter-humaine naturelle, par exemple par l'incarnation du système dans un agent virtuel animé, ou encore par une reproduction de certaines caractéristiques physiques du monde réel. Nous présentons certaines de ces approches dans la section suivante.

3.3 Communication naturelle : les approches existantes

La communication humain-machine naturelle a surtout été étudiée en entrée : la première proposition d'application multimodale, le " mets ça là " combinant parole et geste [Bolt, 1980] va dans ce sens en s'inspirant de la communication humaine. De plus la manipulation directe dans les interfaces graphiques vise aussi le caractère naturel en entrée, cherchant à éviter à l'utilisateur d'apprendre un langage d'interaction et lui proposant de piloter le système d'une façon plus intuitive (reposant sur une métaphore du monde réel) et directe [Shneiderman, 1986]. De même les systèmes de dialogue ont misé sur le langage naturel plus ou moins contraint, en entrée et en sortie, et, pour certains, sur la coopération des systèmes avec les utilisateurs [Sadek, 1999]. Enfin plus récemment, des travaux se sont concentrés sur la prise en compte des comportements non verbaux, tels que les mouvements oculaires [Jacob, 1995], faciaux [Machrouh et Panaget, 2006] ou gestuels [Landragin, 2004a], comme un moyen de détection de l'attention portée par l'utilisateur au système (*attentive interfaces*) ou comme capacité d'action de l'utilisateur.

Toutes ces approches contribuent au caractère naturel de la communication, mais se concentrent sur les entrées du système. Pour les sorties, l'objectif est de rendre le comportement du système plus naturel pour les utilisateurs. Pour cela nous identifions deux approches où la dimension naturelle ne porte pas sur les mêmes aspects : reproduire la multimodalité de la communication inter-humaine par une incarnation du système via un avatar ou exploiter les capacités d'action propres au système pour rendre la réaction plus naturelle ou plus proche d'une perception dans le monde réel. Dans les deux cas, l'utilisation de plusieurs modalités pose la question des références intermodales [Bordegoni *et al.*, 1997].

Les travaux portant sur l'incarnation des systèmes via des avatars [Cassell *et al.*, 2000] cherchent généralement à accroître l'appropriation desdits systèmes par les utilisateurs. Des exemples de ces travaux dans le cas de systèmes d'information ont été proposés dans le cadre du projet SmartKom [Wahlster *et al.*, 2001]. Souvent, un système

personnifié grâce à un avatar est considéré comme un système multimodal et parler de système multimodal induit pour certains sa personnification via un avatar. Pourtant, comme nous l'avons déjà souligné dans le chapitre 2(section 2.2.1), les avatars relèvent d'une multimodalité particulière où les liens, relations, coopérations entre modalité visuelle, *i.e.* la gestuelle corporelle et faciale de l'avatar, et modalité auditive, *i.e.* le message oral émis, sont fortement liées et réciproquement contraintes si l'impression de naturel veut être garantie : ces liens forts nécessitent une coordination accrue entre les sous-messages présentés visuellement et auditivement, d'autant plus si la gestuelle d'accompagnement est sensée sembler aussi naturelle que possible à l'utilisateur [Kopp *et al.*, 2004].

Souvent associés aux travaux sur les avatars, des travaux se concentrent spécifiquement sur les capacités émotionnelles, empathiques ou non, des systèmes. Ils visent à tenir compte des émotions de l'utilisateur mais aussi à en provoquer en lui donnant l'impression que le système est émotionnellement sensible [Picard, 1995, Ochs *et al.*, 2007]. Si l'informatique affective (*affective computing*) se concentre généralement en sortie sur la représentation d'une émotion au niveau d'un avatar, il nous semble intéressant d'étudier aussi la présentation des informations pour engendrer certaines émotions chez l'utilisateur et traduire certaines émotions. En effet si la détection d'émotion semble pouvoir se faire au niveau de la forme de l'entrée émise par l'utilisateur (par exemple, la prosodie en vocal et la typographie en textuel), la forme de la sortie doit sans doute être en mesure d'influencer la perception que se fait l'utilisateur de l'état "émotionnel" du système.

Les systèmes multimodaux en sortie qui n'intègrent pas d'avatar exploitent plusieurs capacités d'action du système afin d'exploiter au mieux les capacités sensorielles et cognitives humaines. Ainsi de nouveaux codes, de nouveaux langages d'interaction non linguistiques [Benali Khoudja *et al.*, 2005], qui peuvent être à portée auditive [Brewster, 1997, Prewett *et al.*, 2006] et/ou tactile [Hoggan et Brewster, 2006, Prewett *et al.*, 2006] sont combinés de façon complémentaire ou redondante, voire simplement assignés en fonction de la situation de communication. Ces modalités de sortie doivent permettre d'exploiter au mieux les capacités sensorielles et cognitives humaines. Nous notons dans ces approches l'importance jouée par les dimensions spatiale et temporelle. La présentation des informations selon ces deux dimensions est alors coordonnée en fonction du but de présentation du système. Les systèmes considérés peuvent être multimodaux ou non. Dans le cas d'une présentation visuelle, il peut s'agir d'organiser spatialement et/ou temporellement les modalités image et texte [Novick et Lowe, 2005, Strothotte, 2007]. Les informations peuvent aussi être disposées spatialement en tenant compte de leur intérêt pour l'utilisateur comme dans le système de réalité augmentée présenté dans [Bell *et al.*, 2005]. Dans le cas d'une présentation auditive, la sonorisation peut être spatialisée pour accroître l'impression de naturel de la perception auditive [Begault, 1994]. L'auditif peut aussi être couplé spatialement et temporellement au visuel de façon à rendre la réalisation de la tâche plus naturelle : par exemple, [Baus *et al.*, 2007] propose une localisation sonore dans un espace de navigation physique.

Certains systèmes multimodaux en sortie visent à une adaptation dynamique du comportement du système à l'utilisateur en particulier ou à la situation en général, que

cette adaptation soit initiée par l'utilisateur (adaptabilité) ou par le système (adaptativité). L'adaptation d'un système multimodal en sortie qui repose sur des modalités pures qualifiées d'équivalentes peut impliquer un comportement mono-modal du système [Rousseau, 2006]. De plus il convient de noter que l'adaptation pour permettre une communication naturelle avec l'utilisateur n'est pas l'apanage des systèmes multimodaux [Thévenin, 2001, Sottet *et al.*, 2007]. En effet, l'adaptation peut porter aussi sur le contenu de la présentation et sa granularité. Ainsi l'adaptation porte à la fois sur le contenu et la forme de la présentation en fonction de la situation de communication qui inclut des caractéristiques de l'utilisateur et des caractéristiques de l'environnement d'utilisation, des dispositifs physiques disponibles répartis éventuellement sur plusieurs terminaux. Par exemple, les travaux sur l'accessibilité sensoriactionnelle ([Weiss, 2005]) reposent parfois sur une multimodalité alternée (d'une intervention de l'utilisateur à l'autre) et visent à garantir l'accès sensoriel aux informations aux utilisateurs handicapés ou utilisateurs placés temporairement dans une situation de communication handicapante (selon les caractéristiques de l'environnement physique ou matériel d'utilisation). De façon plus générale, l'informatique dite ubiquitaire [Weiser, 1994, UBICOMP, 2007] ou encore contextuelle [Moran et Dourish, 2001, CAC, 2007] a pour objectif d'adapter au moins la présentation aux contraintes issues de la situation de communication.

Nous avons présenté plusieurs approches qui visent à augmenter le caractère naturel de la communication humain-machine. Nous constatons des différences importantes entre ces approches qui sont, par conséquent, complémentaires. En particulier, certaines relèvent plus du paradigme dialogique de la communication et d'autres du paradigme actionnel. Dans ce contexte, il convient d'expliciter et de justifier notre approche.

3.4 Approche choisie et ses hypothèses

3.4.1 Caractéristiques et critères visés

Notre approche vise à favoriser le caractère naturel de la communication pour des systèmes multimodaux d'information grand public. Pour ce type de systèmes interactifs, nous avons choisi de cibler nos travaux en nous focalisant sur certaines caractéristiques et certains critères parmi ceux que nous avons introduits pour définir le caractère naturel de la communication (*cf.* la section 3.2). Ces caractéristiques et critères nous semblent les plus prioritaires pour permettre un comportement naturel des systèmes considérés.

Par rapport à la convivialité de ces systèmes d'information, notre objectif est de mettre l'accent sur l'adéquation des formes et styles de réponse ainsi que sur la capacité de négociation et la production de réponses coopératives, en particulier suggestives et sur-informatives. Chacune de ces caractéristiques implique la garantie de critères d'utilisabilité/accessibilité cognitive. En ce qui concerne l'adéquation des formes et styles de réponse, les principaux critères qui y concourent sont les suivants :

- la multiplicité du rendu : si le système n'est pas en mesure de présenter une unité informationnelle donnée de différentes façons, *i.e.* sur différentes modalités ou de façon plus ou moins concise, il ne pourra pas proposer une réponse en adéquation avec la situation de communication, tenant à la fois compte des contraintes sur le

contenu que des contraintes sur la forme ;
- l'adaptation au sens large, qui inclut l'adaptabilité et l'adaptativité : l'adéquation des formes et styles de réponse passe par la capacité du système à adapter son comportement, que ce soit de lui-même ou sur la demande de l'utilisateur.

En ce qui concerne la capacité de négociation et la production de réponses coopératives, nous considérons qu'elles passent principalement par la garantie des critères suivants :
- l'observabilité : la production de réponses coopératives ou l'initiative d'un échange de négociation implique que le système est en mesure de partager son état avec l'utilisateur. Comme nous l'avons noté précédemment (*cf.* la section 3.2), l'observabilité est directement liée à la capacité du système à présenter la quantité adéquate d'informations (maxime de qualité de Grice). Cette maxime intervient plus particulièrement dans le cas des réponses sur-informatives qui peuvent surcharger cognitivement l'utilisateur si elles sont trop nombreuses ;
- l'honnêteté : un système en mesure de produire des réponses coopératives est avant tout un système prédisposé à partager la majeure partie des informations qu'il connaît. L'honnêteté doit toutefois se faire dans le respect de l'espace privé des utilisateurs, par exemple lorsque le système d'information porte sur un partage d'agendas, d'annuaires, de réseaux sociaux, etc. ;
- la migrabilité de tâche, ou initiative mixte : la capacité de négociation passe par la capacité du système à entamer un échange de négociation. Mais la migrabilité de la tâche intervient également quand le système prend l'initiative d'anticiper les demandes à venir de l'utilisateur, en lui proposant des réponses suggestives qui ne correspondent pas exactement à sa requête ;
- l'atteignabilité : la coopération du système passe par la garantie des capacités d'action de l'utilisateur lors de la détermination de sa réponse. Par exemple, si la communication entre le système et l'humain se fait via un téléphone portable qui permet des présentations auditives et visuelles, le système doit toujours laisser la possibilité à l'utilisateur d'interagir par une action sur la présentation visuelle (sélection d'une personne dans une liste grâce au clavier, à la molette, à une pression tactile ou à un stylet) même si la présentation de la réponse est essentiellement auditive (liste des prénoms et noms des personnes qui occupent un bureau donné).

Nos travaux portent donc sur des systèmes multimodaux et coopératifs en sortie pour lesquels l'accent est mis sur certaines caractéristiques et certains critères qui favorisent une communication naturelle.Dans ce cadre, notre approche repose sur plusieurs hypothèses que nous justifions.

3.4.2 Hypothèses de travail

Nous avons souligné précédemment l'importance d'assurer l'accessibilité aux informations et aux moyens d'action de l'utilisateur sur le système aussi bien que l'appropriation du système par l'utilisateur (*cf.* la section 3.1). Nos hypothèses portent sur la façon dont ces deux aspects - accessibilité et appropriation - peuvent être garantis à l'aide des caractéristiques et critères que nous privilégions, ainsi que sur une certaine ap-

préhension de la notion d'"adaptation". Elles nous conduisent à une dernière hypothèse sur la façon dont la relation entre fond et forme doit être traitée dans les systèmes d'information coopératifs multimodaux. Ces différentes hypothèses vont dans le sens d'une anticipation du comportement de l'utilisateur et de la suite de l'interaction. Cette anticipation est inspirée de notre exploration de la notion de "modalité" dans les sciences humaines et expérimentales (*cf.* la section 1.4) et des travaux en ergonomie [Karsenty, 2006, Le Bigot *et al.*, 2006, Fréard *et al.*, 2007]. Grâce à elle, nous prenons en compte la boucle de communication humain-machine même si nous nous concentrons sur la sortie des systèmes.

3.4.2.1 Accessibilités et appropriation

En ce qui concerne les accessibilités sensoriactionnelle et cognitive, nous faisons le pari qu'elles sont garanties par les caractéristiques et les critères sélectionnés. En effet, l'accessibilité sensoriactionnelle est impossible si la multiplicité du rendu n'est pas assurée et si le système n'est pas observable et atteignable. Notons que l'adéquation des formes et styles de réponse repose en partie sur la garantie de l'accessibilité sensoriactionnelle. De plus, l'accessibilité cognitive renvoie directement aux critères d'utilisabilité. Même si nous ne les prenons pas tous en considération, nous considérons que ceux que nous avons sélectionnés sont suffisants, au moins dans un premier temps, dans le cas des systèmes d'information multimodaux.

Par ailleurs, nous pensons que l'accessibilité rhétorique peut être assurée, au moins en partie, grâce à ces mêmes caractéristiques et critères. En particulier, l'accessibilité rhétorique dépend de l'adéquation des formes ou styles de réponse et de l'adaptation du système à la situation de communication en termes de présentation mais aussi de réaction, autrement dit du comportement du système dans son ensemble. De notre point de vue, l'accessibilité rhétorique ainsi envisagée implique notamment que le système est en mesure de choisir le paradigme de communication le plus approprié pour une situation et un état d'interaction donnés : d'une certaine façon, c'est un assouplissement de la migrabilité de tâche ou de l'initiative mixte qui peut aller jusqu'à distinguer une boucle de perception-action parallèle à une boucle de traitement plus longue nécessitant une interprétation sémantique de l'intervention de l'utilisateur [Allen *et al.*, 2001, Gustafson, 2002].

De notre point de vue, l'appropriation du système par l'utilisateur dépend de l'équilibre entre ces trois accessibilités. Ne promouvant pas une prévisibilité entre l'action d'un utilisateur et la réaction du système, nous visons la cohérence des comportements du système dans des situations de communication proches ou semblables.

3.4.2.2 Adaptation

L'adaptation du système qui favorise son caractère naturel est centrale dans nos travaux, en particulier pour assurer l'adéquation des formes ou styles de réponse. Si l'adaptation, initiée ou non par l'utilisateur, a longtemps porté sur le rendu des interfaces graphiques, elle doit, à notre avis, pouvoir porter sur l'ensemble des modalités de

présentation. Par conséquent, un utilisateur doit pouvoir contraindre la présentation, en demandant au système de répondre selon une certaine modalité sensorielle ou en utilisant un dispositif physique particulier. Plus largement, la situation de communication dans son ensemble peut être prise en compte pour déterminer les modalités - sensorielles ou en termes de dispositifs physiques - utilisables (en fonction du bruit ambiant, si l'utilisateur est en réunion ou en train de conduire, etc.). Notons que les systèmes multimodaux tels que nous les considérons doivent donc être en mesure de proposer une présentation monomodale si la situation de communication s'y prête ou si l'utilisateur le demande.

Une telle adaptation peut conduire à des présentations difficiles d'accès :

- d'un point de vue sensoriactionnel (par exemple, si la modalité auditive orale est sélectionnée alors que la présentation porte sur une photographie ou une carte) ;
- d'un point de vue cognitif (par exemple, le texte à présenter visuellement est beaucoup trop long pour tenir sur un écran sans l'ajout d'ascenseurs de navigation) ;
- d'un point de vue rhétorique (par exemple, si la saillance de l'élément important d'une présentation multimodale repose sur la redondance de deux modalités, elle est perdue dans le cas où l'utilisateur demande une présentation exclusivement auditive et que la présentation n'est pas adoptée en conséquence).

Aussi, de façon à pouvoir prendre en compte des contraintes sur la présentation de la réponse du système tout en garantissant les accessibilités sensoriactionnelle, cognitive et rhétorique de l'utilisateur, nous considérons que lesdites contraintes doivent influencer la réaction du système, *i.e.* la sélection du contenu à présenter. Cette approche permet de concilier les contraintes de présentation avec la triple accessibilité aux informations et aux actions proposées par le système, de façon à ce que ces derniers soient perçues comme conviviaux même si la présentation est contrainte. De cette façon, dans le cas de contraintes de présentation explicitées par l'utilisateur, l'adaptation du comportement du système va au-delà de l'adaptabilité, dans la mesure où l'utilisateur exprime une préférence sur la présentation et que la prise en compte de celle-ci a un impact sur la réaction : l'utilisateur ne se prononçant pas par rapport à un souhait de réaction, mais par rapport un souhait de présentation, il n'est qu'indirectement à l'origine de l'adaptation de la réaction. Cette prise en compte des contraintes de présentation issues de l'utilisateur et/ou de l'environnement d'utilisation ne doit pas aller à l'encontre de l'utilisabilité et nécessite donc, à défaut de la prévisibilité du comportement du système, sa cohérence dans des situations équivalentes.

3.4.2.3 Liens entre fond et forme

Notre appréhension de l'adaptation du système, plus précisément de la portée de cette adaptation qui ne se limite pas à la présentation mais s'étend à la réaction, implique une remise en question de la séparation fond-forme courante dans un certain nombre de travaux (*cf.* le chapitre suivant).

Cette séparation fond-forme vise une plus grande modularité du code des systèmes pour augmenter la réutilisabilité, la modifiabilité et l'extensibilité du code (comme l'ajout d'une nouvelle modalité). Bien que séduisante pour ces aspects ingénierie de

l'interaction dans le cadre d'une conception centrée utilisateur, maintenir cette séparation dans notre approche revient à nier qu'au moins certaines contraintes de présentation devraient avoir un impact sur la réaction du système dans le but de respecter des critères d'utilisabilité et/ou de coopération. Nous détaillerons la nécessité que les contraintes de présentation influencent le choix du contenu dans le chapitre 4 et le chapitre 5.

Notons que notre approche s'appuie sur le constat de Berthoz (*cf.* la section 1.3.6) selon lequel corps et cerveau ne peuvent être dissociés : ignorer les liens étroits entre fond et forme revient à prétendre que la perception d'une information au niveau "corporel" n'a aucun impact sur son intégration et sa compréhension au niveau cérébral. Si l'on admet que corps et cerveau sont liés, que le cerveau doit tenir compte du corps, alors il est logique d'admettre que présentation et réaction sont liés et que la réaction doit tenir compte de la présentation.

Notre approche vise également à profiter du recul des sciences expérimentales sur l'intégration sensorielle qui conduit de plus en plus à l'acceptation de stratégies d'intégration qui, pour une part d'entre elles, ne sont pas amodales mais mono ou multimodales (*cf.* la section 2.3) : appliquant ce recul à la conception des systèmes informatiques, nous faisons l'hypothèse que la forme doit, dans certains cas, être prise en compte pour la détermination du fond.

3.5 Limitations du cadre d'étude

Les hypothèses que nous considérons ne restreignent pas le cadre d'étude de façon suffisante pour le temps imparti à un travail de thèse. En effet, bien qu'ayant limité les critères et les caractéristiques pris en compte pour la conception d'une communication naturelle ainsi que le type des systèmes considérés (*i.e.* les systèmes d'information grand public), la problématique à étudier reste vaste. En effet, ces hypothèses peuvent être appliquées à la plupart des approches existantes qui contribuent à une communication naturelle décrites dans la section 3.3. Par exemple, elles peuvent permettre de concevoir aussi bien un système intégrant un avatar, ayant un comportement émotionnel ou pas, qu'un système utilisé en situation de mobilité via un téléphone portable ou encore qu'un système d'information destiné aussi bien à des personnes mal-voyantes qu'à des personnes voyantes, combinant, suivant les cas, visuel et tactile ou visuel et auditif. Ne prétendant pas répondre à toutes les attentes en matière de communication naturelle multimodale en sortie et cherchant plutôt à adopter une approche complémentaire et innovante à celle des travaux existants présentés précédemment, nous limitons notre cadre en nous focalisant sur la conception de systèmes d'information grand public multimodaux dans lesquels l'intention de communication se répercute tant sur le choix de réaction que sur le choix de présentation, tout en assurant l'accessibilité sensoriactionnelle, l'accessibilité cognitive et l'accessibilité rhétorique de l'utilisateur. Nous cherchons, de cette façon, à redonner une place centrale à l'intention de communication qui sous-tend toute communication inter-humaine et qui ne devrait pas être perdue de vue dans la communication humain-machine (*cf.* la section 1.4). Étant donné cet objectif, nous limitons notre cadre d'étude selon les principes suivants.

Tout d'abord, nous avons limité les types de modalités et de combinaisons de modalités prises en compte. Comme nous l'avons justifié dans la section 1.4, en ce qui concerne les modalités, nous nous sommes concentrés sur de l'hypertexte simple visuel (*i.e.* nous n'avons pas pris en compte des cartes interactives par exemple) et de l'oral auditif. Nous rappelons que ce choix est motivé par le fait que seules les modalités visuelles, auditives et tactiles sont susceptibles de transmettre un nombre important d'informations sémantiques d'une part et par le constat que les terminaux grand public utilisés pour accéder aux systèmes d'information ne sont pas équipés de dispositifs physiques tactiles en sortie (*i.e.* haptiques) d'autre part. Toutefois, ce choix n'est pas limitatif et nous avons cherché, durant nos travaux, à faire des choix suffisamment ouverts pour être applicables aux modalités haptiques ou à d'autres modalités innovantes à forte portée sémantique. Par ailleurs, selon les terminologies adoptées au chapitre 1 et au chapitre 2 (*cf.* la section 1.4 et la section 2.3), nous parlons de multimodalité lorsqu'il y a combinaison éventuelle de deux modalités sensorielles, de deux dispositifs physiques et de deux langages d'interaction. Ce choix permet aussi de traiter des cas considérés comme multimodaux à la fois en sciences humaines et en informatique.

Ensuite, de façon à nous focaliser sur la sortie, nous avons choisi de laisser de côté l'interprétation des entrées produites par les utilisateurs et de la situation de communication qui peut engendrer une contrainte de présentation. Nous nous sommes donc limités à des cas où les utilisateurs expriment explicitement leurs contraintes de présentation. Notons toutefois que toute contrainte de présentation que nous considérons initiée par l'utilisateur pourrait tout aussi bien être le résultat d'une analyse contextuelle (par exemple, identification d'un utilisateur au volant et qui n'est pas à même de regarder son téléphone mobile ou son agenda électronique). Ainsi considérons-nous que plusieurs situations de communication différentes peuvent engendrer des contraintes de présentation identiques. Par conséquent, sous couvert de ne prendre en compte que les contraintes de présentation émanant de l'utilisateur, la prise en compte des contraintes émanant de l'environnement physique et/ou matériel voire contextuel de communication n'est pas exclue, les limites d'interaction et de présentation étant finalement les mêmes. Ce choix nous permet de ne focaliser que sur la présentation multimodale, de ne pas traiter l'interprétation des entrées ni la capture de la situation de communication qui constituent des axes de recherche en soi. Par exemple, la thèse de Rey [Rey, 2005] est consacrée exclusivement à la capture du contexte, *i.e.* à la capture d'une partie de la situation de communication.

Enfin, comme nous l'avons expliqué dans la section 2.3, nous avons choisi de privilégier la production de la réponse dans un système multimodal en sortie à des niveaux modalitaire et sémantique. À ces hauts niveaux d'abstraction, nous considérons que seules des fissions sont nécessaires, ce qui n'exclut pas une fusion des informations en amont, au niveau de chacun des dispositifs physiques impliqués. Ces choix évitent l'enchaînement complexe de fusions et de fissions et nous contraint à laisser de côté uniquement un type particulier de multimodalité, celui où un avatar est impliqué. En effet, les différentes modalités que constituent un avatar sont loin d'être indépendantes les unes des autres et nécessitent des fusions à des différents niveaux d'abstraction avec les autres modalités mobilisant les mêmes dispositifs physiques que l'avatar. Notre focalisation sur

les niveaux modalitaire et sémantique de la production d'une réponse multimodale nous ont conduit à concentrer nos travaux sur les contraintes de présentation qui sont à un niveau modalitaire général. Nous laissons donc de côté la prise en compte des contraintes de présentation à un niveau articulatoire ou lexical (par exemple, typographie ou couleurs utilisées en visuel, vitesse d'élocution en auditif). Néanmoins, nous soulignons que de telles contraintes de présentation pourraient modifier la saillance des informations présentées et entraînent, à terme, une nécessaire prise en compte de ces contraintes dites "de bas niveau" pour assurer une accessibilité rhétorique aux informations pertinentes.

Ainsi délimité, notre cadre d'étude porte sur la conception de systèmes d'information grand public multimodaux conviviaux. Plus particulièrement, ces systèmes doivent prendre en compte des contraintes de présentation émanant de l'utilisateur lors de la détermination de leur comportement aux niveaux modalitaire et sémantique. Cette prise en compte doit garantir l'accessibilité sensoriactionnelle, l'accessibilité cognitive et l'accessibilité rhétorique de l'utilisateur en assurant sa coopération, quitte à ce que la présentation mais aussi la réaction du système en soit modifiées. Les systèmes considérés n'incluent pas d'avatar et les modalités sensorielles mobilisées sont visuelles et auditives. Nous ne tenons pas compte d'une éventuelle sauvegarde des informations présentées, par l'utilisateur ou par le système, qui, bien qu'elle ne semble primordiale, nécessite un travail supplémentaire sur la mémorisation de l'historique de la communication et son exploitation(*cf.* la section 1.3.7), qui constitue un champ de recherche à part entière.

Dans ce cadre, nous avons choisi deux axes de contribution. Le premier axe concerne la conception de systèmes d'information : nous proposons une architecture qui intègre un composant dédié au choix conjoint de la forme et du fond. Le deuxième axe porte sur la facilitation de l'étude de l'appropriation de systèmes multimodaux coopératifs où fond et forme sont, au moins en partie, choisis conjointement : nous proposons une interface destinée au paramétrage des choix de réaction et de présentation par des non-informaticiens pour des systèmes respectant l'architecture proposée. Ces deux contributions, ainsi que leurs réalisations logicielles et leurs validations expérimentales, font l'objet de la partie suivante de ce manuscrit. Elles s'inscrivent dans une approche globale de la communication humain-machine naturelle alliant les deux paradigmes dialogique et actionnel de la communication. Cette approche globale fait l'objet du chapitre suivant.

Deuxième partie

Espace-solution : vers le choix conjoint des stratégies de dialogue et de présentation

Chapitre 4

Dialogue et interaction multimodale : vers une approche intégrée

Depuis presqu'une trentaine d'années, il est admis qu'il existe deux métaphores essentielles de l'interaction homme-machine présentées par Hutchins et ses coauteurs dans [Hutchins *et al.*, 1986] : la métaphore du monde (*model world metaphor*) implique une action directe de l'utilisateur sur l'interface alors que la métaphore conversationnelle (*conversation metaphor*) passe par l'utilisation du langage naturel. Communication actionnelle et communication conversationnelle sont reprises par [Frohlich, 1996] avec la notion de mode action et mode langage (*cf.* section 1.2.1) ainsi que par [Gustafson, 2002] avec les termes interface de parole et interface graphique. Dans [Beaudoin-Lafon, 2004], la distinction fondamentale entre ces deux métaphores repose sur le rôle occupé par le système, outil dans un cas et partenaire dans l'autre. Dans la suite de ce manuscrit, nous utilisons les termes interaction et dialogue pour distinguer ces deux paradigmes, en référence aux noms donnés aux systèmes résultants.

La plupart des systèmes informatiques mêlent ces deux approches, ne serait-ce simplement parce que certaines informations et certaines fonctionnalités sont plus adaptées à une métaphore qu'à une autre, comme le souligne Frohlich [Frohlich, 1996]. Mais, malgré le fait que nous constatons dans les systèmes au quotidien un mélange de dialogue (même si il est très simple) et d'interaction, ces deux paradigmes correspondent à des communautés de recherche distinctes avec leurs conférences propres. Nous notons néanmoins des études à l'intersection. Nos travaux sur la multimodalité en sortie s'inscrivent dans cette tendance de rapprocher les deux métaphores : dès le chapitre précédent, nous définissons le caractère naturel de la communication multimodale en considérant des critères de convivialité du dialogue (paradigme dialogique) et des critères d'utilisabilité de l'interaction (paradigme actionnel).

Dans ce chapitre, nous étudions d'abord les résultats issus des dialogues multimodaux puis ceux de l'interaction multimodale en mettant l'accent sur les sorties multimodales (du système vers l'utilisateur). Nous concluons ce chapitre en soulignant les

approches qui allient les deux métaphores. Nous nous limitons à notre cadre d'étude défini dans les chapitres précédents, en excluant notamment les travaux où la multimodalité repose sur les avatars et en restreignant les modalités prises en compte aux modalités visuelles et auditives qui sont les seules à ne pas nécessiter de dispositifs spécifiques pour les systèmes d'information grand-public sur lesquels portent nos travaux. Notons, dans [Foster, 2002], un recensement plus large des travaux sur la multimodalité en sortie.

4.1 Dialogue humain-machine

Dans cette section, nous rappelons d'abord les principes fondateurs du dialogue humain-machine puis nous nous focalisons sur les systèmes de dialogue multimodaux avant de conclure par une synthèse.

4.1.1 Principes et modèles fondateurs du dialogue humain-machine

4.1.1.1 Approches du dialogue humain-machine

L'informatique, dès ses débuts, a été imprégnée de la fascination de l'humain pour lui-même. De là est née l'intelligence artificielle, avec l'espoir de modéliser une machine aussi perfectionnée, aussi intelligente, que l'humain. Parce qu'il est souvent considéré que la communication langagière est une capacité exclusivement humaine, Turing a établi qu'une machine intelligente serait en mesure de faire croire à un sujet que c'est avec un humain qu'il dialogue [Turing, 1950]. Très tôt, des travaux ont été entamés dans ce sens pour permettre aux humains de communiquer en langage naturel avec des machines. Issu de ce courant, ELIZA[1] [Weizenbaum, 1967] est l'un des premiers systèmes de dialogue en langage naturel. Simulant un psychothérapeute, il a été conçu pour réussir le test de Turing. Il est pourtant loin d'être intelligent. En effet, son principe de fonctionnement est simple : reconnaissant des structures de phrases-types utilisées par les sujets, le système ELIZA en extrait des mots-clés qu'il utilise dans des énoncés à trous. Il n'a donc pas accès à la dimension sémantique des messages échangés.

Depuis, le dialogue humain-machine a été envisagé de différentes façons, en fonction, notamment, de l'objectif du système et du type d'interaction souhaité avec l'utilisateur (par exemple, systèmes fournissant des renseignements à des utilisateurs via une communication en question-réponse). Reprenant [Sadek, 1999], nous nous concentrons ici sur les deux approches les plus courantes, à savoir l'approche structurelle d'une part et l'approche différentielle - dite aussi "orientée plan" - d'autre part. Le lecteur intéressé pourra se reporter à [Lehuen, 1997], [Gustafson, 2002] ou à [Bui, 2006] pour une revue plus large. Notons que les deux approches présentées ci-dessous sont souvent mêlées et qu'elles sont clairement distinctes du "dialogue en ligne de commande" où le terme "dialogue" est utilisé abusivement.

La première approche peut rappeler celle adoptée pour ELIZA. C'est en effet la plus ancienne. Selon elle, tout dialogue est structuré. Cette structure peut être extraite

[1] Pour tester ELIZA : http ://library.thinkquest.org/18242/eliza.shtml

en utilisant des informations linguistiques et pragmatiques. En s'appuyant sur ces informations, les systèmes de dialogue conçus selon cette approche visent à identifier la finalité d'un dialogue en cours pour pouvoir répondre à l'utilisateur. De tels systèmes sont généralement implémentés avec des diagrammes à états finis ou avec des grammaires formelles. Ils ne sont pas souples dans la mesure où les énonces de l'utilisateur doivent être formatés. La production de la sortie est proche de celle utilisée dans ELIZA, s'appuyant sur des énoncés à trous. Si, initialement, l'interaction avec des systèmes de dialogue suivant l'approche structurelle était dirigée par les systèmes, l'initiative mixte a été peu à peu introduite, permettant à l'utilisateur une plus grande liberté d'interaction. Les approches probabilistes qui permettent d'optimiser le dialogue dans une certaine optique, par exemple en identifiant par apprentissage automatique les stratégies de dialogue préférées par un utilisateur donné, relèvent aussi de l'approche structurelle.

La deuxième approche est appelée tour à tour planification ou différentiation. Elle part du principe que les actes, dialogiques ou non, sont motivés par des buts et sont planifiés pour atteindre ces buts. En particulier, chaque intervention de l'utilisateur est la réalisation d'actes communicatifs, dits aussi "de dialogue" ou "de langage", produits pour tenter d'atteindre un ou des buts donnés. Le système doit chercher à identifier ces buts sous-jacents de l'utilisateur et à y répondre, en prenant aussi en compte ses propres buts. Utilisateur et système sont considérés comme étant des agents. Ils sont dits rationnels ou encore BDI (pour "*Belief, Desire, Intention*") car ils infèrent leurs buts en fonction de leurs croyances, désirs et intentions à un instant donné. Croyances, désirs et intentions sont des attitudes mentales susceptibles d'évoluer dans le temps. L'ensemble des attitudes mentales d'un agent constitue son état mental. Si le système choisit d'adopter pour siens les buts de l'utilisateur et d'y contribuer au mieux, il est dit coopératif. On retrouve la notion de "système-partenaire" évoquée par [Beaudoin-Lafon, 2004]. Utilisateur et machine peuvent aussi être dans une démarche de négociation [Lehuen, 1997] ou adopter d'autres comportements [Caelen, 2003].

4.1.1.2 Architecture des systèmes de dialogue humain-machine

Les systèmes de dialogue humain-machine comprennent traditionnellement trois ou cinq modules principaux, suivant que le dialogue est oral ou écrit. La figure 4.1 présente l'architecture classique d'un système de dialogue humain-machine en langage naturel oral. Le module de reconnaissance permet de passer du message en langage naturel oral produit par l'utilisateur à un message textuel. Le module d'interprétation permet de passer du message textuel en un message compréhensible par le module de gestion du dialogue. Le module de gestion du dialogue, appelé aussi "gestionnaire du dialogue", "contrôleur de dialogue" ou "*dialogue manager*" permet d'appliquer l'approche du dialogue choisie et produit le message de réponse à présenter à l'utilisateur. Le module de génération permet de passer du message produit par le module de gestion du dialogue à un message textuel. Le module de synthèse permet la production d'un message en langage naturel oral audible par l'utilisateur à partir du message textuel. Les modules de reconnaissance et de synthèse n'apparaissent pas dans le cas des systèmes de dialogue en langage naturel écrit.

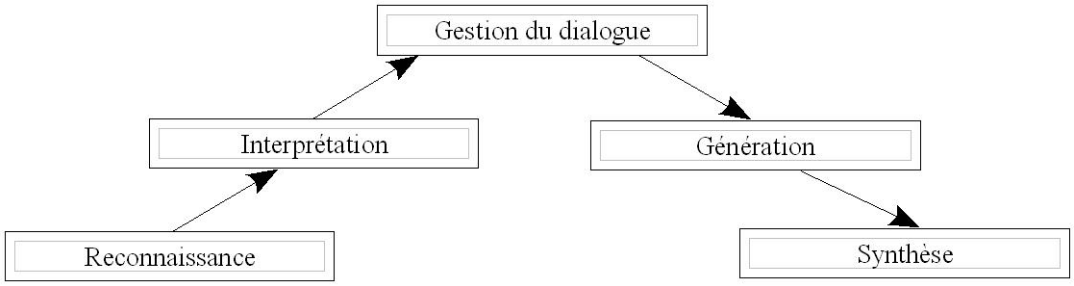

FIG. 4.1 – Architecture des systèmes de dialogue humain-machine en langage naturel oral

L'approche de dialogue choisie par les concepteurs du système est implémentée au niveau du module de gestion du dialogue. Par conséquent, les messages reçus et produits par ce module dépendent de l'approche choisie. Si c'est une approche différentielle coopérative qui est adoptée, ce module reçoit du module d'interprétation le ou les actes communicatifs identifiés dans le message de l'utilisateur et transmet au module de génération le ou les actes communicatifs à présenter à l'utilisateur. C'est ce module qui détermine le but sous-jacent de l'utilisateur ainsi que la réponse à lui donner pour qu'il/elle atteigne ce but. Cette réponse est inférée de l'état dit "mental" du système, de l'état mental supposé de l'utilisateur et de l'historique du dialogue. L'historique, appelé aussi "contexte du dialogue" ou "modèle du contexte", regroupe au minimum l'ensemble des objets évoqués au cours du dialogue. Son maintien est à la charge du module de gestion du dialogue.

Le module de gestion du dialogue intègre le modèle de dialogue correspondant à l'approche choisie. Ce modèle, appelé aussi "modèle d'interaction", peut être implicite ou explicite. Généralement, les systèmes relevant de l'approche structurelle comprennent un modèle de dialogue implicite et ceux orientés plan intègrent un modèle de dialogue explicité au niveau des mécanismes d'inférence.

Les systèmes de dialogue couramment conçus ne sont pas des systèmes dont la finalité est le dialogue - comme c'est le cas d'ELIZA. Ils doivent permettre à l'utilisateur d'accomplir des tâches plus ou moins complexes et d'accéder à des données. Ces données sont stockées dans un modèle du domaine. Celui-ci est aussi parfois appelé "modèle de tâche", bien qu'il ne spécifie pas les différentes étapes nécessaires à l'accomplissement de la tâche pour lequel le système a été conçu. Ces étapes ont longtemps été et sont encore souvent sous-entendues de façon implicite dans le modèle de dialogue et/ou dans le modèle du domaine. La notion de "modèle de tâche" dans l'approche dialogique renvoie donc à deux appréhensions différentes de la tâche pour lequel le système a été conçu : dans un cas, elle est considérée comme étant rattachée aux données ; dans l'autre cas, elle est considérée comme susceptible d'influencer le dialogue. Dans les deux cas, et quelle que soit l'approche du dialogue choisie, elle est plus souvent implicite qu'explicite. On peut donc conclure que la partie générique d'un module de gestion du dialogue comprend le modèle de dialogue alors que la partie applicative comprend le modèle de tâche.

Les systèmes de dialogue humain-machine intègrent traditionnellement un dernier modèle utilisé pour la production de leurs réponses. Il s'agit du modèle de l'utilisateur qui, dans le cas d'une approche différentielle, est sous-entendue dans l'état mental du système. Plus précisément, une partie de l'état mental du système porte sur ce qu'il croit savoir de l'état mental de l'utilisateur. À cela s'ajoute les informations contenues dans l'historique de dialogue sur les éléments évoqués durant les interactions passées.

Les systèmes de dialogue humain-machine ont longtemps privilégié une communication basée sur le langage naturel. Dans ce cadre, la génération d'une réponse passe au minimum par la détermination de la réponse à apporter au niveau d'un module de gestion du dialogue et par la génération de cette réponse en langage naturel au niveau d'un module de génération. S'y ajoute éventuellement une synthèse orale réalisée par un module de synthèse. Ce processus de production d'une réponse nécessite un certain nombre de modèles : un modèle de dialogue qui détermine la façon dont le dialogue est géré ; un modèle du domaine qui comprend les données applicatives connues par le système ; un historique du dialogue qui regroupe les éléments évoqués lors des interactions précédentes par l'utilisateur ou par le système ; un modèle de l'utilisateur qui correspond à ce que le système sait sur l'utilisateur, en particulier sur ses connaissances ; éventuellement un modèle de la tâche qui spécifie les différentes étapes permettant à l'utilisateur d'accomplir la tâche pour laquelle il utilise le système.

Objet de la section suivante, ces modules et leurs modèles ont été adaptés pour permettre aux systèmes de dialogue d'être multimodaux.

4.1.2 Architectures de systèmes de dialogue multimodaux en sortie

L'étude de la multimodalité dans les systèmes de dialogue humain-machine a surtout été centrée sur l'entrée. En effet, l'observation de la communication humaine a conduit à la conclusion que l'humain communique de façon multimodale en combinant geste et oral [Landragin, 2004a]. Par conséquent, une machine doit être capable de comprendre cette multimodalité au niveau de son entrée. En ce qui concerne la sortie, l'intégration de la multimodalité dans les systèmes de dialogue humain-machine a été favorisée ces dernières années par les avatars. Comme nous l'avons expliqué, nous ne considérons pas les systèmes où la multimodalité réside exclusivement dans les avatars. Le lecteur intéressé pourra se référer à l'ouvrage de référence [Cassell *et al.*, 2000], aux travaux du groupe de travail sur les agents conversationnels animés [GT ACA, 2007], ou encore à [Gustafson, 2002] qui détaille l'intérêt des avatars dans la communication humain-machine.

Nous notons aussi des travaux en multimodalité pour la communication humain-robot, comme dans ceux autour de [Nabaztag, 2006] et ceux menés par le laboratoire de robotique mobile et de systèmes intelligents de l'Université de Sherbrooke [LABORIUS, 2007]). Nous laissons également de côté ce type de systèmes car la quantité d'information susceptible d'y être présentée est limitée comparativement aux systèmes d'information qui nous intéressent. De plus, ces systèmes impliquent une prise en compte de la dimension et du positionnement physique du système plus complexe et qui ne nous intéresse pas dans le cadre de nos travaux.

Les systèmes présentés par la suite comprennent donc des présentations auditives et visuelles, avec une dimension langagière forte. S'il n'est pas toujours facile de décréter qu'un système est orienté dialogue plutôt qu'interaction, nous avons tranché en fonction de l'existence d'un modèle de dialogue issu d'une des approches présentées précédemment et de l'importance accordée au langage naturel oral. Nous n'avons pas cherché à proposer une liste exhaustive de ces systèmes et des architectures sous-tendues mais plutôt de présenter des systèmes récents aux architectures différentes et complémentaires.

4.1.2.1 Architecture de TRIPS

Le système TRIPS [Allen *et al.*, 2001] est un système orienté planification qui exploite en sortie des messages auditifs et visuels. Plus précisément, les réponses données à l'utilisateur allient langage naturel oral et affichage à l'écran intégrant des pointages. L'achitecture du système est présentée à la figure 4.2.

Motivés par la mise en œuvre d'un dialogue humain-machine plus naturel, les auteurs ont choisi de rompre avec les trois étapes classiques d'interprétation, de gestion du dialogue et de génération typique des systèmes de dialogue classiques de la figure 4.1. L'architecture proposée distingue bien trois traitements à réaliser par le système, mais elle ne sont pas réalisées de façon séquentielle : le système les réalise parallèlement. Pour que cela soit possible, plusieurs modèles intermédiaires sont introduits entre l'équivalent des modules d'interprétation, de gestion du dialogue et de génération de la figure 4.1. Nous décrivons cette architecture en nous focalisant sur les équivalents des modules de gestion du dialogue et de génération.

Une partie de la gestion du dialogue est à la charge d'un agent qui gère le comportement du système (*behavioral agent*). L'interprétation contextuelle de l'énoncé produit par l'utilisateur lui est transmise par le gestionnaire d'interprétation (*interpretation manager*, équivalent du module d'interprétation). Ceci sous-entend que l'interprétation est entièrement faite en amont de l'agent comportemental et que le gestionnaire d'interprétation accède au contexte du discours (*discourse context*). L'agent comportemental planifie le comportement du système en s'appuyant sur ses buts, ses obligations, les changements de son état mental ainsi que l'énoncé et les actions de l'utilisateur. Il gère la résolution de la requête de l'utilisateur qui détermine le comportement du système. Les actions communicatives ou collaboratives qui en découlent sont envoyées au gestionnaire de génération.

Le gestionnaire de génération (*generation manager*) regroupe une partie du module de gestion du dialogue et une partie du module de génération. En fonction des actions communicatives et collaboratives reçues du composant de dialogue, il planifie le contenu des énoncés et des affichages. Pour cela, il s'appuie sur les obligations dialogiques stockées dans le contexte du discours. Il n'est pas synchronisé au gestionnaire d'interprétation mais à l'agent comportemental et au contexte du discours. Il peut donc y avoir génération alors que l'interprétation n'est pas terminée : ceci rend l'interaction humain-machine plus naturelle en permettant par exemple au système de confirmer des sous-énoncés produits par l'utilisateur avant que ce dernier n'ait terminé son interven-

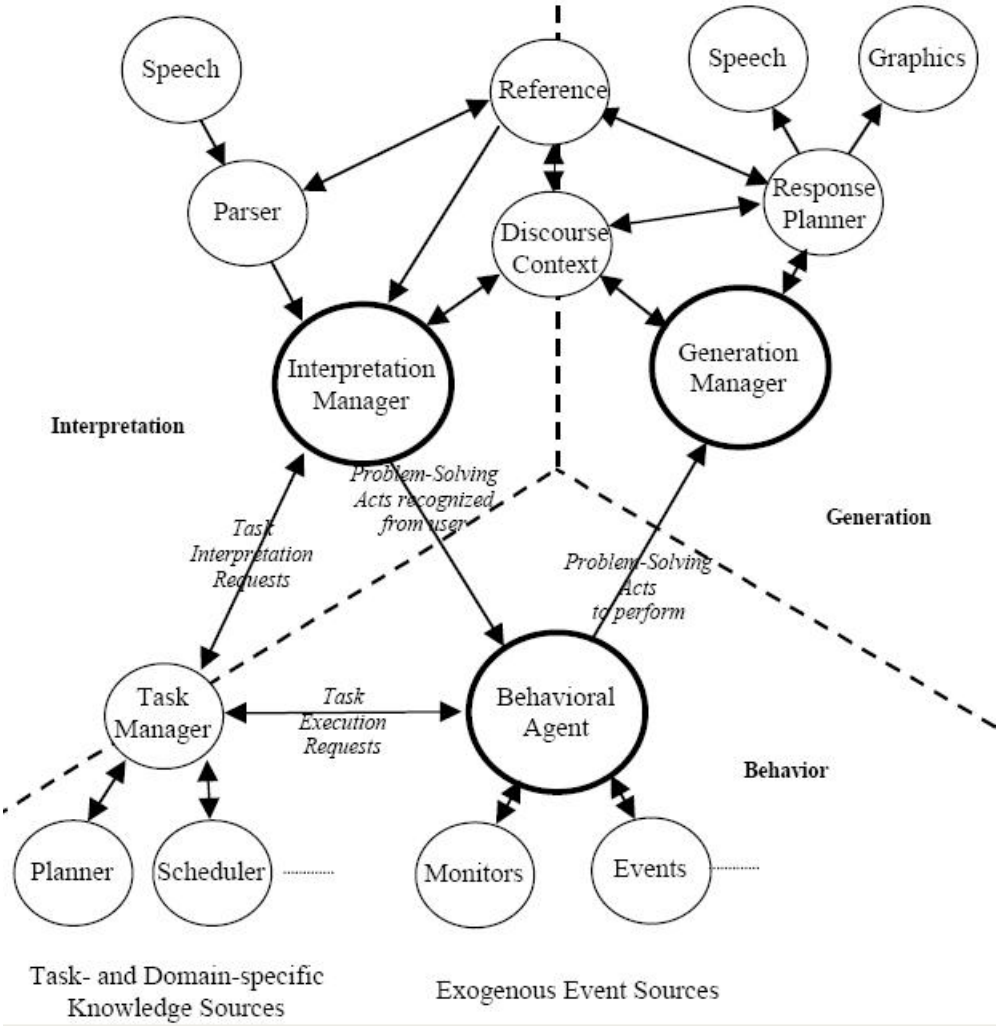

FIG. 4.2 – Architecture de TRIPS (extrait de [Allen *et al.*, 2001])

tion. Le gestionnaire de génération peut aussi ajouter des informations rhétoriques non indiquées par l'agent comportemental dans la réponse à produire : c'est pour cela que nous considérons qu'il remplit partiellement des fonctions propres au module de gestion du dialogue. Les actes qu'il planifie sont envoyés au planificateur de réponses (*response planner*). Il est informé des éléments de réponses concrétisées en aval et il complète le contexte du discours en conséquence.

Le planificateur de réponses (*response planner*) planifie la réponse du système et coordonne sa réalisation à partir des actes à produire envoyés par le gestionnaire de génération. Il intègre une partie du module de génération dans la mesure où l'utilisation d'une autre modalité que le langage naturel oral nécessite une planification supplémentaire. Il correspond aussi au module de synthèse étendu d'une dimension graphique. Pour chaque concrétisation d'acte réalisée, le planificateur de réponses avertit le gestionnaire

de génération qui maintient le contexte du discours.

Concernant les modèles utilisés dans TRIPS, le modèle de tâche(*task manager*) correspond au modèle du domaine. Il détaille les éléments nécessaires à la résolution de requêtes applicatives. Il répond aux requêtes sur les objets du domaine ou de la tâche émises par l'agent comportemental ou par le gestionnaire d'interprétation en appliquant les actes de résolution de requêtes génériques. Il gère la planification des sous-tâches nécessaires à l'accomplissement de la tâche de l'utilisateur. Ces sous-tâches constituent une partie des obligations dialogiques gérées au niveau du contexte du discours.

Le contexte du discours est un historique du dialogue étendu. En effet, en plus des éléments intervenus dans le dialogue antérieur, il gère les tours de dialogue, les obligations dialogiques non accomplies et les sous-dialogues qui en découlent. Plus précisément, il comprend l'état du dialogue, un historique du dialogue et les obligations dialogiques en cours, les éléments saillants [2] dans le dialogue en cours ainsi qu'une représentation de la structure et de l'interprétation du dernier énoncé de l'utilisateur. Notons que le contexte du discours n'est pas mis à jour uniquement par l'agent comportemental : celui-ci étant réduit par rapport au module classique de gestion du dialogue au bénéfice des gestionnaires d'interprétation et de génération, ces derniers ont accès directement au contexte du discours et le maintiennent tout autant.

Cette architecture distingue explicitement dialogue et tâche. Pour les auteurs, cette distinction est nécessaire pour permettre une génération indépendante de l'interprétation. Pour la même raison, la gestion du dialogue est en partie faite par les gestionnaires d'interprétation et de génération. Ces deux caractéristiques permettent au système d'avoir un comportement de type perception-action qui ne fait pas intervenir le module de gestion du dialogue. La gestion du dialogue n'est donc pas centralisée comme c'est classiquement le cas dans les systèmes de dialogue humain-machine. La planification de la réponse est répartie entre l'agent comportemental, le gestionnaire de génération et le planificateur de réponses. Ce dernier étend le module de synthèse classique en traitant aussi bien la synthèse orale que la réalisation graphique.

4.1.2.2 Architecture proposée par Gustafson

Dans [Gustafson, 2002], dix ans de travaux sur les systèmes de dialogue multimodaux orientés plan sont présentés. Ces systèmes incluent des modalités visuelles (un avatar, une carte et le langage naturel écrit) et une modalité auditive (le langage naturel oral). Cinq systèmes - correspondants aux cinq avatars Waxholm, Gulan, August, AdApt et Pixie - sont présentés. Nous nous concentrons ici sur la gestion de la multimodalité en sortie. Pour Gustafson, la production de sorties dialogiques multimodales doit se faire en parallèle des entrées et consiste essentiellement dans le choix des canaux de communication à utiliser et de la répartition du message à présenter sur ces canaux. Pour cela, il propose l'architecture présentée à la figure 4.3.

La particularité de cette architecture est d'être focalisée sur la gestion des entrées-sorties, *i.e.* sur le module d'interprétation et le module de génération dans l'architecture

[2] Terme employé par les auteurs.

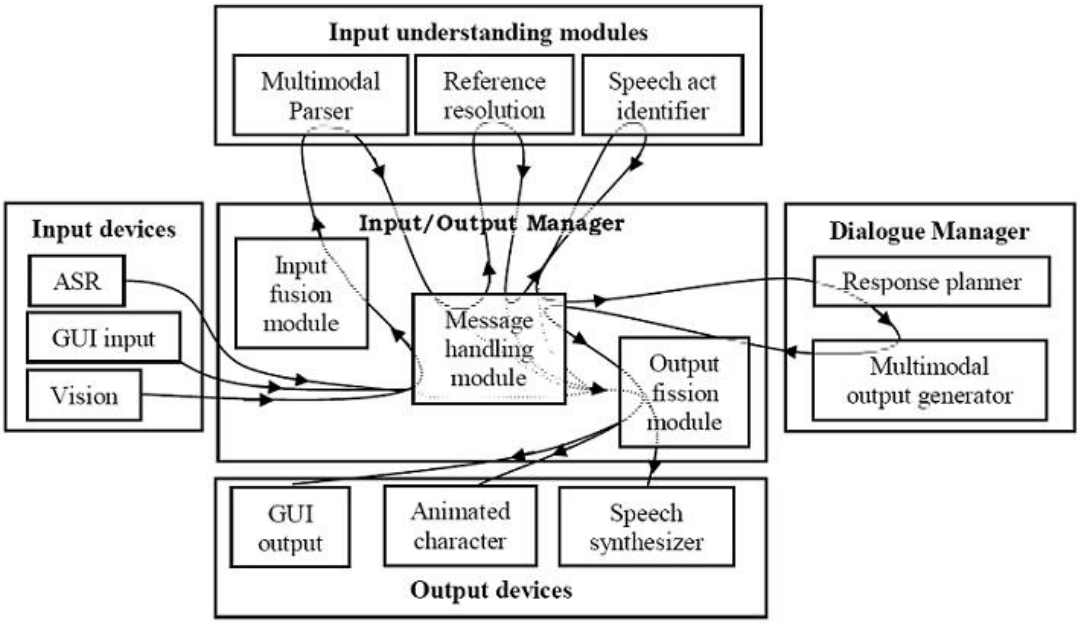

FIG. 4.3 – Architecture des systèmes multimodaux proposées par Gustafson (extrait de [Gustafson, 2002])

classique des systèmes de dialogue humain-machine. Les dispositifs de sorties correspondent au module de synthèse, auquel s'ajoutent la gestion des deux canaux visuels considérés, à savoir l'affichage graphique (carte + texte) et l'animation de l'avatar. C'est parce que l'architecture est centrée sur la gestion des entrées-sorties que, comme dans TRIPS [Allen *et al.*, 2001], des sorties peuvent être produites parallèlement à la perception des entrées. L'essentiel de ces sorties concerne l'animation de l'avatar.

Même si l'architecture est centrée sur la gestion des entrées-sorties, entrées et sorties sont clairement séparées et ne sont pas en contact direct. La gestion du dialogue se fait en partie au niveau du gestionnaire des messages (*message handling module*). Cœur de l'architecture, les messages produits par l'utilisateur ne lui parviennent pas après interprétation, mais c'est lui qui indique que l'interprétation doit avoir lieu. De la même façon, ce n'est pas lui qui détermine le comportement à adopter : une fois identifiés les actes communicatifs exprimés par l'utilisateur, il fait éventuellement appel au planificateur de réponses (*response planner*) et au générateur multimodal (*multimodal output generator*) avant de solliciter le module de fission en sortie (*output fission module*). Le gestionnaire des messages fait appel au planificateur de réponses et au générateur multimodal s'il l'estime nécessaire, *i.e.* si le comportement du système implique une planification de la réponse. En particulier, le gestionnaire des messages ne leur fait pas appel lorsqu'il s'agit de produire une animation de l'avatar destinée à rendre celui-ci plus naturel en lui faisant donner le change à l'utilisateur.

Le gestionnaire du dialogue (*dialogue manager*) défini dans cette architecture correspond aussi au module de la brique de gestion du dialogue et à une partie du module

de génération de l'architecture de la figure 4.1. En effet, y sont inclus le planificateur de réponses et le générateur multimodal. Contrairement à l'architecture de TRIPS [Allen *et al.*, 2001], la planification de la réaction du système en ce qui concerne le choix du contenu est entièrement gérée dans le planificateur de réponses. Le générateur multimodal détermine la répartition multimodale des informations et correspond donc à une partie du module de génération.

La concrétisation du message se fait de la façon suivante. Le gestionnaire des messages indique au module de fission en sortie les messages à concrétiser. Celui-ci informe chacun des dispositifs impliqués dans la concrétisation du message. Au sein de cette architecture, fusion - spécifique à l'entrée - et fission - spécifique à la sortie - s'opèrent au niveau modalitaire (*cf.* la section 2.1.2 et la section 2.1.3.3).

Nous retenons de cette architecture qu'elle déporte la question de la gestion du comportement en sortie du système au niveau du traitement de entrée en amont de gestionnaire de dialogue. Pour cela, le contrôle du système, et donc du dialogue, est centralisé au niveau du gestionnaire des messages. Grâce à cette centralisation, celui-ci connaît explicitement les capacités de tous les autres composants. De plus, cette centralisation doit rendre possible l'extension ou l'adaptation de l'architecture en fonction des dispositifs physiques d'entrée et surtout de sortie considérés. Enfin, les différents modèles classiques des systèmes de dialogue humain-machine ne sont pas explicites ici. En particulier, le modèle du dialogue est réparti au moins entre le gestionnaire d'entrées-sorties et le planificateur de réponses.

4.1.2.3 Architecture de MATCH

Dans le cadre du projet MATCH (acronyme de *Multimodal Access To City Help*) [Johnston *et al.*, 2002], un système de dialogue multimodal mêlant approche structurelle à états finis et approche différentielle coopérative a été proposé. En sortie, ce système combine langage naturel oral et un affichage visuel incluant carte, texte écrit (qui ne relève pas du langage naturel, *i.e.* dont le langage d'interaction n'est pas le langage naturel écrit), déplacement et zoom. Il renseigne l'utilisateur sur le réseau des transports en commun et les restaurants de la ville de New York et lui indique les itinéraires possibles pour se rendre à un endroit ou à un restaurant donné. Comme le montre la figure 4.4, l'architecture de ce système est centralisée. Elle est basée sur des agents. Un module central, appelé MCUBE, sert de relais pour les messages échangés entre ces agents.

L'un de ces agents remplit le rôle du module de gestion du dialogue classique : il s'agit du gestionnaire de dialogue multimodal (*mdm : multimodal dialog manager*). Il gère son propre état mental, l'identification de l'état mental et le profil de l'utilisateur ainsi que l'historique du dialogue, le modèle du domaine et les modalités disponibles. Dans la production de la sortie du système, il est en charge de deux processus, celui de sélection (*selection process*), classiquement accompli par le module de gestion du dialogue, et celui de génération (*generation process*), à la charge du module de génération dans l'architecture de la figure 4.1. Ce processus de génération consiste en la détermination du contenu de la prochaine intervention du système. Si l'utilisateur a formulé une

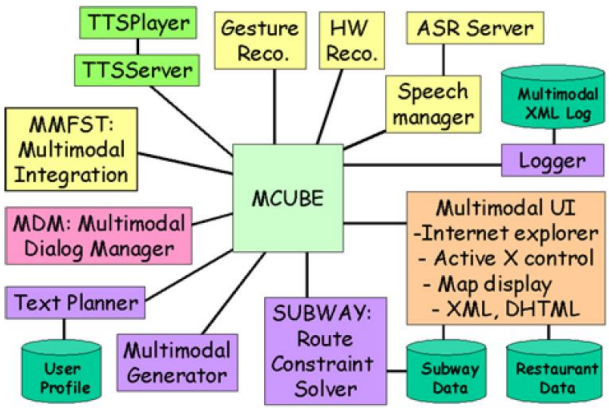

FIG. 4.4 – Architecture de l'application MATCH (extrait de [Johnston *et al.*, 2002])

demande jugée complète par rapport au modèle du domaine, le système y répond par une liste d'informations ou d'actions ; sinon, un sous-dialogue est entamé pour inciter l'utilisateur à compléter sa demande. La détermination de la complétude de la demande de l'utilisateur et le choix de contenu du message à présenter à l'utilisateur sont faits en fonction de règles pré-établies. Le processus de génération prend ensuite le relais et planifie une partie de la présentation auditive en langage naturel oral et la présentation visuelle. En ce qui concerne la présentation auditive en langage naturel oral, il détermine les réponses simples pré-formatées ou, dans le cas de réponses plus complexes, il fait appel au planificateur de texte (*text planner*). Plus précisément, le planificateur de texte est utilisé pour les énoncés de comparaison, de synthèse et de recommandation. Il s'appuie sur un modèle de l'utilisateur (*user profile*) pour tenir compte des préférences de ce dernier. Pour la présentation visuelle, le processus de génération fait appel au modèle du domaine et à un calculateur applicatif d'itinéraires de métro SUBWAY de façon à calculer les itinéraires demandés par l'utilisateur et à déterminer les actions nécessaires à afficher pour relier les deux points de départ et d'arrivée sur l'écran. La liste des énoncés à synthétiser et des affichages graphiques à réaliser est ensuite transmise à un générateur multimodal (*multimodal generator*) et l'historique du dialogue est mis à jour.

Le générateur multimodal (*multimodal generator*) synchronise les affichages à réaliser avec les réponses en langage naturel oral à synthétiser en utilisant des patrons. Le résultat d'énoncés oraux et d'actions graphiques est transmis à l'interface multimodale (*multimodal user interface*). Le générateur organise donc la présentation dans son ensemble mais est déchargé du passage d'un message à transmettre à l'utilisateur à sa forme orale ou multimodale classiquement à la charge du module de génération classique.

En sortie, l'interface multimodale (*multimodal user interface*) réalise l'affichage graphique et fait appel à un synthétiseur pour les énoncés oraux. Si besoin est, elle synchronise la réalisation de l'affichage et la synthèse : le synthétiseur avertit l'interface multimodale à chaque fois qu'un énoncé oral est synthétisé et l'interface s'aligne sur ces

informations pour concrétiser l'affichage adéquat. L'interface multimodale et le synthétiseur combinés correspondent au module de synthèse de la figure 4.1.

Contrairement à l'architecture TRIPS et celle de Gustafson, cette architecture n'est pas générique et dépend du domaine applicatif. En effet, les règles utilisées dans les processus de sélection et de génération sont purement applicatives : elles intègrent implicitement le modèle du dialogue et le modèle de la tâche. Le modèle du domaine est, quant à lui, explicite et réparti entre la base de données du métro, celle des restaurants et le calculateur d'itinéraires SUBWAY. Toutefois, son remplacement par un autre modèle du domaine nécessiterait une adaptation tant au niveau du gestionnaire de dialogue multimodal et du générateur multimodal qu'au niveau des messages gérables par MCUBE. Nous soulignons l'exploitation d'un modèle des modalités, absent des deux architectures présentées précédemment, qui permet la prise en compte de modalités utilisables dans la détermination de la présentation de la réponse du système.

4.1.2.4 Agents rationnels plurimodaux et multimodaux selon Clémente

Les travaux de Clémente [Clémente, 2004] s'inscrivent dans la lignée de la théorie de l'interaction rationnelle proposée par Sadek. Cette théorie défend une approche différentielle du dialogue humain-machine [Sadek, 1999]. Elle permet la conception d'agents rationnels dialoguants dont la particularité est d'adopter systématiquement un comportement coopératif. Ces agents sont mis en œuvre grâce à la technologie ARTIMIS (pour "Agent Rationnel à base d'une Théorie de l'Interaction mise en œuvre par un Moteur d'Inférence Syntaxique"). Les agents résultants sont des systèmes composés de quatre unités principales présentées à la figure 4.5. Ces quatre unités sont :

- une unité de gestion des connaissances qui permet la gestion du modèle du domaine et du modèle de tâche ;
- une unité rationnelle qui constitue le noyau décisionnel de l'agent-système. Elle lui permet de raisonner sur ses connaissances et sur les événements observés et d'y répondre en conséquence. L'unité rationnelle correspond au module de gestion du dialogue dans l'architecture de la figure 4.1 et intègre le maintien de l'historique du dialogue (noté l'historique des objets à la figure 4.5) ;
- une unité de traitement des langages qui fait le lien entre le langage de communication utilisé par l'utilisateur (initialement, le langage naturel oral) et la représentation sémantique interne des connaissances. En sortie et dans le cas du langage naturel comme langage d'interaction, elle est constituée d'un générateur permettant de construire le message à émettre par l'agent-système. Elle correspond au module de génération de la figure 4.1 ;
- une unité de médias gère la concrétisation des messages pour l'interface utilisateur ou pour d'autres agents artificiels. Elle adapte les sorties de l'unité rationnelle et/ou de l'unité de traitement des langages aux standards des dispositifs physiques de l'interface ou des protocoles de communication. Elle correspond à l'adaptation du module de génération de la figure 4.1 au cas des sorties multimédias ou multimodales.

Notons que le module de synthèse, *i.e.* les interfaces utilisateurs dans la figure 4.5, ne fait pas partie de l'agent.

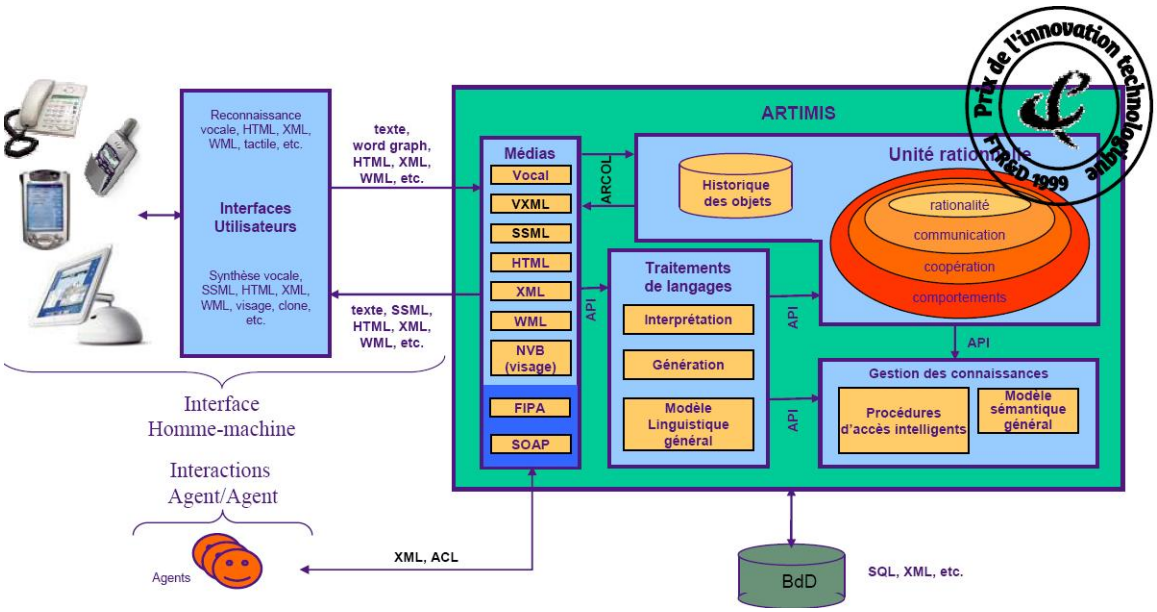

FIG. 4.5 – Architecture des systèmes développés en s'appuyant sur la technologie ARTIMIS

Dans l'approche initiale de la technologie ARTIMIS, la multimodalité n'intervient qu'au niveau mise en forme des messages produits par l'unité rationnelle, *i.e.* au niveau de l'unité des médias et de l'interface utilisateur. Au cœur de l'agent-système, les informations et les actes communicatifs manipulés sont indépendants des modalités et il n'y a pas d'inférence sur la dimension multimodale des informations. Par conséquent la multimodalité en sortie produite par les systèmes basés sur ARTIMIS est statique, *i.e.* dans le sens où le choix de présentation des actes communicatifs est pré-déterminé à la conception des systèmes : ces derniers sont donc incapables de prendre en compte une utilisation particulière pour adapter la forme du contenu présenté. Les travaux de Clémente ont pour objectif de pallier ce manque.

Clémente a donc cherché à étendre la théorie de l'interaction rationnelle de Sadek en proposant un cadre formel pour la définition d'agents plurimodaux - capables de changer de modalité uniquement entre deux tours de parole, chaque énoncé étant monomodal - et multimodaux - combinant les modalités de façon simultanée. Ce cadre formel permet de définir les mécanismes d'inférence qui déterminent les comportements possibles d'un agent rationnel dialoguant en fonction de contraintes considérées comme étant fortes par Clémente. Ces contraintes renvoient aux conditions de communication relatives aux (in-)capacités physiques et mentales de l'utilisateur, aux conditions environnementales d'utilisation (dispositifs physiques, caractéristiques de l'environnement ambiant, etc.) et/ou aux contraintes de présentation explicites de l'utilisateur.

Le cadre formel de définition d'agents plurimodaux proposé par Clémente étend le modèle d'actes communicatifs de la théorie de l'interaction de Sadek par l'ajout d'un argument supplémentaire dans les actes. Cet argument supplémentaire permet de représenter le média et la modalité de l'acte considéré. S'y ajoutent des pré-conditions spécifiques à l'interaction multimodale qui mettent en œuvre les contraintes détaillées ci-dessus. Clémente introduit également les prédicats nécessaires pour traiter ces contraintes et l'axiomatique définissant les comportements fortement contraints qui en résultent. Autrement dit, il formalise les connaissances relatives aux média[3], aux modalités, aux modes et aux processeurs ; il propose les prédicats et les schémas d'axiomes permettant à l'agent de raisonner sur ces connaissances et il étend le modèle des actes communicatifs de Sadek pour y intégrer les média et les modalités. L'agent résultant est donc en mesure de choisir une modalité et sait quelles modalités sont disponibles. Il en est de même pour les média, les modes et les processeurs.

Clémente étend quelque peu ce cadre formel pour permettre la définition d'agents multimodaux. Plus précisément, il affine l'argument qui représente le média et la modalité de façon à associer plusieurs média et modalités à chaque concept et chaque relation entre les concepts impliqués dans un acte communicatif donné. Ainsi les actes sont-ils représentés de façon multimodale.

Le travail de Clémente se concentre donc sur le module de gestion du dialogue qu'il transforme de façon à ce qu'il ne soit pas indépendant des modalités. Si l'on se reporte aux autres systèmes de dialogue multimodaux décrits, ce module a au minimum été adapté dans ce sens, voire a été scindé de façon à définir un sous-module de gestion du dialogue tenant compte des modalités de présentation. Clémente a la particularité d'expliciter la prise en compte de contraintes de présentation pour le choix de la présentation des réactions des agents plurimodaux et multimodaux. De plus, ces agents ont des connaissances explicites sur les notions ayant trait à la multimodalité et sont donc en mesure de répondre à des questions de l'utilisateur sur ces connaissances.

4.1.3 Synthèse : systèmes de dialogue multimodaux

De cette étude sur les systèmes de dialogue multimodaux, nous retenons plusieurs caractéristiques utiles à nos travaux. Tout d'abord, il convient de noter que le paradigme dialogique de communication humain-machine a fait du langage naturel son langage d'interaction dominant en privilégiant les dispositifs physiques auditifs. Les informations visuelles, *i.e.* dont le dispositif physique est l'écran, servent souvent de support à la communication, y compris dans le cas du textuel où le langage d'interaction est le langage naturel. Le visuel sert aussi de support de saillance comme c'est le cas des gestes et pointages déictiques ainsi que les orientations du regard des avatars. Les avatars en tant qu'entités sont généralement couplés au langage naturel oral[4] et sont utilisés pour humaniser la machine afin de faciliter son appropriation par l'humain.

[3]Orthographe utilisée par l'auteur

[4]Les avatars sont aussi utilisés comme "dispositif physique" des langues des signes en tant que langages d'interaction.

La combinaison de modalités dans les systèmes de dialogue est le plus souvent complémentaire, que cette complémentarité soit complète ou partielle (*cf.* section 2.1.3.2 et section 2.1.5). Constatant qu'un système de dialogue qui combine langage naturel oral et langage naturel écrit est considéré comme multimodal, nous en concluons que le paradigme dialogique considère la multimodalité par rapport aux dispositifs physiques impliqués.

Entrée et sortie ne sont pas prises comme des processus exactement inverses : les traitements de l'entrée et de la sortie à un haut niveau d'abstraction, *i.e.* pas au niveau de concrétisation ou de réalisation, sont clairement distincts. En ne considérant pas une boucle entrée-sortie directe, *i.e.* perception-action, la détermination de la sortie peut être décomposée en quatre étapes : (1) identification de l'objectif comportemental du système, (2) détermination du ou des messages à produire en conséquence, (3) allocation des modalités aux différents éléments du ou des messages à produire et (4) concrétisation de ces messages. Nous n'évoquons pas la coordination entre les messages car elle n'est pas toujours traitée au même niveau. L'étape 2 est, suivant les cas, associée à l'étape 1 ou à l'étape 3. Ceci nous amène au constat suivant : l'étape 2 peut être considérée comme partie intégrante de la gestion du dialogue. Or cette étape est en général dépendante des modalités. C'est pour cette raison que la gestion du dialogue dans les systèmes de dialogue multimodaux présentés tient compte des modalités.

Terminons par les modèles évoqués dans ces systèmes et architectures. Le modèle de la tâche n'est que rarement explicite et isolé : il est souvent confondu soit avec le modèle de dialogue soit avec le modèle du domaine. L'historique du dialogue est, quant à lui, explicite et central pour la gestion du dialogue. Il est parfois utilisé pour l'interprétation de l'entrée et la génération de la sortie. Enfin, plusieurs études visent à la prise en compte des préférences de l'utilisateur et des modalités disponibles, voire des caractéristiques du contexte d'utilisation. Ces différentes informations sont mémorisées au niveau d'un modèle de l'utilisateur et/ou d'une version étendue de l'historique du dialogue.

Après avoir étudié l'approche dialogique (métaphore conversationnelle) de la communication multimodale, nous nous tournons vers les travaux relevant de l'interaction (métaphore actionnelle).

4.2 Interaction humain-machine

Comme pour le dialogue humain-machine, nous présentons d'abord les principes fondateurs de l'interaction humain-machine avant d'étudier leurs adaptations au cas de l'interaction multimodale.

4.2.1 Principes fondateurs de l'interaction humain-machine

À la suite de [van Dam, 1997], plusieurs générations d'interfaces peuvent être identifiées. La première génération correspond aux premiers ordinateurs qui n'avaient pas

d'interfaces directes et utilisaient des cartes perforées. La deuxième génération correspond aux premières interfaces directes : la saisie de caractères alpha-numériques permet à l'utilisateur d'indiquer à la machine des commandes et paramètres à exécuter et l'affichage permet à la machine de présenter le résultat d'exécution de ces commandes. On parle d'interaction ou dialogue en lignes ou langages de commande, mais la notion de "dialogue" sous-tendue ici est loin du paradigme dialogique présenté dans la section précédente. En effet, il s'agit d'une forme de dialogue très simple où l'entrée est fortement contrainte par la tâche pour lequel le système a été conçu et où le système est exécutant bien plus que coopératif. La troisième génération, initiée par les travaux du Xerox Parc et utilisée par le grand public depuis plus d'une vingtaine d'années, est celle des interfaces graphiques à manipulation directe. Ces interfaces constituent l'autre paradigme classique de communication humain-machine, à savoir le paradigme actionnel pour reprendre [Hutchins *et al.*, 1986]. Nous étudions dans les paragraphes suivants les principes de ce paradigme et son extension à la multimodalité.

4.2.1.1 Interfaces WIMP et manipulation directe

Dans le courant des années 1970, le premier éditeur graphique est développé au Xerox Parc. Les principes sous-tendus dans cet éditeur, formalisés dans le courant des années 1980, ont fait la popularité de l'informatique grand public. Ce premier éditeur graphique intègre plusieurs éléments qui semblent aujourd'hui banals. Il s'agit des fenêtres, des icônes et des menus ainsi que des pointeurs pour agir sur ces différents objets. Ces quatre éléments sont à l'origine du nom donné aux interfaces de cette génération : elles sont dites WIMP pour "*Windows, Icons, Menus, Pointers*". La fenêtre est l'élément de base de ces interfaces : elle permet le partage de l'écran par plusieurs applications. Les icônes concrétisent les objets et les actions du domaine en fonction de la métaphore utilisée (la plus courante étant celle du bureau). Les menus sont l'adaptation des commandes des interfaces de génération précédente : présentées sous forme de liste, les commandes n'ont plus à être connues par l'utilisateur qui peut les retrouver via l'interface graphique. On parle de *manipulation directe* des objets.

La notion de "manipulation directe", souvent rattachée aux interfaces WIMP, est introduite par Shneiderman dans les années 1980 [Shneiderman, 1986]. Elle s'appuie sur plusieurs principes. Tout d'abord, les objets et les actions sont représentés de façon continue. C'est la base de la manipulation directe, car cette représentation continue permet à l'utilisateur de pouvoir accéder rapidement aux objets et aux actions utiles et lui évite de les mémoriser comme c'était le cas avec les langages de commande. Ensuite, les actions possibles doivent être rapides et incrémentables et toute action doit être réversible. Ceci permet à l'utilisateur de pouvoir contrôler les actions effectuées au mieux, sans crainte que le système ne puisse revenir à un état antérieur. Un autre principe, et non des moindres, est que l'effet des actions effectuées doit être visible immédiatement. Les interfaces à manipulation directe sont donc WYSIWYG (acronyme de "*What you see is what you get*") où l'on voit directement le résultat des actions appliquées sans devoir passer par une transformation ou une vue supplémentaire. Le dernier principe correspond au remplacement des langages de commandes à syntaxe

complexe par la manipulation d'objets. Dans le cas d'une interface graphique, le pointeur est donc primordial.

L'intérêt de ces principes a été validé par la popularité des interfaces graphiques WIMP à manipulation directe. Selon [Shneiderman, 1986], ce succès est dû au fait que l'utilisateur a une impression de transparence, d'accomplissement direct de la tâche[5]. De plus, l'utilisateur a aussi une impression d'engagement direct dans le monde des objets plutôt que le recours à un intermédiaire[6]. D'après [Hutchins *et al.*, 1986], cette impression tient à l'engagement mais aussi à la distance ressentie par l'utilisateur. L'engagement correspond à l'impression d'action directe sur les objets et est donc lié à la métaphore utilisée pour représenter les objets et les actions. Toujours d'après [Hutchins *et al.*, 1986], le paradigme conversationnel ne permet pas un lien direct avec le monde manipulé car l'utilisateur ne peut pas agir directement sur les objets : le langage naturel sert d'intermédiaire entre utilisateur et objets, humain et machine. Le recours à un paradigme actionnel graphique inclut un monde représenté explicitement sur lequel l'utilisateur peut agir, lui donnant cette impression d'engagement direct. L'engagement est renforcé entre autres par les liens entre entrée et sortie permettant à l'utilisateur d'agir en entrée du système sur des objets présentés en sortie, ainsi que par les temps de réponse rapides du système. La distance correspond à une distance entre les pensées de l'utilisateur et les actions physiques requises par le système : plus cette distance est faible, plus le passage des pensées de l'utilisateur aux actions est petite et l'accès (cognitif) aux sorties des systèmes est facile [Norman, 1986]. Cette distance est de deux types : la distance articulatoire entre le sens et la forme d'une expression, et la distance sémantique entre le fond d'une expression et le but. Ces deux distances sont aussi valables en sortie du système.

Le terme d'"interfaces" renvoie souvent aux interfaces graphiques WIMP à manipulation directe, où le dispositif physique dominant est l'écran. Avec le succès des systèmes informatiques qui a découlé de ce type d'interfaces, les travaux se sont concentrés sur les outils d'implémentation, que ce soit des boîtes à outils, des générateurs d'interfaces ou des architectures. Des architectures-types ont ainsi été proposées que nous présentons dans le paragraphe suivant.

4.2.1.2 Architecture des systèmes à manipulation directe

L'élaboration d'architectures pour les systèmes à manipulation directe (paradigme actionnel) s'est faite en plusieurs temps. Initialement, le système désigne exclusivement le noyau fonctionnel de l'application, qui correspond au modèle du domaine et qui est appelé "application". L'utilisateur accède à ce noyau fonctionnel grâce à une interface, typiquement les lignes de commande dans les interfaces de deuxième génération [van Dam, 1997].

[5]Tradution libre de la phrase suivante de Rutkowski citée par Shneiderman : *The user is able to apply intellect directly to the task; the tool itself seems to disappear.*

[6]Traduction libre de "*the feeling of involvement directly with a world of objects rather than of communicating with an intermediary*" [Hutchins *et al.*, 1986].

Modèle de Seeheim La première architecture a été proposée en 1983 lors d'un atelier de travail sur les systèmes de gestion d'interfaces utilisateurs [Pfaff, 1985]. Le nom de cette architecture, *i.e.* le modèle de Seeheim, renvoie à la ville où s'est tenu cet atelier. Comme le montre la figure 4.6, l'apport essentiel de cette architecture consiste en l'ajout d'un contrôleur de dialogue (*dialogue control*) pour gérer les échanges entre l'utilisateur et l'application. Plus précisément, c'est à son niveau que la structure de l'interaction, *i.e.* le modèle du dialogue, entre l'utilisateur et l'application, est défini. Il sert de relais aux unités informationnelles envoyées par l'application à l'utilisateur suite à une requête de ce dernier. De plus, le contrôleur de dialogue maintient un modèle du contexte qui se rapproche d'un historique du dialogue puisqu'il spécifie les changements d'état du dialogue et l'état de la présentation.

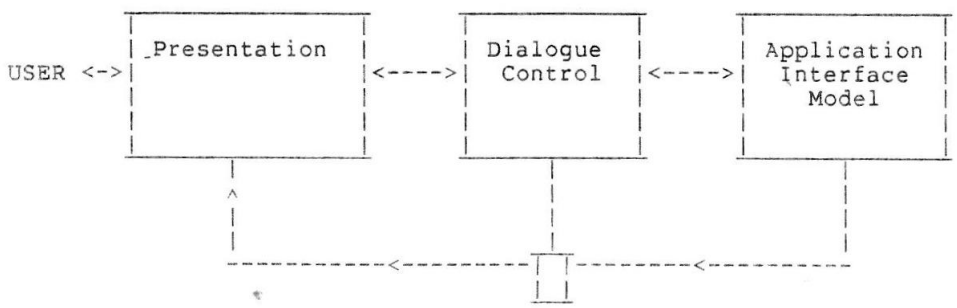

FIG. 4.6 – Modèle de Seeheim (extrait de [Green, 1985])

L'interface, appelée présentation dans le modèle (*presentation*), est explicitement visuelle, voire graphique [Green, 1985]. En sortie, elle est chargée de produire une représentation perceptible affichable à partir d'une représentation abstraite des unités informationnelles à présenter. Pour cela, le module de présentation utilise un dictionnaire faisant le lien entre représentations abstraites et physiques, qui correspond à un modèle du langage ou à un langage d'interaction. Ce dictionnaire n'est modifiable que par le contrôleur de dialogue.

Une interface de l'application (*application interface model*), qui remplit le même rôle que la présentation pour l'utilisateur, a été identifiée entre le contrôle de dialogue et l'application. Elle correspond à une vue de l'application selon l'utilisateur. Elle inclut un modèle du domaine ainsi qu'un modèle des tâches que l'utilisateur peut accomplir. Ce modèle des tâches inclut les contraintes sur les tâches.

Dans le modèle de Seeheim, trois modules sont donc distingués entre l'utilisateur et l'application : la présentation, le contrôle de dialogue et l'interface de l'application. Le contrôle de dialogue permet à l'utilisateur et à l'application d'interagir via des interfaces qui leur sont spécifiques. De plus, modèles du dialogue et des tâches, ainsi qu'un historique du dialogue réduit, sont clairement intégrés.

Méta-modèle Arch L'architecture Arch, aussi appelée "*Seeheim revisited*", constitue une extension du modèle de Seeheim [UIMS, 1992]. Comme le montre la figure 4.7, son principal apport est de décomposer le module de présentation en deux, une partie concrète et une partie abstraite. Plus précisément, Arch comprend les cinq modules suivants, notés composants dans le modèle :

- un composant du domaine (*domain-specific component*), appelé aussi noyau fonctionnel (*functional core*) gère les données du domaine et réalise toutes les fonctions du domaine. Il correspond à l'application dans le modèle de Seeheim ;
- un composant d'adaptation au domaine (*domain-adaptator component*), qui fait le pont entre le composant de dialogue et le composant du domaine et qui intègre les fonctions nécessaires à la réalisation de tâches par l'utiisateur qui n'existent pas au niveau du composant du domaine. Il correspond à l'interface de l'application dans le modèle de Seeheim ;
- un composant de dialogue (*dialogue component*), chargé de gérer le dialogue entre l'utilisateur et l'application. Il détermine les informations à présenter à l'utilisateur en fonction de la tâche identifiée. Il correspond au contrôle de dialogue dans le modèle de Seeheim ;
- un composant de présentation (*presentation component*), médiateur entre le composant de dialogue et celui d'interaction. Il détermine la présentation abstraite qui correspond à la tâche de présentation décidée par le composant de dialogue et que doit concrétiser le composant d'interaction. Il correspond donc au choix du style d'interaction selon la définition de Frohlich [Frohlich, 1996] (*cf.* la section 1.2.1) ;
- un composant d'interaction (*interaction toolkit component*) qui permet de produire l'interface perceptible et manipulable par l'utilisateur, en fonction des capacités des terminaux.

Composants de présentation et d'interaction correspondent au module de présentation du modèle de Seeheim. En sortie du système, sont distinguées la conception de la présentation au niveau du composant de présentation - qui est, par conséquent, parfois appelé "composant de présentation abstraite" - et la réalisation, la concrétisation de la présentation au niveau du composant d'interaction - appelé aussi "composant de présentation concrète". Dans le cadre de nos travaux centrés sur la sortie, nous privilégions les termes de "présentation abstraite" et de "présentation concrète" car ils mettent plus en avant la perception (en sortie) que l'action (en entrée) de l'utilisateur.

Cette architecture distingue donc clairement fond, forme et forme réalisée. Cette modularité permet de réutiliser chacun des éléments implémentés en fonction des besoins, en particulier de changer rapidement la partie visible de l'interface ou encore le style de l'interface. Si la forme d'arche du modèle n'est pas sans rappeler celle de l'architecture classique des systèmes de dialogue humain-machine, le parallélisme entre les arcs gauche et droit ne renvoie pas du tout à la même chose : dans le cas des systèmes de dialogue, il s'agit du parallélisme entre les traitements de l'entrée et de la sortie ; dans le cas des systèmes à manipulation directe, il s'agit du parallélisme entre les traitements pour l'utilisateur et pour l'application.

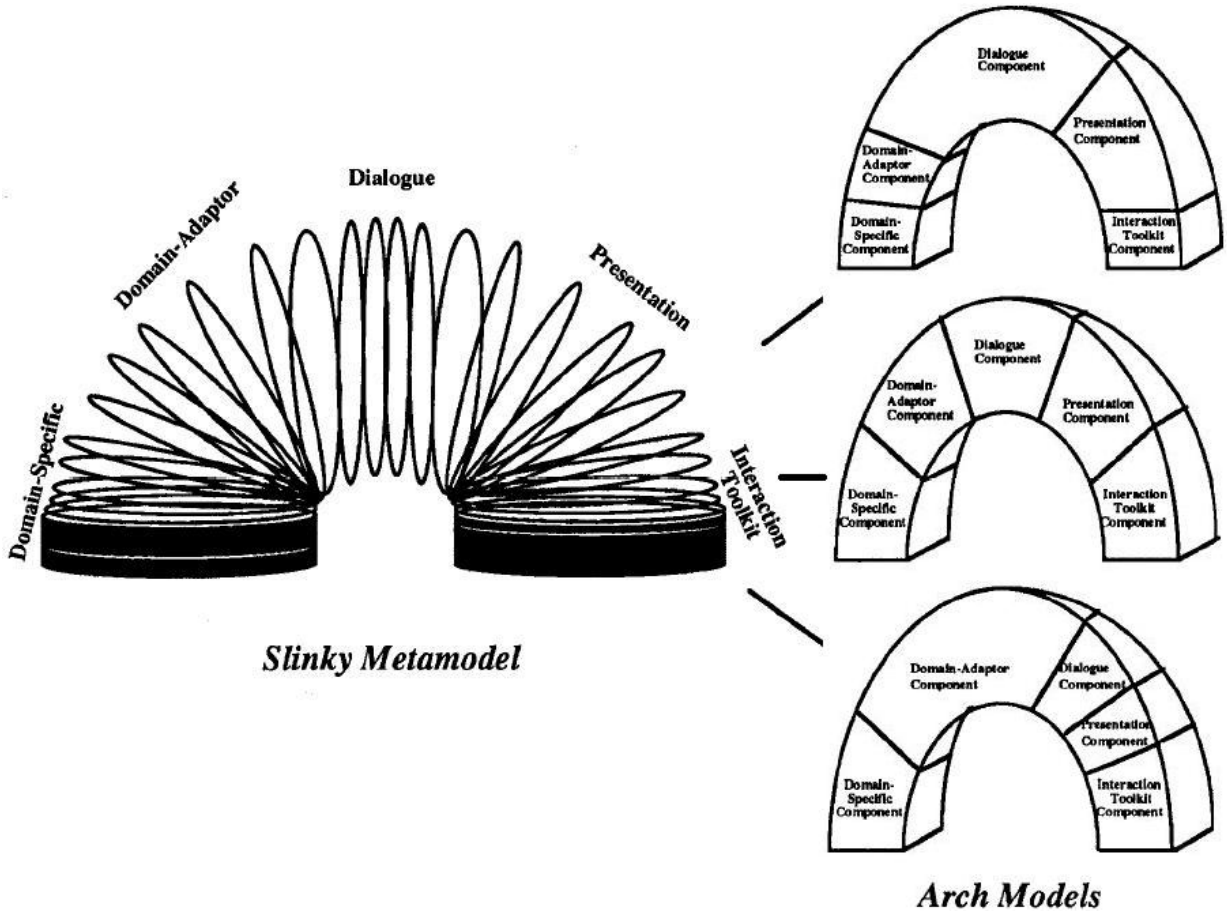

FIG. 4.7 – Méta-modèle Slinky et les modèles Arch dérivés (extrait de [UIMS, 1992])

L'architecture Arch est parfois considérée en tant que méta-architecture car elle est associée à l'approche Slinky en référence au nom du long ressort servant de jouet [UIMS, 1992]. Selon cette approche, la place accordée aux différents composants de l'architecture de base peut être différente en fonction des besoins. En particulier, les deux composants d'adaptation du domaine et de présentation qui constituent les interfaces logicielles sont optionnels. Notons toutefois que le composant de dialogue reste le composant central, la clef de voûte de l'arche.

Seeheim et Arch sont des modèles intéressants du point de vue du génie logiciel car ils identifient différents composants réutilisables. Toutefois, ils sont limités en ce qui concerne la décomposition fonctionnelle d'un système et son explicitation. Cette faiblesse est l'atout des modèles multi-agents, que nous présentons ci-dessous.

Modèles multi-agents Les modèles multi-agents sont des modèles qui structurent un système interactif non pas par rapport à une architecture conceptuelle de référence

mais par rapport aux fonctionnalités offertes par le système. Chaque agent peut, en réaction à des événements qu'il acquiert, provoquer un ou plusieurs événements. Il est donc spécialisé pour un type ou plusieurs types d'événements. Il possède généralement un état qui est modifié lorsqu'il acquiert ou produit des événements. Nous décrivons succintement les principes des deux modèles multi-agents les plus connus, à savoir PAC et MVC.

PAC [Coutaz, 1987], acronyme de "Présentation, Abstraction, Contrôle", est un modèle qui structure un système interactif de façon récursive sous forme d'une hiérarchie d'agents. Chaque agent est défini grâce à trois facettes : une facette présentation, une facette abstraction et une facette contrôle. La facette présentation définit le comportement perceptible de l'agent par un agent humain. Elle permet l'interprétation des événements de l'utilisateur sur l'interface et la concrétisation des événements produits par l'agent au niveau de l'interface. Cette facette renvoie à la partie présentation dans Seeheim et dans Arch. La facette abstraction correspond aux compétences de l'agent, *i.e.* aux événements auxquels il est capable de réagir. Cette facette correspond au noyau fonctionnel dans Seeheim ou dans Arch. La facette contrôle assure à la fois la communication entre les facettes présentation et abstraction et la communication de l'agent avec les autres agents. Cette facette est à rapprocher du contrôle de dialogue dans Seeheim, respectivement du composant de dialogue dans Arch. Chaque agent peut traiter d'un niveau d'abstraction plus ou moins important de l'interaction humain-machine. La hiérarchie des agents PAC permet d'exprimer les relations entre les agents et reflète l'existence de différents niveaux d'abstraction depuis l'application jusqu'aux éléments fins de l'interaction.

MVC [Krasner et Pope, 1988], acronyme de "Model, View, Controller", structure aussi un agent en trois composantes, le modèle (*model*), la vue (*view*) et le contrôle (*controller*). Le modèle décrit le comportement interne de l'agent. Il correspond à la facette abstraction de PAC. La vue indique l'état du modèle et recouvre la fonction de restitution de l'agent. Elle est donc spécifique à la sortie de l'agent. Le contrôle permet d'interpréter les actions de l'utilisateur sur la vue. Il ne s'occupe donc que de l'entrée de l'agent. Vue et contrôle correspondent à la facette présentation d'un agent PAC. Le modèle MVC n'impose pas de contrainte sur la structuration globale d'un système interactif en agents MVC. La notion de contrôle de PAC est inexistante dans MVC. Notons que si, à l'origine, les composantes d'un agent MVC étaient réalisables sous forme d'un objet Smalltalk, leur principe a été repris pour un certain nombre de langages orientés objet, dont Java, en faisant à l'heure actuelle l'approche la plus courante de conception d'applications interactives.

L'intérêt des modèles multi-agents est qu'ils permettent une décomposition fine et modulaire des différentes fonctionnalités d'un système interactif, ce qui autorise l'exécution de traitements en parallèle et la modification d'un comportement sans remettre en cause l'ensemble du système.

Soulignons que Nigay a proposé l'architecture PAC-Amodeus [Nigay, 1994] de façon à combiner les avantages du modèle conceptuel Arch et ceux du modèle à agents PAC. Plus précisément, le composant de dialogue d'Arch y est affiné sous forme d'une hiérarchie d'agents PAC. Dans ce cadre, la facette abstraction d'un agent PAC modélise sa

compétence et renvoie à des objets conceptuels du composant d'adaptation au domaine ; la facette présentation, qui peut être reliée à un ou plusieurs objets de présentation, permet la récupération des commandes de l'utilisateur et définit le rendu à concrétiser ; la facette contrôle assure la liaison entre les facettes présentation et abstraction et intervient dans les processus d'abstraction et de concrétisation de la hiérarchie des agents PAC permettant de passer des objets de présentation aux objets conceptuels connus du composant d'adaptation au domaine.

4.2.2 Architectures de systèmes multimodaux à manipulation directe

Les premiers travaux sur les interfaces multimodales remontent à Bolt et à son célèbre "mets ça là" [Bolt, 1980]. Les principes de la manipulation directe n'étaient pas encore formalisés qu'associer l'oral au visuel était déjà envisagé. Il convient de noter que, bien que le système de Bolt était multimodal en sortie, couplant synthèse de parole et animation graphique, ces travaux fondateurs en multimodalité restent connus pour l'interface en entrée et la référence multimodale combinant langage naturel oral et gestes de désignation.

Presque trente ans plus tard, la multimodalité en sortie n'a été que peu étudiée. En effet, elle a été initialement assimilée au multimédia, souvent considéré comme étant à un plus bas niveau que le multimodal. Les choix de modalités qu'un système multimodal en sortie peut effectuer reste une problématique peu étudiée. Par conséquent, peu de systèmes centrés sur la multimodalité en sortie, ou simplement la traitant à hauteur de la multimodalité en entrée, ont été proposés. Nous retenons, de la littérature, trois architectures qui ont été proposées récemment.

4.2.2.1 Architecture du W3C pour l'interaction multimodale

Dans le cadre du groupe de travail du W3C sur l'interaction multimodale, une méta-architecture à préciser en fonction des besoins a été proposée [Bodell *et al.*, 2003]. Pour des raisons de simplicité du discours, nous parlerons simplement d'architecture par la suite. Cette architecture est présentée à la figure 4.8. La figure 4.9 détaille la production de la sortie du système.

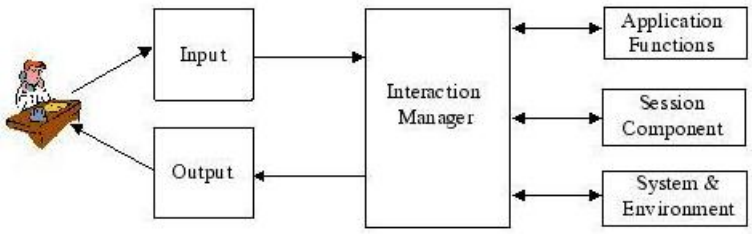

FIG. 4.8 – Composants principaux d'un système multimodal selon le W3C (extrait de [Bodell *et al.*, 2003])

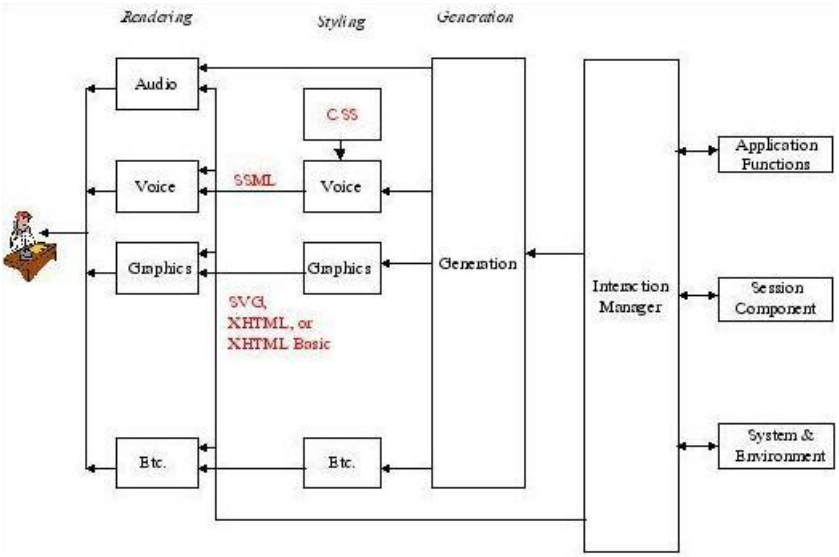

FIG. 4.9 – Composants intervenant dans la sortie d'un système multimodal selon le W3C (extrait de [Bodell *et al.*, 2003])

La classification de cette architecture par rapport au paradigme sous-tendu n'a pas été aisée, ne serait-ce que parce qu'en proposant une méta-architecture des systèmes d'*interaction* multimodale, les auteurs ne revendiquent leur rattachement à aucun paradigme (actionnel ou conversationnel). De plus, aucun héritage d'une architecture-mère non-multimodale n'est évoqué. Nous avons finalement que cette architecture s'applique plus à des interfaces multimodales à manipulation directe pour les raisons suivantes. Tout d'abord, l'interprétation des messages de l'utilisateur n'est pas utilisée pour extraire les buts ou les intentions de l'utilisateur comme c'est le cas dans les systèmes de dialogue multimodaux présentés précédemment. Ensuite, cette architecture est destinée à des applications Web pour lesquelles le graphique est la modalité dominante et le langage naturel oral, notamment, joue un rôle moindre.

L'architecture en question distingue, à l'image des architectures orientées dialogue, entrée et sortie. Nous nous concentrons sur les composants et modèles qui interviennent dans la production de la sortie du système (*cf.* figure 4.9).

Le gestionnaire de l'interaction (*interaction manager*) correspond au composant de dialogue dans Arch. C'est un composant qui détermine le contenu de la sortie du système en fonction de son entrée, des données applicatives et des informations sur l'interaction et sur les sessions. Il peut être unique ou composé de plusieurs composants spécialisés. Les données sont conservées dans un modèle du domaine maintenu par le composant noté (*application functions*). Outre ce composant, le gestionnaire de l'interaction échange des informations avec deux autres composants : le composant de session et le composant d'interaction.

Le composant de sessions (*session component*) permet au gestionnaire de dialogue

de gérer les changements d'états de l'application lors d'une session ou entre les sessions. Il est particulièrement utilisé lorsque l'accès à l'application peut se faire via différents dispositifs ou par différents utilisateurs. Ce composant maintient un historique de l'interaction que nous assimilons à une version simple d'historique de dialogue (section4.1.1.2).

Le contexte d'interaction (*system and environment*) comprend aussi bien des informations sur l'environnement matériel d'utilisation (*i.e.* les dispositifs physiques), sur l'environnement physique d'utilisation (localisation, bruit ambiant, etc.) que sur le nombre et les préférences statiques ou dynamiques des utilisateurs. Ce composant maintient donc un modèle de l'utilisateur étendu au contexte d'utilisation.

Une fois la sortie à présenter définie par le gestionnaire de l'interaction, le composant de génération (*generation*) en est informé. Celui-ci détermine la ou les modalités à utiliser pour présenter les informations transmises par le gestionnaire d'interaction ainsi que, s'il y a lieu, les relations entre ces modalités. Il n'intervient pas dans les cas où c'est un fichier ayant un format bien particulier qui est exécuté, comme les sons pré-enregistrés. De tels fichiers sont donc directement envoyés par le gestionnaire d'interaction aux composants de réalisation, *i.e.* aux dispositifs physiques, adéquats. Selon les cas, le composant de génération transmet les messages à présenter à ces mêmes composants de réalisation (*rendering*) ou les envoie aux composants de style lorsque la forme de la présentation doit être complétée avant concrétisation. Le composant de génération correspond au composant de présentation dans Arch.

Si cela est pertinent pour une modalité donnée, des composants de style (*styling*) sont définis. Leur rôle est de compléter le message envoyé par le composant de génération avec des informations sur le style, *i.e.* la forme, du message à produire avant de l'envoyer au composant de réalisation adéquat pour concrétisation. En fonction des types de styles appliqués, ces composants peuvent être aussi bien au niveau du composant de présentation abstraite (par exemple, onglets versus liste pour présenter un ensemble de solutions) ou du composant de présentation concrète (par exemple, feuilles de style pour l'affichage d'une liste de solutions) dans Arch.

En ce qui concerne les sorties multimodales, les contributions de cette architecture par rapport à l'architecture de référence Arch restent faibles. En effet, même si elle détaille les différents niveaux de concrétisation des sorties (style à appliquer et concrétisation à réaliser), l'architecture du W3C ne donne pas d'informations supplémentaires sur ce qu'implique l'utilisation de plusieurs modalités de présentation sur la production de la sortie du système ni sur la façon dont ces modalités sont combinées, ou non. Ceci est sans doute dû au fait que le groupe de travail du W3C sur l'interaction multimodale se concentre avant tout sur le format des messages échangés entre les composants des systèmes multimodaux.

4.2.2.2 Architecture générique des interfaces multimodales selon Carbonell

Carbonell [Carbonell, 2005] propose une architecture générique pour les interfaces multimodales. Comme le montre la figure 4.10, cette architecture s'appuie sur Arch et s'applique donc tant en entrée qu'en sortie. Nous nous concentrons sur la production d'une réaction multimodale par le système.

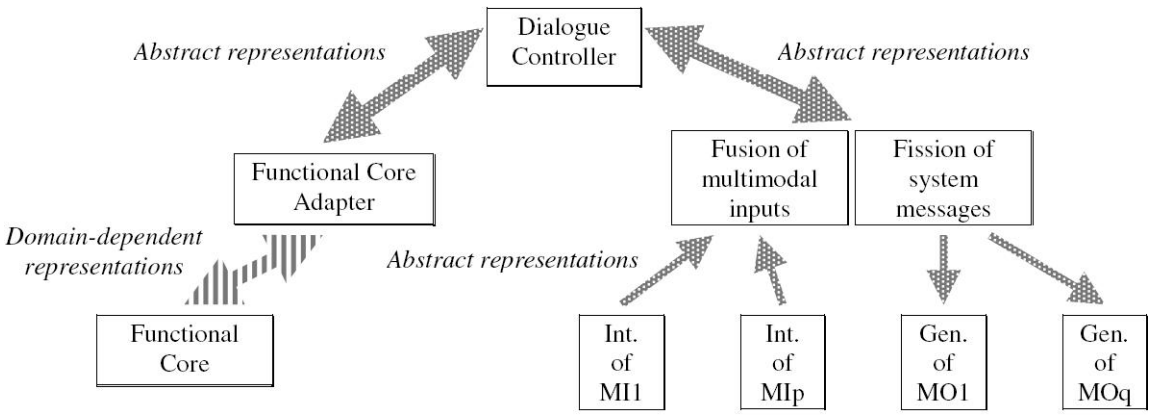

FIG. 4.10 – Adaptation d'Arch aux interfaces multimodales selon Carbonell (extrait de [Carbonell, 2005] - "Int." et "Gen." désignent respectivement les interpréteurs et les générateurs dédiés aux entrées monomodales (MIi) et aux sorties monomodales (MOi))

Selon l'architecture présentée ici, et en se référant à Arch, la génération multimodale permet de passer de la représentation abstraite d'un contenu informationnel produit par le composant du domaine (*functional core*) à des unités informationnelles concrètes présentées grâce à différents dispositifs physiques. Pour cela, le composant de dialogue (*dialogue controller*) récupère le contenu informationnel disponible au niveau du noyau fonctionnel en fonction de l'entrée fournie par l'utilisateur. Ce contenu informationnel est annoté et modifié selon le profil de l'utilisateur et le contexte d'interaction. Il est composé d'unités informationnelles indépendantes des modalités qui sont ensuite transmises au composant de présentation.

Le composant de présentation joue le rôle de générateur multimodal (*multimodal generator*). Dans le cas de multimodalité simultanée, il y a fission modalitaire des différentes unités informationnelles (*fission of system messages*) en fonction de critères sémantiques et ergonomiques pré-définis. Dans le cas d'une multimodalité séquentielle, cette fission n'a pas lieu et le composant de présentation se contente d'allouer des modalités à chaque unité informationnelle.

Les unités informationnelles monomodalement allouées sont transmises à des générateurs monomodaux (*monomodal generators*) qui constituent le composant d'interaction : ces générateurs sont chargés de concrétiser, de réaliser ces unités informationnelles dont l'ensemble constitue le message du système.

Carbonell précise que le modèle du contexte utilisé pour la sélection des unités informationnelles à présenter à l'utilisateur est accessible et modifiable uniquement par le composant de dialogue. Il contient en particulier l'historique des interactions, le profil de l'utilisateur, l'état courant de l'application et de l'affichage et l'environnement matériel d'utilisation. La façon dont ces différents (sous-)modèles peuvent influencer la sélection de la réponse du système n'est pas précisée. Notons toutefois que ce modèle du contexte se rapproche plus de l'historique du dialogue utilisé dans les systèmes de dialogue humain-machine qu'aucun des modèles évoqués dans le cadre de l'architecture

proposée par le W3C. Toutefois, si le composant de dialogue est indépendant des modalités comme dans Arch, cela signifie que les informations sur l'environnement matériel d'utilisation sont transmises au composant de présentation par le composant de dialogue sans que celui-ci en maîtrise le contenu. De même, l'historique des interactions et l'état courant de l'application sont maintenus de façon indépendante des modalités et de leurs usages.

Ce modèle précise néanmoins que les choix de génération multimodale dépendent de critères ergonomiques et sémantiques. Mais la façon dont ces choix sont faits n'est pas définie. Enfin, les relations entre les modalités, en particulier redondance ou complémentarité, ne sont pas traitées.

4.2.2.3 Modèle conceptuel WWHT selon Rousseau

Les travaux présentés dans [Rousseau, 2006] sont dédiés à la production de présentations multimodales adaptées au contexte d'utilisation. Le modèle conceptuel WWHT (*What Which How Then*) définit le processus de production de ces présentations. Nous en détaillons les trois premières étapes, présentées à la figure 4.11, qui correspondent aux trois questions "*what ?*", "*which ?*" et "*how ?*. La dernière étape du processus traite de l'évolution de la présentation produite en fonction du contexte et se rapproche plus des travaux sur la plasticité des interfaces [Thévenin, 2001, Sottet *et al.*, 2007].

La première étape de production d'une réponse multimodale consiste à répondre à la question "*what ?*" : il s'agit de déterminer le contenu à présenter. Plus précisément, le contenu sélectionné en amont, en l'occurrence par le composant de dialogue dans Arch, est décomposé en unités informationnelles élémentaires. Comme défini dans [Nigay, 1994], une fission sémantique peut être effectuée. Cette fission est généralement pré-définie par les concepteurs car son automatisation lors de l'utilisation du système nécessiterait une analyse sémantique du contenu qui est difficilement réalisable. Par rapport à Arch, la fission sémantique se fait au niveau du composant de dialogue ou celui de présentation abstraite.

La deuxième étape de production d'une réponse multimodale consiste à répondre à la question "*which ?*" : il s'agit de déterminer l'allocation des modalités aux unités informationnelles définies à l'étape précédente. Plus précisément, chaque unité informationnelle élémentaire est associée à une présentation multimodale adaptée au contexte d'interaction puis toutes les unités informationnelles multimodalement allouées sont regroupées pour former une seule présentation en distinguant les modalités et les médias mobilisés et les relations de complémentarité ou de redondance entre ces modalités. Il y a donc fusion en sortie, comme définie dans [Coutaz *et al.*, 1993] et [Nigay, 1994]. L'allocation multimodale repose sur un modèle comportemental : celui-ci spécifie les composants d'interaction jugés compatibles avec un contexte d'interaction donné. Le formalisme choisi pour ce modèle est une base de règles. Trois types de règles sont distinguées : les règles contextuelles qui ciblent des composants d'interaction, les règles critérielles qui ciblent un critère de modalité (*i.e.* une caractéristique comme celles proposées par Bernsen [Bernsen, 1994, Bernsen, 1997]) et les règles de composition qui ciblent une combinaison de modalités. L'allocation à proprement parler se fait en deux

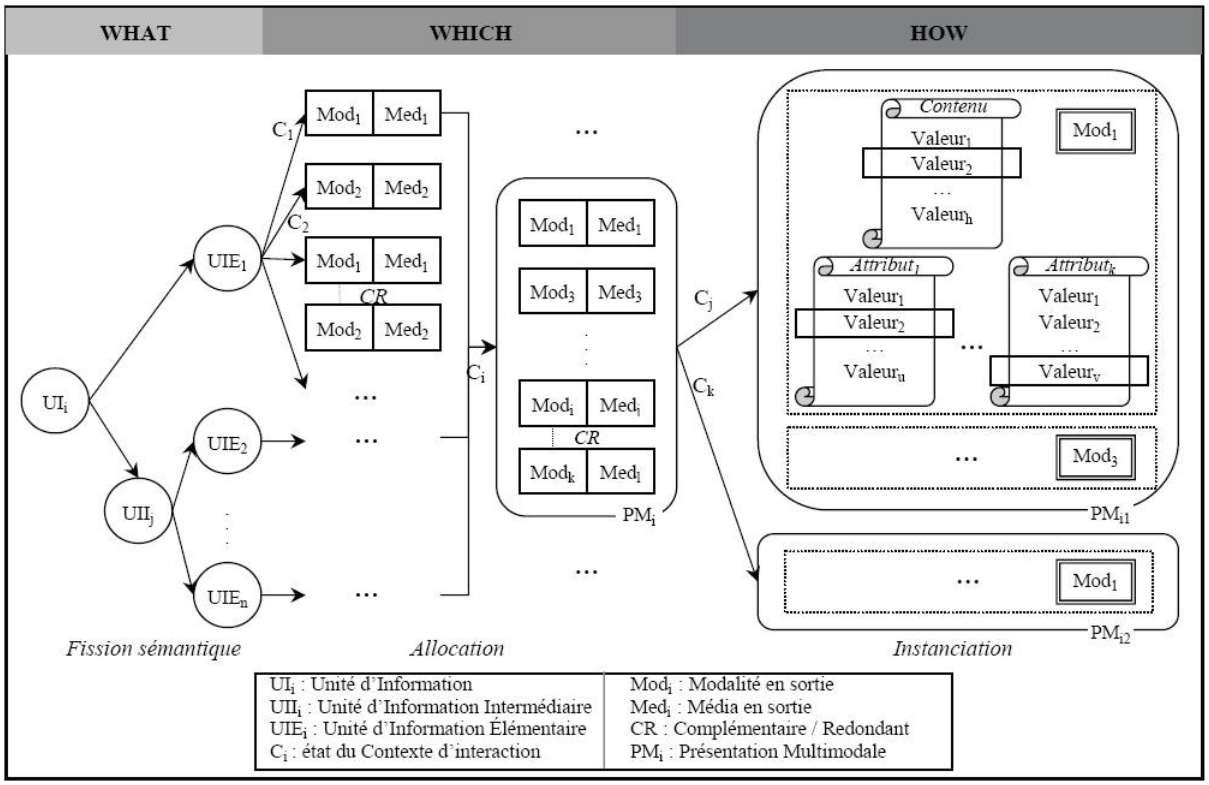

FIG. 4.11 – Étapes de conception d'une présentation multimodale adaptée au contexte d'interaction selon Rousseau (extrait de [Rousseau, 2006])

temps, l'élection pure puis l'élection composée. L'élection pure consiste en la sélection du meilleur couple <média, modalité> pour réaliser une unité informationnelle donnée, étant donné le contexte d'interaction. Un poids d'adéquation avec le modèle comportemental est attribué à chaque couple. L'élection composée sélectionne de nouveaux couples complémentaires ou redondants avec le premier. Le poids des couples intervient dans cette nouvelle sélection. Ce deuxième jeu de couples ainsi sélectionné sert de recommandation qui pourra ne pas être suivie, ou de façon partielle, lors de l'étape de réalisation (celle qui répond à la question "*how ?*"). Par rapport à Arch, l'allocation des modalités et des médias se fait au niveau du composant de présentation (abstraite). L'ensemble des unités informationnelles associées à un ou plusieurs couples <média, modalité> correspond à une présentation abstraite transmise au composant d'interaction.

La troisième étape de production d'une réponse multimodale consiste à répondre à la question "*how ?*" : après la fission sémantique et l'allocation des modalités, il s'agit d'instancier la présentation abstraite produite. Plus précisément, pour chaque modalité et chaque unité informationnelle, une présentation concrète est choisie et les attributs de présentation sont fixés. Ce processus repose sur un modèle d'instanciation. Celui-ci référence l'ensemble des instances possibles pour une modalité donnée et les contextes

d'interaction avec lesquels ces instances sont jugées compatibles. Un modèle d'instanciation d'une modalité spécifie une stratégie d'instanciation en fonction d'un contexte d'interaction donné, d'une unité informationnelle à concrétiser et d'un média (*i.e.* dispositif physique) considéré. L'application du modèle d'instanciation conduit à la sélection de la modalité répondant à ces critères. Pour une présentation multimodale, *i.e.* pour un ensemble d'unités informationnelles multimodalement allouées, sont sélectionnés les modèles d'instanciation dont l'union correspond à l'instantiation de la présentation, *i.e.* des unités informationnelles, à réaliser. Par rapport à Arch, cette étape se fait au niveau du composant d'interaction.

Notons que Rousseau précise que chacune de ces étapes pourraient être automatisée et constituer chacune un sujet de recherche à part entière.

Parce que ses travaux sont focalisés sur les sorties multimodales, Rousseau propose un modèle conceptuel qui détaille les différentes étapes de la génération d'une sortie multimodale et donc les différents choix que doit faire le système. Gardant à l'esprit qu'une présentation doit toujours être adaptée au contexte d'utilisation au sens large, tout élément perturbateur (Shannon parlerait de bruit [Shannon, 1948]) est pris en compte. En particulier, Rousseau intègre dans son modèle du contexte un modèle de l'utilisateur, de l'environnement physique et de l'environnement logiciel d'utilisation. De plus, les étapes consistant à déterminer la présentation à réaliser, que ce soit à un niveau abstrait (*i.e.* l'allocation en réponse à la question "*which ?*") ou concret (*i.e.* l'instanciation en réponse à la question "*how ?*") prennent en compte deux autres modèles que sont le modèle comportemental et le modèle d'instanciation. Le modèle comportemental, respectivement le modèle d'instancitation, permet de décrire l'adéquation entre une présentation abstraite donnée *i.e.* un couple <média,modalité> (Nigay dirait <dispositif physique, langage d'interaction>), respectivement une présentation concrète donnée *i.e.* une modalité caractérisée par ses attributs, et une unité informationnelle donnée.

4.2.3 Synthèse

De notre étude de l'interaction multimodale (paradigme actionnel), nous retenons que peu de travaux sont dédiés aux sorties. Comme le souligne le modèle WWHT, la difficulté de l'interaction multimodale en sortie réside dans les choix que doit effectuer le système interactif, choix qui revient à l'utilisateur pour le cas des entrées.

De plus, alors que les relations entre modalités ont été introduites dans des travaux s'inscrivant dans un paradigme actionnel de l'interaction et ont donné lieu à plusieurs mécanismes de fusion pour les entrées multimodales (*cf.* la section 2.1.3.2), ces relations ne sont que rarement exploitées dans les moteurs de fission. Notons néanmoins les travaux de Rousseau (*cf.* la section 2.1.4.2 et la section 4.2.2.3) et Vernier (*cf.* la section 2.1.3.3) qui exploitent les propriétés CARE. Dans les autres travaux, comme l'architecture proposée dans (*cf.* la section 4.2.2.2), nous faisons l'hypothèse que la fission effectuée par le générateur multimodal détermine si les unités informationnelles doivent être présentées de façon redondante, complémentaire ou assignée en fonction des critères sémantiques et ergonomiques pris en compte.

Enfin, nous constatons que le processus de génération de présentation multimodale

exploite de nombreux modèles, certains communs avec l'approche dialogique. Il s'agit :

- du modèle des tâches qui organise les tâches que l'utilisateur peut/doit réaliser ;
- du modèle de dialogue, le plus souvent implicite au niveau du composant de dialogue ;
- du modèle du contexte qui inclut au minimum les préférences de l'utilisateur, voire la caractérisation des environnements logiciel et physique d'utilisation ;
- de l'historique du dialogue qui précise au minimum les changements d'état du dialogue et l'état de la présentation rémanante (généralement visuelle) mais peut aussi inclure un historique des interactions ;
- du modèle du domaine qui regroupe les données applicatives.

S'y ajoutent les modèles de comportement et d'instanciation des modalités introduits par Rousseau qui permettent de formaliser les stratégies de présentation abstraite et concrète utilisées par le système. Ces modèles sont à rapprocher des comportements fortement contraints identifiés par Clémente (*cf.* la section 4.1.2.4).

Toutefois, l'uniformité des noms dans les approches dialogiques et actionnelles ne révèle pas toujours une uniformité des concepts et encore moins de l'approche sous-tendue. Nous détaillons cela dans la partie suivante qui ne porte pas à proprement parler sur un nouveau paradigme de communication humain-machine mais plutôt sur un rapprochement des deux paradigmes que nous venons de présenter.

4.3 Vers une approche intégrée de la communication humain-machine

Les deux paradigmes, dialogique et actionnel, de la communication humain-machine ont été étudiés et développés parallèlement, par des communautés qui se sont longtemps ignorées. Or, ces deux paradigmes ne sont pas antinomiques et leurs études respectives peuvent au contraire s'enrichir au contact l'un de l'autre. Ce constat motive l'objet de ce paragraphe qui est dédié aux travaux alliant les deux paradigmes.

4.3.1 Paradigmes dialogique et actionnel : des approches complémentaires

Basé sur le constat que les paradigmes dialogique et actionnel sont étudiés par deux communautés de recherche distinctes, nous avons structuré ce chapitre en présentant indépendamment les travaux sur les sorties multimodales de ces deux communautés. Néanmoins il convient de noter des rapprochements entre ces deux communautés, qui ont mené à des études alliant les deux paradigmes. Avant de présenter les résultats principaux de ces études au sein de systèmes ou de modèles d'architecture, nous montrons comment le rapprochement des deux communautés a commencé à s'opérer dans chacune des communautés initiales.

4.3.1.1 Du dialogue vers les interfaces à manipulation directe

Les systèmes de dialogue humain-machine, essentiellement oraux initialement et parfois textuels, se sont dotés d'éléments typiques des interfaces multimodales à manipulation directe. Notamment, les cartes, support visuel servent de support au dialogue. Le texte écrit, descriptif et informatif plus que dialogique, et pas nécessairement basé sur le langage naturel à proprement parler (mots-clés, liste d'items, etc.), a aussi été intégré en plus des langages naturels oral et écrit : comme la carte, il est utilisé comme support au dialogue.

De plus, avec l'utilisation exclusive du langage naturel oral, l'historique du dialogue n'avait pas à tenir compte de la rémanence des informations présentées. Les modalités visuelles au sein des systèmes de dialogue ont entraîné la prise en compte de la rémanence des informations dans l'historique du dialogue.

Par ailleurs, si le modèle de tâche était classiquement implicite, que ce soit au niveau du modèle de dialogue ou au niveau du modèle du domaine, il est de plus en plus explicité et considéré comme un modèle à part entière. Ceci permet de distinguer la tâche principale pour laquelle a été conçu le système et l'appréhension de cette tâche par l'utilisateur comme autant de sous-tâches ou de tâches plus élémentaires. Le modèle de tâche est peu à peu adapté pour devenir un modèle *des* tâches qui prend plus en compte le point de vue de l'utilisateur.

Enfin, l'architecture classique des systèmes de dialogue humain-machine distingue les modules de traitement en fonction de type de messages échangés, à savoir un énoncé oral, un énoncé textuel ou une représentation logique de cet énoncé. Dans le modèle Arch, le plus proche de l'architecture classique des systèmes de dialogue humain-machine bien que dédié aux interfaces de type actionnel, les composants correspondent à des niveaux d'abstraction dans la chaîne de traitement des messages perçus ou produits, qui sont le niveau articulatoire ou lexical, le niveau syntaxique et le niveau sémantique. Dans les systèmes de dialogue, ces trois niveaux sont traités par le module de génération. Certaines architectures de systèmes de dialogue multimodaux (*cf.* la section 4.1.2.1, la section 4.1.2.2 et la section 4.1.2.3) partitionnent le module de génération (ainsi que celui d'interprétation) de façon à ce que ces différents niveaux d'abstraction soient explicites. Il en résulte parfois une multiplicité des composants rendant l'architecture résultante complexe.

Si l'on constate une certaine tendance des systèmes de dialogue humain-machine à aller dans le même sens ou à s'inspirer des systèmes à interfaces à manipulation directe, certaines différences demeurent. Nous en listons quelques-unes ci-après.

Comme nous l'avons souligné à plusieurs reprises, les architectures de systèmes de dialogue distinguent, en général, le traitement de l'entrée et celui de la sortie. Dans le cas d'une interaction en langage naturel oral, l'entrée du système peut sembler décorrélée de sa sortie d'un point de vue articulatoire/lexical, car l'utilisateur n'utilise pas les moyens d'expression - ou capacités d'actions - du système pour agir sur ce dernier. Or cette corrélation est possible avec l'utilisation du visuel en sortie et du gestuel en entrée. Ce constat met en évidence un fait simple trop souvent sous-estimé dans les systèmes de dialogue humain-machine : les capacités d'action de l'utilisateur sont conditionnées

par la sortie du système. En effet, les études ergonomiques montrent que, quelle que soit la modalité, le message produit par le système peut inciter l'utilisateur à utiliser une capacité d'action plutôt qu'une autre - c'est l'effet de modalité [Lieury, 2005] - ou limiter (si l'utilisateur ignore certaines commandes vocales par exemple) voire interdire l'utilisation d'une capacité d'action (comme c'est le cas de la saisie de texte s'il n'y a pas un champ de saisie affiché) [Karsenty, 2006, Le Bigot *et al.*, 2006, Fréard *et al.*, 2007]. Les systèmes multimodaux devraient donc garantir les capacités d'action de l'utilisateur, quelles que soient les restrictions de présentation existantes.

Par ailleurs, les systèmes de dialogue multimodaux présentés n'ont pas une gestion du dialogue complètement amodale : une partie de calcul de réaction sur les intentions et buts de l'utilisateur, est amodale et une autre, qui se fait parfois au niveau du module de génération, tient compte des modalités. Si ce choix permet de prendre en compte les caractéristiques des modalités mobilisées pour la sélection du contenu à présenter, la réutilisabilité de la gestion du dialogue dans le cas de modalités différentes est compromise. L'amodalité de composant de dialogue dans les architectures comme une architecture Arch garantit sa réutilisation et son extension à de nouvelles modalités.

De plus, le choix de modalités, *i.e.* de répartition des informations sur les modalités considérées, se fait en général à un seul niveau, ce qui n'est pas le cas dans les interfaces de type actionnel. Ainsi est-il classique de distinguer au moins deux niveaux dans le choix des modalités au sein d'une architecture comme Arch : ces niveaux correspondent à ceux de composition des modalités proposés par Vernier (*cf.* la section 2.1.3.3), aux étapes d'allocation et d'instanciation utilisés par Rousseau (*cf.* la section 4.2.2.3) ou tout simplement à la distinction entre présentation abstraire et concrète. Cette remarque va de pair avec la précédente, car si la gestion du dialogue n'est pas complètement amodale dans les systèmes de dialogue, c'est parce qu'il y a centralisation des décisions au niveau du module de gestion de dialogue. L'organisation de la gestion du dialogue de façon non centralisée, avec notamment une répartition des choix de traitements à réaliser, pourrait ainsi permettre d'accroître la réutilisabilité et l'extensibilité du dialogue.

Enfin, si la notion de "stratégie de dialogue" est plutôt bien explicitée dans les systèmes de dialogue, celle de "stratégie de présentation" reste encore à affiner. En effet, elles sont généralement définies de façon implicite et corrélées à la stratégie de dialogue. Les travaux de Clémente (*cf.* la section 4.1.2.4) vont dans ce sens avec la prise en compte de différentes contraintes liées à la présentation pour le choix de ladite présentation.

4.3.1.2 Des interfaces à manipulation directe vers le dialogue

Les premiers appels explicites de rapprochement entre paradigme dialogique et actionnel ont émanés de la communauté des interfaces. Le plus ancien est sans doute celui de Nielsen [Nielsen, 1993] qui part du constat que les interfaces graphiques sont basées sur des commandes dont la complexité est cachée par la manipulation directe et les principes WIMP. Il appelle à une interaction qui ne soit pas basée sur des commandes à exécuter par la machine et propose pour cela un accroissement du recours au langage naturel non contraint. Pour Beaudoin-Lafon [Beaudoin-Lafon, 2004], il s'agit d'aller

plus loin que de concevoir des interfaces en cherchant à concevoir l'interaction (*"Designing interaction, not interfaces"*). Le terme "interaction" tel qu'employé par Beaudoin-Lafon est à rapprocher de ce que nous appelons la communication humain-machine au sens large. Distinguant les paradigmes existants de communication humain-machine, Beaudoin-Lafon déclare qu'il ne pense pas qu'un de ces paradigmes est meilleur que les autres et croit que ces paradigmes doivent être intégrés dans une seule et unique vision de la communication humain-machine[7].

Ces appels au rapprochement du dialogue et des interfaces (entre autres) sont à l'origine même d'une nouvelle génération d'interfaces, la quatrième selon [van Dam, 1997]. Pour van Dam, l'appel de Nielsen a marqué le lancement des interfaces post-WIMP qui cherchent à dépasser les limites de la communication à base de commandes dans les interfaces à manipulation directe. Pour cela, elles étudient des tâches plus complexes à réaliser (*e.g.* les interfaces zoomables), en intégrant éventuellement de nouvelles modalités informatiques (*e.g.* la réalité augmentée ou les interfaces intégrant des illusions auditives) et/ou sensorielles (*e.g.* les interfaces haptiques) et admettant l'utilisation conjointe de plusieurs utilisateurs.

Par ailleurs, le traitement de la sortie des systèmes relevant du paradigme actionnel a longtemps été considéré comme parallèle à celui de l'entrée. Les travaux de Rousseau montre qu'il commence à y avoir une différenciation entre traitement de l'entrée et traitement de la sortie et une prise en compte des problématiques de choix spécifiques à la sortie. Cette prise en compte des caractéristiques du traitement de la sortie est à l'origine d'autres travaux que la multimodalité, tels que la plasticité, la visualisation à grande échelle, etc.

De plus, le modèle des tâches définissant l'espace de résolution de l'utilisateur peut être rapproché d'une identification des objectifs de l'utilisateur basée sur ses actions. Il est ainsi possible d'aller plus loin en anticipant non seulement les actions mais aussi les attentes à venir de l'utilisateur. Ceci reviendrait à déterminer le modèle des tâches effectives en fonction des actions de l'utilisateur qui sous-tendraient ses objectifs, *i.e.* ses buts récurrents sans les formaliser en tant que tels. La mémorisation des interactions passées dans un historique du dialogue, comme cela est déjà fait dans certains systèmes de type actionnel, est un premier pas vers l'extraction de comportements récurrents pour accomplir une tâche donnée par un utilisateur et correspond à l'identification de buts récurrents de ce dernier. Cette anticipation des attentes et actions de l'utilisateur contribue à ce que ces systèmes soient plus considérés comme des partenaires que comme outils [Beaudoin-Lafon, 2004].

Malgré ces rapprochements, nous constatons aussi des divergences. Tout d'abord, la notion de "stratégie de dialogue" n'est pas explicitée dans les interfaces à manipulation directe. Le système étant vu comme un outil, il n'est pas en mesure de choisir la façon dont il va interagir avec l'utilisateur autrement que par des réactions pour répondre aux actions de ce dernier. La stratégie de dialogue n'a de sens que s'il est admis que la machine est en mesure d'influencer le comportement de l'utilisateur, ce qui sous-entend

[7] Traduction libre de "*no paradigm can subsume the others and I believe that, ultimately, all three paradigms must be integrated into a single vision*".

un rôle moins réduit que celui d'un outil réactif. Pourtant, la notion de "stratégie de présentation" est déjà admise dans Arch où elle est choisie au niveau de la présentation abstraite et au niveau de la présentation concrète.

De plus, nous avons reproché aux systèmes de dialogue de ne pas avoir une gestion du dialogue strictement amodale. Le reproche inverse est possible pour les systèmes relevant du paradigme actionnel. En effet, la définition d'un composant de gestion du dialogue indépendant des modalités permet aussi bien la réutilisation que l'extension de ce composant. Toutefois, il nous semble qu'une partie de la gestion du dialogue pourrait tenir compte des modalités, en particulier si la possibilité est laissée à l'utilisateur de contraindre la présentation du système ou d'interroger ce dernier sur ses capacités de présentation. Le premier cas est généralement traité au niveau des composants de présentation (abstraite et concrète) mais le type des données peut poser problème pour un tel traitement : quel traitement est-il possible si l'utilisateur demande une photo sous format auditif ? Le système devrait alors au moins être en mesure d'expliquer pourquoi cette présentation est impossible, autrement dit devrait être en mesure de communiquer sur ses capacités de présentation. Or une telle réponse relève de la gestion du dialogue. Plus précisément, l'objet du dialogue est l'interaction elle-même.

Le début de ce chapitre est bâti sur le constat de l'existence de deux communautés de recherche, l'une focalisant sur le paradigme dialogique et l'autre sur le paradigme actionnel. Nous soulignons une synergie encore timide entre les travaux des deux communautés. Il convient néanmoins de mettre en avant des travaux qui allient effectivement les deux paradigmes et dans la lignée desquels nous nous inscrivons, avant de pouvoir expliciter les fondements de notre contibution à cette convergence.

4.3.2 Des interfaces multimédias intelligentes à la présentation multimodale d'information

Très tôt, certains travaux ont compris l'intérêt qu'il pouvait y avoir à combiner les paradigmes de dialogue et d'interface et à tenir compte des notions et modèles identifiés par chacun de ces paradigmes. Ainsi, dès 1986, Norman [Norman, 1986] souligne que les buts et les intentions de l'utilisateur interviennent sur la réalisation de la tâche et, par conséquent, le séquencement des tâches en sous-tâches. Frohlich [Frohlich, 1996] constate que les paradigmes dialogique et actionnel ne sont pas adaptés aux mêmes tâches et que le choix de l'un ou l'autre dépend donc du type de tâche ou de sous-tâche que l'utilisateur est en train d'accomplir. De même, Colineau et Paris [Colineau et Paris, 2003] signalent que dans certains cas, en l'occurrence la génération de documents multimédias incluant graphiques et texte, le couplage d'approches différentes peut s'avérer nécessaire. Les systèmes centrés sur la génération multimédia sont couramment appelés IMMPS pour Intelligent MultiMedia Presentation Systems. La notion de "génération", ou de "présentation intelligente" renvoie à l'automatisation des présentations produites et celle de "multimédia" n'est pas antinomique avec celle de "multimodalité" telle que nous l'avons définie (*cf.* la conclusion du chapitre 2).

4.3.2.1 Architecture conceptuelle des interfaces intelligentes multimédias selon Roth et Hefley

Faisant un point sur les travaux sur les systèmes intelligents de présentation multimédia[8], Roth et Hefley [Roth et Hefley, 1993] précisent les objectifs de tels systèmes. Pour eux, ces systèmes doivent servir d'intermédiaires entre l'utilisateur et les informations sauvegardées dans le ou les modèles du domaine. De plus, les présentations produites par ces systèmes doivent aider l'utilisateur à atteindre ses objectifs et à réaliser ses tâches. Ces systèmes sont donc à la fois outil (d'accès) et soutien. Les auteurs identifient les processus impliqués dans la génération d'interfaces multimédias intelligentes ainsi qu'une architecture. Processus et architecture sont détaillés ci-dessous.

Sans tenir compte de la réalisation effective de la présentation, plusieurs processus sont impliqués dans la production automatique de présentations multimédias. Il s'agit de :

- la sélection du contenu (*content planning*) : elle doit se faire en considérant la tâche de l'utilisateur ;
- la sélection des modalités (*technique selection*) : plus précisément, il s'agit de sélectionner les modalités à utiliser pour la présentation (*i.e.* langages d'interaction et dispositifs physiques), de déterminer la répartition des unités informationnelles sur ces modalités et de coordonner les modalités entre elles pour chaque unité informationnelle considérée ;
- la conception de la présentation (*presentation design*) : l'utilisation précise des modalités pour les unités informationnelles à présenter est définie ;
- la coordination (*coordination*) : la présentation est composée et organisée en tenant compte des conflits éventuels et en veillant au maintien de sa cohérence globale.

Plusieurs précisions sont apportées sur ces processus. Tout d'abord, il ne constituent pas une architecture de génération multimodale (ou multimédia intelligente) à proprement parler. Ces processus peuvent être récursifs à différentes étapes de la génération. Par exemple, le processus de sélection des modalités peut être affiné, par la sélection du sens de perception, puis par la sélection du langage d'interaction et enfin par la sélection du dispositif physique. Il en va de même pour les autres processus. Ensuite, et pour corollaire, ces processus ne peuvent être exécutés de façon séquentielle. La génération multimodale doit être le résultat d'une collaboration entre ces processus ou, à défaut, de retours et de mises à jours en conséquence. Pour reprendre les termes des auteurs, le processus de production d'une présentation comporte plusieurs sous-processus parallèles en interaction : ceci implique des échanges entre les composants impliqués dans le choix et l'utilisation des modalités plutôt qu'un processus séquentiel et hiérarchisé[9].

Ce constat amène Roth et Hefley à proposer l'architecture conceptuelle pour les systèmes intelligents de présentation multimédia présentée à la figure 4.12. On y re-

[8] Traduction libre d'IMMPS.

[9] Traduction libre de : "*an effective presentation process must involve processes which are parallel and interacting. This suggests extensive feedback between components making decisions about media and modalities, rather than hierachical, sequential decision-making process.*"

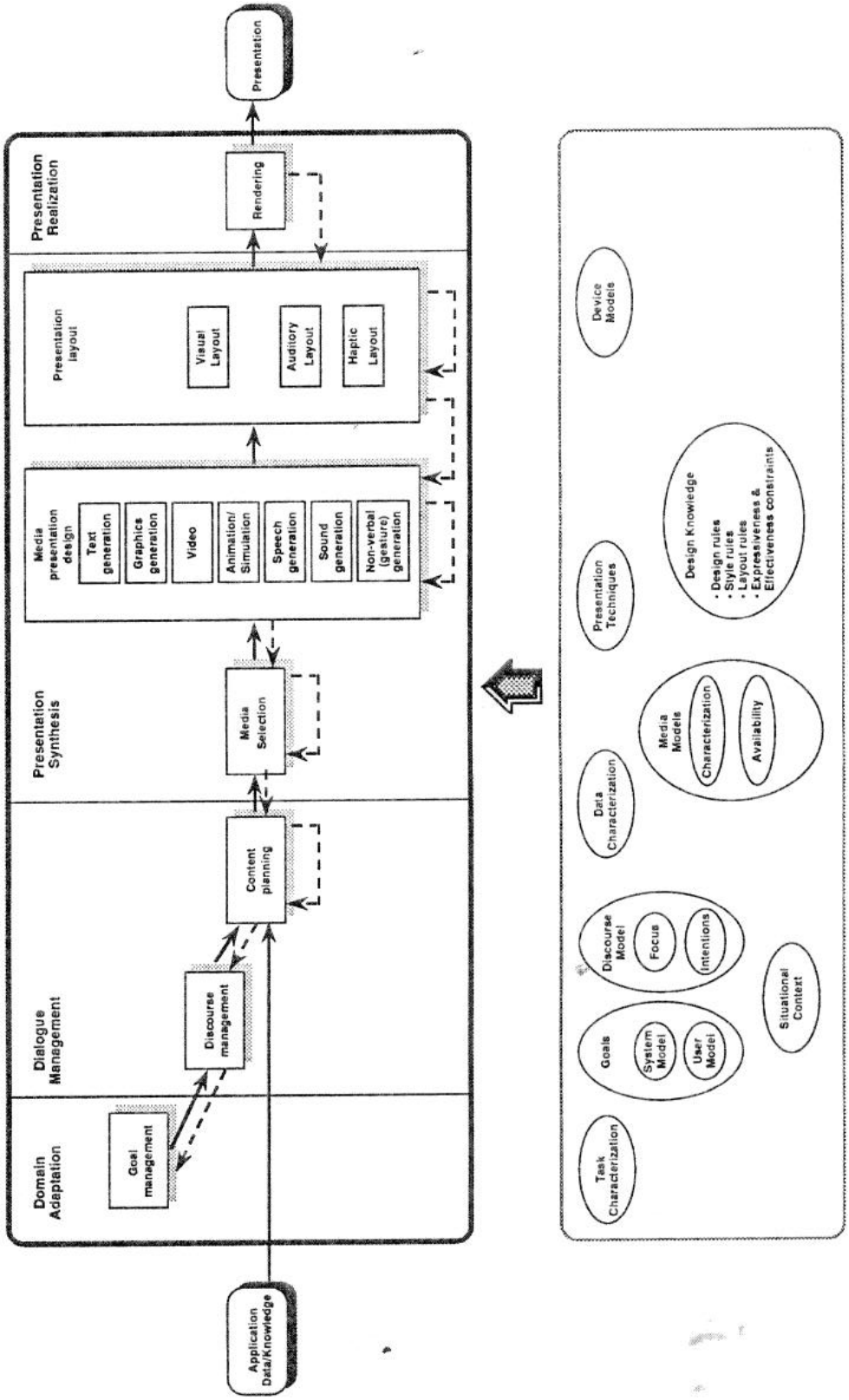

FIG. 4.12 – Architecture conceptuelle pour les systèmes intelligents de présentation multimédia (extrait de [Roth et Hefley, 1993])

trouve les composants d'Arch. Les données et les connaissances applicatives (*application data/knowledge*), à gauche de la figure, correspondent au composant du domaine, l'adaptation du domaine (*domain adaptation*) à celui d'Arch, la gestion du dialogue (*dialogue management*) au composant de dialogue, la synthèse de la présentation (*presentation synthesis*) au composant de présentation et la réalisation de la présentation (*presentation realization*) au composant d'interaction. Toutefois, les flèches en pointillé apportent une nuance de taille à cette architecture séquentielle : elles correspondent aux retours, mises à jour et coordination entre les différents composants classiques. Le processus de coordination est donc géré en partie, en ce qui concerne la gestion des conflits et la cohérence, par ces retours et mises à jours. Les auteurs ne nient pas que la séparation entre la sélection du contenu et la sélection de la présentation, classique dans les systèmes de type actionnel en général et établie comme principe dans Arch, est extrêmement utile pour garantir la réutilisation et l'extension du composant de dialogue. Mais ils considèrent que c'est finalement plus un mal qu'un bien car les choix dépendants du type, de la quantité, du niveau de détail, de la combinaison et de la répartition des unités informationnelles dans un énoncé unique ou dans plusieurs énoncés devraient se faire en fonction des limitations d'affichage, des capacités des dipositifs physiques, des obligations de présentation imposées par des conventions applicatives ou par l'utilisateur ainsi que de la complexité cognitive résultant des différents contenus présentés[10]. Roth et Hefley défendent donc des interactions entre composants durant la génération et vont jusqu'à prôner la prise en compte de contraintes liées à la présentation finale pour le choix et la répartition des informations.

Par ailleurs, Roth et Hefley plaident pour la prise en considération des buts de l'utilisateur pour la détermination de la réponse du système. Plus précisément, ils constatent que la réussite de l'appréhension des informations présentées selon une certaine combinaison de modalités dépend des tâches de l'utilisateur et du contexte interactionnel. Par conséquent, la génération d'une présentation multimédia intelligente ne peut se faire qu'en fonction du modèle du domaine et doit aussi prendre en compte les buts, les tâches et les mécanismes cognitifs de l'utilisateur ainsi que les buts du système et l'état courant de l'interaction. Pour caractériser ces éléments et les faire intervenir lors de la génération de la présentation, modèle des tâches (*task characterization*) et/ou modèle des actes communicatifs (*discourse model*) sont utilisés dans les systèmes intelligents de présentation multimédia. Ces systèmes adoptent donc un comportement plus anticipatif que réactif, et s'ils restent des outils, sont des outils coopératifs.

[10] Traduction libre de : "*Ultimately, however, il will be a mistake to strictly isolate application content selection from IMMPS components. Decisions about the type, quantity, level of detail, combinations and distribution of information within a single, or across a sequence of discourse segments should be coordinated with and strongly influenced by display space limitations, hardware capabilities [...], design commitments made previously conveyed information (possibly by user-imposed design choices or domaine conventions), and by a consideration of the cognitive complexity resulting from attempts to convey different content alternatives.*"

4.3.2.2 Modèle de référence pour les systèmes intelligents de présentation multimédia

Le modèle de référence pour les IMMPS proposé dans [Bordegoni *et al.*, 1997] est une architecture générique conceptuelle qui décrit le processus de la génération multimodale. Comme le montre la figure 4.13, il est composé de couches abstraites qui correspondent aux grandes étapes de la génération multimodale. Ces couches sont au nombre de cinq :

- la couche de contrôle (*control layer*) : elle détermine les buts de présentation ;
- la couche de contenu (*content layer*) : à partir des buts de présentation produits par la couche de contrôle, la couche de contenu affine les buts (*goal refinement*), sélectionne le contenu (*content selection*), alloue les modalités (*media allocation*) et ordonne les informations (*ordering*). Il en résulte des unités informationnelles associées à des modalités (ou au moins à des langages d'interaction) constituant une représentation abstraite de la sortie multimodal. La couche de contenu intègre les processus de sélection du contenu et de sélection des modalités (*cf.* la section 4.3.2.1) ;
- la couche de conception (*design layer*) : partant de la représentation abstraite de la sortie multimodale produite par la couche de contenu, la couche de conception transforme les unités informationnelles en objets de la modalité associée et spécifie les liens spatio-temporels entre ces objets modalement alloués. Le résultat correspond à des plans de réalisation (*realization plans*). La couche de conception est en charge du processus de conception de la présentation (*cf.* la section 4.3.2.1) à un niveau abstrait ;
- la couche de réalisation (*realization layer*) : utilisant les plans de réalisation produits par la couche de conception, la couche de réalisation génère une spécification des objets concrets à générer, ainsi que leurs organisations spatiales et temporelles. Elle correspond au processus de conception de la présentation (*cf.* la section 4.3.2.1) à un niveau concret ;
- la couche de production de la présentation (*presentation display layer*) : elle convertit la spécification produite par la couche de réalisation en une présentation perceptible par l'utilisateur, en respectant la coordination spatio-temporelle.

Le processus de coordination identifié par Roth et Hefley (*cf.* la section 4.3.2.1) est traité au niveau de plusieurs couches du modèle de référence. Ceci correspond à l'approche de Roth et Hefley qui considèrent que ces processus s'appliquent de façon récurrente à différentes étapes de la génération multimodale. Toutefois, notons que, bien que la coordination spatio-temporelle soit largement intégrée dans le modèle de référence, la génération d'expressions référentielles (*referring expressions*) - qui explicitent la coordination via un lien entre les modalités - n'y est absolument pas prise en compte. Tout comme pour la coordination spatio-temporelle, la génération des expressions référentielles devrait être répartie sur les différentes couches du modèle de référence [André, 2000, Foster, 2002].

Les étapes du modèle de référence s'intégrent assez facilement dans Arch [UIMS, 1992]. En effet :

- le composant de dialogue d'Arch correspond à la couche de contrôle et à une

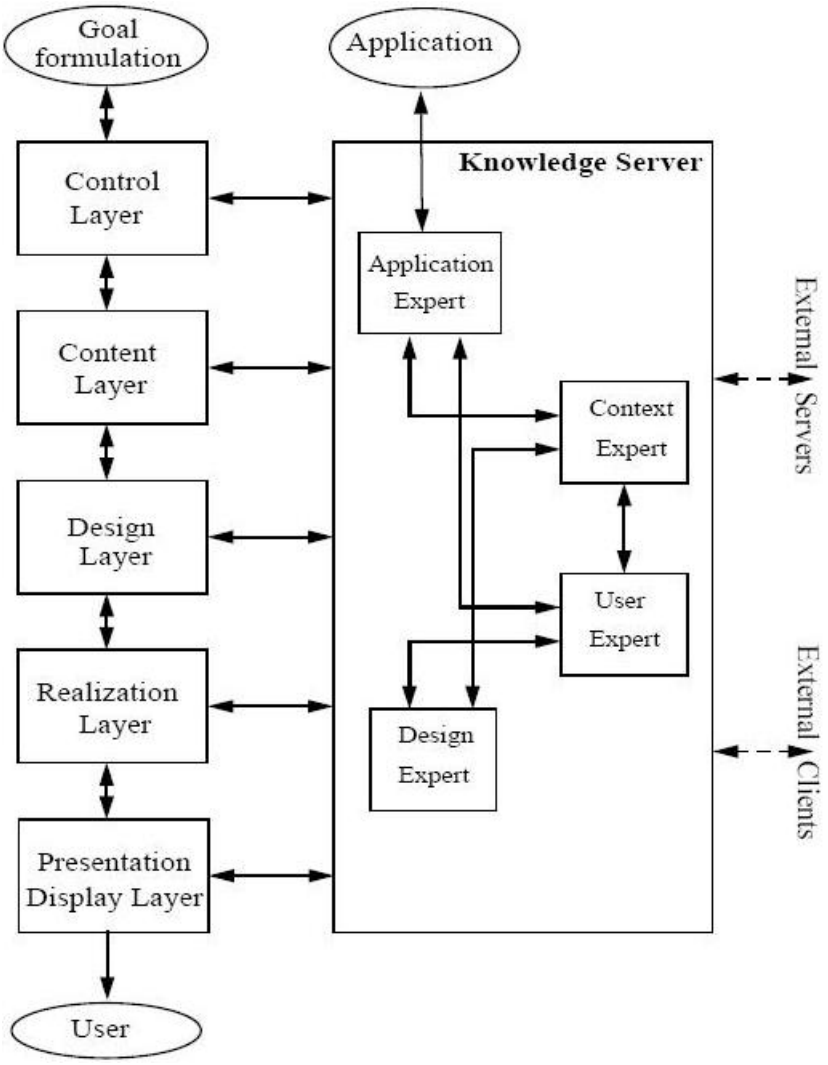

FIG. 4.13 – Modèle de référence pour les systèmes intelligents de présentation multimédia (extrait de [Bordegoni *et al.*, 1997])

partie de la couche de contenu (raffinement des buts, sélection du contenu et ordonnancement) du modèle de référence ;
- le composant de présentation d'Arch intègre une partie de la couche de contenu (allocation des modalités) et la couche de conception ;
- le composant d'interaction d'Arch correspond à la couche de réalisation et à la couche de production de la présentation.

Les couches de modèle de référence sont donc plus détaillées que les composants d'Arch. De plus, les réajustements entre les couches/composants, nécessaires à la génération multimodale tenant compte des contraintes de présentation et qui ne sont pas explicités dans Arch, sont clairement définis dans le modèle de référence. La séparation

entre le contenu et la forme, nette dans Arch, est ainsi modulée. Cette modulation est particulièrement importante lorsque les décisions prises à un niveau de traitement donné dépendent de décisions à prendre à un niveau subséquent.

Les travaux de Roth et Hefley (*cf.* la section 4.3.2.1) ainsi que le modèle de référence (*cf.* la section 4.3.2.2) sont considérés comme fondateurs pour les systèmes intelligents de présentation multimédia. Nous avons souligné dans ces travaux la synergie entre les travaux issus du paradigme dialogique et ceux du paradigme actionnel. Nous présentons maintenant des travaux récents qui s'inscrivent dans cette tendance et reposent sur ces modèles de référence.

4.3.2.3 Architecture proposée dans le cadre du projet COMIC

L'application étudiée dans le cadre du projet COMIC (pour "*COnversational Multimodal Interaction with Computers*") est un logiciel d'aide à la configuration de salles de bain [Foster *et al.*, 2005]. La sortie du système est composée d'une tête parlante (avatar) associée à un affichage graphique de la salle de bain configurée et à des pointages déictiques. Une architecture simplifiée valable pour tous les systèmes de dialogue combinant avatar parlant et affichage graphique est proposée. Cette architecture est présentée dans la figure 4.14. On y retrouve les modules de l'architecture des systèmes de dialogue adaptés à la multimodalité en fonction de certains principes d'Arch. C'est ce rapprochement, explicité par la suite, qui nous a poussé à ne rattacher cette architecture ni au paradigme dialogique ni au paradigme actionnel.

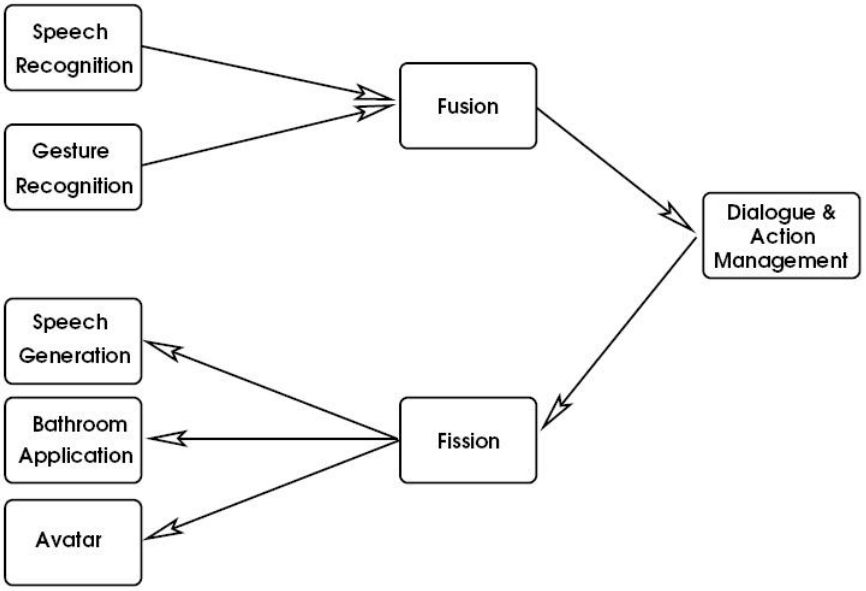

FIG. 4.14 – Architecture simplifiée de l'application COMIC (extrait de [Catizone *et al.*, 2003])

Le gestionnaire de dialogue et d'action (*dialogue and action manager*) décide du contenu à présenter. Comme le préconise Arch, il est indépendant des modalités. Il permet à la fois une interaction dirigée par l'utilisateur, par le système ou en initiative mixte. Il s'appuie sur le modèle du domaine, le modèle de l'utilisateur, l'historique du dialogue et l'ontologie du système pour choisir le contenu à présenter. Le modèle du domaine est constitué de réseaux de transition augmentés, intégrant une représentation du déroulement général du dialogue et des sous-tâches : il comprend donc le modèle de tâche. L'ontologie du système permet de sélectionner les modèles de salle de bain correspondant à la caractéristique demandée par l'utilisateur. Il correspond au composant d'adaptation au domaine dans Arch. Le modèle de l'utilisateur et l'historique du dialogue sont utilisés pour restreindre le nombre des modèles de salle de bain : il correspond donc au modèle du domaine classique. Le gestionnaire de dialogue et d'action envoie au module de fission une spécification de haut niveau des unités informationnelles à présenter, *i.e.* une représentation abstraite du contenu à présenter.

À partir de cette représentation abstraite, le module de fission conçoit la présentation à réaliser. Il structure la présentation en répartissant les unités informationnelles sur les différentes modalités. Plus précisément, il détermine le contenu graphique applicatif lié à la conception de la salle de bain et le contenu en langage naturel (en précisant éventuellement la tournure langagière à utiliser) grâce à l'ontologie du système, le modèle de l'utilisateur et le modèle du domaine. Il établit en conséquence le comportement de l'avatar (accents prosodiques, expressions et directions du regard) et les pointages déictiques. Il définit la coordination entre ces éléments et contrôle la concrétisation de la présentation. Le module de fission remplit les attributions du composant de présentation abstraite d'Arch. Mais son rôle est plus étendu dans la mesure où il gère aussi la concrétisation de la présentation : c'est lui qui synchronise la réalisation concrète des différents éléments de la présentation. Notamment, le synthétiseur de parole prépare le message en langage naturel oral, mais c'est le module de fission qui lui indique quand l'émettre. Le module de fission est donc aussi en charge de la synchronisation concrète correspondant au composant de présentation concrète dans Arch.

Le gestionnaire de dialogue et d'action (*dialogue and action manager*) correspond au module de gestion du dialogue de l'architecture des systèmes de dialogue. Le nom donné à ce gestionnaire est révélateur de la prise en compte des deux paradigmes d'interaction dialogique et actionnel. Il intègre une approche structurelle de l'interaction orientée par la tâche applicative. Le composant de fission, quant à lui, couvre à la fois le composant de présentation abstraite et une partie du composant de présentation concrète d'Arch : il conçoit la présentation et gère sa réalisation par les composants logiciels des dispositifs physiques. Dans l'architecture classique des systèmes de dialogue, il correspond à une adaptation multimodale du module de génération. Au module de synthèse classique s'ajoutent des modules permettant de concrétiser les autres formats de présentation : l'ensemble de ces modules correspond exclusivement à la partie concrétisation du composant d'interaction dans Arch. Par ailleurs, notons que le composant de fission accède aux mêmes modèles que le gestionnaire de dialogue et d'action, permettant d'affiner la présentation en cohérence avec le contenu sélectionné.

4.3.2.4 Architecture de référence pour les systèmes de présentation d'informations interactifs multimodaux

[Bunt *et al.*, 2005] propose une architecture de référence destinée à faciliter la compréhension et la description des systèmes interactifs multimodaux de présentation d'informations. S'appuyant sur l'étude de systèmes existants, l'architecture obtenue, présentée à la figure 4.15, distingue, comme l'architecture COMIC, entrée et sortie et trois niveaux de traitement. En tant qu'architecture de référence, elle est toutefois plus détaillée.

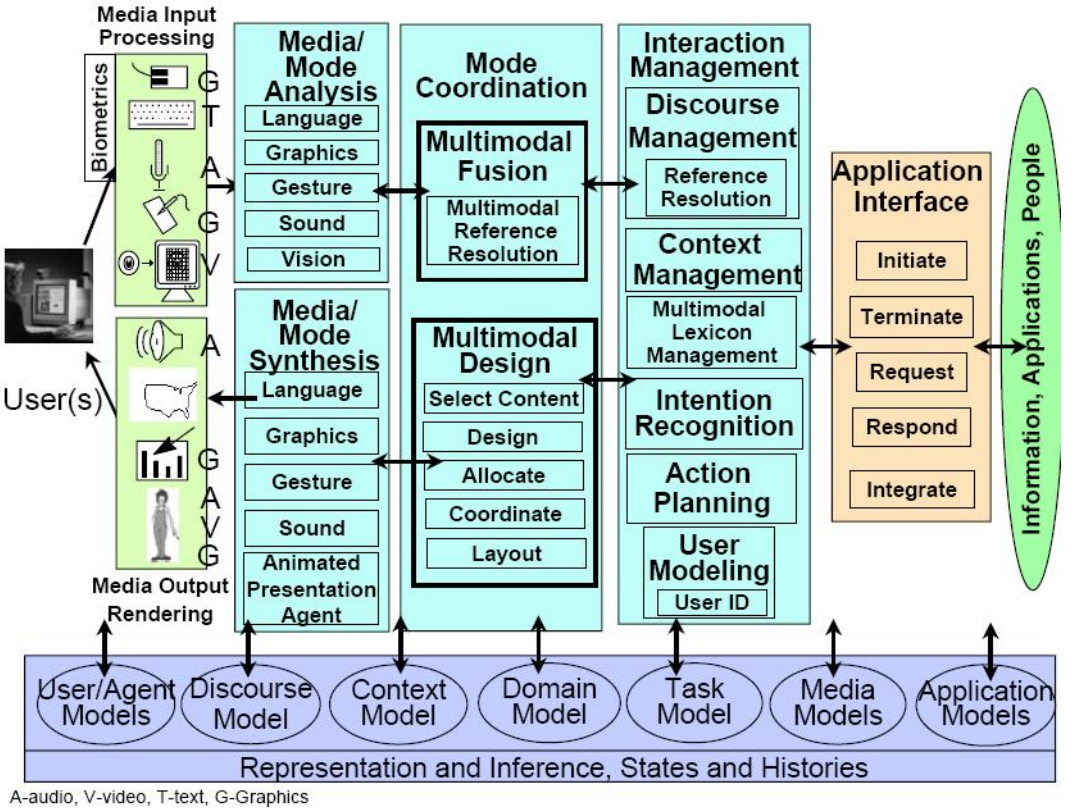

FIG. 4.15 – Architecture de systèmes de présentation d'informations interactifs multimodaux (extrait de [Bunt *et al.*, 2005])

Les processus impliqués dans la génération d'une présentation multimodale sont regroupés en trois niveaux de traitement : la gestion de l'interaction (*interaction management*), la conception multimodale (*multimodal design*) et la réalisation (*media/mode synthesis*).

Cherchant à contribuer à l'intention de l'utilisateur telle qu'elle a été identifiée en amont, le composant de gestion de l'interaction planifie la réponse du système en accédant éventuellement aux informations applicatives. Pour cela, il passe par une interface d'application (*application interface*) telle que définie dans le modèle de Seeheim ou en-

core dans Arch (sous le nom de composant d'adaptation au domaine). Il manipule et produit des actes de langage multimodaux. Ce processus est à rapprocher de l'unité rationnelle adaptée par Clémente pour les agents rationnels multimodaux et plurimodaux (*cf.* la section 4.1.2.4)). Ce composant ne respecte pas Arch dans la mesure où il n'est pas indépendant des modalités.

Les actes de langage multimodaux produits par le processus de gestion de l'interaction sont transmis au processus d'ébauche multimodale. S'il y a lieu, ce processus conçoit la présentation multimodale. Pour cela, le contenu est sélectionné, l'ébauche de l'utilisation des médias est établie, les modalités sont allouées, la coordination est prévue et le maquettage est spécifié. On retrouve les processus identifiés par Roth et Hefley (*cf.* la section 4.3.2.1), à savoir la sélection du contenu et des modalités, la conception de la présentation et la coordination. Le maquettage correspond à une anticipation de la réalisation qui, dans Arch, serait réalisée au niveau du composant d'interaction.

La présentation conçue est ensuite réalisée en faisant appel aux dispositifs physiques et langages d'interaction adéquats.

Chacun de ces processus exploite un ensemble de modèles correspondant aux modèles identifiés précédemment pour les systèmes de type dialogique et de type actionnel. Les modèles utilisés sont un modèle du domaine, un modèle de tâche, un modèle du contexte (réduit à l'état temporel et spatial de l'interaction), un modèle du discours et un modèle des agents impliqués dans l'interaction à commencer par l'utilisateur et le système (précisant leurs identités, leurs capacités, leurs croyances et leurs intentions). S'y ajoutent un modèle des applications et un modèle des modalités (précisant leurs propriétés et leurs codes, *i.e.* leur langage d'interaction). Ce dernier modèle est généralement sous-entendu dans l'étape de répartition ou d'allocation multimodale. L'historique de dialogue est absent de l'architecture en tant que modèle car il est maintenu pour chacun des modèles utilisés, accessibles par tous les composants qui interviennent dans la génération de la réponse multimodale du système. C'est un historique distribué selon les différents niveaux d'abstraction qui permet au système d'avoir une vision plus large de l'évolution de l'interaction et constitue une bonne base pour permettre une interaction plus naturelle, plus proche du comportement humain. Notons que cette distribution peut poser des problèmes d'homogénéité dans la gestion des historiques locaux.

4.3.2.5 Architecture SmartKom

Le projet SmartKom [Wahlster, 2006, Herzog et Reithinger, 2006] a pour but de proposer un cadre conceptuel et logiciel pour les systèmes d'informations multimodaux combinant un avatar parlant et un affichage visuel qui inclut du texte en langage naturel. Si l'architecture issue de ces travaux, et présentée dans la figure 4.16, se positionne dans une approche dialogique, certains éléments de la génération multimodale s'inscrivent plutôt dans une approche actionnelle. Nous détaillons ces éléments au fur et à mesure de notre description de l'architecture.

Un système qui s'appuie sur cette architecture est à initiative mixte et coopératif. Plus précisément, il est un partenaire : l'utilisateur lui confie une mission qu'il va chercher à réaliser. Wahlster [Wahlster, 2006] parle de délégation de tâche car le système va

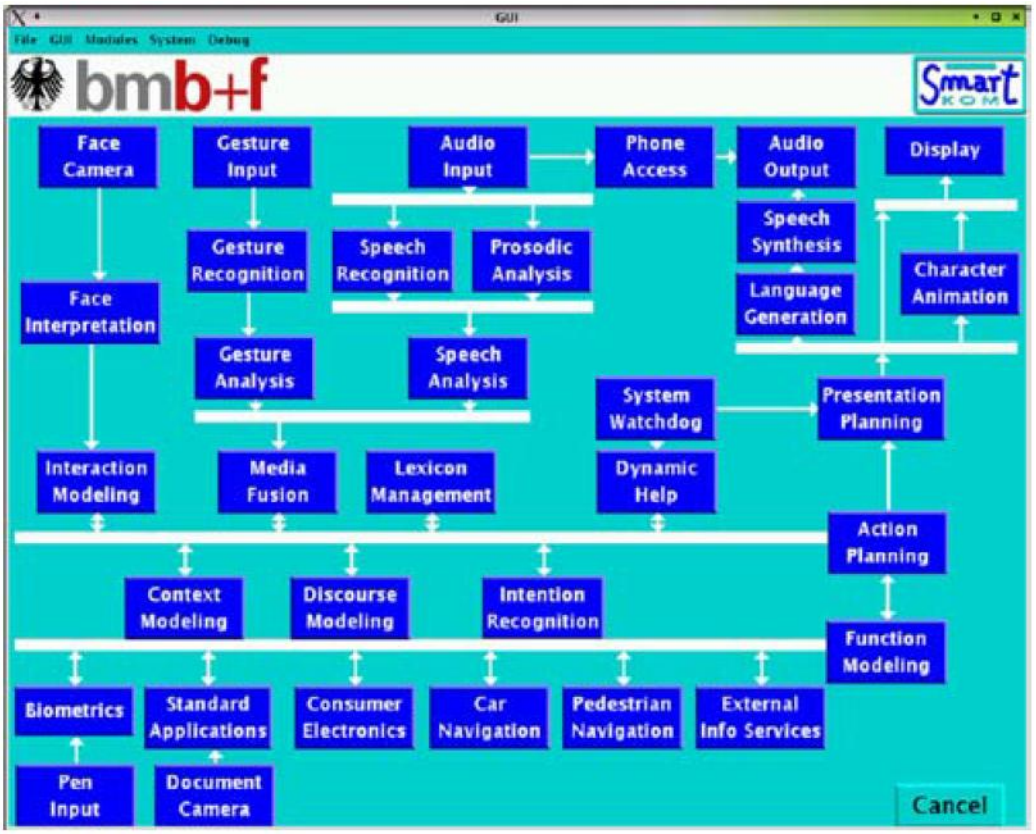

FIG. 4.16 – Architecture SmartKom (extrait de [Herzog et Reithinger, 2006])

chercher à connaître les buts de l'utilisateur - selon une approche dialogique - en termes de tâche que celui-ci cherche à accomplir - selon une inspiration du paradigme actionnel. L'interprétation des entrées de l'utilisateur consiste donc à reconnaître les buts, *i.e.* la tâche que cherche à accomplir l'utilisateur. Cette identification faite, une planification de l'action (*action planning*) conduit à l'identification d'un but de présentation. Ce but de présentation est amodal et peut être rapproché du contenu à présenter défini par le composant de dialogue dans Arch. Les composants en aval doivent réussir à réaliser le but de présentation, *i.e.* à présenter le contenu associé.

Les composants impliqués dans la fission multimodale sont présentés à la figure 4.17. Les étapes et éléments de l'architecture générale qui permettent la fission multimodale y sont détaillés. L'objectif de la fission multimodale est de passer d'une intention communicative, *i.e.* un but de présentation, indépendant des modalités à une présentation multimodale coordonnée et cohérente.

Au plus haut niveau d'abstraction, la fission multimodale est pilotée par un planificateur de présentation (*presentation planner* à la figure 4.17 correspondant à l'étape de *presentation planning* de figure 4.16). La planificateur de présentation décompose le but de présentation amodal qui lui parvient en tâches de présentation élémentaires.

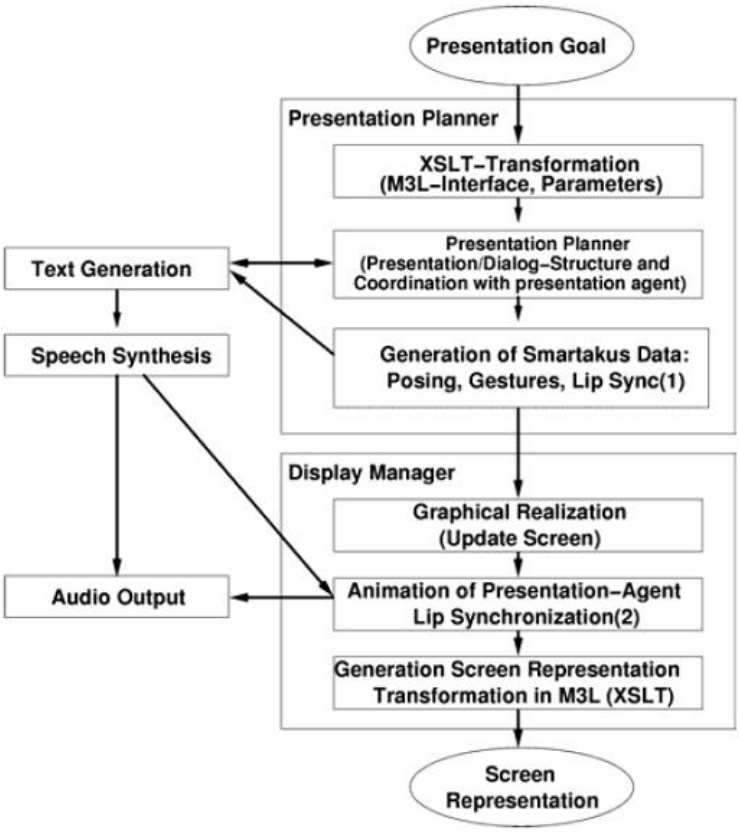

FIG. 4.17 – Fission multimodale dans SmartKom (extrait de [Herzog et Reithinger, 2006])

Pour cela, il utilise plus d'une centaine de stratégies de présentation et tient compte du contexte du discours, du modèle de l'utilisateur et des conditions d'utilisation (*i.e.* du contexte, des médias et des modalités). Ces stratégies de présentation sont adaptées en fonction des scénarios applicatifs considérés, en modifiant les paramètres qui encodent les préférences de l'utilisateur, sa langue maternelle, les dispositifs de sortie existants ou utilisés, etc. Outre les tâches de présentation élémentaires, les stratégies de présentation définissent aussi la répartition de ces tâches sur les modalités de sortie possibles, les styles et les objets concrets à utiliser ainsi que les comportements, *i.e.* les animations, de l'affichage visuel. En ce qui concerne les modalités dont le langage d'interaction est le langage naturel (*i.e.* oral ou textuel), les tâches de présentation correspondantes sont transformées en texte en dehors du planificateur de présentation (*text generation* dans la figure 4.17 et *language generation* dans la figure 4.16). Dans le cas du langage naturel écrit, le texte produit est transmis au composant de gestion de l'affichage. Toutes les tâches de présentation élémentaires visuelles sont ensuite transférées au composant aval dédié à la concrétisation de l'affichage. Dans le cas du langage naturel oral, le texte produit est directement synthétisé (*speech synthesis*).

Comme une partie de la fission multimodale, la production de la présentation concrète distingue tâches de présentation auditives et visuelles. Les tâches de présentation élémentaires auditives étant nécessairement en langage naturel oral, les textes synthétisés sont envoyés aux haut-parleurs qui les concrétisent (*audio output*). Les différentes tâches de présentation élémentaires visuelles, incluant texte, images et avatar, sont concrétisées par le gestionnaire d'affichage (*presentation planner/manager*). Celui-ci organise et synchronise les tâches de présentation visuelles, y compris par rapport aux tâches de présentation auditives, avant de les concrétiser.

Les modalités de sortie ne sont pas toutes traitées de la même manière. En particulier, la génération de langage naturel oral est plus rudimentaire que la génération visuelle. De plus, l'architecture intègre un historique du dialogue (*discourse model*) assez complexe tenant compte des modalités. Plus précisément, cet historique mémorise toutes les informations présentées en tenant compte de la modalité impliquée. Pour cela, il comprend une facette "domaine", une facette "dialogue" et une facette "modalité". La facette "modalité" comprend les objets linguistiques, graphiques et gestuels présentés en indiquant les objets du discours correspondants : un même objet du discours peut donc avoir plusieurs réalisations au niveau de l'aspect "modalité". L'aspect "domaine" fait le pont entre les objets du discours et les objets du modèle de domaine. Si le modèle du domaine est indépendant des modalités, ce n'est donc pas le cas de l'historique du dialogue qui est plus large qu'un historique du dialogue classique et prend en compte la forme des informations présentées. Un dernier élément qui tendrait à classer cette architecture comme étant dédiée à des systèmes de dialogue est que la machine y est vue comme un partenaire ou du moins comme un collaborateur.

D'autres éléments montrent sa prise en compte des enseignements tirés du paradigme actionnel. Tout d'abord, la planification de l'action, qui se situe au niveau du composant de dialogue dans Arch, est indépendante des modalités. Ensuite, au niveau de la fission multimodale, sont bien distinguées, à la suite de l'architecture Arch, présentation abstraite et présentation concrète. Enfin, les modalités visuelles sont dominantes : [Herzog et Reithinger, 2006] précise qu'au niveau du planificateur de présentation, les informations textuelles et graphiques sont choisies en premier, l'oral, respectivement l'avatar (*i.e.* ses gestes), complétant ensuite la présentation en tant que commentaire, respectivement en tant que déictique, des informations affichées.

4.3.3 Synthèse et positionnement

Dans la présente section, nous avons commencé par identifier les éléments de rapprochement entre paradigmes actionnel et dialogique ainsi que les limites de cette synergie. En mettant l'accent sur le dépassement de ces limites, nous avons ensuite présenté des travaux relevant des IMMPS, où les deux paradigmes de la communication humain-machine sont plus franchement combinés. Nous reprenons brièvement la façon dont ces limites sont dépassées et les apports des IMMPS avant de positionner la suite de notre contribution par rapport aux manques qui subsistent.

Tout d'abord, nous avons souligné, dans la section 4.3.1.1, l'intérêt d'une gestion du dialogue amodale et dans la section 4.3.1.2 celui d'une prise en compte, au minimum,

des contraintes de présentation dans cette gestion du dialogue et pour le choix du contenu. Dans un cas, modularité, réutilisabilité et répartition de choix de comportement du système sont répartis ; dans l'autre, les contraintes de présentation peuvent être prises en compte et le choix du comportement du système, tant sa réaction que la présentation de cette réaction, est centralisé, ce qui lui assure une maîtrise globale dudit comportement. Cette double critique trouve plusieurs réponses dans les travaux sur les IMMPS présentés. D'une part, l'architecture conceptuelle de Roth et Hefley (*cf.* la section 4.3.2.1), le modèle de référence pour les IMMPS (*cf.* la section 4.3.2.2) et l'architecture de référence pour les systèmes de présentation d'informations interactifs multimodaux (*cf.* la section 4.3.2.4) proposent une collaboration (par une concertation ou une négociation) entre les différents processus/couches qui interviennent dans la génération de la sortie. Ce choix permet de conserver une gestion du dialogue amodale, mais qui peut être redéfinie en fonction des retours des couches dédiées à la présentation. D'autre part, dans l'application COMIC (*cf.* la section 4.3.2.3), les mêmes modèles sont utilisés par le gestionnaire de dialogue et d'action et le composant de fission, servant d'intermédiaire entre les deux niveaux de détermination de la sortie.

Ensuite, nous avons évoqué, toujours dans la section 4.3.1.1, l'intérêt de décomposer l'étape de production de la présentation, comme c'est le cas dans les architectures des systèmes à manipulation directe (*cf.* la section 4.2.1.2). Cette décomposition existe dans l'architecture conceptuelle de Roth et Hefley (*cf.* la section 4.3.2.1), dans le modèle de référence (*cf.* la section 4.3.2.2) qui va jusqu'à affiner Arch et dans l'architecture SmartKom (*cf.* la section 4.3.2.5). L'architecture de référence pour les systèmes de présentation d'informations interactifs multimodaux (*cf.* la section 4.3.2.4) s'appuie sur le modèle de référence, même si elle place les différentes couches de ce modèle au même niveau.

Enfin, plusieurs travaux donnent aux systèmes intelligents de présentation multimédia un rôle de partenaire qui cherche à satisfaire les buts de l'utilisateur et donc, indirectement ou directement, à anticiper les actions et demandes de l'utilisateur (*cf.* l'une des limites évoquées dans la section 4.3.1.1). Plus précisément, l'architecture de Roth et Hefley (*cf.* la section 4.3.2.1) intègre les modèles de but (*goals model*), notamment de l'utilisateur, de discours (*discourse model*) - à rapprocher d'un historique du dialogue - et du contexte environnant (*situational context*). L'architecture de référence pour les systèmes de présentation d'informations interactifs multimodaux (*cf.* la section 4.3.2.4) utilise des modèles des agents, humain et artificiel (user/agent models), impliqués dans la communication humain-machine, un modèle du discours (*discourse model*) et un modèle du contexte (*context*). La génération de la sortie dans l'architecture SmartKom (*cf.* la section 4.3.2.5) s'appuie sur un but de présentation et utilise un historique du dialogue, un modèle du domaine et un modèle des modalités.

Malgré ces apports des IMMPS pour une convergence des paradigmes actionnel et dialogique, des différences subsistent sur lesquelles reposent la suite de nos contributions présentées dans les chapitres suivants. Tout d'abord, si nous adhérons à la distinction faite entre entrée et sortie des systèmes et à l'identification des problématiques liées à la génération de la sortie, les travaux présentés semblent négliger la corrélation à garantir entre sortie et entrée (*cf.* la limite présentée dans la section 4.3.1.1). En particulier, il

nous semble primordial de garder à l'esprit que, d'une part, la sortie d'un système a un impact sur l'utilisateur et sur la poursuite de la communication et que, d'autre part, le lien entre sortie et entrée doit être explicite via la garantie, en sortie, des moyens d'action disponibles pour l'utilisateur. Par exemple, même si le système présente une réponse (par exemple, une liste de solutions) auditivement, il ne doit pas contraindre l'utilisateur à interagir oralement avec le système et il doit lui laisser la possibilité d'utiliser la manipulation directe (par exemple, grâce à des champs de saisie ou des zones cliquables).

De plus, les notions de "stratégie de dialogue" et de "stratégie de présentation" ne sont pas vraiment explicitées. La notion de stratégie de dialogue n'intervient pas dans la génération de la présentation : elle est décidée en amont (voire même en amont de la sélection du contenu dans le modèle de référence - *cf.* la section 4.3.2.2 - et dans l'architecture des systèmes de présentation d'informations interactifs multimodaux - *cf.* la section 4.3.2.4) et son choix n'est pas évoqué. La stratégie de présentation n'intervient explicitement que dans l'architecture SmartKom (*cf.* la section 4.3.2.5). Il est possible que la dimension de concertation entre les différentes couches ou étapes rende caduque l'explicitation de ces stratégies, en tant que résultat de cette concertation.

Ceci nous amène à un point particulièrement flou dans les travaux sur les IMMPS. Si la collaboration entre processus/couches se retrouve dans les trois principales architectures conceptuelles, celle de Roth et Hefley (*cf.* la section 4.3.2.1), celle de Bordegoni et de ses collègues (*cf.* la section 4.3.2.2) et celle de Bunt et de ses collègues (*cf.* la section 4.3.2.4), la façon dont se fait cette collaboration n'est pas explicite. La solution que nous proposons dans le prochain chapitre est une solution pour réaliser cette collaboration tout en explicitant les choix de stratégies de dialogue et de présentation. Cette explicitation permettra de dépasser un dernier manque des travaux sur les IMMPS, celui de la prise en compte des contraintes de présentation. Roth et Hefley [Roth et Hefley, 1993] avaient déjà évoqué la nécessité de prendre en compte de telles contraintes (taille et complexité de la présentation, quantité d'informations, taille de l'écran s'il y en a un, cohérence et complétude de la présentation) dans des travaux ultérieurs mais sans proposer une solution qui permette effectivement cette prise en compte.

C'est sur ces différents constats que repose la suite de notre contribution présentée dans les chapitres suivants. Considérant que l'aspect naturel de la communication des systèmes d'information passe par leur intégration des paradigmes dialogique et actionnel, et constatant que les principaux points faibles des travaux existants concernent l'explicitation des stratégies de dialogue et de présentation et leur choix dans l'architecture des systèmes, nous proposons, dans le chapitre suivant, une caractérisation des notions de "stratégie de dialogue" et de "stratégie de présentation" ainsi qu'une architecture incluant un composant dédié au choix de ces deux stratégies.

4.4 Conclusion

Dans ce chapitre, nous avons souligné l'existence de deux communautés de recherche distinctes travaillant sur la communication humain-machine du point de vue informa-

tique, l'une privilégiant le paradigme dialogique et l'autre le paradigme actionnel. Notre contribution a été d'analyser en détail les résultats de chacun des domaines en nous focalisant sur la multimodalité en sortie, puis de mettre en exergue les points de synergie et les différences. Nous avons également identifié les éléments relevant d'un des deux paradigmes et exploités dans l'autre paradigme. Enfin, nous avons présenté les résultats principaux du domaine de la présentation multimédia intelligente, qui adopte une approche mixte en alliant les deux paradigmes et dans la lignée de laquelle nous nous inscrivons. L'identification des limites de ces travaux sur les IMMPS et notre conviction que le naturel de la communication des systèmes d'information dépend de la capacité de ces derniers à combiner les deux paradigmes dialogique et actionnel nous conduit, dans le chapitre suivant, à caractériser les notions de "stratégie de dialogue" et de "stratégie de présentation" et à proposer une architecture qui comprend un composant dédié au choix des stratégies de dialogue et de présentation.

Chapitre 5

Stratégies de dialogue et de présentation : un choix conjoint

Notre étude des paradigmes de communication humain-machine (*cf.* le chapitre précédent) nous a conduits à constater l'existence d'une convergence entre les deux principaux paradigmes, à savoir le paradigme dialogique et le paradigme actionnel. Ce rapprochement est nécessaire pour une communication humain-machine naturelle, ne serait-ce que pour que les systèmes soient en mesure de passer d'un mode d'interaction à un autre, comme le suggérait déjà Frohlich (*cf.* la section 2.1.1). Les travaux sur les systèmes intelligents de présentation multimédia qui se sont focalisés sur cette synergie suggèrent, pour la plupart, que le choix du comportement du système, *i.e.* de sa réaction et de la présentation de cette réaction, devrait être déterminé de façon concertée entre les différents composants qui leur sont dédiés dans l'architecture des systèmes. Nous constatons, toutefois, que, dans ces travaux, les notions de "stratégie de dialogue" et de "stratégie de présentation", sous-tendues au comportement du système, ne sont que rarement explicitées et que le déroulement de la concertation n'est pas évoqué.

De plus, notre premier objectif au sein d'une approche globale de la communication naturelle intégrant paradigmes dialogique et actionnel est de définir une architecture permettant aux systèmes multimodaux de prendre en compte les contraintes de présentation imposées par l'utilisateur et les contraintes inhérentes aux modalités. Ces contraintes constituent de façon plus générale des contraintes issues de la situation de communication. Pour cela, nous avons fait le choix de nous appuyer sur l'architecture Arch (*cf.* la section 4.2.1.2) dont la modularité assure la réutilisabilité des composants ainsi que la modifiabilité et l'extensibilité du code (pour l'ajout ou le changement d'une modalité par exemple) et qui peut être appliquée à la plupart des architectures (comme le modèle de référence pour les IMMPS - *cf.* la section 4.3.2.2).

Nous inspirant du paradigme dialogique, plus précisément de certains systèmes multimodaux de dialogue présentés dans le chapitre précédent, nous introduisons un composant, intermédiaire entre le composant de dialogue et le composant de présentation abstraite, destiné à tenir compte des contraintes de présentation évoquées dans la détermination du comportement du système. Ce composant, dédié au choix des stratégies

de dialogue et de présentation, concrétise la concertation suggérée dans les IMMPS entre les composants en charge du choix du contenu et ceux en charge du choix de la présentation.

Avant de détailler le composant proposé et son intégration dans Arch, nous exposons nos motivations et les concepts de stratégie de dialogue et de stratégie de présentation, manipulés par ce composant. Nous terminons ce chapitre par une revue des limites de notre proposition et des perspectives possibles d'amélioration. Une implémentation du composant de choix est décrite dans le chapitre 6. Nos travaux se concentrant sur les systèmes d'information grand public, nous illustrons nos propos en nous appuyant sur une recherche d'informations dans un annuaire multimodal ou dans un annuaire de restaurants.

5.1 Motivations et existant

Par nécessité de restreindre la problématique étudiée, nous avons choisi de nous concentrer sur une adaptation de la présentation initiée par l'utilisateur. Toutefois, défendant un impact de l'adaptation de la présentation sur la sélection du contenu et la détermination de la réaction du système, cette adaptabilité de la présentation induit une adaptativité du choix du contenu à présenter et de la réaction du système. Il s'agit d'un cas d'adaptativité dans la mesure où elle est entièrement décidée par le système et qu'elle ne résulte pas d'une demande explicite de l'utilisateur. Nous détaillons, dans les paragraphes qui suivent, nos motivations à ce que l'adaptabilité de la présentation déclenche une adaptativité de la réaction.

5.1.1 Tenir compte des contraintes de présentation pour répondre aux situations particulières de communication

Toutes les informations ne se prêtent pas à être présentées par la même modalité sensorielle ou par le même langage d'interaction. Comme le signale déjà Martin [Martin, 1995], certaines informations sont mieux transmises par l'image et d'autres par le langage (en particulier oral). Ce constat s'appuie sur les travaux en psychologie qui mettent en évidence que chaque modalité sensorielle est plus compétente pour une dimension donnée perceptible (chapitre 1). Cette appropriation d'une modalité pour un type d'information donné est généralement prise en compte dans les systèmes multimodaux dont les présentations multimodales sont adaptées à un ou plusieurs types standards d'utilisateurs. Au niveau de la caractérisation des modalités, il s'agit de déterminer celles qui sont assignées à des types d'information spécifiques et/ou celles qui sont meilleures - *i.e.* qui ne sont pas équivalentes entre elles - pour ces types d'information.

Le problème est plus complexe quand, donnant une place centrale à l'utilisateur dans la communication avec le système, celui-ci est autorisé à contraindre la présentation de la réaction du système. Si la crainte de la désorientation de l'utilisateur pousse à refuser une telle adaptabilité, donner la possibilité à l'utilisateur de spécifier cette adaptation de la présentation multimodale va dans le sens d'une exigence grandissante des utilisateurs de plus en plus familiers avec les systèmes d'information. Même si les

concepteurs cherchent à concevoir des systèmes utilisables et d'appropriation simplifiée par les utilisateurs, ces derniers sont les plus à même à déterminer ce qui leur convient le mieux. Ceci se vérifie par les usages que font les utilisateurs des systèmes informatiques, au même titre que de toutes les technologies, parfois en opposition totale avec ce qui avait été envisagé ou prédit. Ainsi le premier mythe d'Oviatt [Oviatt, 1999] évoque-t-il l'utilisation multimodale des systèmes multimodaux. Si les travaux d'Oviatt portent sur l'entrée, ce mythe, et plusieurs autres, peuvent être appliqués à la sortie. Tout comme il ne faut pas contraindre l'utilisateur à interagir avec le système multimodalement, il ne faut pas l'empêcher d'avoir des présentations monomodales s'il le souhaite : il s'agit ni plus ni moins d'une adéquation des formes et styles de réponses à la demande de l'utilisateur. Et si cette demande peut être motivée par des préférences personnelles, elle peut aussi résulter de la situation de communication de façon plus large.

Ce constat nous amène à préciser quelques situations de communication où cette adaptabilité est souhaitable, et à mettre en avant la nécessité que cette adaptabilité engendre une adaptativité de la réaction du système.

5.1.1.1 Pourquoi la présentation devrait influencer la réaction du système ?

Le cas le plus évident où la présentation devrait pouvoir être adaptée à la demande de l'utilisateur est celui où ce dernier présente un handicap. Si la présentation de la réaction du système est conçue pour être multimodale, chaque modalité présentant les informations pour lesquelles elle est la plus appropriée, la présentation est complémentaire. Par conséquent, un utilisateur mal-voyant n'aura pas accès aux informations présentées visuellement et un utilisateur mal-entendant n'aura pas accès aux informations présentées auditivement. De même, dans une situation où l'utilisateur ne peut pas avoir accès à une des modalités de présentation (par exemple, sa vue est occupée parce qu'il est en train de conduire ou ses haut-parleurs sont coupés ou inutilisables parce qu'il est en réunion), il se retrouve privé, partiellement ou complètement des informations présentées sur lesdites modalités : on peut alors parler de situation handicapante. Dans de telles situations ou dans des cas d'handicaps physiques, permettre à l'utilisateur de contraindre la présentation favoriserait son accessibilité sensoriactionnelle. Or, comme nous l'avons expliqué dans le chapitre 3, un comportement adapté à une situation de communication est le résultat d'un équilibre entre l'accessibilité sensoriactionnelle, l'accessibilité cognitive et l'accessibilité rhétorique. Modifier la présentation pour garantir l'accessibilité sensoriactionnelle entraîne un déséquilibre au niveau des deux autres accessibilités.

Typiquement, si l'utilisateur ou la situation de communication impose la nécessité d'une présentation auditive, une liste d'informations présentée de façon auditive doit généralement être plus courte qu'une liste des mêmes informations présentée visuellement pour une même charge mentale de l'utilisateur. Par exemple, si l'utilisateur mal-voyant ou en train de conduire utilise un système d'information dont il peut influencer la présentation en termes de modalités de présentation et qu'il cherche à accéder à des informations répondant à certaines caractéristiques, il risque de ne pas pouvoir appréhender ces informations si elle sont trop nombreuses. Nous ne détaillons pas la

problématique de la détermination du nombre d'informations présentables sur une modalité donnée sans surcharge cognitive car ceci constitue un sujet de recherche à part entière. Ce nombre dépend de nombreux facteurs incluant la granularité de l'information considérée, sa portée sémantique, sa complexité propre et la familiarité de l'utilisateur (avec les systèmes d'information en général ou avec un système étudié en particulier). Nous soulignons seulement que l'adaptabilité de la présentation dans un souci d'accessibilité sensoriactionnelle peut avoir des conséquences, en l'occurrence négatives, sur l'accessibilité cognitive.

Par ailleurs, la redondance peut être utilisée dans une présentation multimodale pour souligner l'importance d'une information donnée en la présentant deux fois, de façon à augmenter les chances d'attirer l'attention de l'utilisateur sur cette information. Or, l'adaptabilité de la présentation ne doit pas engendrer la perte de cette mise en avant de l'information considérée. À un autre niveau, un choix rhétorique peut consister, dans le cas d'une présentation multimodale, à recourir à des informations supplémentaires pour répondre au mieux à l'utilisateur et éventuellement anticiper ses questions suivantes en présentant ces informations sur une modalité différente de celle utilisée pour les informations principales. Ce choix ne s'avérera sans doute pas adapté dans le cas d'une contrainte de présentation auditive, au risque de surcharger l'utilisateur d'informations secondaires par rapport à sa requête stricte. De plus, les études en psychologie et en ergonomie centrées sur l'appréhension des systèmes informatiques par les utilisateurs montrent que le recours à une modalité plutôt qu'à une autre pour présenter la réaction du système peut influencer l'assimilation et l'intégration des informations par l'utilisateur (notamment en termes de mémorisation) et son comportement [Tabbers *et al.*, 2001, Karsenty, 2006, Le Bigot *et al.*, 2006, Fréard *et al.*, 2007]. Les choix de modalités relèvent donc aussi d'une dimension rhétorique dans la mesure où ces choix ont un impact direct sur l'utilisateur, ainsi que sur la poursuite de la communication humain-machine.

La garantie de l'accessibilité sensoriactionnelle est donc susceptible d'entraîner un déséquilibre au niveau des accessibilités cognitive et rhétorique. Pour rétablir ce déséquilibre, le contenu de la réponse du système peut être amené à être modifié, autrement dit sa réaction doit être ré-évaluée en tenant compte des contraintes de présentation exprimées par l'utilisateur - ou à terme identifiées au niveau de la situation de communication. En outre, les choix liés à la présentation relèvent certes des accessibilités sensoriactionnelle et cognitive, mais aussi de l'accessibilité rhétorique. C'est pourquoi il est inévitable de choisir de façon concertée fond et forme, réaction et présentation, dans les systèmes d'information multimodaux offrant une communication naturelle. Notons que ces liens forts entre fond et forme, en particulier la prise en compte nécessaire de la forme pour le choix du fond, a déjà été soulignée dans le chapitre précédent. En particulier, les travaux décrits dans [Roth et Hefley, 1993] (*cf.* la section 4.3.2.1) et [Bordegoni *et al.*, 1997] (*cf.* la section 4.3.2.2) insistent sur la nécessité d'allers-retours entre les différents modules correspondant aux étapes de détermination du comportement du système en termes de réaction et de présentation.

Pour pouvoir prendre en compte des éléments de la présentation dans la détermination de la réaction du système, il est nécessaire de définir ces éléments.

5.1.1.2 Contrainte de présentation

Comme nous l'avons exposé dans le chapitre 3, nous avons restreint le cadre de notre étude et nous ne prenons pas en compte les contraintes de présentation au niveau de la concrétisation (par exemple, la typographie pour le langage naturel écrit, la prosodie pour le langage naturel oral ou encore les palettes de couleurs pour les cartes et les schémas). Nous concentrant sur un plus haut niveau, celui de la modalité, nous identifions trois types de contraintes de présentation.

Le premier type identifié est celui des contraintes explicitement imposées par l'utilisateur : par exemple, dans le cas d'une formulation orale, l'utilisateur peut avoir recours à des verbes d'action renvoyant à des capacités d'action du système, tels que "dis-moi" pour une présentation auditive orale ou "affiche" pour une présentation visuelle.

Le deuxième type correspond aux contraintes de présentation émanant de la situation de communication, en particulier du profil de l'utilisateur ou de l'environnement matériel ou physique de communication. Concernant l'environnement, les terminaux actuels sont en mesure d'identifier les dispositifs qu'ils proposent ainsi que certaines de leurs caractéristiques, comme par exemple l'existence de haut-parleurs ou la taille de l'écran. De plus, des travaux visent à l'intégration de capteurs dans ces terminaux pour caractériser l'environnement physique de communication en termes de bruit, de luminosité, de présence de tierces personnes, etc. S'y ajoutent également les régularités d'usage observées pour un utilisateur donné : par exemple, dans une situation donnée, l'utilisateur demande toujours qu'une information déjà présentée lui soit représentée sur une autre modalité particulière : l'anticipation de cette demande incluant une contrainte de présentation explicite de l'utilisateur est aussi une contrainte de présentation, implicite et émanant de la situation de communication. Par rapport aux contraintes de présentation explicitement exprimées par l'utilisateur, la principale difficulté dans la prise en compte de ces contraintes de présentation implicites est d'identifier correctement ces contraintes dans la situation de communication : par exemple, le système pourrait considérer que l'absence de bruit est une condition nécessaire et suffisante pour avoir recours aux modalités auditives, ce qui serait inopportun si l'absence de bruit est due à la participation de l'utilisateur à une réunion.

Le troisième type de contraintes de présentation est quelque peu différent des deux premiers. Il s'agit des contraintes inhérentes aux modalités, que ce soit du point de vue des sensibilités, des dispositifs physiques, des langages d'interaction ou des modalités en tant que couples <dispositif physique ; langage d'interaction> (*cf.* la terminologie adoptée en fin de chapitre 1). Elles renvoient aux caractéristiques des modalités en termes d'accessibilité sensoriactionnelle, d'accessibilité cognitive et d'accessibilité rhétorique. Elles regroupent les contraintes imposées par le fonctionnement humain et identifiées par des études psychologiques ou ergonomiques (par exemple, ne pas présenter plus de x items sur un écran et pas plus de y de façon orale, ne pas présenter sur un écran plus d'informations que l'écran ne peut en contenir sans recours à un ascenseur, ou encore ne pas présenter deux tâches de présentation auditives parallèlement) ainsi que les contraintes propres à la représentation des informations (par exemple, une photographie est, par définition, visuelle). Les critères de caractérisation des modalités [Bernsen,

1994, Bellik, 1995, Bernsen, 1997, Clémente, 2004, Ratzka, 2006], ainsi que ceux de coopération entre modalités s'il y a lieu (*cf.* la section 2.1.2), peuvent donc tenir lieu de contraintes de présentation.

La prise en compte explicite des contraintes de présentation dans le comportement du système fait déjà l'objet de travaux que nous présentons dans la section suivante.

5.1.2 L'existant

Tout d'abord, l'application MATCH, décrite dans le chapitre 4, permet à l'utilisateur d'indiquer ses préférences de présentation [Johnston *et al.*, 2002] mais ces contraintes de présentation explicites n'influent pas sur la détermination de la réaction du système.

Par ailleurs, les travaux sur la plasticité des interfaces [Thévenin, 2001, Sottet *et al.*, 2007] visent en particulier à adapter la présentation en fonction des caractéristiques des dispositifs physiques utilisés, qui peuvent varier en cours d'interaction. L'adaptation de la présentation influence le choix de la réaction du système : néanmoins, la réaction du système reste la même et c'est la granularité, *i.e.* le détail, des informations présentées, qui est modifiée.

Enfin, les travaux menés à IBM sur l'application RIA (pour *Responsive Information Architect*) [Zhou et Aggarwal, 2004, Zhou *et al.*, 2005] ont des objectifs proches des notres à double titre. D'une part, la sélection du contenu dans RIA s'appuie sur des métriques de pertinence (*relevance*) dont la pertinence des modalités (*media relevance*) et le coût de présentation (*presentation cost*). Plus précisément, la pertinence des modalités tient à la fois compte de l'ordonnancement de l'adéquation des modalités pour une dimension donnée à présenter (par exemple, les informations numériques sont mieux perçues (1) par du texte visuel, (2) par du texte oral et (3) par du graphique) et de la disponibilité des dispositifs physiques associés à cette modalité. Le coût de présentation est calculé en fonction du temps et de l'espace car RIA présente les informations visuellement et auditivement. Le coût de présentation en termes de temps, *i.e.* le temps nécessaire à l'énonciation, est déterminé par rapport au nombre moyen de mots nécessaires pour exprimer une dimension informationnelle donnée (par exemple, trois mots pour décrire le style d'une maison) qui est converti en secondes (considérant que 160 mots par minute sont oralisés). Le coût de présentation en termes d'espace correspond au nombre de pixels nécessaires pour afficher une image ou un texte. La pertinence des modalités dépend donc de contraintes inhérentes aux modalités et de contraintes inhérentes à l'environnement matériel de communication. D'autre part, l'utilisateur peut contrôler la modalité de présentation d'une information particulière en exprimant explicitement qu'il veut que le système *affiche* ou *énonce* l'information en question. Les contraintes de présentation de l'utilisateur sont donc prises en compte mais uniquement au niveau de la concrétisation de la réaction du système. Ces travaux proposent donc une solution basée sur des métriques de pertinence pour adapter la présentation aux contraintes de présentation issues de la situation de communication. Vis à vis de ces travaux, notre objectif est complémentaire en visant non seulement une adaptation de la présentation mais aussi de la réaction du système. Cette double adaptation repose sur deux notions-clefs que sont la "stratégie de dialogue" et la "stratégie de présentation",

notions que nous définissons dans la section suivante.

5.2 Stratégie de dialogue et stratégie de présentation

La détermination du comportement d'un système multimodal implique la détermination de la réaction du système d'une part et la détermination de la présentation de cette réaction d'autre part. Considérant des systèmes qui doivent être en mesure de faire eux-même ces choix, sans se contenter de concrétiser un message pré-conçu, nous considérons respectivement le choix de la "stratégie de dialogue" et le choix de la "stratégie de présentation". Nous définissons ces deux types de stratégies avant de justifier leur choix conjoint dans le processus de génération multimodale.

5.2.1 Stratégie de dialogue

Le choix de la stratégie de dialogue correspond à la détermination de la réaction du système. Elle revient à privilégier une tâche communicative, qui est susceptible d'orienter la suite du dialogue et qui conditionne, voire contraint, le choix du contenu à présenter. Cette définition de la notion de "stratégie de dialogue" est différente de celle adoptée par Caelen [Caelen, 2003]. Pour lui, une stratégie de dialogue est une façon de chercher à atteindre un but par le dialogue. Plus précisément, un dialogue implique deux interlocuteurs et s'intègre dans un cadre défini par :

1. un but conversationnel qui correspond à la finalité du dialogue et qui est partagé par les interlocuteurs. Quatre types de buts conversationnels, *i.e.* de dialogue, sont possibles : les dialogues portant sur l'état des objets du monde (description, information, etc.), les dialogues impliquant un engagement (prise de décision, négociation, etc.), les dialogues avec une double direction d'ajustement (théorisation, séance de travail, etc.) et les dialogues exprimant des attitudes mentales (complainte, prière, etc.). Le but conversationnel est à distinguer du but initial propre à chaque interlocuteur qui pousse chacun d'eux à dialoguer en partageant un même but conversationnel ;
2. un thème ;
3. un arrière-plan (rôles sociaux, contexte environnant, etc.) ;
4. un déroulement.

Pour un but conversationnel, un thème et un arrière-plan donnés, le déroulement du dialogue dépend des stratégies de dialogue utilisées pour atteindre le but conversationnel. Le choix de la stratégie de dialogue par un des interlocuteurs vise à sélectionner la meilleure direction d'ajustement des buts à un moment donné. Cette direction d'ajustement détermine donc la stratégie de dialogue de l'interlocuteur et peut varier au cours du déroulement du dialogue. Caelen identifie cinq directions d'ajustement, et donc cinq stratégies de dialogue, qui sont :

- la stratégie réactive, quand l'interlocuteur considéré abandonne son but au profit de celui de l'autre intervenant ;

- la stratégie directive, quand l'interlocuteur impose son but au détriment de celui de l'autre intervenant ;
- la stratégie de négociation, quand chaque interlocuteur conserve son propre but ;
- la stratégie coopérative, quand chaque interlocuteur tient compte du but de l'autre intervenant ;
- la stratégie constructive, quand les deux interlocuteurs abandonnent chacun leur but au profit d'un troisième but.

Selon l'approche de Caelen, les systèmes que nous considérons ont un but conversationnel portant sur l'état des objets du monde et le système adopte systématiquement une stratégie coopérative. À chaque intervention, le système ne va donc pas décider de la meilleure direction d'ajustement des buts mais de la meilleure façon d'aider l'utilisateur à atteindre son but. Dans ce contexte de stratégie coopérative comme définie par Caelen, nous affinons les différents comportements mis en œuvre que nous nommons dans nos travaux "stratégies de dialogue". Ces stratégies de dialogue sont donc des affinements de la stratégie coopérative définie par Caelen. Pour cela, nous assimilons la stratégie de dialogue adoptée par le système pour une réponse donnée à la tâche communicative qu'il privilégie pour cette réponse.

C'est ce que nous appelons "stratégie de dialogue". La stratégie de dialogue définie par Caelen se rapproche de l'état d'esprit de l'interlocuteur, en l'occurrence du système, alors que ***la stratégie de dialogue telle que nous la définissons correspond au comportement qui met en œuvre cet état d'esprit***. C'est pourquoi nous définissons la stratégie de dialogue adoptée par le système à une intervention donnée comme la tâche communicative qu'il privilégie.

Dans le cas des systèmes d'information coopératifs, et en ne tenant pas compte du méta-dialogue (messages de bienvenue, d'incompréhension, d'aide ...), nous identifions trois stratégies de dialogue de base :

- ***la relaxation*** : le système suggère des solutions alternatives ou, à défaut, des critères de recherche alternatifs, dans le cas où il ne trouve pas de solution correspondant à la requête exacte de l'utilisateur. Les réponses présentées sont des réponses suggestives [Sadek, 1999]. La tâche communicative privilégiée est la transmission de solutions approchées ou de critères de recherche moins restrictifs ;
- ***l'énumération*** : le système présente une liste de 1 à n solutions si le nombre de solutions n'est pas trop important. Des informations supplémentaires (on parle de sur-informations en référence aux réponses sur-informatives [Sadek, 1999]) peuvent être présentées. C'est la stratégie de dialogue la plus récurrente dans les systèmes d'information. La tâche communicative privilégiée est la transmission de l'information recherchée ;
- ***la restriction*** : le système suggère des critères afin de restreindre le nombre de solutions car celui-ci est jugé trop important. Le système peut aussi proposer des réponses conditionnelles [Sadek, 1999]. La tâche communicative privilégiée est la transmission de conditions pour restreindre l'ensemble des solutions.

Prenons l'exemple d'un annuaire d'entreprise. Si l'utilisateur demande le numéro de téléphone de Jack Bauer et qu'il n'y a aucune personne de ce nom travaillant dans l'entreprise, une stratégie de relaxation possible serait d'indiquer à l'utilisateur qu'il n'y

a pas de Jack Bauer mais qu'il y a un John Bauer. Une autre stratégie de relaxation serait de lui indiquer qu'il n'y a pas de Jack Bauer mais qu'il y a trois personnes dont le nom de famille est Bauer et de lui présenter les trois numéros de téléphone correspondants. Si l'utilisateur demande le numéro de téléphone de John Bauer, une stratégie d'énumération possible serait de présenter le numéro de téléphone fixe, mais aussi le numéro de téléphone portable et l'adresse courriel. Si l'utilisateur demande le numéro de téléphone d'une personne nommée Bauer, une stratégie d'énumération serait de présenter la liste des personnes portant ce nom ainsi que leur numéro de téléphone. Enfin, si l'utilisateur demande le numéro de téléphone d'une personne prénommée Jack, une stratégie de restriction serait d'indiquer à l'utilisateur qu'il y a 50 Jack, et que l'utilisateur devrait préciser le nom de la personne ou son équipe.

Prenons un autre exemple, celui d'un annuaire de restaurants à Paris. Si l'utilisateur demande un restaurant brésilien près de la Gare de Lyon et qu'il n'y en a pas, une stratégie de relaxation possible serait d'indiquer à l'utilisateur qu'il n'y a pas de restaurant brésilien près de la Gare de Lyon, mais qu'il y a un restaurant mexicain près de cette gare ou qu'il y a un restaurant brésilien près de la Gare d'Austerlitz. Si l'utilisateur demande un restaurant gastronomique près de la Tour Eiffel pour moins de 15 euro, une stratégie de relaxation serait de lui indiquer qu'il n'y a pas de restaurant correspondant à ses critères, mais qu'il y a des restaurants rapides pour la même gamme de prix et le même secteur ; une autre stratégie de relaxation serait de lui indiquer qu'il y a des restaurants gastronomiques dans ce secteur, mais à plus de 15 euro, et de lui présenter le moins cher. Les stratégies d'énumération seront moins riches en terme de sur-informations dans un annuaire de restaurants que dans un annuaire d'entreprise, car les informations à présenter sont plus limitées (à moins de pouvoir indiquer les plats servis ...). Enfin, une stratégie de restriction serait d'indiquer à un utilisateur qui cherche un restaurant près de la Tour Eiffel de préciser le type ou la gamme de prix.

Ces trois stratégies de dialogue tiennent compte de l'état de la communication à travers la précision et la complétude admises pour la requête de l'utilisateur. Mais le choix d'une stratégie de dialogue plutôt que d'une autre, en particulier le choix de la restriction plutôt que de l'énumération, dépend directement de la situation de communication et des contraintes de présentation prises en compte. Ce sont ces deux éléments qui permettent de faire la distinction entre le fait qu'il y ait *plusieurs* solutions ou le fait qu'il y ait *trop de* solutions. Dans un cas, l'utilisateur a accès à l'information demandée, voire à des informations supplémentaires ; dans l'autre cas, il doit préciser sa requête pour accéder à l'information recherchée.

Le choix de la stratégie de dialogue correspond à la détermination de la réaction du système. Cette réaction précise le contenu informationnel à présenter. ***Elle peut être rapprochée d'un ensemble de buts/tâches communicatifs à réaliser, chacun renvoyant à une unité informationnelle élémentaire ou composée.*** Notons que, pour un système donné, la granularité d'une unité informationnelle élémentaire est du ressort des concepteurs. Par exemple, il est possible de considérer que l'unité informationnelle "une liste de solutions" est une unité informationnelle élémentaire. Mais il est aussi possible de considérer que c'est une unité informationnelle composée de plusieurs unités informationnelles élémentaires "une solution". Ce choix

de granularité va avoir un impact sur le choix d'un autre type de stratégie, la stratégie de présentation.

5.2.2 Stratégie de présentation

La stratégie de présentation correspond à la configuration mono ou multimodale choisie pour présenter la réaction du système, *i.e.* le contenu sélectionné. Par exemple, dans l'expérimentation présentée [Fréard *et al.*, 2007, Horchani *et al.*, 2007b] sur un service de prise de rendez-vous médicaux, la stratégie de présentation est d'allouer une modalité à chaque type de tâche : dans cette expérimentation, une réponse du système est composée d'un feedback (*e.g.* "Vous avez demandé un rendez-vous avec le Docteur Pasteur mercredi après-midi"), d'une réponse à proprement parler ("Il y a trois rendez-vous disponibles : 14h, 15h45 et 18h30") et d'une relance ("Lequel désirez-vous ?") et le feedback et la relance sont considérés comme relevant de la tâche interactive de l'utilisateur avec le système, alors que la réponse sert la tâche applicative qui pousse l'utilisateur à utiliser le système. Par conséquent, la même modalité est allouée aux feedbacks et aux relances. Dans l'architecture Arch (*cf.* la section 4.2.1.2), la stratégie de présentation est du ressort des composants de présentation (abstraite et concrète) : si le composant de dialogue indique qu'il faut présenter qu'il n'y a pas de solutions à la requête sur le numéro de téléphone de Jack Bauer, mais qu'il y a trois personnes dont le nom de famille est Bauer et dont les numéros de téléphones sont "1234", "4567" et "8910", une stratégie de présentation serait de présenter auditivement grâce à la synthèse du message "Il n'y a pas de Jack Bauer mais il y a trois personnes nommées Bauer" et d'afficher les prénom, nom et numéro de téléphone des trois personnes sous forme de liste. Une autre stratégie de présentation serait de tout afficher à l'écran, ou encore de tout présenter sous forme d'un message synthétisé. ***La stratégie de présentation correspond simplement au choix de présentation pour un contenu, une réaction donnée.*** Nous focalisant sur le niveau sémantique et modalitaire de la génération du comportement du système (*cf.* la section 2.3), nous nous limitons au choix de la présentation abstraite, *i.e.* au niveau du composant de présentation abstraite dans Arch.

Si l'on se réfère à l'architecture conceptuelle de Roth et Hefley (*cf.* la section 4.3.2.1), le choix de la stratégie de présentation inclut le choix des modalités et la conception de la présentation. Par rapport au modèle de référence pour les IMMPS (*cf.* la section 4.3.2.2), le choix de la stratégie de présentation inclut non seulement l'allocation des modalités et l'ordonnancement des informations, mais aussi les différentes étapes qui permettent de passer d'unités informationnelles à des objets à concrétiser de façon mono ou multimodale. Étant donné que la stratégie de présentation est déterminée par la machine, les termes "monomodal" et "multimodal" renvoient dans la suite de cette section à la notion de modalité du point de vue système, *i.e.* comme le couple <dispositif physique, langage d'interaction>, que nous considérons comme conditionnant une modalité sensorielle (*cf.* la section 1.4).

De façon à clarifier une stratégie de présentation sans contraindre les paramètres pris en compte pour la choisir, nous la définissons par rapport au résultat de son choix. ***Le choix d'une stratégie de présentation produit une spécification de la pré-***

sentation à réaliser. Une spécification de présentation est constituée d'au moins une tâche de présentation. Une tâche de présentation correspond à une unité informationnelle sémantiquement cohérente et allouée mono ou multimodalement : c'est, en quelque sorte, un but communicatif mono ou multimodalement alloué. Une unité informationnelle pouvant être élémentaire ou composée, il est est de même pour une tâche de présentation. Par exemple, si "une liste de solutions" est une unité informationnelle élémentaire, alors "présenter une liste de solutions multimodalement en combinant langage naturel oral auditif et hypertexte visuel" est une tâche de présentation élémentaire ; par contre, si "une liste de solutions" est une unité informationnelle composée de plusieurs unités informationnelles "une solution", alors, "présenter une liste de solutions multimodalement en combinant langage naturel oral auditif et hypertexte visuel" est une tâche de présentation composée. Nous rappelons que la granularité d'une unité informationnelle, et donc d'une tâche de présentation, dépend des concepteurs.

Le choix d'une stratégie de présentation consiste donc en l'allocation d'une ou plusieurs modalités à chacune des unités informationnelles qui constituent la réaction du système. En résulte une coopération ou des relations non seulement entre modalités (lorsqu'elles sont allouées à une seule tâche ou lorsque plusieurs tâches allouées à différentes modalités sont combinées), mais aussi entre tâches de présentation et plus largement entre spécifications de présentation.

Nous appuyant sur les coopérations entre modalités, sur lesquelles nous nous concentrons, appliquées à un haut niveau d'abstraction (*cf.* la section 2.3), nous définissons trois types de relations au niveau sémantique entre des tâches ou des spécifications de présentation :

- ***deux spécifications de présentation sont équivalentes si elles constituent deux réponses possibles du système et que, en tenant compte des contraintes de présentation, une seule est appliquée.*** Par exemple, la spécification de présentation qui comprend une présentation visuelle du nombre de solutions, une invitation visuelle à préciser la requête et une présentation visuelle de la liste des solutions est considérée équivalente à une spécification de présentation qui comprend une présentation auditive du nombre de solutions, une présentation auditive des critères de restriction et une invitation visuelle à préciser la requête (*cf.* la figure 5.5 page 171). Elles sont distinguées par la stratégie de dialogue choisie (énumération versus restriction) et la contrainte de présentation prise en compte (visuelle versus auditive) ;
- ***une spécification de présentation inclut deux tâches de présentation complémentaires si celles-ci constituent deux éléments de réponse complémentaires et que, en tenant compte des contraintes de présentation, toutes deux sont nécessaires pour présenter la réaction du système***, *i.e.* elles font partie de la concrétisation d'une même stratégie de dialogue. Par exemple, une présentation visuelle du nombre de solutions, une invitation visuelle à préciser la requête et une présentation visuelle de la liste des solutions sont des tâches de présentation complémentaires qui peuvent constituer le comportement du système dans le cas où la stratégie de dialogue adoptée est l'énumération et

qu'il y a une contrainte de présentation visuelle (*cf.* la figure 5.5 page 171) ;
- ***une spécification de présentation, respectivement une tâche de présentation, est assignée si elle est la seule réponse possible du système à une requête de l'utilisateur, respectivement si elle est la seule tâche constitutive de la spécification de présentation considérée.*** Par exemple, la présentation auditive du nombre de solutions, l'invitation visuelle à préciser la requête et la présentation visuelle de la liste des solutions peuvent constituer la spécification de présentation assignée dans le cas où il y a plus d'une solution et où aucune contrainte de présentation n'est prise en compte (*cf.* la figure 5.4 page 170).

Nous insistons sur le fait que notre définition de l'équivalence entre deux spécifications de présentation ne porte pas sur le contenu sémantique de ces spécifications de présentation, *i.e.* sur les unités informationnelles à présenter, mais sur la relation entre ces spécifications du point de vue de la stratégie de présentation. Deux spécifications de présentation équivalentes se distinguent par la stratégie de dialogue qu'elles mettent en œuvre. Cette définition de l'équivalence ne peut donc donner lieu à des spécifications de présentation redondantes d'un point de vue sémantique même si les deux sont appliquées. Soulignons, par ailleurs, que l'équivalence entre deux unités informationnelles sémantiques ne donnerait d'ailleurs pas lieu à deux tâches de présentation équivalentes en ce qui concerne l'impact sur l'utilisateur, à cause de l'effet de modalité (*cf.* chapitre 3).

Chaque tâche de présentation peut être monomodalement ou multimodalement allouée : c'est le niveau modalitaire de coopération entre modalités. Au niveau d'une tâche de présentation donnée, nous considérons que les coopérations entre modalités exploitables en sortie des systèmes sont l'assignation, la complémentarité et la redondance. L'équivalence n'est pas envisagée dans la mesure où nous estimons que le choix de la stratégie de présentation doit aboutir à un choix de coopération et que, en sortie, l'équivalence n'est pas un choix, mais un constat. D'autres travaux abordent l'équivalence différemment, l'admettant en sortie et faisant appel à l'utilisateur pour faire le choix adéquat [Mansoux, 2005]. Dans notre cas, les coopérations possibles entre modalités pour une tâche de présentation sont donc les suivantes (*cf.* la section 2.3) :
- une modalité peut être assignée à une tâche de présentation lorsque cette modalité est utilisée exclusivement pour présenter l'unité informationnelle constitutive de la tâche de présentation ;
- deux modalités sont redondantes pour présenter une tâche de présentation lorsque chacune est utilisée pour présenter l'unité informationnelle constitutive de la tâche de présentation ;
- deux modalités sont complémentaires pour présenter une tâche de présentation lorsqu'elles sont toutes deux utilisées pour présenter chacune une partie de l'unité informationnelle constitutive de la tâche de présentation. Dans ce cas, un choix doit être fait de l'allocation des modalités pour chaque partie de l'unité informationnelle considérée.

Étant donnés ces deux niveaux possibles de relations admises, la multimodalité de la sortie d'un système peut découler de spécifications de présentation combinant des

tâches de présentation monomodales complémentaires ou de spécifications de présentation combinant des tâches de présentation multimodales.

La complexité de la stratégie de présentation, *i.e.* des relations entre tâches de présentation et entre modalités, et donc de l'allocation des modalités correspondantes, dépend du choix des concepteurs. Plus précisément, plus la notion d'unité informationnelle aura une granularité fine, plus les tâches de présentation seront élémentaires et plus les relations entre modalités seront réduites à l'assignation ou à la redondance. Toutefois, moins la granularité sera fine, plus la stratégie de présentation sera répartie entre le choix des tâches de présentation et le choix de la coopération entre modalités pour une même tâche de présentation. C'est aux concepteurs du système de fixer la granularité en fonction de leurs besoins.

Ayant défini précisément les stratégies de dialogue et de présentation, nous les situons dans le processus de génération multimodale.

5.2.3 Stratégies de dialogue et de présentation au sein du processus de génération multimodale

De façon schématique, la détermination du comportement d'un système multimodal se fait en trois étapes : (1) sélection du contenu, (2) sélection des modalités et (3) allocation des modalités aux différents éléments du contenu. Comme le montre la figure 5.1, la stratégie de dialogue est décidée durant l'étape (1) et la stratégie de présentation durant les étapes (2) et (3).

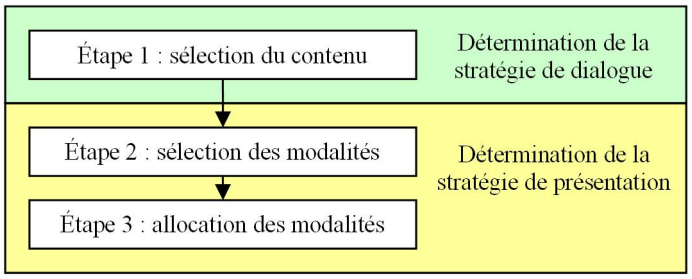

FIG. 5.1 – Stratégies de dialogue et de présentation par rapport aux principales étapes de détermination du comportement d'un système multimodal

Si l'on se réfère à l'architecture Arch, l'étape (1) est du ressort du composant de dialogue et les étapes (2) et (3) sont du ressort du composant de présentation abstraite, voire du composant de présentation concrète si celui-ci fait des choix de concrétisation et ne se contente pas d'identifier les éléments de présentation concrète pré-définis qui correspondent à la présentation abstraite à réaliser. La stratégie de dialogue est donc déterminée par le composant de dialogue et la stratégie de présentation par les composants de présentation abstraite et concrète. Par rapport aux couches du modèle de référence pour les systèmes intelligents de présentation multimédia de la figure 4.13 (page 135), la sélection du contenu est répartie entre la couche de contrôle et les étapes de raffinement du but et de sélection du contenu de la couche de contenu alors que la sélection

et l'allocation des modalités est faite au niveau des étapes d'allocation des modalités d'ordonnancement de la couche de contenu et sont affinées par les couches de conception et de réalisation. La stratégie de dialogue est donc déterminée au niveau des couches de contrôle et de contenu tandis que la stratégie de présentation est choisie au niveau des couches de contenu, de conception et de réalisation. Que ce soit dans ces architectures de référence ou dans les architectures des systèmes présentés dans le chapitre 4, la stratégie de dialogue est choisie avant la stratégie de présentation et la stratégie de présentation doit permettre de présenter le contenu sélectionnée par la stratégie de dialogue.

Il est généralement considéré que la coopérativité ou la convivialité d'un système dépend en grande partie de la stratégie de dialogue adoptée alors que son accessibilité dépend de la stratégie de présentation. Or, comme nous l'avons explicité dans le chapitre 3, la coopérativité d'un système devrait s'étendre à la garantie de son accessibilité, que ce soit d'un point de vue sensoriactionnel, d'un point de vue cognitif ou d'un point de vue rhétorique. De plus, comme nous l'avons souligné en début de ce chapitre, des études sur les stratégies de présentation ont montré que le choix d'une modalité ou d'une combinaison multimodale plutôt qu'une autre a un impact tant sur la perception et sur la charge mentale de l'utilisateur que sur sa réaction. Plus précisément, les modalités imposent ou tolèrent un rythme d'interaction qui n'a pas toujours le même impact sur les utilisateurs, et entraînent un comportement différent. Par conséquent, la stratégie de présentation a un impact sur l'utilisateur, mais aussi sur la suite de la communication. Il est donc délicat de décréter qu'elle est indépendante de la stratégie de dialogue. Nous pensons qu'elle devrait au contraire y être intégrée : stratégie de dialogue et stratégie de présentation sont les deux faces d'un même problème. Stratégie de dialogue et stratégie de présentation devraient donc être choisies de façon conjointe ou concertée, comme le suggèrent déjà [Roth et Hefley, 1993] et [Bordegoni *et al.*, 1997] (*cf.* la section 4.3.3). Bien que ceci aille à l'encontre des architectures classiques qui séparent généralement fond et forme, contenu/réaction et présentation, stratégie de dialogue et stratégie de présentation, nous proposons d'intégrer un composant qui soit dédié au choix des stratégies de dialogue et de présentation dans les architectures des systèmes d'information coopératifs multimodaux en sortie.

5.3 Composant de choix de stratégies de dialogue et de présentation

Nous définissons une architecture de systèmes d'information coopératifs et multimodaux en sortie. Cette architecture consiste en une extension de l'architecture de référence Arch (*cf.* la section 4.2.1.2) par l'ajout d'un composant dédié au choix conjoint de la stratégie de dialogue et de la présentation. Cette architecture a été privilégiée car sa modularité assure la réutilisabilité des composants ainsi que la modifiabilité et l'extensibilité du code (pour l'ajout ou le changement d'une modalité par exemple). De plus, elle peut être appliquée à la plupart des architectures et donc permettre de réutiliser des composants d'architecture existants. Nous justifions d'abord l'ajout de ce composant, puis nous expliquons comment il s'intègre au sein d'une architecture Arch. Nous expli-

quons et illustrons ensuite son mécanisme avant de conclure sur des limites de notre solution et des perspectives d'extension et d'implémentation de ce composant avec des outils de réalisation concrète de l'interface en sortie.

5.3.1 Pourquoi un composant dédié au choix conjoint de stratégies de dialogue et de présentation ?

Comme nous l'avons souligné à plusieurs reprises, stratégies de dialogue et de présentation sont difficilement dissociables par rapport à leur impact sur l'utilisateur et sur la communication à venir. De plus, comme nous l'avons montré, la coopérativité et l'accessibilité - incluant accessibilité sensoriactionnelle, mais aussi accessibilités cognitive et rhétorique - peuvent intervenir aussi bien sur le choix de la stratégie de dialogue que sur celui de la stratégie de présentation. C'est pourquoi nous avons choisi d'aborder la définition du comportement du système comme un tout. Nous nous sommes restreint à travailler sur cette définition du choix des stratégies de dialogue et de présentation à un haut niveau d'abstraction. En particulier, nous ne considérons pas la stratégie de présentation au niveau de la concrétisation d'une tâche de présentation mono ou multimodalement allouée. Nous nous focalisons sur l'allocation des modalités aux tâches de présentation. Le choix des stratégies de dialogue et de présentation, *i.e.* le choix du comportement du système, à haut niveau d'abstraction considéré correspond à la détermination (1) des unités informationnelles à présenter grâce à des tâches de présentation (et les éventuelles relations entre ces tâches de présentation) et (2) des modalités utilisées, ainsi que des coopérations entre ces modalités (complémentarité ou redondance), pour présenter chacune de ces tâches de présentation. Nous fixons la granularité des éléments considérés pour l'allocation des modalités au niveau d'une tâche de présentation. Aussi, même si une tâche de présentation consiste en la présentation d'une unité informationnelle composée en combinant plusieurs modalités, nous ne traitons pas de la répartition des modalités sur les unités informationnelles élémentaires qui est donc à réaliser en aval, au niveau syntaxique des langages d'interaction.

Idéalement, à l'image de ce qui est proposé dans [Roth et Hefley, 1993] (*cf.* la section 4.3.2.1)et dans [Bordegoni *et al.*, 1997] (*cf.* la section 4.3.2.2), le choix conjoint de la stratégie de dialogue et de la stratégie de présentation dans Arch devrait se faire par concertation entre composant de dialogue et composant de présentation abstraite. Pour réaliser ce choix dans le respect du mécanisme Slinky lié à Arch, nous avons deux solutions : le choix est effectué par le composant de dialogue ou par les composants de présentation. La première solution consistant à affecter ce choix des stratégies au sein du composant de dialogue nous semble impossible car l'une des caractéristiques de ce composant est d'être indépendant des modalités. Or, même si nous ne considérons la stratégie de présentation qu'à un haut niveau d'abstraction, elle génère une spécification de présentation qui inclut des tâches de présentation mono ou multimodalement allouées, et qui n'est donc pas indépendante des modalités. De plus, nous souhaitons que les stratégies de dialogue et de présentation soient choisies en tenant compte des contraintes de présentation, de façon à assurer la triple accessibilité de l'utilisateur aux informations présentées et aux capacités d'actions offertes. Par conséquent, le choix des

stratégies de dialogue et de présentation n'est pas indépendant des modalités et tient compte d'informations sur la présentation. Il ne peut donc faire partie du composant de dialogue.

Considérant la deuxième solution qui consiste à étendre le composant de présentation abstraite pour qu'il intègre le choix des stratégies de dialogue et de présentation ne nous semble pas souhaitable non plus. En effet, le composant de présentation n'est classiquement pas en mesure de choisir la réaction du système, *i.e.* le contenu à présenter. Il ne peut pas déterminer la stratégie de dialogue et n'accède pas à certains modèles tels que le modèle de(s) tâche(s) ou l'historique du dialogue. Il nous semble donc important, dans le respect de l'architecture Arch, de ne pas rattacher le choix de stratégies de dialogue et de présentation au composant de présentation abstraite.

Ne souhaitant pas remettre en question l'indépendance du composant de dialogue par rapport aux modalités et aux informations de présentation d'une part, ni l'indépendance du composant de présentation (abstraite) par rapport à la détermination de la réaction du système d'autre part, nous avons opté pour un nouveau composant de choix à part entière, étendant ainsi le modèle Arch. L'architecture résultante, le rôle de chaque composant, les informations échangées entre les composants et le mécanisme de détermination du comportement d'un système d'information coopératif multimodal selon l'architecture proposée sont décrits ci-après.

5.3.2 Architecture globale : extension d'Arch

Au sein d'une architecture Arch, nous définissons un composant dédié au choix de stratégies de dialogue et de présentation, appelé dans la suite du document "composant de choix de stratégies de dialogue et de présentation" ou - plus simplement - "composant de choix". Comme le montre la figure 5.2, le composant de choix simule la concertation et la coordination entre le composant de dialogue et le composant de présentation abstraite (voire composant de présentation concrète si des éléments de présentation à un plus bas niveau d'abstraction sont pris en compte). Le composant de choix fait donc office d'intermédiaire entre le composant de dialogue et le composant de présentation abstraite tout en reprenant une partie des attributions de chacun de ceux-ci. La figure 5.2 détaille les informations échangées et les modèles au sein de l'architecture résultante en ne considérant que les sorties du système et en ignorant le traitement des entrées de l'utilisateur. Les flèches entre les composants renvoient aux sens d'échange d'informations entre ces composants, la double flèche correspondant à une consultation de la base de données spécifiée par le composant concerné. Focalisant sur le choix de stratégies et donc le comportement du système, par mesure de simplicité, nous appelons "composants du domaine" le composant du domaine et le composant d'adaptation au domaine et "composants de présentation" le composant de présentation abstraite et le composant de présentation concrète.

Nous avons choisi d'étendre le modèle de référence Arch pour deux raisons : tout d'abord notre solution permet de coupler notre composant de choix à des composants de présentation existants tels des boîtes à outils ou environnements pour la multimodalité (comme ICARE [Mansoux, 2005, Bouchet, 2006, Mansoux *et al.*, 2006]). Ces outils

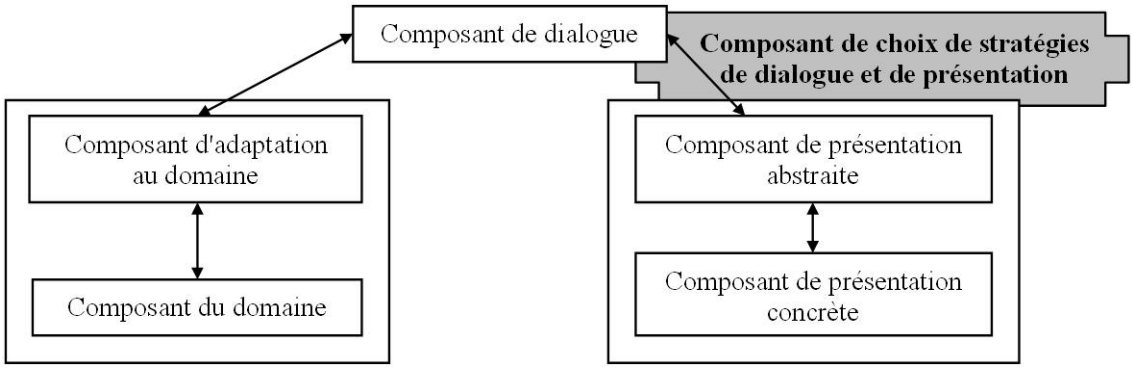

FIG. 5.2 – Le composant de choix de stratégies de dialogue et de présentation au sein d'une architecture Arch

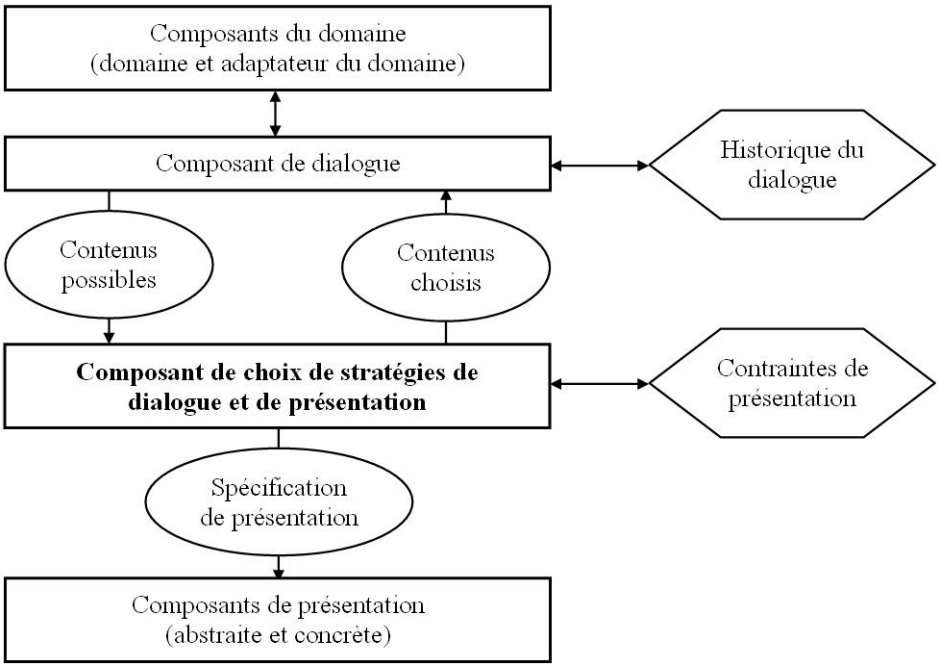

FIG. 5.3 – Multimodalité en sortie : détails des informations échangées

permettent de compléter le composant de choix qui reste au niveau de la présentation abstraite en fournissant une solution pour la concrétisation de la présentation multimodale définie. D'autre part, notre solution permet d'affiner le choix de stratégies de dialogue et de présentation en fonction des contraintes de présentation dans des systèmes déjà développés selon le modèle Arch.

Dans le cadre d'une communication multimodale naturelle, centrée sur la coopérativité de la machine et l'accessibilité des informations, nous avons apporté des modifications aux éléments constitutifs du modèle Arch ainsi étendu. Ces modifications sont

détaillées dans la section suivante.

5.3.2.1 Modifications des composants de dialogue et de présentation

Pour définir le comportement en sortie du système, le composant de dialogue dans Arch calcule la réaction du système et la transmet aux composants de présentation. La conception de la présentation de cette réaction est alors réalisée par le composant de présentation abstraite résultant en une présentation indépendante des dispositifs utilisés. Cette présentation abstraite est ensuite transmise au composant de présentation concrète chargé de la concrétiser en éléments réalisables par les dispositifs physiques, de façon à ce qu'ils soient perceptibles par l'utilisateur.

Dans le cas des systèmes d'information, nous décomposons en deux étapes la détermination de la réaction de la machine à une requête de l'utilisateur : dans une première étape, des solutions à la requête de l'utilisateur et, s'il y a lieu, des informations supplémentaires telles que les solutions approchées, les éventuels critères de restriction de l'ensemble des solutions, etc. sont définies ; lors d'une deuxième étape, la réaction du système à proprement parler est déterminée et conduit à la sélection des informations à présenter. La détermination des solutions, en particulier des diverses informations supplémentaires qui ne correspondent pas à la requête stricte de l'utilisateur, dépendent certes du modèle de dialogue sous-tendu au composant de dialogue, mais aussi, et pour une part importante, des composants du domaine, en particulier du composant d'adaptation au domaine qui s'appuie sur un modèle des tâches. En effet, les composants du domaine peuvent calculer la distance sémantique entre deux données [Zhou *et al.*, 2005] et déterminer ainsi quelles données font partie des solutions, constituent des sur-informations à la solution exacte de la requête stricte de l'utilisateur, représentent des critères de restriction ou de relaxation de l'ensemble des solutions, etc. La deuxième étape correspond au choix de la stratégie de dialogue telle que nous l'avons définie, dans la mesure où elle détermine si la réponse du système consiste en une relaxation, une énumération ou une restriction et spécifie les informations à présenter de façon effective. Dans l'architecture que nous proposons, la première étape - de détermination des solutions possibles à la requête de l'utilisateur - reste sous la responsabilité du composant de dialogue. Par contre, la deuxième étape - de sélection de la stratégie et des informations présentées - revient au composant de choix. De plus, nous inspirant des architectures des systèmes de dialogue, nous intégrons dans l'architecture proposée un historique du dialogue. Le composant de dialogue l'utilise éventuellement pour déterminer des régularités d'usage susceptibles d'enrichir les informations supplémentaires, par exemple quand l'utilisateur sollicite (presque) systématiquement, après une information demandée ou présentée, une même autre information.

Au sein d'une architecture Arch, le composant de présentation abstraite est en charge de la conception de la présentation à un haut niveau d'abstraction. Par conséquent, il sélectionne les modalités, les alloue aux unités informationnelles et spécifie le type de présentation à appliquer. Défendant un choix conjoint de la stratégie de dialogue et de la stratégie de présentation à un haut niveau d'abstraction, le composant de choix se charge de la sélection et de l'allocation globale des modalités aux tâches de présentation, laissant

aux composants de présentation le soin de concevoir de façon précise la présentation abstraite à produire en termes d'allocation des modalités aux unités informationnelles élémentaires s'il y a lieu et en termes d'objets à concrétiser. Par exemple, si une liste de solutions est considérée comme une seule unité informationnelle, le composant de choix indique au composant de présentation abstraite qu'il doit présenter cette liste de façon multimodale audio-visuelle en combinant de façon complémentaire langage naturel oral et hypertexte. La seule contrainte imposée par le composant de choix au composant de présentation abstraite se résume au type de coopération entre les modalités à utiliser. L'exploitation de ces modalités reste alors à la charge du composant de présentation abstraite, comme c'est déjà le cas dans les systèmes multimodaux en sortie basés sur Arch.

Le composant de choix de stratégies de dialogue et de présentation est donc chargé (1) de déterminer la stratégie de dialogue à adopter en définissant les unités informationnelles à présenter, (2) de choisir les modalités à utiliser pour présenter ces unités informationnelles et (3) de déterminer la ou les modalités allouées à chaque unité informationnelle à présenter. La répartition multimodale faite par le composant de choix permet ainsi la prise en compte des contraintes de présentation explicitement exprimées par l'utilisateur à un haut niveau d'abstraction. La section suivante détaille comment le composant de choix remplit ces fonctions ainsi que les informations qu'il échange avec les autres composants.

5.3.2.2 Fonctionnement du composant de choix de stratégies de dialogue et de présentation

Le composant de dialogue envoie au composant de choix l'ensemble des solutions à la requête stricte de l'utilisateur et les autres informations associées. Cet ensemble des solutions est plus ou moins riche en fonction de l'élaboration du composant de dialogue et des composants du domaine. Cet ensemble correspond à l'ensemble des contenus possibles, assimilable à des buts de communication potentiels dans le modèle de référence pour les systèmes intelligents de présentation multimédia [Bordegoni *et al.*, 1997]. Le message envoyé du composant de dialogue au composant de choix ne contient aucune information ayant trait aux modalités ou aux contraintes de présentation.

À partir de ces contenus possibles, le composant de choix sélectionne l'un des comportements possibles du système, qui correspond à un couple <stratégie de dialogue, stratégie de présentation>. Cette sélection se fait en fonction des contraintes de présentation prises en compte. Les contraintes de présentation émanant de la situation de communication sont mémorisées dans une base de données à part. Ainsi ces contraintes et leurs évolutions ne sont-elles pas mêlées aux informations propres à l'état de la communication dans l'historique du dialogue. Celui-ci reste donc indépendant des modalités. La base des contraintes de présentation peut être vue comme un modèle de l'utilisateur étendu à toute la situation de communication. Le comportement choisi par le composant de choix à partir de cette base se traduit par une spécification de présentation telle que définie précédemment. Cette spécification comprend au moins une tâche de présentation mono ou multimodalement allouée et, le cas échéant, ces tâches de présentation

sont ordonnées. Précisons que les tâches correspondent à des buts de communication possibles transmis par le composant de choix, mais aussi à d'autres actes communicatifs qui relèvent généralement du méta-dialogue et contribuent à l'aspect coopératif du système : il peut s'agir, par exemple, de relances invitant l'utilisateur à préciser sa requête ou à en formuler une nouvelle.

La spécification de présentation est transmise au composant de présentation abstraite chargé de concevoir la présentation abstraite correspondante. Les tâches de présentation qui la constituent doivent donc être compréhensibles, connues, de ce composant. De plus, les tâches de présentation étant ordonnées, cet ordre peut être traité différemment au niveau du composant de présentation abstraite mais aussi au niveau du composant de présentation concrète. Par exemple, dans le cas de tâches de présentation auditives, l'ordre correspond à l'ordre séquentiel d'énonciation ; par contre, si les tâches de présentation sont visuelles, l'ordre peut correspondre à un ordre d'affichage horizontal, d'onglets, etc. Cette dimension d'ordre est un des aspects des stratégies de dialogue et de présentation, car elle est susceptible d'influencer l'interprétation des informations présentées par l'utilisateur ainsi que son comportement. Tout comme pour l'allocation des modalités, nous avons fait le choix, en limitant la prise en compte conjointe des stratégies de dialogue et de présentation à un haut niveau d'abstraction, d'en laisser une partie à la charge des composants de présentation.

Les tâches de présentation transmises aux composants de présentation peuvent être vues comme des actes communicatifs mono ou multimodalement alloués. En effet, elles correspondent à une intention de communication d'un contenu informationnel, avec, en plus, la définition des modalités utilisées pour les présenter. Si le composant de dialogue, indépendant des modalités, n'a pas besoin de connaître précisément la spécification de présentation choisie, il doit toutefois être informé du contenu sémantique effectivement présenté. Ceci lui est nécessaire pour maintenir la cohérence de l'historique du dialogue. Le composant de choix lui transmet donc les actes communicatifs correspondant à la spécification de la présentation sélectionnée, sans préciser les modalités allouées (*cf.* "contenus choisis" dans la figure 5.2).

Il convient de noter que les stratégies de dialogue et de présentation ne sont pas explicites dans le composant de choix. Nous considérons que les comportements possibles du système sont définis à la conception par les concepteurs : si les stratégies de dialogue et de présentation sous-tendues ne sont pas explicites dans le système, elles doivent toutefois être explicitées au moment de la conception. Lorsqu'une contrainte de présentation inhérente aux modalités possibles est utilisée pour déterminer le comportement du système, elle n'est pas non plus explicitée, mais elle doit être identifiée en tant que telle par le concepteur. Par exemple, les critères de caractérisation des modalités [Bernsen, 1994, Bellik, 1995, Bernsen, 1997, Clémente, 2004, Ratzka, 2006] (*cf.* chapitre 1) ou de coopération entre ces modalités [Martin, 1995, Clémente, 2004] (*cf.* chapitre 2) peuvent constituer des critères de choix lors de la conception mais ne sont pas mémorisés dans la base des contraintes de présentation. Par contre, elles sont utilisées pour le choix du comportement du système. Par exemple, dans la réalisation logicielle de composant décrite au chapitre 7 et basée sur des règles, les contraintes de présentation, issues de la base ou inhérentes aux modalités, ainsi que certains éléments génériques des

contenus possibles, tiennent lieu de conditions. Les spécifications de présentation constituent les conclusions. Le choix du comportement du système, sous-tendant le choix de la stratégie de dialogue et de la stratégie de présentation, consiste à l'identification de la règle qui s'applique en fonction des conditions vérifiées. La spécification de présentation résultante et les actes communicatifs correspondants sont respectivement envoyés aux composants de présentation et au composant de dialogue.

5.3.2.3 Illustration des modifications des rôles fonctionnels des composants d'Arch

Afin de donner une vision plus claire du rôle fonctionnel attribué au composant de choix de stratégies de dialogue et de présentation au sein d'une architecture Arch étendue, nous comparons la production du comportement d'un système qui respecte l'architecture Arch classique avec la production du comportement du système dont l'architecture intègre un composant de choix tel que nous l'avons défini. Tous les choix du composant de choix ont été spécifiés par les concepteurs. Ce ne sont que des exemples qui n'ont pas valeur de vérité mais qui visent à favoriser une communication naturelle selon les principes que nous avons présentés dans le chapitre 3.

Considérons le cas d'un annuaire multimodal d'entreprise consultable via un téléphone mobile. Cet annuaire est conçu pour adopter un certain comportement dans le cas où il trouve un certain nombre de solutions à une requête de l'utilisateur. Par exemple, si l'utilisateur cherche le numéro de téléphone d'un certain Bauer et que le système trouve 3 personnes portant ce nom, son comportement est conçu pour indiquer par un message en langage naturel oral auditif à l'utilisateur qu'il y a trois personnes nommées Bauer, lui laisser la possibilité de préciser la requête par un champ de saisie présenté à l'écran et lui présenter la liste des prénom, nom et numéro de téléphone des trois personnes trouvées. Dans une architecture basée sur Arch, la génération du comportement du système se déroule comme suit :

1. le composant de dialogue détermine, avec l'aide du composant du domaine et du composant d'adaptation au domaine, qu'il y a trois personnes nommées Bauer. Il détermine que la réponse du système consistera en une information sur le nombre de solutions, une information sur la liste des solutions et une invitation à préciser la requête. Il envoie des tâches communicatives au composant de présentation abstraite ;
2. le composant de présentation abstraite décide de l'allocation des modalités. Il détermine donc que la présentation du nombre de solutions se fera auditivement alors que la présentation de la liste de solutions et l'invitation à préciser la requête se feront visuellement. Il envoie cette liste de tâches de présentation au composant de présentation concrète ;
3. le composant de présentation concrète concrétise, réalise la spécification de présentation qui lui a été envoyée.

Imaginons à présent que l'utilisateur demande explicitement au système de lui répondre visuellement (par exemple, en lui disant "*affiche-moi* le numéro de téléphone de

Bauer") ou, au contraire, auditivement (par exemple, en lui disant "*dis-moi* le numéro de téléphone de Bauer")[1]. Dans le meilleur des cas, seule la présentation est adaptée : la présentation visuelle consiste à présenter les tâches de présentation visuellement et la présentation auditive consiste à présenter les tâches de présentation auditivement. Les tâches communicatives envoyées du composant de dialogue au composant au composant de présentation abstraite sont strictement les mêmes. La contrainte de présentation émise par l'utilisateur n'influence pas la stratégie de dialogue.

L'intégration d'un composant de choix de stratégies de dialogue et de présentation dans l'architecture du système a pour objectif de permettre la prise en compte de la contrainte de présentation émanant de l'utilisateur, mais aussi de contraintes de présentation liée au choix de la présentation, dans le choix de la stratégie de dialogue. Comme nous l'avons souligné dans les sections précédentes, le rôle des composants de dialogue et de présentation sont modifiés. Dans le cas par défaut, *i.e.* où il n'y a aucune contrainte de présentation émanant de l'utilisateur, la production du comportement du système se fait comme suit :

1. le composant de dialogue détermine, avec l'aide du composant du domaine et du composant d'adaptation au domaine, qu'il y a trois personnes nommées Bauer. Son rôle s'arrête là. Il envoie l'ensemble des solutions à l'utilisateur, en l'occurrence la liste des trois personnes avec leur prénom et leur numéro de téléphone, ainsi que la liste des critères de restriction éventuels transmis par le composant d'adaptation au domaine ;
2. le composant de choix choisit à la fois la stratégie de dialogue et la stratégie de présentation à un haut niveau d'abstraction. Ceci signifie qu'il détermine la stratégie de dialogue à appliquer (relaxation, énumération ou restriction) et qu'il détermine les tâches/buts communicatifs à produire en conséquence, mais aussi qu'il alloue, à chaque tâche/but communicatif une ou plusieurs modalités de concrétisation. En l'absence de contrainte de présentation, il peut considérer que l'existence de plus d'une solution entraîne une stratégie de dialogue d'énumération, qui se traduit par une information sur le nombre de solutions, une information sur la liste des solutions et une invitation à préciser la requête. L'absence de prise en compte de contrainte de présentation permet l'application d'une stratégie de présentation multimodale, où la modalité auditive est utilisée pour le nombre de solutions et la modalité visuelle est utilisée pour la liste des solutions (qui serait difficile d'accès cognitivement en raison de la mémoire humaine limitée pour tout ce qui est numérique) et l'invitation à préciser la requête (qui est implicite auditivement car un message oral incite l'utilisateur à répondre oralement). Les tâches de présentation résultantes sont envoyées au composant de présentation abstraite ;
3. le composant de présentation abstraite n'a donc pas à décider de l'allocation des modalités. Il se contente de spécifier le comportement de l'utilisateur s'il y a

[1] Nous rappelons que nous avons précisé, dans la section 3.5, que nous ne nous pré-occupons pas de la façon dont les contraintes de présentation explicites de l'utilisateur sont identifiées. Ces deux phrases sont de simples exemples et l'identification des contraintes de présentation, émanant de l'utilisateur ou de la situation d'utilisation, est un objet de recherche à part entière.

lieu (c'est le cas quand une tâche de présentation composée est multimodale) et de transformer la spécification de présentation en instances interprétables par le composant de présentation concrète ;

4. le composant de présentation concrète concrétise, réalise la spécification de présentation qui lui a été envoyée.

La prise en compte d'une contrainte de présentation émanant de l'utilisateur ne provoque aucune modification dans le message envoyé du composant de dialogue au composant de choix. Par contre, le composant de choix ne va pas se comporter de la même manière (s'il a été défini dans ce sens, bien-sûr). Dans le cas d'une contrainte de présentation visuelle, respectivement auditive, il ne se contente pas d'allouer une modalité visuelle, respectivement auditive, à un ensemble de tâches communicatives prédéfinies. Par exemple, dans le cas d'une présentation visuelle, la même stratégie de dialogue et les mêmes tâches communicatives vont être conservées avec une présentation visuelle : ce choix ne contraint pas les capacités d'action (*i.e.* d'expression) de l'utilisateur qui peut toujours, s'il le souhaite, interagir oralement avec le système. Par contre, dans le cas d'une présentation auditive, une stratégie de dialogue différente est adoptée car la présentation de trois numéros de téléphone de façon auditive est jugée lourde cognitivement. Aussi, le composant de choix décide d'appliquer une stratégie de restriction et, par conséquent de présenter le nombre de solutions, de présenter la liste des critères de restriction et d'inviter l'utilisateur à préciser sa requête. La stratégie de présentation associée consiste à présenter les informations auditivement mais à garantir les capacités d'action de l'utilisateur en affichant l'invitation à préciser la requête : cette dernière n'est pas présentée auditivement car elle est implicite à l'énonciation des critères de restriction.

Ayant présenté notre solution architecturale et détaillé le traitement réalisé par le composant de choix, nous considérons son application dans le cadre d'un exemple, le système @mie, dont la réalisation logicielle est décrite au chapitre 7.

5.3.3 Exemple : @mie

Le système considéré est intitulé @mie (pour Annuaire Multimodal Intelligent d'Entreprise). Ce système permet aux salariés d'une entreprise d'avoir accès à des informations sur leurs collègues (en particulier, leurs prénoms, noms, photos, adresses courriels, numéros de téléphones fixes et portables, numéros de bureau, etc.). Les informations sont présentées en combinant - ou non - la modalité auditive <haut-parleurs, langage naturel oral> et la modalité visuelle <écran, hypertexte (incluant des photos)> d'un téléphone mobile. La figure 1 (page 3) présente des exemples de sorties proposées par le système @mie.

À titre d'exemple, nous avons choisi de considérer l'impact possible de la prise en compte des contraintes de présentation émanant de l'utilisateur sur la distinction entre deux stratégies de dialogue que sont l'énumération et la restriction. Autrement dit, nous considérons qu'à une requête de l'utilisateur à laquelle correspondent plus d'une solution, la distinction entre "*plusieurs* solutions" et "*trop* de solutions" - à l'origine

du choix de la stratégie de dialogue d'énumération ou de la stratégie de dialogue de restriction - dépend des contraintes de présentation. Nous proposons trois niveaux de sophistication du composant de choix, en fonction des contraintes de présentation prises en compte. Dans chacun de ces cas, nous supposons que suite à une requête de l'utilisateur, le composant de dialogue envoie au composant de choix une liste de solutions incluant, pour chaque solution, des informations supplémentaires à la requête stricte de l'utilisateur ainsi que des critères de restriction pour restreindre l'ensemble des solutions. Ces informations correspondent aux buts de communication possibles. Les choix de stratégies de dialogue et de présentation utilisés pour cet exemple n'ont pas été validés expérimentalement : ils servent uniquement à illustrer et à mettre en avant l'intérêt du composant de choix proposé. Précisons toutefois que les choix de stratégie de présentation utilisés s'appuient sur les choix faits pour le prototype originel d'@mie et qu'ils visent une communication naturelle. Dans ce prototype, le comportement multimodal du système s'appuie sur l'hypothèse que l'utilisateur utilise l'annuaire pour accéder aux informations concernant une personne particulière. Aussi, tant que *la* personne-solution n'est pas identifiée, le système ne donne pas l'information oralement. L'information donnée oralement a donc pour objectif de permettre à l'utilisateur de mieux cerner l'ensemble des solutions trouvées par le système : c'est une forme de réponse coopérative. Par ailleurs, autant que faire se peut, les informations sont affichées de façon à mettre à profit la rémanence du visuel et d'assurer la disponibilité des informations présentées, ce qui est particulièrement utile si l'utilisateur souhaite préciser sa requête en ayant des éléments présentés sur les solutions trouvées. Enfin, de façon à ne pas contraindre les capacités d'action de l'utilisateur, nous lui laissons toujours la possibilité d'interagir via le clavier, la molette ou le stylet de son téléphone en présentant un affichage même s'il y a une contrainte auditive. Cette précaution n'est pas nécessaire pour la capacité d'action orale, qui est toujours disponible.

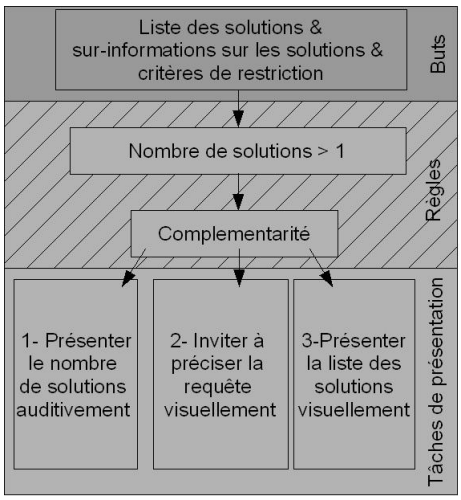

FIG. 5.4 – Exemple de règle par défaut (*i.e.* sans prise en compte de contraintes de présentation) du composant de choix dans le cas où il y a plusieurs solutions

Le premier cas correspond au comportement multimodal par défaut du système dans le cas où il y a plus d'une solution. La sortie du système correspond à l'exemple de gauche de la figure 5.4. Elle inclut un message oral précisant le nombre de solutions et un message visuel listant l'ensemble des solutions. Comme le montre la règle présentée en figure 5.4, la réponse du système à la requête de l'utilisateur se fait grâce à trois tâches de présentation monomodalement allouées. Il y a une tâche de présentation pour le nombre de solutions et une autre pour la liste des solutions, auxquelles s'ajoute une invitation à l'utilisateur pour qu'il précise éventuellement sa requête : cette tâche de présentation permet de laisser à l'utilisateur le choix des moyens d'expression en entrée du système (saisie au clavier ou clic avec un pointeur). Ces trois tâches de présentation sont complémentaires car leurs contenus sémantiques sont différents et qu'elles sont toutes trois nécessaires à la construction du comportement du système en réponse à l'utilisateur. De plus, les tâches de présentation sont numérotées selon l'ordre de présentation à réaliser. Comme nous l'avons déjà souligné, la concrétisation d'une telle organisation spatio-temporaire, même simpliste, dépend des capacités des composants de présentation (abstraite et concrète).

Dans ce premier cas illustratif, le composant de choix est réduit à un rôle minimal : à partir des réponses possibles fournies par le composant de dialogue, le composant de choix ne fait qu'appliquer l'unique règle correspondant au nombre de solutions connues du système et se contente d'envoyer au composant de présentation abstraite la spécification de présentation à générer.

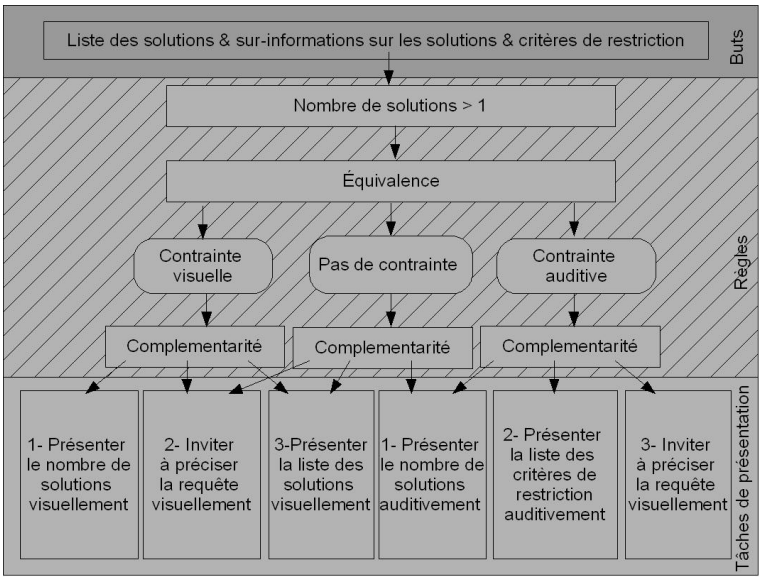

Fig. 5.5 – Exemple de règles du composant de choix dans le cas où il y a plusieurs solutions et une prise en compte de la contrainte de présentation auditive ou visuelle de l'utilisateur

Le deuxième cas illustratif étend le premier en tenant compte d'une contrainte de présentation imposée par l'utilisateur. Nous considérons que celui-ci peut préciser si le

système doit lui répondre oralement via les hauts-parleurs ou visuellement via l'écran. Cette contrainte de présentation est sauvegardée dans la base des contraintes de présentation lors de l'interprétation de la requête de l'utilisateur. Comme le montre la figure 5.5, trois spécifications de présentation sont possibles :

- le comportement multimodal par défaut qui correspond au premier cas illustratif ;
- un comportement monomodal visuel qui consiste à afficher sur l'écran du terminal utilisé une présentation visuelle du nombre de solutions, de la liste de solutions et d'une invitation à préciser éventuellement la requête. Ce comportement est valable quelle que soit la taille de l'écran ;
- un comportement monomodal auditif qui consiste à oraliser via les haut-parleurs du terminal utilisé le nombre de solutions et les différents critères de restriction pertinents, ainsi qu'une invitation visuelle de précision garantissant à l'utilisateur la possibilité de saisir manuellement sa requête. Ce comportement est valable quelque soit le nombre de solutions trouvées.

Dans ce deuxième cas illustratif, le composant de choix est à même de remplir la fonction pour laquelle il a été conçu, à savoir choisir la règle correspondant au choix de stratégies de dialogue et de présentation, en fonction de l'état de la communication (il y a plusieurs solutions) et de la contrainte de présentation explicite de l'utilisateur. Les trois spécifications de présentation correspondant aux trois contraintes de présentation possibles sont considérées comme étant équivalentes dans la mesure où une seule d'entre elles suffit à répondre à la requête de l'utilisateur. C'est l'identification de la contrainte de présentation qui détermine la spécification de présentation. Dans les faits, et bien qu'elles soient équivalentes pour les solutions à présenter, ces spécifications de présentation n'en demeurent pas moins différentes dans la mesure où les stratégies de dialogue et de présentation qu'elles mettent en œuvre sont différentes. Par conséquent, en fonction de la spécification de présentation sélectionnée par le composant de choix, celui-ci ne va pas envoyer en retour au composant de dialogue les mêmes actes communicatifs présentés dans les trois cas. Par exemple, en considérant que l'invitation à la précision n'est pas une unité informationnelle pertinente pour le maintien de l'historique du dialogue, le composant de choix indique au composant de dialogue que le nombre de solutions et la liste des solutions ont été présentés dans le cas du comportement multimodal par défaut ou dans le cas du comportement visuel et que le nombre de solutions et la liste des critères de restriction ont été présentés dans le cas du comportement auditif.

Le troisième cas illustratif prend en compte, en plus de la contrainte de présentation spécifiée par l'utilisateur, une contrainte de présentation inhérente aux modalités utilisées. Afin d'assurer les accessibilités cognitive et rhétorique et tenant compte de la taille limitée d'un écran de téléphone mobile, nous définissons que pour une présentation multimodale ou visuelle, un maximum de X solutions peuvent être affichées sans recourir à un ascenseur. De même, pour une présentation auditive, un maximum de Y solutions, X étant inférieur à Y, est présentable oralement. Par conséquent, comme le montre la figure 5.6, trois nouveaux comportements suivants s'ajoutent aux trois déjà définis dans le cas précédent. Les six règles résultantes sont les suivantes :

- s'il n'y a pas de contrainte de présentation de l'utilisateur, deux cas sont différenciés en fonction du nombre de solutions présentables visuellement :

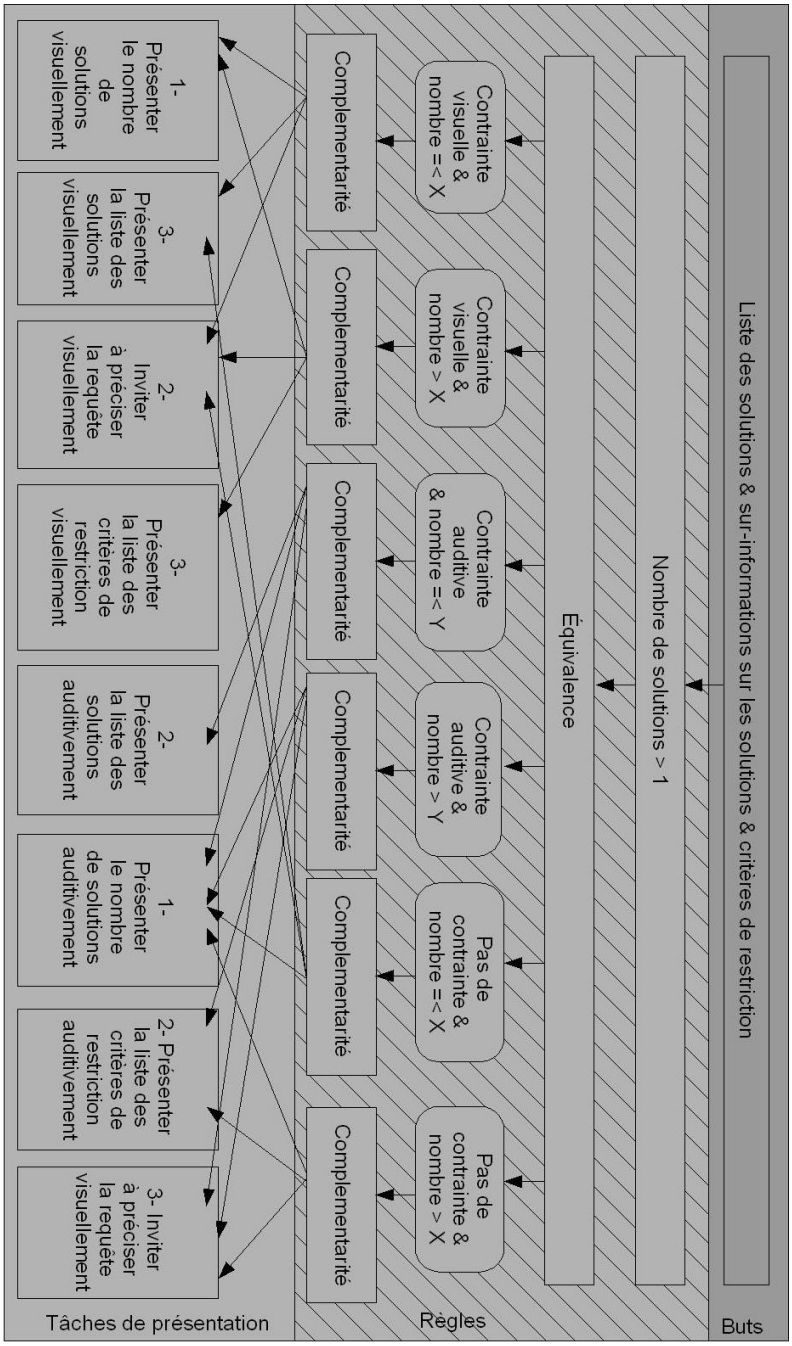

FIG. 5.6 – Exemple de règles du composant de choix dans le cas où il y a plusieurs solutions, une prise en compte de la contrainte de présentation auditive ou visuelle de l'utilisateur et une prise en compte du nombre d'informations visualisables et audibles

  - il y a X solutions ou moins, le comportement multimodal par défaut présenté pour le premier cas illustratif est appliqué ;
  - il y a plus de X solutions, le nombre de solutions et les critères de restriction possibles sont donnés auditivement. S'y ajoute une invitation de précision présentée visuellement ;
- si l'utilisateur signifie une présentation auditive, deux cas sont différenciés en fonction du nombre de solutions présentables auditivement :
  - il y a Y solutions ou moins, le système présente alors le nombre et la liste de solutions auditivement. S'y ajoute une invitation de précision présentée visuellement ;
  - il y a plus de Y solutions : le comportement auditif décrit pour le deuxième cas est appliqué ;
- si l'utilisateur signifie une présentation visuelle, deux cas sont différenciés en fonction du nombre de solutions présentables visuellement :
  - il y a X solutions ou moins, le comportement visuel décrit pour le deuxième cas est appliqué ;
  - il y a plus de X solutions : le système présente visuellement le nombre de solutions et la liste des critères de restriction possibles. S'y ajoute une invitation de précision présentée visuellement.

La figure 5.6 peut sembler compliquée dans la mesure où elle inclut toutes les tâches de présentation pour toutes les contraintes considérées. Toutefois, si on omet les modalités allouées et l'ordre de présentation, il n'y a que quatre actes communicatifs ("présenter le nombre de solutions", "présenter la liste des solutions", "inviter à préciser la requête", et "présenter les critères de restriction possibles"). Chaque acte se décline en tâches de présentation visuelle et/ou auditive. En résulte sept tâches de présentation. Dans la figure 5.6, certaines de ces tâches de présentation sont dupliquées pour spécifier des ordres de présentation distincts.

Ces trois cas montrent la façon dont le composant de choix peut être utilisé en modifiant, en fonction des besoins, les comportements possibles du système. Pour changer le comportement multimodal en sortie de la machine, le concepteur doit déterminer les contraintes prises en compte, identifier les tâches de présentation nécessaires et établir les règles permettant de faire le lien entre les contraintes et les tâches de présentation, autrement dit les stratégies de dialogue et de présentation à adopter par le système. Dans une démarche de conception itérative [Boehm, 1986], le composant de choix d'un système donné est stabilisé au fur et à mesure de l'affinement des comportements du système, par l'identification progressive des stratégies de dialogue et de présentation à appliquer ainsi que des tâches de présentation nécessaires pour que le système communique naturellement avec l'utilisateur, malgré les contraintes de présentation considérées.

5.4 Discussion et perspectives

Pour clore ce chapitre, nous résumons notre contribution à un comportement naturel des systèmes puis nous identifions trois limites faisant l'objet de perspectives d'amélioration.

5.4.1 Synthèse de la contribution

Résumé en une phrase, l'objectif principal à l'origine de cette première contribution est le suivant : donner les moyens à des systèmes d'information d'avoir un comportement coopératif, en tenant compte de contraintes de présentation imposées par l'utilisateur, voire, plus largement, par la situation de communication, et en garantissant l'accès sensoriactionnel, cognitif et rhétorique aux informations et aux capacités d'action proposées. Pour atteindre cet objectif, et partant de la distinction entre réaction/contenu et présentation, ou entre fond et forme, nous avons identifié et défini les notions de "stratégie de dialogue" et de "stratégie de présentation". Pour les stratégies de dialogue, nous avons mis en évidence trois principales stratégies possibles lors de l'étape de réponse à la requête de l'utilisateur dans le cas des systèmes d'information. Pour les stratégies de présentation, nous avons, en nous limitant à un haut niveau d'abstraction, proposé la notion de "spécification de présentation" combinant plusieurs tâches de présentation qui correspondent à des actes communicatifs mono ou multimodalement allouées. En nous appuyant sur les coopérations entre modalités présentées au chapitre 2 [Nigay, 1994, Martin, 1995], nous avons également introduit les relations d'équivalence, respectivement de complémentarité, entre deux tâches ou deux spécifications de présentation qui constituent des réponses possibles, respectivement qui forment des éléments de réponse, du système à la requête de l'utilisateur.

Partant du constat des études en psychologie expérimentale et en ergonomie selon lequel la stratégie de présentation a un impact sur l'utilisateur, sur son appréhension des informations et sur la poursuite de la communication, nous avons choisi de considérer conjointement les stratégies de dialogue et de présentation dans le processus de détermination du comportement du système. Cherchant, par ailleurs, tant à réutiliser les briques logicielles existantes qu'à favoriser la réutilisabilité à venir, nous avons choisi de nous appuyer sur l'architecture de référence Arch [UIMS, 1992] pour y intégrer une prise en compte des contraintes de présentation dans le choix des stratégies de dialogue et de présentation. De façon à garantir l'indépendance du composant de dialogue par rapport aux modalités et aux contraintes de présentation et l'indépendance du composant de présentation par rapport aux choix de réaction du système (*i.e.* par rapport au modèle de dialogue et au modèle des tâches sous-tendues), nous avons décidé d'introduire un composant intermédiaire entre le composant de dialogue et celui de présentation abstraite. Ce nouveau composant est dédié au choix des stratégies de dialogue et de présentation. Le rôle de ce composant est de déterminer la spécification de présentation à concevoir par le composant de présentation abstraite à partir (1) de l'ensemble des réponses possibles du système à la requête de l'utilisateur identifiées par le composant de dialogue et (2) en fonction des contraintes de présentation prises en compte. Sont

donc distingués (1) les réponses possibles du système produites par le composant de dialogue, (2) la réponse choisie, déterminée par le composant de choix et dont est informé le composant de dialogue pour un maintien de l'historique du dialogue et (3) la réponse choisie mono ou multimodalement allouée qui correspond à la spécification de présentation que doivent concevoir et réaliser les composants de présentation.

Le composant de choix comme extension du modèle Arch poursuit la logique de ce modèle, car il permet de réduire la trop grande distance entre le contrôleur de dialogue, qui est indépendant des modalités, et les composants de présentation (abstraite et concrète), qui en sont dépendants. De plus, dans le cas d'un contrôleur de dialogue simpliste qui ne permet pas la production de l'ensemble des réponses possibles (incluant des réponses approchées, des informations supplémentaires, des critères de restriction ou de relaxation de la requête de l'utilisateur), le composant de choix n'a aucune décision à prendre et se contente de transmettre la réponse décidée par le composant de dialogue au composant de présentation abstraite. Dans le cas d'un composant de dialogue plus élaboré, le composant de choix prend tout son sens. Enfin, respectant le principe de modularité du génie logiciel qui contribue à la modifiabilité/extensibilité et à la réutilisabilité des composants, le composant de choix est dédié à la sélection des stratégies de dialogue et de présentation en fonction des réponses possibles et des contraintes de présentation. Ainsi, la prise en compte de nouvelles contraintes de présentation et de nouveaux types d'informations implique-t-elle de modifier le composant de choix uniquement par l'ajout de nouvelles règles.

Si ce composant de choix est un premier pas vers la prise en compte des contraintes de présentation tant au niveau des stratégies de dialogue que des stratégies de présentation dans les systèmes d'information, notre contribution a ses propres limites que nous proposons de dépasser en proposant plusieurs perspectives.

5.4.2 Limites et perspectives

5.4.2.1 Couplage du composant de choix de stratégies de dialogue et de présentation à des composants de présentation

Une première limite de notre proposition est que nous l'avons restreint à la détermination du comportement du système à un haut niveau d'abstraction, n'allant pas jusqu'à la concrétisation de la présentation résultante. Cette limite est double : d'une part, la répercussion des choix de stratégies de dialogue et de présentation sur la sortie du système n'est pas observable sur des systèmes complets, et, surtout, ne peut être testée en conditions réelles de communication ; d'autre part, les choix de stratégies de dialogue et de présentation à un plus bas niveau d'abstraction, *i.e.* sur des aspects plus fins, ne sont pas pris en compte. Par exemple, la stratégie de présentation à un niveau plus fin peut porter sur le choix des couleurs utilisées pour présenter les unités informationnelles, ce choix de couleur ayant un impact sur la perception des informations par l'utilisateur.

Pour répondre à cette double limite, la solution consiste à coupler le composant de choix avec des composants de présentation (abstraite et concrète) à même de traiter les

choix de stratégies de dialogue et de présentation d'une part, et d'affiner ces choix à un bas niveau d'abstraction d'autre part. Nous envisageons donc de combiner le composant de choix avec les composants ICARE [Mansoux, 2005] pour la production effective de la spécification de présentation. L'intégration du composant de choix dans Arch a entre autres été motivée par cette possibilité. Nous avons étudié et défini conceptuellement ce couplage dans [Horchani *et al.*, 2007b], mais sa réalisation logicielle n'a pas été faite et constitue une perspective à court terme.

Il est aussi intéressant de considérer le modèle conceptuel WWHT proposé par Rousseau [Rousseau, 2006] et décrit dans le chapitre 4. Notre composant de choix serait alors utilisé pour l'aspect "What ?" et les processus "Which ? How ? Then ?" permettraient d'affiner le comportement du système et à sa concrétisation. Les stratégies de dialogue et de présentation seraient alors étendues à l'évolution du comportement au cours du temps, dans l'attente d'une nouvelle intervention de l'utilisateur ou dans le cas d'informations susceptibles d'évoluer dans un laps de temps assez court (trafic, cours de bourse, données météorologiques, etc.), accentuant la dimension rhétorique de mise en avant de l'évolution de l'information. La coopération et l'accessibilité du système mettraient alors l'accent non seulement sur les informations à un instant donné mais aussi sur leurs évolutions au cours du temps.

5.4.2.2 Récupération, sauvegarde et utilisation des contraintes de présentation

Une deuxième limite de notre composant tient au fait que nous nous sommes concentrés sur la sortie, proposant une première prise en compte des contraintes de présentation émanant de la situation de communication. Nous avons donc négligés le traitement de l'entrée. En particulier, nous n'avons pas étudié la collecte et la sauvegarde des contraintes de présentation susceptibles d'influencer le comportement du système. Travailler sur l'identification des contraintes de présentation dans la situation de communication permettrait de mieux prendre en compte leur impact sur le comportement du système, renforçant ainsi la boucle de communication entre l'utilisateur et la machine. Travailler à la sauvegarde et à la mémorisation des contraintes de présentation contribuerait à tenir compte des régularités d'usage des modalités mais aussi des capacités cognitives mnésiques humaines. Nous détaillons ces deux perspectives dans les paragraphes suivants.

En ce qui concerne l'identification des contraintes de présentation, une première perspective concerne les contraintes de présentation de l'utilisateur. Dans le cadre de notre étude, nous avons restreint ces contraintes à celles exprimées explicitement avec des syntagmes tels que "dis-moi" pour une présentation auditive ou "affiche(-moi)" pour une présentation visuelle. En laissant de côté les problèmes liés à la reconnaissance vocale, il serait intéressant d'étudier la façon dont l'utilisateur s'exprime pour imposer - ou pas - un format de présentation et la modalité qu'il utilise pour ce faire. Des liens sont sans doute à faire ou à confirmer entre les capacités d'action utilisées par l'utilisateur et les capacités d'action qu'il demande à la machine d'utiliser, autrement dit entre les capacités de perception souhaitées par rapport aux capacités d'actions mobilisées par

l'utilisateur. En particulier, des liens ont été fait entre geste et vision, entre vocal et auditif. Mais ces liens sont peut-être plus dépendants des paradigmes communicationnels utilisés et des informations et tâches mobilisées (qui relèvent plus des contraintes de présentation inhérentes aux modalités) que d'une réelle adéquation à un comportement naturel - *i.e.* spontané - de l'utilisateur. L'identification de couples <capacités d'action, capacités de perception> permettrait une prise en compte des contraintes implicites de l'utilisateur en s'appuyant sur une analyse des capacités d'action qu'il a utilisées.

Outre les contraintes de présentation explicites et implicites de l'utilisateur, une autre perspective possible est de tenir compte de la multiplicité des terminaux disponibles à un instant donné, souvent traitée en plasticité des interfaces [Thévenin, 2001, Sottet *et al.*, 2007]. Les stratégies de dialogue et de présentation pourraient alors être affinées par rapport à la répartition des informations à présenter sur les différents terminaux, et donc les différents dispositifs physiques disponibles. Seraient alors privilégiées les contraintes de présentation issues de la situation de communication, plus précisément des évolutions des terminaux disponibles, ainsi que de potentielles contraintes de présentation inhérentes à la multiplicité de ces terminaux.

Plus généralement, une dernière perspective concernant l'extension des contraintes de présentation prises en compte serait de combiner les informations se rapportant à différentes caractéristiques de la situation de communication pour identifier au mieux d'éventuelles contraintes de présentation sous-tendues. Cette perspective se rapporte à la capture de contexte, vaste axe de recherche particulièrement actif [Rey, 2005]. Par exemple, on pourrait considérer qu'un niveau de bruit ambiant bas autorise, voire encourage, une présentation auditive. Or, ce niveau de bruit ambiant peut être dû au fait que l'utilisateur est en réunion, en conférence ou dans le wagon "silence" d'un train à grande vitesse : le choix d'une présentation auditive serait alors inopportun. Ce type de situation pourrait peut-être être identifié en tenant compte d'autres informations sur l'environnement de communication : par exemple, la présence de nombreux terminaux téléphoniques dans l'environnement physique proche de l'utilisateur pourraient correspondre à une situation où l'utilisateur est en public, et ce même si le niveau de bruit ambiant est bas. Des travaux sur l'identification de données de la situation de communication susceptibles d'être recoupées pour en déduire des contraintes de présentation pourraient donc être menés.

Ces trois perspectives qui consistent à considérer des contraintes de présentation non explicitées par l'utiliateur soulèvent néanmoins le problème de la prévisibilité du comportement multimodal du système, en particulier de sa présentation : il convient de garantir un comportement cohérent et surtout de maintenir l'utilisateur dans la "boucle" en lui donnant la possibilité d'observer et de contrôler le mécanisme de choix.

Par rapport à la sauvegarde des contraintes de présentation émanant de la situation de communication, nous avons considéré que les contraintes de l'utilisateur n'étaient valables que pour une intervention du système. Elles ne sont pas mémorisées au-delà. Dans le cas où l'utilisateur ne souhaite pas obtenir une information en particulier sur une modalité donnée, mais que sa contrainte de présentation est motivée par la situation de communication dans laquelle il se trouve, par une préférence globale ou par un handicap, la mémorisation de ses contraintes de présentation est souhaitable. Les régularités

d'usage, qui sous-tendent une préférence de l'utilisateur ou une accessibilité à une information donnée jugée meilleure sur une certaine modalité, pourraient aussi être prises en compte. Elles impliquent une mémorisation et une analyse des informations demandées sur une modalité spécifique, ainsi que des informations présentées et redemandées immédiatement par l'utilisateur. Parallèlement à l'historique du dialogue qui maintient les informations présentées à l'utilisateur, la base de contraintes de présentation que nous avons proposée pourrait être étendue à un historique des modalités utilisées. De plus, un tel historique des modalités couplé à l'historique du dialogue permettrait, par observation et analyse des régularités d'usage, d'anticiper les contraintes de présentation de l'utilisateur ainsi que les capacités d'action qu'il a tendance à utiliser. Par exemple, nous avons fait le choix, dans le système @mie, de toujours laisser la possibilité à l'utilisateur d'interagir en manipulation directe même dans le cas d'une contrainte de présentation auditive. Il serait sans doute plus judicieux de ne pas systématiser ces capacités d'action, mais de pouvoir les proposer en fonction des préférences de l'utilisateur qui transparaissent à travers l'usage fait des modalités lorsqu'il a posé certaines contraintes de présentation. Nous rejoignons donc ici l'étude des liens entre capacités d'action et capacités de perception précédemment évoquée. Toutes les prises en compte des régularités d'usage impliquent non seulement un recours à une mémorisation des modalités mais aussi à la mise en place de mécanismes d'apprentissage automatique réajustables en fonction des usages effectifs des utilisateurs. Cette perspective est exposée dans la section suivante.

5.4.2.3 Scalabilité, dynamicité et complétude du composant de choix de stratégies de dialogue et de présentation

Dans la première version du composant de choix présentée dans ce mémoire, nous avons opté pour un formalisme de mécanisme de choix à base de règles. Si ce formalisme a le mérite d'être facile à utiliser pour faire une preuve de concept il a aussi des inconvénients. L'un d'eux est la rigidité du mécanisme mis en œuvre. Tout d'abord, les comportements souhaités doivent tous être spécifiés : ils ne peuvent évoluer en fonction des régularités d'usage. De plus, la complétude des situations de communication prises en compte n'est pas assurée. Enfin, le passage à l'échelle de l'approche par règles pose des problèmes pour des systèmes d'information complets et réels car le nombre de règles risquent de devenir rapidement trop important pour devenir gérable : dans nos exemples, nous n'avons pas tenu compte du type de l'information sur lequel porte la requête de l'utilisateur (numéro de téléphone, nom, adresse, personne, etc.) mais il est aisé de voir que sa prise en compte dans les stratégies de dialogue et de présentation engendrera un nombre de règles d'autant plus important qu'il y a de types de données dans le système d'information.

Pour pallier cette rigidité, une première possibilité réside dans des règles de logique floue. Ceci permettrait d'assouplir les conditions de choix et de ne pas avoir à s'assurer de la complétude stricte des règles proposées. Une autre possibilité serait d'attribuer aux règles des poids qui seraient réajustés en fonction de la validation ou pas des comportements adoptés : ceci nécessiterait d'identifier une validation implicite du comportement

du système par l'utilisateur, par exemple quand il ne redemande pas une information présentée. Une dernière possibilité envisagée serait de permettre l'intégration de nouvelles règles ou de nouvelles sous-règles basées sur les régularités d'usage : si l'analyse de la communication de l'utilisateur avec la machine met en évidence que, après une information i, une information i' liée est demandé sur une modalité m, cette information i' pourrait être intégrée dans la réponse du système sur l'information i, avec l'allocation de la modalité m.

Outre ces perspectives qui visent à enrichir le composant de choix de stratégies de dialogue et de présentation, un point qui nous semble crucial est la définition de stratégies appropriées à la situation de communication. Ceci fait l'objet de notre deuxième contribution : un éditeur pour les concepteurs non-informaticiens destiné à paramétrer le composant de choix de stratégies de dialogue et de présentation.

Chapitre 6

Spécifier les choix de stratégies de dialogue et de présentation

Dans le cadre de la conception de systèmes d'information multimodaux capables d'avoir un comportement naturel, nous avons proposé l'intégration d'un composant dédié au choix conjoint de stratégies de dialogue et de présentation dans l'architecture Arch. Cherchant à valider les choix réalisés par ce composant, nous avons voulu nous appuyer sur les résultats d'études en psychologie expérimentale ou en ergonomie pour proposer des choix de stratégies de dialogue et de présentation en adéquation avec le fonctionnement humain dans une situation de communication donnée. Or, la validité de telles stratégies est mal connue et les paramètres de choix de ces stratégies, en particulier les contraintes de présentation inhérentes aux modalités utilisées, sont loin d'être maîtrisés. Pourtant, les travaux allant dans ce sens se multiplient [Le Bigot *et al.*, 2006, Fréard *et al.*, 2007] mais des outils pratiques manquent aux psychologues et aux ergonomes pour permettre une étude systématique de choix relevant des stratégies de dialogue et de présentation. Ces constats sont à l'origine de notre deuxième contribution. Celle-ci consiste en la proposition d'une démarche de conception intégrant concepteurs informaticiens et non-informaticiens qui s'appuie sur un éditeur graphique permettant de spécifier le composant de choix de stratégies de dialogue et de présentation par les concepteurs non-informaticiens. Ainsi, les psychologues et les ergonomes disposent d'un outil qui leur permet d'étudier les paramètres de choix des stratégies de dialogue et de présentation à des niveaux de composition sémantique et modalitaire en facilitant la conception de différentes versions du composant de choix.

Le chapitre est organisé comme suit. Nous commençons par exposer nos motivations et l'existant. Ensuite, nous décrivons une expérimentation qui est à l'origine de notre proposition d'un éditeur de spécification du choix de stratégies de dialogue et de présentation et qui nous a permis d'extraire les premiers éléments nécessaires aux ergonomes/psychologues dans ledit éditeur. Puis, nous détaillons les concepts manipulés par l'éditeur et le processus de conception proposé. Nous illustrons l'éditeur et son utilisation dans le cadre d'une collaboration entre le concepteur non-informaticien et un informaticien grâce à l'exemple de l'annuaire d'entreprise @mie. Nous concluons le

chapitre par une synthèse de notre contribution et par des perspectives d'amélioration de l'éditeur de spécification.

6.1 Motivations et existant

6.1.1 Expertises en sciences humaines pour la conception

Dans une approche naturelle de la communication humain-machine, les choix de stratégies de dialogue et de présentation, s'ils dépendent des contraintes de présentation, doivent être centrés sur l'utilisateur de façon à garantir l'accessibilité sensoriactionnelle, l'accessibilité cognitive et l'accessibilité rhétorique de l'utilisateur aux informations et aux capacités d'action disponibles. Cette garantie implique une meilleure connaissance de l'impact des choix de stratégies de dialogue et de présentation sur l'utilisateur. Or, si la perception humaine à un bas niveau est aujourd'hui assez bien connue, l'intégration des informations perçues est moins maîtrisée. De plus, comme souligné dans le chapitre 1, le chapitre 2 et le chapitre 4, l'étude des entrées des systèmes informatiques a longtemps été privilégiée au détriment de l'étude de leurs sorties. Cette remarque est valable y compris en sciences humaines : un nombre plus important de travaux a porté sur les capacités d'action de l'utilisateur et sur leur utilisation dans la communication que sur les capacités de perception de l'utilisateur et sur leur mise à profit pour assurer l'aboutissement de la communication. Certes, comme cela est souligné dans [Horchani *et al.*, 2007a], certains travaux se sont penchés sur l'étude de certaines modalités considérées de façon isolée [Le Bigot *et al.*, 2006], mais les travaux sont quasiment inexistants pour l'étude de l'impact des combinaisons des modalités. S'y ajoute le fait que les études en ergonomie ou en psychologie expérimentale (1) ne peuvent, pour pouvoir comparer ce qui est comparable, faire varier trop de paramètres à la fois et (2) ont des résultats applicatifs car elles portent sur l'évaluation de systèmes bien spécifiques. De plus, comme cela a été mis en avant dans la section 1.3.5, les paramètres participant à l'attention, à l'intégration sensorielle et, par extension, à la saillance des informations, ne sont pas encore maîtrisés. Aussi, il n'existe pas d'études absolues ou de "tutoriaux" de choix d'une stratégie de dialogue ou de présentation plutôt qu'une autre.

Par conséquent, si les ergonomes/psychologues sont en mesure d'interdire certains choix de stratégies de dialogue et de présentation en l'absence de contraintes de présentation émanant de la situation de communication ou de recommander vivement certaines tâches de présentation garantissant la transparence du système [Karsenty, 2006, Fréard *et al.*, 2007], ils ne sont pas à même d'indiquer les choix de stratégies de dialogue et de présentation les plus adéquats dans une situation de communication contraignante donnée. Notamment, sans étude préalable du cadre applicatif considéré, ils sont démunis pour indiquer les critères de relaxation ou les caractéristiques à prendre en compte pour sélectionner les solutions approchées à une requête donnée dans le cas d'une stratégie de dialogue de relaxation. Ils sont également démunis pour déterminer les sur-informations susceptibles d'intéresser l'utilisateur dans le cas d'une stratégie de dialogue d'énumération. Et ils sont encore plus démunis pour suggérer, lorsque le système connaît plus d'une solution à une certaine requête de l'utilisateur portant sur un type d'informa-

tion donné, le nombre de solutions qui conditionne le choix d'une stratégie de dialogue d'énumération ou de restriction.

Face à l'absence de résultats issus d'expérimentation, l'implication de psychologues et d'ergonomes pour la conception de systèmes d'information multimodaux capables de communiquer naturellement avec les utilisateurs est essentielle. Ils doivent intervenir dans le processus de conception le plus tôt possible, et ce pour deux principales raisons. D'une part, pour mener à bien l'étude des paramètres de choix de stratégies de dialogue et de présentation pour un système d'information donné, les sujets doivent être placés dans des situations aussi proches que possible des situations de communication réelles. Ceci implique non seulement de définir les contraintes de présentation issues des situations de communication, mais aussi de recourir à un système de test - maquette-prototype ou interface pilotée par un magicien d'Oz - qui soit le plus proche possible du système final, que ce soit par rapport à ses caractéristiques physiques ou à son comportement. D'autre part, pour un système en cours de conception, les ergonomes/psychologues doivent identifier les caractéristiques utilisées pour poser les hypothèses servant à évaluer l'adéquation des stratégies de dialogue et de présentation aux situations de communication envisagées. Si ces hypothèses sont validées, les caractéristiques peuvent tenir lieu de contraintes de présentation inhérentes aux modalités dans le système d'information final. Il est donc préférable que ces contraintes soient identifiées le plus tôt possible, de façon à ce qu'une place correcte leur soit accordée dans le comportement du système final. Pour cela, nous adhérons à une approche de conception où ergonomes/psychologues sont impliqués dès le début du processus et travaillent de pair avec les informaticiens. Nous détaillons la façon dont nous concevons cette collaboration dans la section suivante.

6.1.2 Processus incrémental : affinement du choix des stratégies de dialogue et de présentation

Étant donné le manque de connaissances sur l'impact des stratégies de dialogue et de présentation et donc la nécessité d'intégrer les ergonomes/psychologues tout au long du processus de conception, nous adoptons une démarche de conception incrémentale.

En effet, si les éléments de la situation de communication susceptibles de limiter l'accessibilité sensoriactionnelle de l'utilisateur sont faciles à identifier et sont relativement génériques à tous les cadres applicatifs, ceux limitant les accessibilités cognitive et rhétorique et impliquant la prise en compte de contraintes de présentation inhérentes aux modalités utilisées semblent être plus applicatives et dépendent du type d'information présentées. Par exemple, le nombre d'informations présentables en maximisant l'appréhension de ces informations, classiquement fixé entre cinq et sept, est nettement inférieur pour des numéros de téléphones. Il est donc difficile, voire impossible, d'anticiper toutes les contraintes de présentation qu'il est nécessaire de prendre en compte. C'est pourquoi, pour un cadre applicatif donné, la prise en compte des contraintes de présentation doit se faire progressivement, en commençant par les contraintes de présentation génériques puis en sélectionnant les contraintes de présentation applicatives en fonction des résultats des étapes de test. De plus, étant donné que la conception

de systèmes d'information multimodaux capables de communiquer naturellement est encore très exploratoire, la prise en compte des contraintes de présentation par le choix d'un jeu donné de stratégies de dialogue et de présentation n'est pas forcément valide. Avant de proposer un système final, les concepteurs doivent donc envisager différents choix de stratégies de dialogue et de présentation, ainsi que différentes conditions de choix. L'adéquation entre les conditions et les choix devra être testée, de façon à améliorer progressivement le comportement du système. L'affinement tant des contraintes de présentation que des choix de stratégies de dialogue et de présentation en fonction de ces contraintes par succession de phases de conception, d'implémentation et de test d'adéquation constitue notre démarche de conception incrémentale.

Dans cette démarche incrémentale, l'intervention de l'informaticien est nécessaire tout au long du processus. Néanmoins, donner aux ergonomes/psychologues des outils pour élaborer les premiers pré-tests à la conception d'un système, pour tester différents comportements dans le cas de systèmes exploratoires (*i.e.* en l'occurrence, permettant d'étudier le paradigme de communication naturelle) et finalement pour pouvoir efficacement apporter des modifications aux comportements du système final, faciliterait la conception et ce processus incrémental. Les premières esquisses de système conçues pour les pré-tests constitueraient une première version du système final qui serait modifiée et complétée tout au long du processus de conception incrémental.

Notre objectif est donc de fournir un outil aux ergonomes/psychologues pour étudier les choix de stratégies de dialogue et de présentation en adoptant un processus de conception de systèmes multimodaux en sortie incrémental. Avant de présenter notre solution, nous décrivons les outils existants dédiés à la conception de tels systèmes.

6.1.3 Outils existants

Les outils dédiés à la conception des systèmes multimodaux sont rares. Nous présentons ceux qui sont les plus proches de notre objectif.

[Sinha et Landay, 2003] propose un outil de prototypage rapide d'interfaces multimodales en entrée et en sortie noté Crossweaver. Cet outil permet aux concepteurs de décrire le fonctionnement de l'interface sous forme de storyboard. Plus précisément, les concepteurs peuvent décrire les scènes ou états possibles de l'interface - combinant éventuellement présentations visuelle et auditive - et les événements déclenchant des transitions entre états de l'interface. Cet outil de prototypage est complété d'un outil de simulation pour tester l'interface. Il peut être utilisé par des non-informaticiens pour des maquettes de systèmes multimodaux. Néanmoins, Crossweaver est dédié exclusivement à des systèmes du paradigme actionnel.

La plate-forme ICARE (pour Interaction CARE) permet le développement rapide d'interfaces multimodales par réutilisation et assemblage de composants. [Bouchet, 2006] définit des composants de type "dispositif physique", "langage d'interaction" et "combinaison des modalités" qui peuvent être combinés pour spécifier le traitement des entrées d'une interface multimodale. Bouchet propose également un outil permettant d'assembler graphiquement ces composants. L'outil est étendu aux cas des sorties [Mansoux, 2005, Mansoux *et al.*, 2006] et permet alors de prendre en compte une contrainte de

présentation spécifiée par l'utilisateur dans le cas de présentations jugées équivalentes. Toutefois, cette contrainte de présentation n'influence pas le choix du composant de dialogue, *i.e.* de la réaction du système, en amont. ICARE en sortie permet donc le choix de la stratégie de présentation et l'outil est plus dédié aux systèmes relevant du paradigme actionnel puisqu'appliqué au cas de la chirurgie augmentée. De plus, l'outil ICARE, en entrée comme en sortie, n'a jamais fait l'objet de validation avec des utilisateurs ergonomes/psychologues.

Enfin, cherchant à faciliter le processus de conception d'interfaces multimodales, Rousseau [Rousseau, 2006] propose un cycle de conception et y associe deux outils. Ce cycle, adapté du modèle en spirale, est itératif et comprend trois étapes : l'analyse, la spécification et la simulation. L'analyse permet l'identification des éléments impliqués dans la sortie multimodale selon le modèle WWHT. La spécification permet la formalisation de ces éléments grâce à un premier outil, MOSTe (pour Multimodal Output Specification Tools editors). MOSTe comprend cinq éditeurs : un pour les composants d'interaction, un pour le contexte d'interaction, un pour les unités d'information, un pour le modèle comportemental et un pour le modèle d'instanciation. La simulation permet le jeu et la validation - ou non - par le concepteur du résultat de la spécification, grâce au deuxième outil, MOSTs (pour Multimodal Output Specification Tool simulator). Des deux outils proposés, MOSTe, en tant qu'outil de spécification, est le plus proche de nos objectifs. Même s'il n'est pas dédié à des non-informaticiens, l'éditeur du modèle comportemental est assez facile à appréhender. Comme ICARE, l'outil MOSTe se concentre sur la stratégie de présentation et les contraintes de présentation identifiées n'influencent pas la stratégie de dialogue.

Outre que leur nombre est limité, les outils existants se concentrent rarement sur la multimodalité en sortie et sont quasi-inexistants pour un public de concepteurs non-informaticiens. L'outil dont nous décrivons les concepts constitutifs dans la section suivante est dédié à des non-informaticiens et se concentre sur les stratégies de dialogue et de présentation. En cela, l'outil est complémentaire à d'autres outils existants comme ICARE ou MOSTe.

6.2 Expérimentation de référence sur le service Santiago

Notre proposition d'un éditeur pour des non-informaticiens destiné à la spécification de choix de stratégies de dialogue et de présentation s'appuie sur une expérimentation réalisée à France Télécom R&D. Cette expérimentation s'inscrit dans le cadre de travaux portant sur les systèmes de recherche d'information grand public qui allient présentations visuelles (essentiellement langage naturel écrit et hypertexte) et présentations auditives (langage naturel oral). Si l'impact des modalités considérées isolément a déjà été étudié [Hone et Baber, 2001, Le Bigot *et al.*, 2006], les travaux sur l'impact de la combinaison des modalités pour un même message semblent inexistants. L'étude résumée ici part du principe que les modalités imposent ou tolèrent un rythme d'interaction qui n'a pas toujours le même impact sur les utilisateurs. Il s'agit donc de pouvoir induire chez eux un comportement qui améliore la qualité globale de leurs interactions avec le

système. Aussi, nous avons cherché à identifier des configurations (i.e. des stratégies de présentation) pertinentes de façon à les comparer du point de vue de la charge cognitive ressentie par les utilisateurs, du comportement qu'elles induisent chez eux, ainsi que de la performance qu'elles permettent d'atteindre (rapidité, atteinte des buts, mémorisation des informations et apprentissage d'utilisation du système). Nous nous focalisons ici (1) sur l'analyse préalable qui nous a permis d'identifier des stratégies de présentation pertinentes et (2) sur les résultats de l'expérimentation qui confirment l'existence d'un impact de la stratégie de présentation sur l'utilisateur et sur le dialogue sur les expérimentations envisageables à la suite de ces résultats et sur les besoins identifiés des ergonomes en ce qui concerne un outil de spécification des stratégies de dialogue et de présentation. Une présentation plus détaillée des résultats peut être trouvée dans [Fréard *et al.*, 2007].

6.2.1 Caractérisation des informations présentées

Afin de déterminer des stratégies de présentation cohérentes et comparables, les expérimentateurs ont cherché à caractériser les informations à présenter. C'est sur cette caractérisation que nous nous appuyons pour affiner certains concepts intervenant dans le composant de choix de stratégies de dialogue et de présentation. L'étude des dialogues humain-machine a permis d'identifier deux axes de caractérisation des informations présentées : la distinction entre les types de tâches favorisées d'une part, et entre les types d'informations présentées d'autre part.

Concernant les types de tâches, il est admis [Sweller, 1999, Oviatt et Lunsford, 2004] qu'un utilisateur en interaction avec un système réalise deux tâches parallèlement : la tâche d'utilisation du système - dite tâche interactive - et la tâche applicative à proprement parler. La tâche interactive se rapporte à l'utilisation des fonctionnalités et des procédures du système. Elle permet à l'utilisateur de se repérer, de naviguer et d'interagir avec le système. La tâche applicative est à l'origine de l'utilisation du système. Elle permet à l'utilisateur d'atteindre son ou ses buts, le minimum étant l'accès aux informations applicatives. Cette distinction permet la catégorisation des actions du système en fonction de la tâche qu'elles favorisent : les informations présentées peuvent donc être interactives ou applicatives. La répartition (multi)modale doit faciliter ces deux tâches.

Concernant les types d'informations, les expérimentateurs s'appuient sur une typologie des informations fournies à l'utilisateur qui distingue principalement (1) les informations se rapportant aux actions passées (*trails*), (2) les informations cibles (*sites*) et (3) les informations renvoyant aux actions à venir (*modes*) [Nievergelt et Weydert, 1980]. Dans le cas des systèmes de dialogue pour la recherche d'informations sur lesquels se concentrent l'expérimentation, le premier type est appelé "feedback", le deuxième "réponse" et le troisième "relance". Les feedbacks permettent au système d'informer l'utilisateur de ce qu'il a compris et de l'inviter implicitement à corriger toute incompréhension ou erreur. Les réponses correspondent aux informations qui répondent à la requête de l'utilisateur. Les relances l'informent des choix et actions disponibles et l'invitent à poursuivre le dialogue. Comme le montre la figure 6.1, les expérimentateurs ont

choisi que le système réponde systématiquement à une demande de l'utilisateur par un feedback, une réponse et une relance.

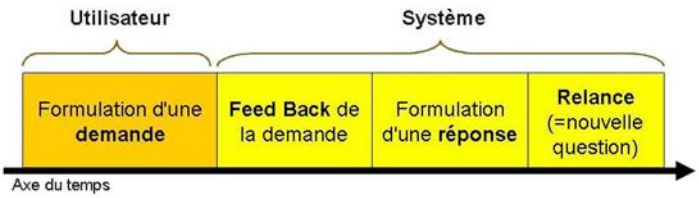

FIG. 6.1 – Structure d'une interaction (extrait de [Horchani *et al.*, 2007a])

Combinant ces deux niveaux de caractérisation, une information à présenter est donc caractérisée par la tâche qu'elle sert et par son type. Pour l'expérience, nous avons fait l'hypothèse que les feedbacks et les relances relèvent de la tâche interactive et les réponses de la tâche applicative. Par ailleurs, les expérimentateurs ont alloué une modalité (visuelle, *i.e.* langue naturelle écrite et hypertexte, ou auditive, *i.e.* langue naturelle synthétisée) à chaque tâche (interactive ou applicative). Ainsi les informations relatives à une tâche (*i.e.* feedbacks et relances pour la tâche interactive et réponse pour la tâche applicative) sont-elles toujours présentées sur la même modalité pendant tout le dialogue. Comme le synthétise la figure 6.2, les expérimentateurs obtiennent quatre stratégies de présentation (*i.e.* configurations), deux monomodales (AAA et VVV) et deux bimodales (AVA et VAV).

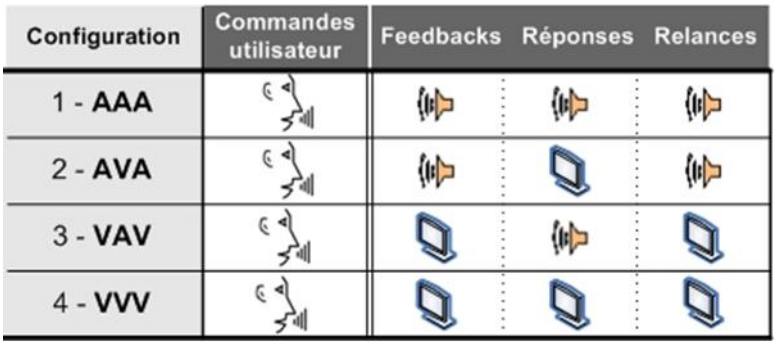

FIG. 6.2 – Configurations testées (A = auditif ; V = visuel) (extrait de [Horchani *et al.*, 2007a])

Afin d'étudier l'impact de ces stratégies de présentation sur l'utilisateur et sur le dialogue, une expérimentation en Magicien d'Oz a été réalisée sur Santiago, service en développement dédié à la prise de rendez-vous médicaux.

6.2.2 Expérience sur le service Santiago

Aucune plateforme n'étant disponible au moment de l'expérience, un matériel expérimental *ad hoc* a été développé. Ce matériel permettait de reproduire le comportement

du service sous le contrôle du magicien d'Oz. L'interface utilisée par les sujets était composée d'un microphone en entrée et d'un écran et de haut-parleurs en sortie. Chacun des 80 sujets disposait d'une consigne incluant deux scénarios, ainsi qu'un planning récapitulant les contraintes horaires et les noms des médecins. Chaque scénario correspondait à une prise de rendez-vous. Les sujets devaient réaliser les scénarios en "parlant" au "système". Les messages vocaux des sujets étaient directement entendus par le magicien d'Oz qui sélectionnait une touche de son clavier en fonction de l'état du dialogue et du scénario. Cette sélection de touche entraînait l'application d'un script et la diffusion de messages auditifs et visuels préexistants. Lorsque la requête du sujet était complète (*i.e.* que le nom du médecin et l'horaire étaient indiqués), le "système" présentait la liste des solutions en trois temps : le feedback reprenait la demande du sujet ; la réponse à proprement parler comprenait une liste de cinq propositions répondant à la requête (une proposition correspondant à un jour et un horaire) ; la relance incitait le sujet à faire un choix (*e.g.* "Laquelle de ces propositions vous convient ? Je vous écoute.").

Durant l'expérience, deux variables étaient manipulées : la configuration (chaque sujet testait une seule configuration sélectionnée par le magicien d'Oz en début de test) et une erreur de reconnaissance vocale (le magicien d'Oz simulait une erreur pour l'un des deux scénarios). Les variables observées étaient la performance (nombre de tours de dialogue, durée du dialogue), l'apprentissage (mémorisation des informations applicatives et interactives), le comportement verbal des utilisateurs (nombre de mots, hésitations, précision du vocabulaire, expressions d'humeur) et l'évaluation subjective de la charge cognitive liée à la tâche (questionnaire à remplir en fin de test).

6.2.3 Résultats, discussion et besoins

En ce qui concerne la charge cognitive, les configurations bimodales (AVA et VAV) sont équivalentes, la configuration VVV est significativement plus coûteuse et la configuration AAA est intermédiaire. Par contre, les configurations bimodales sont opposées en termes de performance : VAV augmente la verbosité des sujets et le taux d'hésitation tout en diminuant la mémorisation des informations applicatives (dates et heures des rendez-vous) ; AVA, au contraire, favorise la performance verbale et la mémorisation et réduit la durée du dialogue.

Ces résultats mettent en évidence que la réduction de la charge cognitive et la performance ne sont pas toujours corrélées. Dans l'expérience, la présentation visuelle des informations invite à la lecture. Elle permet la révision et la comparaison des informations et favorise ainsi la mémorisation. La présentation auditive incite, au contraire, à réagir à chaque proposition pour réajuster la demande. L'auditif conduit donc à un comportement réactif, alors que le visuel permet un plus grand recul. Ceci montre l'importance du choix de la répartition (multi)modale et confirme l'intuition selon laquelle la stratégie de présentation adoptée influence le déroulement de l'interaction.

Mais cette confirmation entraîne d'autres questionnements tant par rapport aux stratégies de présentation que par rapport au lien entre stratégies de présentation et de dialogue : les résultats sur les configurations étudiées sont-ils valables quel que soit le nombre de solutions trouvées ? Comment les aides - qui peuvent être interactives ou ap-

plicatives - doivent-elles être considérées ? Dans quelle mesure est-il souhaitable d'avoir systématiquement un feedback, une réponse et une relance ? La relance n'est-elle pas implicite pour certaines modalités (typiquement, pour la langue naturelle synthétisée) ? Le contenu de la réponse (liste des solutions exactes, liste des solutions approchées, nombre de solutions ...) ne devrait-il pas modifier le choix de la stratégie de présentation ? Des expériences supplémentaires sont nécessaires pour clarifier l'impact des choix de stratégies de présentation et de dialogue. Elles permettraient d'affiner les critères de choix en tenant notamment compte des modalités disponibles et/ou préférées.

Un autre problème se pose, celui de la gestion de la surcharge cognitive du magicien d'Oz due au temps de réaction qu'il doit garantir pour simuler au mieux le service [Fraser et Gilbert, 1990]. Cela conduit généralement à réduire la richesse des services testés : pour l'expérience décrite, le nombre de messages possibles en entrée du "système" a été réduit et le nombre de solutions présentées a été fixé à cinq. Or, si l'on veut vraiment tester l'impact des choix de stratégies de présentation et de dialogue sur l'utilisateur et sur l'interaction, il faut pouvoir tester un nombre plus important de scénarios. Et le recours à un outil est alors souhaitable pour ne pas surcharger le magicien d'Oz.

L'analyse de cette expérimentation, que ce soit de sa réalisation et ou de ses résultats, nous conduit à identifier plusieurs besoins. Tout d'abord, comme nous l'avons déjà signalé, les connaissances manquent quant à l'impact exact des stratégies de présentation, *a fortiori* d'un couple <stratégie de dialogue, stratégie de présentation>. Des études supplémentaires doivent être menées. Mais nous constatons que, au moins dans le cas de la création d'un service, les ergonomes/psychologues partent de rien et élaborent tant bien que mal un matériel "logiciel" nécessaire à la réalisation des expérimentations. D'une part, leur travail serait facilité par un outil qui leur évite de devoir s'improviser développeurs tout en leur garantissant un minimum d'autonomie dans la conception des tests et, d'autre part, le matériel expérimental utilisé pourrait servir de base pour le développement ultérieur du prototype. Un outil de spécifications de choix de stratégies de dialogue et de présentation trouverait donc parfaitement sa place dans une équipe de conception ergonomique de systèmes d'information ou focalisée sur l'étude de l'intégration des informations par les utilisateurs.

Par ailleurs, certains éléments identifiés dans le chapitre précédent semblent pouvoir être affinés par l'expérimentation présentée et par des expérimentations futures. En particulier, l'expérimentation met en avant plusieurs types d'informations présentées, *i.e.* de tâches de présentation présentées qui devraient intervenir dans un éditeur de spécification du composant de choix de stratégies de dialogue et de présentation. Les concepts impliqués dans cet éditeur, ainsi que le processus de conception que nous préconisons pour utiliser un tel éditeur, font l'objet de la partie suivante.

6.3 Éditeur de spécification du composant de choix

La nécessité d'étudier l'impact des stratégies de dialogue et de présentation et sur la communication humain-machine dans une situation de communication donnée et la volonté que ces études soient parties intégrantes du processus de conception de sys-

tèmes d'information multimodaux nous ont poussés à proposer un outil permettant à des non-informaticiens de spécifier le composant de choix de stratégies de dialogue et de présentation. L'outil proposé a donc une portée limitée à ce composant au sein de l'architecture Arch étendue, composant présenté dans le chapitre précédent. En particulier, il se concentre sur la génération multimodale en sortie à des niveaux modalitaire et sémantique (*cf.* la section 2.3), laissant de côté les aspects liés au méta-dialogue (interruptions, reprises, changement de terminal, etc.) et au style (présentation pressante ou apaisante, par exemple). Notons que le traitement de la problématique du style dans l'approche de communication naturelle qui est la nôtre nécessiterait de prendre en compte d'autres niveaux de génération (en particulier les niveaux syntaxique et articulatoire/lexical) ainsi qu'une prise en compte plus large de la situation de communication, que ce soit du contexte ambiant [Thévenin, 2001, Rey, 2005, Rousseau, 2006, Sottet *et al.*, 2007] ou de l'état mental de l'utilisateur [Ochs *et al.*, 2007], et sous-entend de mettre l'accent sur l'accessibilité rhétorique. Or, les travaux présentés dans ce mémoire se concentrent plus sur l'accessibilité sensoriactionnelle (quitte à ce que la présentation résultante soit monomodale et non multimodale) et sur l'accessibilité cognitive (en permettant la prise en compte de contraintes de présentation propres aux modalités utilisées et un impact des contraintes de présentation sur le choix de la stratégie de présentation mais aussi de la stratégie de dialogue). Nous définissons donc une interface graphique capable d'éditer et de générer un composant de choix tel que défini dans le chapitre précédent. Cela nous conduit à affiner certaines notions impliquées dans ce composant. Nous explicitons les motivations d'opter pour un éditeur graphique, avant de détailler les notions qu'il manipule. Une implémentation de l'éditeur est décrite dans le chapitre 7.

6.3.1 Choix d'un éditeur graphique

L'outil étant destiné à des utilisateurs, nous avons choisi de concevoir un éditeur graphique : suivant les principes de la manipulation directe, le recours à des objets manipulables permet une représentation directe des notions impliquées dans la définition du comportement du composant de choix de stratégies de dialogue et de présentation. De plus ces notions sont connues des utilisateurs ergonomes/psychologues comme le souligne l'expérimentation décrite dans [Horchani *et al.*, 2007a] et menée par un ergonome dans le cadre d'un projet pluridisciplinaire. Dans ce chapitre, nous décrivons les concepts manipulés par l'éditeur graphique, tandis que le chapitre suivant en propose une réalisation logicielle en présentant l'éditeur développé.

Par ailleurs, notre objectif est que l'outil puisse être utilisé tout au long du cycle de conception incrémentale, de façon à ce que les premières ébauches du comportement du système d'information multimodal ne soient pas perdues et puissent, au contraire, être enrichies tout au long du cycle de conception. Aussi le code du composant de choix de stratégies de dialogue et de présentation est généré à partir de sa spécification au sein de l'éditeur graphique. L'intervention de l'informaticien est limitée à la configuration de la génération, en particulier pour faire le lien entre les unités informationnelles manipulées dans l'éditeur graphique et les données du programme du système.

Comme l'éditeur graphique permet la spécification du composant de choix de straté-

gies de dialogue et de présentation, plusieurs notions présentées dans le chapitre 5 font partie intégrante de l'éditeur proposé. Nous détaillons dans le paragraphe suivant les concepts manipulés par l'éditeur graphique.

6.3.2 Concepts manipulés par l'éditeur

Tel que nous l'avons proposé dans le chapitre 5, le composant de choix de stratégies de dialogue et de présentation est mis en œuvre grâce à des règles. Les unités informationnelles applicatives correspondant à la requête de l'utilisateur et les contraintes de présentation tiennent lieu de conditions. Les tâches de présentation constituent les conclusions. Ces trois notions - règles, unités informationnelles applicatives et contraintes de présentation, tâches de présentation - sont explicites dans l'éditeur graphique.

6.3.2.1 Notion d'"unité informationnelle"

La notion d'"unité informationnelle" regroupe à la fois les unités informationnelles applicatives envoyées par le composant de dialogue pour répondre à la requête de l'utilisateur et les contraintes de présentation issues de la situation de communication. Nous identifions trois sources possibles pour une unité informationnelle :

- le système : c'est le cas des unités informationnelles applicatives fournies par le composant de dialogue, mais aussi de l'analyse de ces unités informationnelles, comme par exemple le nombre de solutions trouvées ;
- l'utilisateur : il s'agit des contraintes de présentation qu'il a explicitement indiquées, mais aussi, s'il y a lieu, des préférences de l'utilisateur ou des données le concernant sauvegardées dans un modèle de l'utilisateur ;
- l'environnement : les unités informationnelles issues des dispositifs physiques et des capteurs du terminal utilisé sont considérées comme faisant partie de l'environnement.

Cette liste n'est pas exhaustive. Par exemple, si nous considérions les régularités d'usage ou la caractérisation du contexte de communication par d'autres moyens que par le terminal, d'autres types d'unités informationnelles seraient à considérer dans l'éditeur.

Chaque unité informationnelle a un nom choisi par les utilisateurs-concepteurs. Les informaticiens doivent donc relier chacune d'entre elles avec la donnée correspondante connue du système. Pour que ceci soit possible, il est nécessaire de formater les données auxquelles le composant de choix peut avoir accès, en particulier les messages qu'il peut recevoir du composant de dialogue. Celles que nous avons définies sont les suivantes :

- la description qui définit la requête de l'utilisateur : elle correspond à une caractérisation des unités informationnelles recherchées par l'utilisateur qui est définie à partir de l'interprétation de sa requête. Par exemple, si l'utilisateur demande le numéro de téléphone de Carole, la description correspondante se résume à un prénom dont la valeur est "Carole" ; s'il précise qu'il cherche le numéro de téléphone d'une Carole à Grenoble, la description porte sur le prénom (Carole) et sur la localisation géographique (Grenoble). Le système est la source de cette unité

informationnelle car elle dépend de son interprétation de la requête de l'utilisateur ;
- le centre d'attention de l'utilisateur : il renvoie à l'unité ou aux unités informationnelles principales recherchées par l'utilisateur. Dans le cas d'une requête portant sur le numéro de téléphone de Carole, le centre d'attention est le numéro de téléphone. Le système est la source de cette unité informationnelle car elle dépend de son interprétation de la requête de l'utilisateur ;
- la modalité de la requête : elle précise la ou les modalités utilisées par l'utilisateur pour exprimer sa requête. Le système étant à l'origine de l'identification des modalités utilisées en entrée, c'est lui la source de cette unité informationnelle ;
- la modalité de la réponse : elle indique l'existence d'une éventuelle contrainte de présentation explicitement exprimée par l'utilisateur. La source de cette unité informationnelle est considérée comme étant l'utilisateur ;
- le nombre de solutions : elle correspond au nombre de solutions exactes trouvées à la requête stricte de l'utilisateur. La source de cette unité informationnelle est le système ;
- les caractéristiques des solutions : les caractéristiques décrivant un élément pouvant faire l'objet d'une requête de l'utilisateur doivent être définies comme unités informationnelles au niveau de l'éditeur. Par exemple, dans le cas d'un annuaire d'entreprise, l'entité de base est une personne : elle est caractérisée par un prénom, un nom, une photo, un numéro de téléphone, une équipe, etc. Ces caractéristiques sont purement applicatives. Elles nécessitent toutefois d'être reliées aux données connues du système - en termes de nom attribué, en particulier - de façon à être correctement prises en compte et traitées dans le composant de choix généré. La source de ces unités informationnelles est le système ;

Parmi ces unités informationnelles, les cinq premières sont génériques dans le sens où elles sont valables pour tous les systèmes, même si les valeurs qu'elles peuvent prendre sont applicatives (notamment pour les deux premières dont les valeurs sont issues des caractéristiques applicatives des solutions). D'autres unités informationnelles génériques sont possibles, en fonction des contraintes de présentation prises en compte. Si les unités informationnelles génériques servent surtout pour établir les règles de comportement du système définissant le composant de choix, les unités informationnelles applicatives sont utilisées pour définir les tâches de présentation. Nous détaillons cette deuxième notion de base de l'éditeur dans la section suivante.

6.3.2.2 Notion de "tâche de présentation"

Dans le chapitre 5, nous avons défini la notion de "tâche de présentation" comme une unité informationnelle sémantiquement cohérente mono ou multimodalement allouée. L'unité informationnelle impliquée doit préalablement être définie en tant qu'unité informationnelle par les utilisateurs-concepteurs via l'éditeur. Si les tâches de présentation concerne essentiellement les unités informationnelles applicatives, elles peuvent aussi concerner des unités informationnelles génériques : par exemple, un rappel de la requête de l'utilisateur est une tâche de présentation sur la description qui définit cette

requête. C'est aussi le cas quand le nombre de solutions est présentée.

Telle que définie dans le chapitre 5, une tâche de présentation est forcément associée à une modalité. Par conséquent, une même unité informationnelle présentée de trois façons nécessite la déclaration de trois tâches de présentation. Par exemple, "présenter le nombre de solutions auditivement", "présenter le nombre de solutions visuellement" ou "présenter le nombre de solutions multimodalement de façon complémentaire" sont trois tâches de présentation différentes. La conception de l'éditeur nous a fait prendre conscience que cette définition de la notion de tâche de présentation s'avère rapidement fastidieuse. Ce constat nous a amenés à distinguer deux types de tâches de présentation, les tâches de présentation abstraites et les tâches de présentation concrètes. Les tâches de présentation abstraites correspondent aux actes communicatifs envoyés par le composant de choix au composant de dialogue pour l'informer des unités informationnelles effectivement présentées. Elles ne sont pas mono ou multimodalement allouées. Une tâche de présentation concrète est une tâche de présentation abstraite dont la modalité ou la combinaison allouée est précisée. Elle correspond exactement à la notion de "tâche de présentation" telle que définie dans le chapitre 5. Selon ces définitions, le composant de présentation abstraite reçoit du composant de choix une spécification de présentation consistant en plusieurs tâches de présentation concrètes synchronisées.

De plus, la collaboration avec l'ergonome [Horchani *et al.*, 2007a], suite à l'expérimentation présentée dans la section 6.2, nous a amené à distinguer trois types de tâches de présentation. Une tâche de présentation peut être :

- un feedback : il permet l'expression de l'interprétation faite par le système de la requête de l'utilisateur, lui laissant ainsi la possibilité de corriger une mauvaise interprétation ou de rectifier sa requête ;
- une réponse : elle correspond à la réponse à proprement parler apportée par le système à la requête de l'utilisateur ;
- une relance : elle invite explicitement l'utilisateur à poursuivre la communication.

Nous appuyant sur les trois stratégies de dialogue identifiées dans les systèmes d'information pour répondre à une requête de l'utilisateur (*i.e.* la restriction, l'énumération et la relaxation, *cf.* section 5.2.1), nous proposons trois sous-types aux tâches de présentation de type "réponse" :

- réponse-relaxation : il s'agit de présenter une liste de solutions approchées à la requête exacte de l'utilisateur ;
- réponse-énumération : il s'agit de présenter une liste de solutions exactes à la requête de l'utilisateur ;
- réponse-restriction : il s'agit de présenter une liste de critères de restriction pour restreindre l'ensemble des solutions trouvées à la requête de l'utilisateur.

De plus, nous distinguons deux types de relances, celle invitant l'utilisateur à une nouvelle requête et celle invitant l'utilisateur à préciser sa requête. Une fois concrétisées par les composants de présentation (abstraite et concrète), ces deux tâches de présentation pourront être identiques (par exemple, dans le cas d'une présentation auditive, un simple "c'est à vous") ou clairement différentes (par exemple, respectivement "souhaitez-vous formuler une nouvelle requête ?" et "vous devez préciser votre requête ?"

Un dernier type de tâche de présentation proposé est le type "aide". Il permet de

présenter à l'utilisateur une aide sur le fonctionnement du système. Nous avons choisi de ne pas considérer ce type comme un sous-type "réponse", de façon à distinguer clairement les tâches de présentation applicatives qui permettent de répondre à une requête de l'utilisateur d'une part, des tâches de présentation interactives liées à l'utilisation du système et à la communication de façon plus large d'autre part. Les tâches interactives regroupent les tâches de présentation de type "aide", mais aussi de type "feedback" et de type "relance". Les tâches de présentation interactives contribuent, tout comme les tâches de présentation applicatives, au bon déroulement de la communication humain-machine mais en mettant l'accent sur le pilotage et le contrôle de la communication (*dialogue control*) [Bunt, 1994].

En conclusion, une tâche de présentation peut être de type "feedback", de type "réponse-relaxation, de type "réponse-énumération", de type "réponse-restriction", de type "relance-nouvelle requête", de type "relance-précision" ou de type "aide".

Des modalités ou des combinaisons de modalité applicables doivent être définies pour chaque tâche de présentation abstraite. Les modalités applicables considérées dans le cadre de ce mémoire sont les suivantes : la modalité visuelle, la modalité auditive, une complémentarité des modalités auditive et visuelle, une redondance - sous-entendue totale - des modalités auditive et visuelle et une combinaison des modalités auditive et visuelle (sans coopération imposée qui peut donc être redondante ou complémentaire et qui doit être déterminée par les composants de présentation en aval du composant de choix). Nous nous sommes restreints à ces modalités pour l'instant car ce sont celles que les ergonomes/psychologues avec lesquels nous travaillons cherchent à étudier. Leur choix est motivé par le fait que ce sont les deux modalités sensorielles mobilisées à l'heure actuelle dans le cadre des systèmes d'information grand public utilisées pour des messages sémantiquement riches. De plus, nous rappelons que, parmi les coopérations identifiées entre modalités, nous avons retenus la complémentarité et la redondance, auxquelles s'ajoute l'assignation qui peut être à l'origine d'une complémentarité, car nous les considérons suffisantes pour les niveaux d'abstraction que nous considérons, un affinement étant possible lors de la production du comportement du système en aval (*cf.* la section 2.3). En particulier, dans le cas où la multimodalité appliquée à une tâche de présentation est complémentaire, la façon dont la complémentarité est appliquée aux unités informationnelles élémentaires constituant l'unité informationnelle considérée par la tâche de présentation n'est pas du ressort du composant de choix généré mais du ressort des composants de présentation (abstraite et concrète) qui lui sont associés.

En résumé, pour chaque tâche de présentation concrète qui est utilisée pour déterminer le comportement du système, il convient de définir la tâche de présentation abstraite correspondante en spécifiant son type, l'unité informationnelle impliquée et les modalités ou combinaison de modalités applicables. Les tâches de présentation concrètes sont définies au sein des règles.

6.3.2.3 Notion de "règle"

Comme nous l'avons expliqué dans le chapitre 5, nous avons choisi de déterminer le comportement du système dans le composant de choix grâce à des règles. Malgré les

inconvénients évoqués dans la fin du chapitre 5, les règles ont au moins un avantage, celui d'être facile à manipuler par des non-informaticiens. Dans l'éditeur, l'utilisateur-concepteur peut créer ces règles de comportement à partir des unités informationnelles et des tâches de présentation qu'il a créées.

Plus précisément, les unités informationnelles permettent de composer les conditions des règles. Pour cela, l'unité informationnelle considérée est comparée à une valeur dite de comparaison. Par exemple, le nombre de solutions peut être comparé à un nombre et la modalité de réponse peut être comparée à une valeur de référence représentant une des modalités possibles de présentation (par exemple, "visuel" ou "oral"). Bien évidemment, chaque valeur de référence utilisée dans l'éditeur doit être associée à une constante lors de la génération du composant de choix (par exemple, "HYPERTEXTE" ou "ORAL"). Une condition portant sur une unité informationnelle donnée peut être combinée avec une autre condition, en utilisant des opérateurs d'union (*i.e.* OU), d'intersection (*i.e.* ET) ou d'exclusion (*i.e.* NON).

Un ensemble de conditions est relié à un ensemble de tâches de présentation. Ces tâches de présentation sont nécessairement concrètes, aussi une modalité parmi les modalités applicables doit être associée à chaque tâche de présentation sélectionnée. De plus, les tâches de présentation sont synchronisées. Les relations temporelles suivantes entre tâches de présentation, issues de l'espace de composition temporelle des modalités [Vernier, 2001], sont proposées :

- "commence avant" : permet de déterminer qu'une tâche de présentation commence avant une autre et, éventuellement, de spécifier le décalage entre le début des deux tâches de présentation ;
- "commence en même temps" : permet de déterminer que deux tâches de présentation commencent en même temps ;
- "commence après" : permet de déterminer qu'une tâche de présentation commence après une autre ;
- "termine en même temps" : permet de déterminer que deux tâches de présentation terminent en même temps. Cette relation sous-entend que les composants de présentation sont en mesure d'évaluer le temps de chaque tâche de présentation et donc la différence temporelle entre l'exécution ou le lancement des deux tâches de présentation pour garantir qu'elles s'achèvent en même temps ;
- "commence et termine en même temps" : permet de fixer un temps de réalisation égal entre deux tâches de présentation. Cette relation temporelle ne peut être gérée de la même façon lorsque des tâches de présentation - en partie ou complètement - visuelles et auditives sont impliquées. En effet, il peut être problématique d'imposer une certaine durée à la réalisation d'une tâche auditive, qui peut devoir alors être accélérée ou ralentie ;
- "indifférent" : aucune indication n'est donnée sur les relations temporelles entre tâches de présentation. Un cas par défaut est alors appliqué.

Les différentes tâches de présentation organisées temporellement en conclusion d'une règle constituent la spécification de présentation correspondant à un comportement possible du système, étant donné un ensemble de conditions vérifiées. Nous retrouvons les deux niveaux de complémentarité multimodale identifiés dans le chapitre 5, une complé-

mentarité multimodale résultant de la combinaison à un niveau sémantique de tâches de présentation monomodales sur différentes modalités d'une part et une complémentarité multimodale résultant de tâches de présentation combinant plusieurs modalités à un niveau modalitaire d'autre part.

Les notions d'"unité informationnelle", de "tâche de présentation" et de "règle" sont manipulées par l'ergonome/psy-chologue utilisateur de l'éditeur graphique. Afin de pouvoir générer le composant logiciel correspondant, l'informaticien doit définir la correspondance entre les unités informationnelles, utilisées pour définir les deux autres types de notion, et les variables du programme dans lequel s'insère le composant généré. Nous détaillons la collaboration entre l'informaticien et le concepteur non-informaticien dans la section suivante.

6.3.3 Processus de conception avec l'éditeur graphique

Nous décrivons la collaboration entre l'informaticien et le concepteur ergonome/psychologue lors du processus de conception incrémentale avec l'éditeur graphique. Cette collaboration est envisagée comme suit :

1. Étape de définition des unités informationnelles : elle consiste en la déclaration des unités informationnelles utilisables ensuite pour la définition des tâches de présentation abstraite et pour la détermination des conditions des règles de comportement du système ;
   - Mission de l'utilisateur-concepteur : déclaration des unités informationnelles souhaitées en spécifiant le nom attribué et leur source ;
   - Intervention de l'informaticien : vérification de la validité de la déclaration des unités informationnelles et spécification de la correspondance entre les unités informationnelles déclarées par l'utilisateur-concepteur et les données connues par le système ;
2. Étape de définition des tâches de présentation abstraites : elle correspond à la déclaration des tâches de présentation nécessaires à la détermination des spécifications de présentation dans les règles. À tout moment, le retour à l'étape 1 est possible pour définir de nouvelles unités informationnelles ;
   - Mission de l'utilisateur-concepteur : déclaration des tâches de présentation abstraites souhaitées en spécifiant le nom attribué, leur type, l'unité informationnelle concernée et les modalités applicables ;
   - Intervention de l'informaticien : vérification de la validité de la déclaration des tâches de présentation ainsi que de leur génération conformément aux attentes de l'utilisateur-concepteur. Le cas échéant, intégration de nouvelles caractéristiques pour les tâches de présentation ou de nouvelles tâches de présentation possibles et implémentation des processus nécessaires à la génération conforme aux attentes de l'utilisateur-concepteur des nouvelles tâches de présentation définies ;
3. Étape de définition des règles : elle permet le choix implicite des stratégies de dialogue et de présentation à travers la déclaration des règles qui régissent le

comportement multimodal du système. À tout moment, le retour aux étapes 1 et 2 est possible pour définir de nouvelles unités informationnelles ou de nouvelles tâches de présentation ;

- Mission de l'utilisateur-concepteur : pour chaque comportement souhaité, construction d'une règle en définissant un ensemble de conditions portant sur des unités informationnelles ainsi que la spécification de présentation associée consistant en un ensemble de tâches de présentation concrètes (*i.e.* mono ou multimodalement allouées) synchronisées ;
- Intervention de l'informaticien : a priori aucune, si ce n'est la définition éventuelle de constantes correspondant à des valeurs de référence (par exemple, le nombre de solutions maximales présentables sur une modalité donnée) ;

4. Étape de génération : elle consiste en la génération du composant de choix de stratégies de dialogue et de présentation intégrant les règles définies par l'utilisateur-concepteur. À tout moment, il est possible de revenir aux étapes 1, 2 ou 3 pour apporter des modifications au composant de choix généré.

Ce processus de conception implique donc une intervention relative de l'informaticien dans la mesure où le fonctionnement de composant de choix tel que nous l'avons proposé n'est pas modifié. Si l'utilisateur-concepteur novice dans l'utilisation de l'éditeur a besoin de se familiariser avec les notions qu'il peut manipuler avec l'éditeur, elles correspondent à ses besoins pour définir le comportement multimodal du système étudié ou à venir. Au fur et à mesure des itérations, des pré-tests au système final, l'intervention de l'informaticien est de plus en plus limitée, les unités informationnelles et les tâches de présentation nécessaires à la détermination du fonctionnement du composant de choix étant peu à peu complétées. Bien évidemment, le cycle de conception décrit ici est valable uniquement pour le composant de choix et n'intègre pas la conception des autres composants du système final.

Pour illustrer l'utilisation de l'éditeur graphique et la collaboration entre l'informaticien et le concepteur, nous considérons le même exemple que dans le chapitre 5, le système @mie.

6.3.4 Exemple : @mie

L'exemple considéré s'appuie sur l'annuaire multimodal d'entreprise @mie. Ce système permet au personnel d'une entreprise de trouver des informations sur leurs collègues. Les unités informationnelles peuvent être présentées visuellement sur l'écran d'un téléphone mobile avec la modalité <écran, hypertexte (incluant des photos)> et/ou auditivement grâce aux haut-parleurs du terminal avec la modalité <haut-parleurs, langage naturel oral>. La figure 1 (page 3) présente des exemples de comportement du système @mie. Nous considérons que la conception de comportement multimodal, *i.e.* des choix de stratégies de dialogue et de présentation, est spécifiée grâce à l'éditeur graphique. Notons que, à titre d'exemple, nous reprenons les règles utilisées dans le chapitre précédent. Nous rappelons que la validité de ces règles n'a pas été établie mais que celles-ci sont cohérentes avec les principes suivants (*cf.* la section 5.3.3) : (1) partant

du principe que l'utilisateur cherche à accéder à une personne-solution ou des informations la concernant, le système ne donne pas oralement une propriété-cible tant qu'une personne-solution unique n'a pas été identifiée ; (2) dans la mesure du possible, une liste d'informations susceptible de contenir la solution unique à la requête de l'utilisateur ou de l'y mener est affichée de façon à exploiter la rémanence du visuel asynchrone (*i.e.* qui ne change pas au cours du temps, comme c'est le cas d'une vidéo, par exemple) ; (3) la possibilité d'interagir via les capacités d'action gestuelles est toujours laissée à l'utilisateur grâce à un affichage visuel, même s'il a contraint la présentation pour qu'elle soit auditive.

Pour commencer, l'utilisateur-concepteur définit les unités informationnelles qui seront utilisées comme conditions de choix du comportement du système et qui seront présentées grâce aux tâches de présentation définissant le comportement multimodal du système. Reprenant le premier cas de l'exemple illustratif du chapitre 5, deux unités informationnelles doivent être définies : la liste des solutions et le nombre de solutions. Soulignons que ces deux unités informationnelles nécessitent la définition des unités informationnelles auxquelles elles peuvent correspondre, à savoir, dans le cas d'@mie, toutes les propriétés susceptibles de caractériser un employé (*i.e.* son prénom, son nom, son numéro de téléphone fixe, son numéro de téléphone portable, son bureau, sa photo, son fax, sa localisation, son adresse courriel, sa fonction et son équipe) : si ces propriétés ne servent pas à la définition des règles utilisées en exemple, elles pourraient cependant l'être (en particulier, si l'on avait considéré le cas où une seule solution était trouvée, la propriété-cible aurait déterminé les sur-informations associées, comme le numéro de mobile si la requête porte sur le numéro de téléphone et *vice et versa*). Toutefois, dans les règles utilisées en exemple, elles n'interviennent pas directement et l'ergonome/psychologue peut très bien les définir ultérieurement (lors de l'affinement du comportement du système).

L'informaticien doit alors faire le lien entre ces deux unités informationnelles et les données connues du composant de choix. Ce lien dépend directement des données auxquelles le composant de choix peut accéder. Il consiste en une table de correspondance entre le nom donné par l'utilisateur-concepteur à une unité informationnelle et le nom de la donnée correspondante utilisée dans le système. Dans le cas considéré, la table comprend deux entrées, une pour l'unité informationnelle "la liste des solutions" (identifiée dans le système par une variable "solutionsSet") et une pour l'unité informationnelle "le nombre de solutions" (identifiée dans le système par une variable "nbSol").

L'utilisateur-concepteur peut alors définir les tâches de présentation dont il pense avoir besoin pour spécifier le comportement par défaut du système. Dans le cas considéré, il déclare trois tâches de présentation abstraites. La première est nommée "présenter la liste des solutions" : elle est de type "réponse-énumération", elle porte sur l'unité informationnelle "la liste des solutions" et les modalités applicables sont la modalité visuelle et la modalité auditive. La deuxième tâche de présentation abstraite est nommée "présenter le nombre de solutions", elle est de type "réponse-énumération", elle porte sur "le nombre de solutions" et les modalités applicables sont la modalité visuelle et la modalité auditive. La troisième tâche de présentation abstraite définie est nommée "inviter à préciser la requête", elle ne porte sur aucune unité informationnelle, elle est

de type "relance-précision" et les modalités applicables sont la modalité visuelle et la modalité auditive. Dans ce cas précis, l'informaticien n'a pas besoin d'intervenir car les tâches de présentation abstraites que souhaite définir l'utilisateur-concepteur sont des tâches de présentation standard à celles qui ont déjà été spécifiées dans l'éditeur.

L'utilisateur-concepteur peut alors déterminer les règles qui régissent le comportement multimodal du système. Pour le comportement par défaut du système lorsqu'il y a plusieurs solutions, il n'a besoin de définir qu'une condition : elle porte sur l'unité informationnelle "le nombre de solutions", qui est comparée (opérateur $>$) à la valeur de référence 1. Cette condition est reliée à la spécification de présentation correspondante, qui regroupe trois tâches de présentation concrètes synchronisées indifféremment (*i.e.* il n'y a pas de synchronisation imposée entre les tâches de présentation, partant du principe que les tâches de présentation sont concrétisées dans l'ordre de la spécification de présentation). La première tâche de présentation est "présenter le nombre de solutions" à laquelle la modalité auditive est appliquée. La deuxième tâche de présentation est "inviter à préciser la requête" à laquelle la modalité visuelle est appliquée. La troisième tâche de présentation est "présenter la liste des solutions" à laquelle la modalité visuelle est appliquée. La spécification de présentation résultante est multimodalement complémentaire, bien que les tâches de présentation concrètes impliquées sont monomodales. De façon plus concise, la règle définie est la suivante :

si le nombre de solutions est supérieur à 1, alors appliquer le comportement suivant (présenter le nombre de solutions auditivement, inviter à une nouvelle requête visuellement, présenter la liste des solutions auditivement) sans synchronisation particulière.

À ce stade, l'utilisateur peut générer une première version du composant de choix qui ne traite que le cas du comportement multimodal par défaut quand il y a plus d'une solution.

Considérons à présent le deuxième cas de l'exemple illustratif présenté dans le chapitre 5. L'utilisateur-concepteur souhaite que le comportement du système tienne compte de la contrainte de présentation imposée par l'utilisateur. Pour cela, il va définir une unité informationnelle supplémentaire : la modalité de réponse. De plus, il veut qu'une tâche de présentation permette de présenter les critères de restriction pertinents pour une requête donnée, identifiés en amont du composant de choix et faisant partie des sur-informations envoyées par le composant de dialogue au composant de choix. Il définit donc l'unité informationnelle "critères de restriction".

L'informaticien complète la table de correspondance avec l'unité informationnelle "modalité de réponse" (qui, par exemple, correspond à la variable "answerMode") et l'unité informationnelle "critères de restriction" (qui, par exemple, correspond à la variable "restrictionSeq").

L'utilisateur-concepteur peut alors définir la seule tâche de présentation abstraite supplémentaire dont il a besoin, celle qui permet de présenter les critères de restriction permettant de restreindre la requête de l'utilisateur. Il la nomme "présenter la liste des critères de restrictions". Elle est de type "réponse-restriction", porte sur l'unité informationnelle "critères de restriction" et a pour modalités applicables la modalité visuelle et la modalité auditive. Il pourrait également compléter les modalités applicables des tâches de présentation abstraites déjà définies, mais cela est inutile ici car il a déjà

indiqué la modalité auditive et la modalité visuelle comme étant applicables à toutes les tâches de présentation abstraites, et il n'a pas besoin d'autres modalités applicables pour le cas considéré.

L'utilisateur-concepteur peut à présent spécifier les règles qui correspondent au comportement du système décrit dans le cas 2 de l'exemple illustratif du chapitre 5. La règle par défaut déjà existante est modifiée : une condition est ajoutée à la condition sur le nombre de solutions, portant sur l'unité informationnelle "modalité de réponse". En résulte la règle suivante :

si le nombre de solutions est supérieur à 1 ET si la modalité de réponse n'est pas définie, alors appliquer le comportement suivant (présenter le nombre de solutions auditivement, inviter à une nouvelle requête visuellement, présenter la liste des solutions auditivement) sans synchronisation particulière.

Une deuxième règle considère les mêmes conditions mais dans le cas où la modalité de réponse est visuelle. La spécification de présentation correspondante comprend les tâches de présentation concrètes "présenter le nombre de solutions" avec l'application de la modalité visuelle, "inviter à préciser la requête" avec l'application de la modalité visuelle et "présenter la liste des solutions" avec l'application de la modalité visuelle. La règle correspondante est la suivante :

si le nombre de solutions est supérieur à 1 ET si la modalité de réponse est visuelle, alors appliquer le comportement suivant (présenter le nombre de solutions visuellement, inviter à une nouvelle requête visuellement, présenter la liste des solutions visuellement) sans synchronisation particulière.

Une troisième règle correspond aux mêmes conditions dans le cas où la modalité de réponse est auditive. La spécification de présentation associée inclut les tâches de présentation concrètes "présenter le nombre de solutions" avec l'application de la modalité auditive, "présenter la liste des critères de restriction" avec l'application de la modalité auditive et "inviter à préciser la requête" avec l'application de la modalité auditive. La règle correspondante est la suivante :

si le nombre de solutions est supérieur à 1 ET si la modalité de réponse est auditive, alors appliquer le comportement suivant (présenter le nombre de solutions auditivement, présenter la liste des critères de restriction auditivement, inviter à une nouvelle requête auditivement) sans synchronisation particulière.

Le composant de choix de stratégies de dialogue et de présentation correspondant au deuxième cas de l'exemple illustratif du chapitre 5 peut alors être généré.

Nous ne détaillons pas le troisième cas de l'exemple illustratif du chapitre 5. Nous précisons juste que pour ce cas, il n'y a pas besoin de rajouter de nouvelles unités informationnelles car les conditions portant toujours sur la modalité de réponse et sur le nombre de solutions et aucune nouvelle tâche de présentation abstraite n'est nécessaire. Les conditions des trois règles définies pour le deuxième cas doivent être adaptées de façon à prendre en compte les nombres de solutions maximal et minimal pour lesquelles ces règles sont valables et trois nouvelles règles doivent être définies.

Pour conclure ce chapitre, nous résumons notre contribution avant d'en étudier les limites qui peuvent faire l'objet de perspectives.

6.4 Discussion et perspectives

Notre deuxième contribution au choix conjoint de stratégies de dialogue et de présentation se concentre sur la conception des systèmes d'information s'appuyant sur une communication humain-machine naturelle. Nous résumons cette contribution, avant de proposer des perspectives d'amélioration.

6.4.1 Synthèse de la contribution

La prise en compte des contraintes de présentation, qu'elles soient inhérentes aux modalités utilisées ou qu'elles soient issues de la situation de communication, au niveau du choix des stratégies de dialogue **et** de présentation est aujourd'hui réduite. L'une des raisons est sans doute que l'impact des stratégies de dialogue et de présentation sur l'utilisateur et sur la poursuite de la communication reste mal connu. Ce constat est à l'origine de notre éditeur graphique pour faciliter la conception de systèmes d'information multimodaux. Notre outil est destiné à des non-informaticiens tels que les ergonomes et psychologues chargés des études sur l'adéquation du comportement du système aux attentes des utilisateurs et à la réussite de la communication humain-machine. De la spécification faite avec l'éditeur, le composant de choix doit être généré. De cette façon, les premières maquettes utilisées pour l'étude des comportements d'un éventuel nouveau système d'information avant le début de la conception à proprement parler peuvent directement être ré-exploitées dans une mise au point incrémentale du composant de choix du système final.

La conception de cet éditeur nous a conduit à détailler la notion de "tâche de présentation". Plus précisément, nous avons distingué les tâches de présentation abstraites qui incluent des modalités applicables et les tâches de présentation concrètes pour lesquelles la ou les modalités appliquées sont spécifiées. Nous avons également distingué plusieurs types de tâche de présentation, en fonction de la fonction de la tâche de présentation dans la réaction globale du système et de la stratégie de dialogue adoptée. La notion d'"unité informationnelle" a également été précisée. En particulier, nous avons identifié des informations génériques valables pour tous les cadres applicatifs qui peuvent tenir lieu de conditions dans tous les systèmes d'information. Les informations applicatives propres aux composants du domaine sont généralement utilisées pour définir les tâches de présentation.

Une implémentation logicielle de l'éditeur graphique est décrite dans le chapitre suivant. Avant de traiter de sa réalisation logicielle, nous identifions trois limites à notre proposition dans son état actuel, qui donnent lieu à trois perspectives de travail.

6.4.2 Limites de la contribution et perspectives

D'un point de vue purement conceptuel, *i.e.* sans tenir compte des aspects de réalisation logicielle, il conviendrait que la notion d'"unité informationnelle" d'une part et certains éléments des spécifications de présentation d'autre part soient mieux caractérisés. Nous justifions cette remarque et proposons des pistes de travail allant dans ce sens.

6.4.2.1 Caractérisation de la notion d'"unité informationnelle"

Nous avons regroupé sous la notion d'"information" à la fois les unités informationnelles présentables grâce aux tâches de présentation et les contraintes de présentation prises en compte. De par les premiers retours que nous avons eus de la part d'utilisateurs-concepteurs, ce regroupement prête à confusion. En effet, cette notion regroupe différents types d'unités informationnelles définis du point de vue du système. Or, les ergonomes/psychologues qui utilisent l'éditeur considèrent les unités informationnelles du point de vue de l'utilisateur. Par conséquent, ils n'assimilent la plupart des contraintes de présentation à des unités informationnelles. Nous avions commencé à identifier différents types d'unités informationnelles en fonction de leurs sources mais cette distinction ne suffit pas et la détermination de la source n'est pas toujours consensuelle (par exemple, pour les utilisateurs-concepteurs, les unités informationnelles ayant trait à la requête de l'utilisateur, comme sa description et son centre d'attention, ont pour source l'utilisateur).

Il est donc nécessaire de mieux formaliser la notion d'"unité informationnelle" et d'approfondir les différents types possibles. Il convient également d'identifier d'autres informations génériques susceptibles de servir de condition de choix des stratégies de dialogue et de présentation. Ceci implique une analyse des situations de communication contraignantes et des caractéristiques des modalités ainsi qu'une étude de l'impact des stratégies de dialogue et de présentation sur l'utilisateur et la communication. Si le deuxième aspect est facilité par l'éditeur proposé, le premier aspect repose sur la caractérisation des modalités existantes (*cf.* entre autres [Bellik, 1995, Bernsen, 1994, Bernsen, 1997, Clémente, 2004, Ratzka, 2006, Vernier, 2001]) et les critères de coopération entre modalités [Martin, 1995, Clémente, 2004].

6.4.2.2 Saillance comme moyen d'assouplissement des règles

Par ailleurs, les spécifications de présentation peuvent être améliorées. En effet, telle que proposée dans l'éditeur, une spécification de présentation consiste essentiellement en la composition de tâches de présentation, qui ont toutes le même poids. Par conséquent, les composants de présentation (abstraite et concrète) sont obligés de toutes les réaliser. Or, au moins dans une optique d'accessibilité rhétorique, il serait souhaitable de pouvoir déterminer les tâches de présentation les plus importantes, de façon à les privilégier dans les cas où les contraintes de présentation issues de l'environnement matériel d'utilisation ne sont pas prises en compte au niveau du composant de choix. De plus, nous avons fait le choix de considérer la stratégie de présentation à un haut niveau d'abstraction, laissant notamment de côté la mise en avant - par exemple, typographique pour les présentations visuelles ou prosodique pour les présentations auditives - de certaines unités informationnelles. Sans remettre en cause le fait que cet aspect de la stratégie de présentation est traité par les composants de présentation (abstraite et concrète), le composant de choix pourrait donner des indications sur les unités informationnelles à mettre en avant, de façon à ce que la stratégie de présentation dans sa dimension rhétorique soit cohérente.

Pour ces deux limites, il conviendrait que l'utilisateur-concepteur puisse déterminer,

pour une spécification de présentation donnée, les tâches de présentation à mettre en avant. Landragin [Landragin, 2004b] distingue plusieurs types de saillance en fonction des facteurs qui les caractérisent. Nous pensons les intégrer pour préciser les tâches de présentation - et donc les unités informationnelles - à mettre en avant, en indiquant le facteur de saillance impliqué. Pour que cela soit possible, il est nécessaire d'étudier plus finement, pour chaque facteur de saillance considéré et pour chaque modalité (en tant que couple <dispositif physique, langage d'interaction>) les différentes dimensions intéressantes pour la stratégie de présentation et leurs exploitations possibles.

6.4.2.3 Synchronisation des tâches de présentation

Dans l'éditeur proposé, les relations temporelles au sein d'une spécification de présentation se limitent à des indications sur le début et la fin de deux tâches de présentation considérées. Or ces relations temporelles peuvent être interprétées différemment au moment de la concrétisation en fonction des modalités considérées. Par exemple, les synchronisations considérées ne précisent pas la rémanence des tâches de présentation visuelles. De plus, ces synchronisations peuvent aller à l'encontre de l'accessibilité cognitive dans certains cas : la simultanéité de deux tâches de présentation orales limitent fortement l'accès de l'utilisateur aux unités informationnelles. Enfin, la synchronisation temporelle peut entraîner une synchronisation spatiale pour certains modalités. Ainsi, la synchronisation "commence après" pour deux tâches de présentation visuelles peut signifier qu'une première unité informationnelle est affichée, puis une autre, mais aussi qu'une unité informationnelle est affichée en premier sur un écran, l'autre étant affichée en dessous. Les synchronisations que nous avons prises en compte sont donc réductrices. Elles sont pourtant omniprésentes dans la communication humaine, ne serait-ce que des gestes déictiques permettant une mise en saillance d'une information visuelle donnée [Zhou *et al.*, 2005, Wahlster, 2006].

Nous appuyant sur les travaux centrés sur la composition temporelle et/ou spatiale [Feiner *et al.*, 1993, Vernier, 2001], nous pouvons affiner la synchronisation des tâches de présentation. Par exemple, les tâches de présentation pourraient être synchronisées selon les schémas de composition temporelle identifiés par Vernier (*cf.* figure 2.1, page 65). Il serait alors nécessaire d'identifier les critères de description permettant de caractériser au mieux ces schémas, et donc la synchronisation résultante entre les tâches de présentation. Par exemple, l'utilisateur-concepteur devrait être en mesure de déterminer les temps minimal et maximal pouvant s'écouler entre deux tâches de présentation séquentielles. Ces schémas de composition devront être affinés en fonction des modalités allouées aux tâches de présentation : par exemple, deux tâches de présentation partiellement ou complètement auditives ne pourront pas être coïncidentes à moins de remplir certaines conditions ; ou encore, il faut définir ce que l'utilisateur-concepteur entend précisément par la simultanéité de tâches de présentation visuelles simultanées (leurs affichages sont progressifs, commençant et terminant en même temps, ou sont complets en une seule fois de façon parallèle). Ces schémas devront également pouvoir prendre en compte des tâches de présentation multimodales complémentaires.

Signalons que la synchronisation des tâches de présentation au niveau du composant

de choix est susceptible d'entraîner des difficultés de réalisation de la spécification de présentation qu'il faut anticiper au niveau du composant de choix. Par exemple, déclarer qu'une tâche de présentation visuelle commence et termine en même temps qu'une tâche de présentation auditive nécessite, s'il s'agit bien d'un affichage progressif, une coordination entre les vitesses des présentations. Or ces vitesses (d'élocution de la synthèse et d'affichage) ne sont pas toujours contrôlables, et donc pas toujours réalisables. C'est pour éviter de tels problèmes de concrétisation de la synchronisation qu'il est nécessaire de définir finement la synchronisation des tâches de présentation au niveau du composant de choix et de ne pas déléguer leur gestion aux composants de présentation.

Ayant expliqué les concepts manipulés par l'éditeur graphique, nous en décrivons une réalisation logicielle dans le chapitre suivant.

Chapitre 7

Réalisations logicielles

Pour nos deux contributions conceptuelles que sont le composant de choix de stratégies de dialogue et de présentation et l'éditeur graphique permettant la spécification du composant de choix, nous proposons une réalisation logicielle. Ces deux réalisations sont présentées dans ce chapitre. Nous commençons par décrire la réalisation logicielle du composant de choix puis celle de son éditeur de spécification. Ensuite, nous présentons l'intégration du composant de choix, implémenté ou généré, dans une plate-forme de simulation. Enfin, nous illustrons ces réalisations logicielles dans le cadre deux applications, @mie et Santiago.

7.1 Composant de choix de stratégies de dialogue et de présentation

La réalisation logicielle du composant de choix de stratégies de dialogue et de présentation décrit au chapitre 5 s'appuie sur la plate-forme JADE, plus précisément sur son extension JSA. Cette plate-forme permet la création et la gestion d'agents qui communiquent en utilisant le standard FIPA-ACL pour les actes communicatifs [FIPA, 2002a]. Après avoir présenté la plate-forme et les agents JADE-JSA, nous explicitons l'intérêt d'utiliser cette technologie pour la réalisation logicielle du composant de choix, avant d'en décrire les grandes lignes qui seront détaillées dans la section 7.4 pour chacun des exemples implémentés.

7.1.1 Agents JADE-JSA : principes

La plate-forme JADE (pour *Java Agent DEvelopment framework*) [JADE,] est une plate-forme de développement de systèmes multi-agents. Les agents qu'elle permet de concevoir se conforment aux standards de la FIPA (*Foundation for Intelligent Physical Agents*). Ils respectent notamment le standard FIPA-ACL (ACL pour *Agent Communication Language*) pour la communication entre agents. Ce standard formalise les actes communicatifs utilisables par des agents en s'appuyant sur la théorie de l'interaction rationnelle définie par Sadek [Sadek, 1999] (*cf.* la section 4.1.2.4). La plate-forme, en-

tièrement implémentée en JAVA, inclut une librairie pour le développement d'agents JADE, un environnement d'exécution permettant aux agents de "vivre" et d'interagir, ainsi qu'un ensemble d'outils graphiques pour gérer et contrôler l'activité des agents dans l'environnement d'exécution. Nous utilisons la dimension multi-agent pour la réalisation logicielle du composant de choix, plus précisément la possibilité de faire communiquer le composant de choix avec d'autres composants simulés par des agents.

Les agents JADE manquent de flexibilité pour ce qui est de l'interprétation des messages reçus, en particulier du point de vue de leur interprétation sémantique [Louis et Martinez, 2005, Louis et Martinez, 2007]. Par exemple, considérons une même requête qui peut être traduite sous la forme de différents actes communicatifs en fonction de la formulation utilisée par l'utilisateur (de requête ou d'intention d'être informé, par exemple) : si un agent JADE reçoit cette requête sous la forme d'actes communicatifs différents, il n'identifie pas que la finalité de l'agent-interlocuteur est la même. La plate-forme JSA (pour *Jade Semantic Add-on*) permet de pallier cette faiblesse en proposant d'étendre JADE de façon à automatiser l'interprétation des messages reçus par les agents. Grâce à cette plate-forme, la conception d'agents dits "sémantiques" est simplifiée par l'adjonction d'un ensemble de classes spécifiques à l'interprétation sémantique des messages reçus. Le processus d'interprétation sémantique des actes FIPA-ACL, proposé dans le cadre de la plate-forme et des agents JSA, implique une production de sens à partir des messages reçus et une consommation de ce sens pour générer de nouvelles croyances et de nouvelles actions. Pour rendre cette production et cette consommation de sens possibles, deux notions sont introduites : la représentation sémantique et le principe d'interprétation sémantique. La figure 7.1 présente la place de ces deux notions (SR étant l'acronyme anglais "*Semantic Representation*" de "représentation sémantique", et SIP l'acronyme anglais "*Semantic Interpretation Principle*" de "principe d'interprétation sémantique") dans le processus d'interprétation des agents JSA.

Une représentation sémantique est une formule en FIPA-SL (SL renvoyant à *Semantic Language*) [FIPA, 2002b] qui représente le sens attribué par l'agent à un message reçu. La production de sens, *i.e.* de représentations sémantiques, passe par la production de croyances qui portent notamment sur un état du monde et sur les intentions d'un autre agent (l'utilisateur, par exemple). Ces croyances sont stockées dans une base de croyances propre à chaque agent. La consommation de sens, *i.e.* de représentations sémantiques, est réalisée par les principes d'interprétation sémantique.

Un principe d'interprétation sémantique permet la consommation de représentations sémantiques. Cette consommation conduit à la production de nouvelles croyances ajoutées à la base de croyances, de nouvelles représentations sémantiques qui pourront être consommées à leur tour et/ou de nouvelles actions. Pour que ces productions soient possibles, un principe d'interprétation sémantique est associé à une condition d'application (concernant son état mental par exemple) et à une forme générale de représentation sémantique à laquelle il s'applique. Pour qu'il y ait consommation d'une représentation sémantique donnée par un principe d'interprétation sémantique particulier, cette représentation sémantique doit respecter la forme générale associée à ce principe d'interprétation sémantique et la condition doit être vérifiée. C'est l'application du principe d'interprétation sémantique qui est à l'origine de la production de nouvelles croyances,

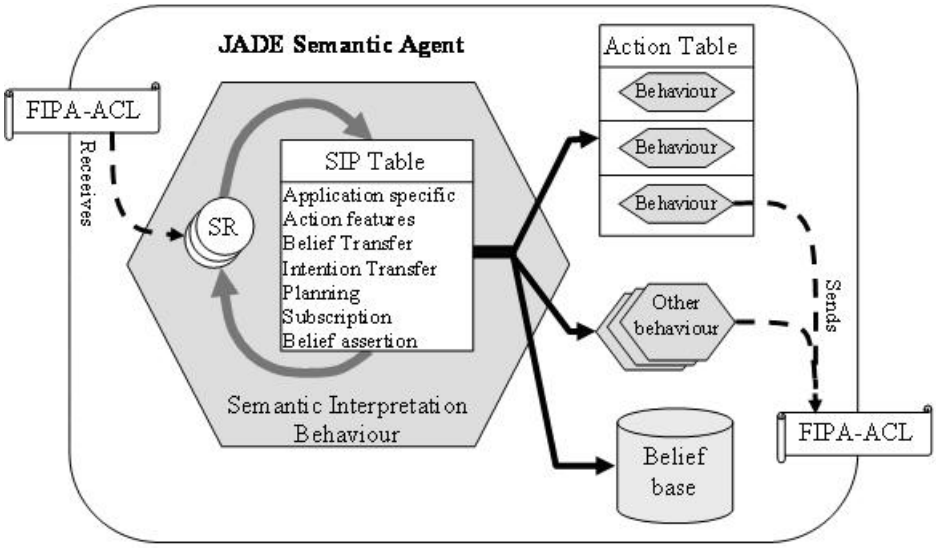

FIG. 7.1 – Processus d'interprétation des agents JSA (extrait de [Louis et Martinez, 2007])

de nouvelles représentations sémantiques et/ou de nouvelles actions.

Un agent JSA se distingue d'un agent JADE par des principes d'interprétation sémantique génériques qui lui permettent d'identifier les actes communicatifs sémantiquement identiques et de les consommer en conséquence. De plus, la plate-forme JSA permet de construire un agent JSA en spécialisant - selon l'approche orientée objet - des principes d'interprétation sémantique génériques ou en ajoutant de nouveaux principes d'interprétation sémantique spécifiques voire applicatifs. La consommation des principes d'interprétation sémantique est assimilable à un moteur de règles. C'est cet aspect de la plate-forme JSA que nous exploitons pour la réalisation logicielle du composant de choix de stratégies de dialogue et de présentation à base de règles. L'utilisation de ce moteur de règles est détaillée dans la section suivante.

7.1.2 Intérêt du choix d'un agent JADE-JSA pour implémenter un composant de choix de stratégies de dialogue et de présentation

Le formalisme du choix des stratégies de dialogue et de présentation pour lequel nous avons opté est à base de règles. Pour un ensemble de contenus possibles et de contraintes de présentation, ces règles déterminent un ensemble de tâches de présentation à réaliser et de contenus choisis à indiquer en retour au composant de dialogue. L'ensemble des contenus possibles et des contraintes de présentation peut être exprimé sous forme de messages en FIPA-ACL. Le composant de choix peut alors, s'il est un agent JSA, consommer ces messages grâce à des principes d'interprétation sémantique applicatifs de façon à produire un ou des actes communicatifs qui correspondent aux tâches de présentation à réaliser.

Un agent JSA joue donc le rôle de moteur de règles et la plate-forme JSA facilite l'implémentation d'un tel moteur. En effet, la création de nouveaux principes d'interprétation sémantique , *i.e.* de nouvelles règles, pour consommer des représentations sémantiques identifiées, *i.e.* des contenus possibles et des contraintes de présentation connues, est simplifiée par l'existence de différentes méthodes propres aux agents JSA. Parmi ces méthodes, certaines sont dédiées à l'identification des représentations sémantiques issues des messages reçus. Ces représentations sémantiques correspondent à la forme générale de représentation sémantique associée à un principe d'interprétation sémantique donné. La reconnaissance des messages, *i.e.* des contenus possibles et des contraintes de présentation, pour lesquels un principe d'interprétation sémantique s'applique est facilitée par le mécanisme de consommation des représentations sémantiques dans un agent JSA.

Les principes d'interprétation sémantique mettent alors en œuvre les règles de choix de stratégies de dialogue et de présentation adoptées. Le cas échéant, il est possible de considérer qu'une règle est générique et d'en faire un principe d'interprétation sémantique générique pour tous les agents JSA qui sont des composants de choix. De plus, dans le cas où un principe d'interprétation sémantique s'applique, c'est-à-dire où une règle est valable pour un ensemble de contenus et de contraintes de présentation, la production de tâches de présentation sous forme d'actes communicatifs est déjà instrumentée.

La plate-forme JSA étant une extension de JADE, l'environnement d'exécution permet de mettre en place un système multi-agent incluant un agent composant de choix et d'autres agents pour les autres composants de notre solution architecturale qui étend Arch (*cf.* le chapitre 5). C'est l'approche que nous utilisons dans l'architecture logicielle décrite dans la section suivante. Néanmoins, il est important de noter que le composant de choix développé par un agent JSA peut tout à fait s'intégrer dans un système existant où les composants et en particulier le contrôleur de dialogue et les composants de présentation ne sont pas des agents JSA. Dans ce cas, l'envoi de message à destination ou en provenance du composant de choix se fait grâce à une bibliothèque logicielle (API) qui transforme les messages reçus en messages au format FIPA-ACL et *vice et versa*. Plus généralement, l'intégration d'un agent JSA au sein d'un système par l'ajout de traducteurs FIPA-ACL dans les agents JSA en amont de l'interprétation sémantique des messages fait l'objet d'études en cours.

7.1.3 Principes d'implémentation d'un composant de choix avec un agent JADE-JSA

Nous décrivons ici les principes d'implémentation d'un composant de choix de stratégies de dialogue et de présentation avec un agent JADE-JSA. Ces principes seront détaillés pour les applications-exemples utilisées dans la section 7.4.

Les prémisses des règles de choix de stratégies de dialogue et de présentation peuvent être implémentées au niveau des représentations sémantiques ou au sein de principes d'interprétation sémantique. Les conclusions de ces règles, *i.e.* la définition des spécifications de présentation correspondantes, sont déterminées au sein des principes d'interprétation sémantique. Dans notre cas, l'application d'un principe d'interprétation sémantique se solde par l'envoi d'un message aux autres agents, message qui contient la

spécification de présentation produite. Nous détaillons, dans les paragraphes suivants, la façon dont les règles possibles du composant de choix peuvent être implémentées dans un agent JADE-JSA.

Qu'elles soient implémentées au niveau des représentations sémantiques ou des principes d'interprétation sémantique, les règles de choix de stratégies de dialogue et de présentation s'appuient sur l'identification d'un ensemble donné de contenus possibles donné par le composant de dialogue et d'éventuelles contraintes de présentation. Pour que cette identification soit possible, l'ensemble des contenus possibles est exprimé sous forme d'une représentation sémantique. La représentation sémantique utilisée comme patron-type dans le cadre de l'implémentation du composant de choix du système @mie est présentée dans la section 7.4.2.2 (*cf.* page 234). Ce patron peut n'indiquer que la structure générale de la représentation sémantique ou préciser la valeur prise par certains paramètres qui caractérisent cette représentation sémantique : dans le premier cas, une ou plusieurs règles pourront être définies au niveau du principe d'interprétation sémantique associé ; dans le deuxième cas, une seule règle est déterminée au niveau de la représentation sémantique. Cette deuxième possibilité nécessite qu'il n'y ait aucune opération à réaliser sur la valeur d'un paramètre (*e.g.* l'extraction du nombre d'éléments contenus dans un ensemble, comme c'est le cas pour le nombre solution) et de connaître les valeurs que peut prendre un paramètre donné (ce qui est le cas pour les contraintes de présentation émanant de l'utilisateur : par exemple, "visual", "aural" ou "null" dans le cas du système @mie). Notons que, de façon à limiter les accès et le maintien à une base de données extérieure dédiée aux contraintes de présentation, nous avons choisi d'intégrer les contraintes de présentation émanant de l'utilisateur comme faisant partie de la formule FIPA-SL envoyée par le composant de dialogue au composant de choix et qui comprend par ailleurs l'ensemble des contenus possibles pour une requête donnée de l'utilisateur ainsi que la caractérisation de cette requête (*cf.* page 234).

L'identification d'une représentation sémantique, *i.e.* d'un ensemble de contenus possibles, permet l'extraction des valeurs prises par les paramètres qui caractérisent cette représentation sémantique. À l'exécution, ces valeurs correspondent à la caractérisation de la requête (incluant la contrainte de présentation de l'utilisateur) et l'ensemble des contenus possibles. La récupération de ces valeurs dans le cas de l'application @mie est présentée dans la section 7.4.2.2 (*cf.* page 234).

Lorsqu'une règle ne peut être caractérisée par une représentation sémantique unique, ou si des opérations sur les valeurs des paramètres de la représentation sémantique sont nécessaires pour exprimer les conditions d'une ou de plusieurs règles, les valeurs extraites lors de l'identification de la représentation sémantique sont utilisées dans le principe d'interprétation sémantique associé comme conditions des règles de choix et correspondent aux conditions d'envoi d'un message aux autres composants de l'architecture. Par exemple, considérons l'exemple d'@mie dans le cas des règles présentées dans la figure 5.6 (*cf.* page 173) et dont l'extraction (décrite *cf.* page 234) conduit à obtenir, entre autres, la contrainte de présentation émanant de l'utilisateur ("answerModeInput") et l'ensemble des solutions ("solutionSetInput"). La règle dont les conditions sont la limitation à 3 du nombre de solutions (3 renvoyant à Y et correspond au nombre maximal de solutions pour lequel une stratégie d'énumération s'applique au-delà duquel le système

adopte une stratégie de restriction) et sur une contrainte de présentation auditive est définie en tant que condition d'envoi du message au sein du principe d'interprétation sémantique associé à la représentation sémantique identifiée de la façon suivante :

```
// contrainte de présentation auditive ?
if ( answerModeInput.stringValue().equals("aural") )
{
// récupération du nombre de solutions
int cardSolutionSet = solutionSetInput.size();
// examination du nombre de solutions : cas où il y a plus de X solutions
if ( cardSolutionSet > 3 )
{
// construction des différents tâches de présentation qui constituent
la spécification de présentation en utilisant les valeurs des paramètres
récupérés en entrée
}
else // cas où il y a moins de X solutions = nouvelle règle
{
// construction des différents tâches de présentation qui constituent
la spécification de présentation en utilisant les valeurs des paramètres
récupérés en entrée
}

// ajout d'une action qui correspond à un envoi de message aux autres agents
contenant la spécification de présentation produite
}
```

Dans cet exemple, le principe d'interprétation sémantique comprend deux règles où la contrainte de présentation est auditive, l'une où le nombre de solutions est supérieur à 3 et l'autre où le nombre de solutions est égal ou inférieur à trois. Une autre possibilité aurait été de ne pas inclure d'instruction "sinon" et de créer une nouvelle représentation sémantique incluant un principe d'interprétation sémantique incluant les conditions suivantes :

```
// contrainte de présentation auditive ?
if ( answerModeInput.stringValue().equals("aural") )
{
// récupération du nombre de solutions
int cardSolutionSet = solutionSetInput.size();
// examination du nombre de solutions : cas où il y a plus de X solutions
if ( cardSolutionSet <= 3 )
{
// construction des différents tâches de présentation qui constituent
la spécification de présentation en utilisant les valeurs des paramètres
récupérés en entrée
```

```
}
// ajout d'une action qui correspond à un envoi de message aux autres agents
contenant la spécification de présentation produite
}
```

Les spécifications de présentation associées à chaque règle sont définies au sein des principes d'interprétation sémantique. Pour cela, les valeurs des paramètres extraits de la représentation sémantique identifiée sont également utilisées. En reprenant la règle du système @mie où il y a une contrainte de présentation auditive et où le nombre de solutions est supérieur à 3, nous donnons deux exemples de construction de tâches de présentation utilisées pour la spécification de présentation résultante. La première est construite à partir de l'ensemble des critères de restriction ("restrictionSequenceInput") récupéré à partir de la représentation sémantique identifié :

```
// construction de la tâche de présentation de restriction auditive
grâce à la fonction restrictionConstruction
partieRep2 = restrictionConstruction(restrictionSequenceInput,ORALMOD);
```

La tâche de présentation résultante sera, à l'exécution, de la forme suivante (les points de suspension correspondant à la liste de restriction récupérée) :

```
(restriction :restrictionSet ... :modality oral)
```

La deuxième tâche de présentation donnée en exemple ne réutilise pas de valeurs de paramètres de représentation sémantique identifié. Il s'agit d'une invitation à préciser la requête construite de la façon suivante :

```
partieRep3 = (Term)SL.fromTerm("(invit :type precision
:modality hypertext)");
```

Les mêmes principes ont été appliqués pour générer le composant de choix au sein de l'éditeur graphique dont les concepts ont été présentés au chapitre 6. Nous décrivons la réalisation logicielle de cet éditeur graphique dans la section suivante.

7.2 Éditeur graphique de spécification du composant de choix

La réalisation logicielle d'un composant de choix de stratégies de dialogue et de présentation par un agent JSA a servi de modèle pour la réalisation logicielle de l'éditeur graphique décrit dont les concepts et l'utilisation ont été décrits au chapitre 6. À partir de la spécification du composant de choix qu'il aide à concevoir, cet éditeur doit générer un agent JSA composant de choix. Par simplicité, nous avons opté pour un éditeur de génération qui, tout comme les agents JSA, est implémenté en JAVA. Comme nous

allons le voir plus en détail, des formulaires sont utilisés pour la saisie des unités informationnelles et des tâches de présentation abstraites. Pour la gestion des règles de choix, nous avons utilisé la libraire JGraph. Nous présentons l'organisation générale de l'éditeur avant de détailler les trois principaux onglets utilisés pour la conception du composant de choix de stratégies de dialogue et de présentation.

7.2.1 Organisation générale de l'éditeur

Comme le montre la figure 7.2, l'éditeur est constitué de six onglets. Deux vues sont distinguées : la "vue ergonome" et la "vue informaticien". La "vue ergonome" permet l'accès aux trois onglets suivants :

- un onglet "Informations" qui permet la gestion des unités informationnelles utilisées par le système ;
- un onglet "Tâches de p." qui permet la gestion des tâches de présentation abstraites ;
- un onglet "Règles" qui permet la gestion des règles constitutives du composant de choix généré.

La "vue informaticien" permet l'accès, en plus de ces trois onglets, à trois autres onglets qui sont :

- l'onglet "Informations techniques" qui permet de visualiser des informations liées à la génération du composant de choix (plus précisément, le nom de la classe du composant de choix, les classes importées, les constantes existantes et leur valeur, et la structure de la spécification de présentation dans l'acte communicatif de sortie) ;
- l'onglet "Message" qui permet de compléter, en fonction des unités informationnelles applicatives, la structure générale de la formule FIPA-SL de l'acte communicatif que reçoit le composant de choix ;
- l'onglet "Génération" qui permet la génération du composant de choix et, à terme, le lancement d'une simulation intégrant un simulateur de composant de dialogue en entrée du composant de choix et un simulateur des composants de présentation en sortie du composant de choix.

Parmi ces six onglets, quatre sont indispensables pour la conception du composant de choix grâce à l'éditeur graphique de génération : l'onglet "Informations", l'onglet "Tâches de p.", l'onglet "Règles" et l'onglet "Message". L'onglet "Informations techniques" sert uniquement de support à l'informaticien et l'onglet "Génération" est utilisé en fin de conception pour générer le composant de choix de stratégies de dialogue et de présentation conçu avec l'éditeur, voire lancer une simulation pour l'utiliser. Nous détaillons les quatre onglets principaux dans leur ordre d'utilisation pour la conception d'un composant de choix.

7.2.2 Définition des unités informationnelles

L'onglet "Informations" permet à l'utilisateur-concepteur de définir les unités informationnelles utilisées pour le choix des stratégies de dialogue et de présentation. Les

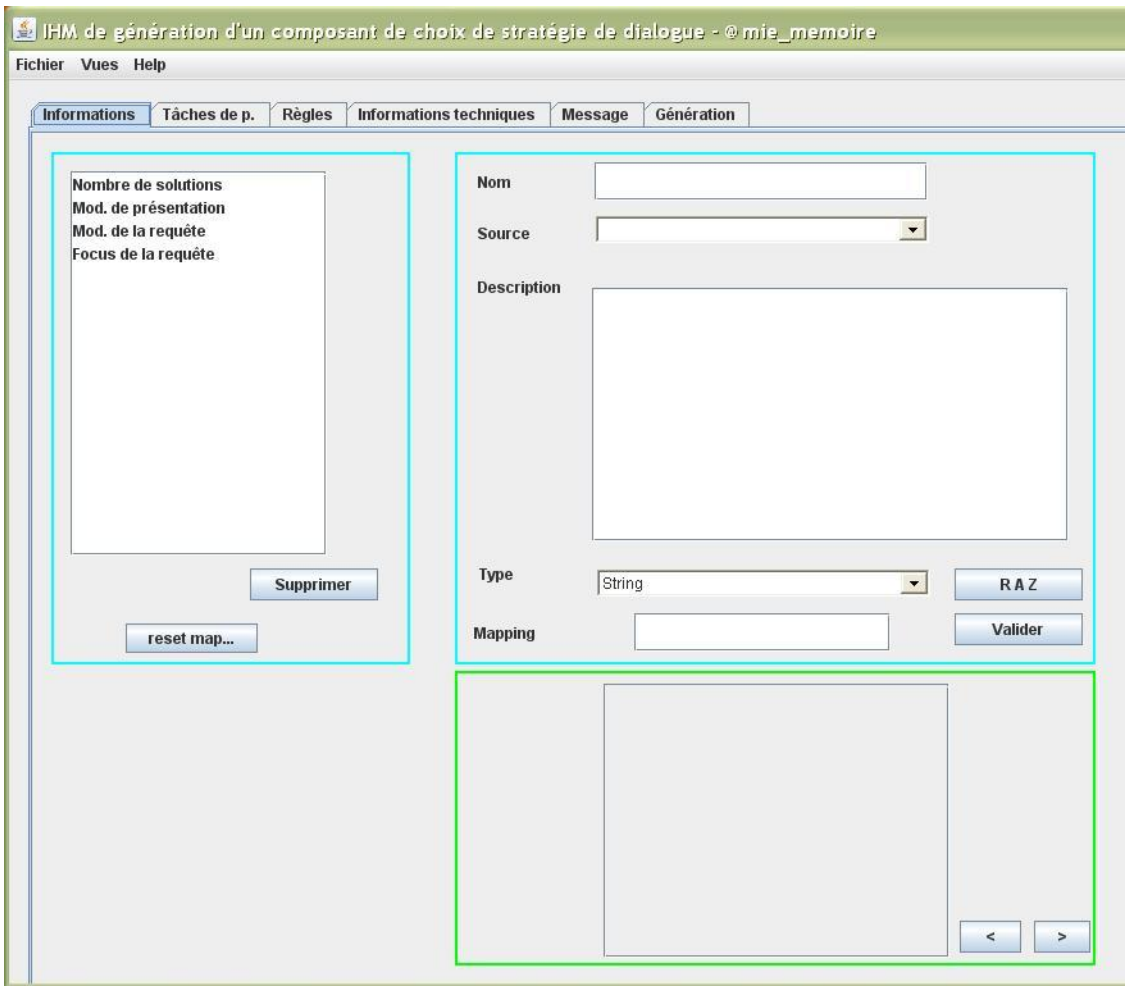

FIG. 7.2 – Éditeur graphique pour spécifier le composant de choix

principales caractéristiques renseignées sont celles identifiées dans le chapitre 6 pour la notion d'"unité informationnelle", auxquelles s'ajoutent des attributs utilisées par l'éditeur ou lors de la génération du composant de choix. Comme le montre la figure 7.3, l'utilisateur-concepteur doit donc spécifier le nom de l'unité informationnelle (zone A) et sa description optionnelle (zone C). Lui ou l'informaticien doivent aussi déterminer la source de l'unité informationnelle (zone B), son type (zone D) et le nom de la propriété correspondante dans le système (zone E). La source est l'une des trois sources possibles identifiées, à savoir l'utilisateur, le système ou le contexte. Le type permet de proposer les opérateurs de comparaison adéquats lorsque l'unité informationnelle est utilisée en condition dans les règles de choix. Dans la figure 7.3, l'unité informationnelle appelée "prénom" a pour source le système, est une chaîne de caractères ("String") et est connue dans le système sous la variable ou le paramètre "firstname". Notons qu'une unité informationnelle peut correspondre à une composition de plusieurs unités infor-

mationnelles (zone F) : par exemple, une personne est définie par son prénom, son nom et sa photographie[1] : nous pensons, à terme, utiliser cette possibilité pour définir des tâches de présentation non pas sur une unité informationnelle mais sur plusieurs unités informationnelles qui font sens.

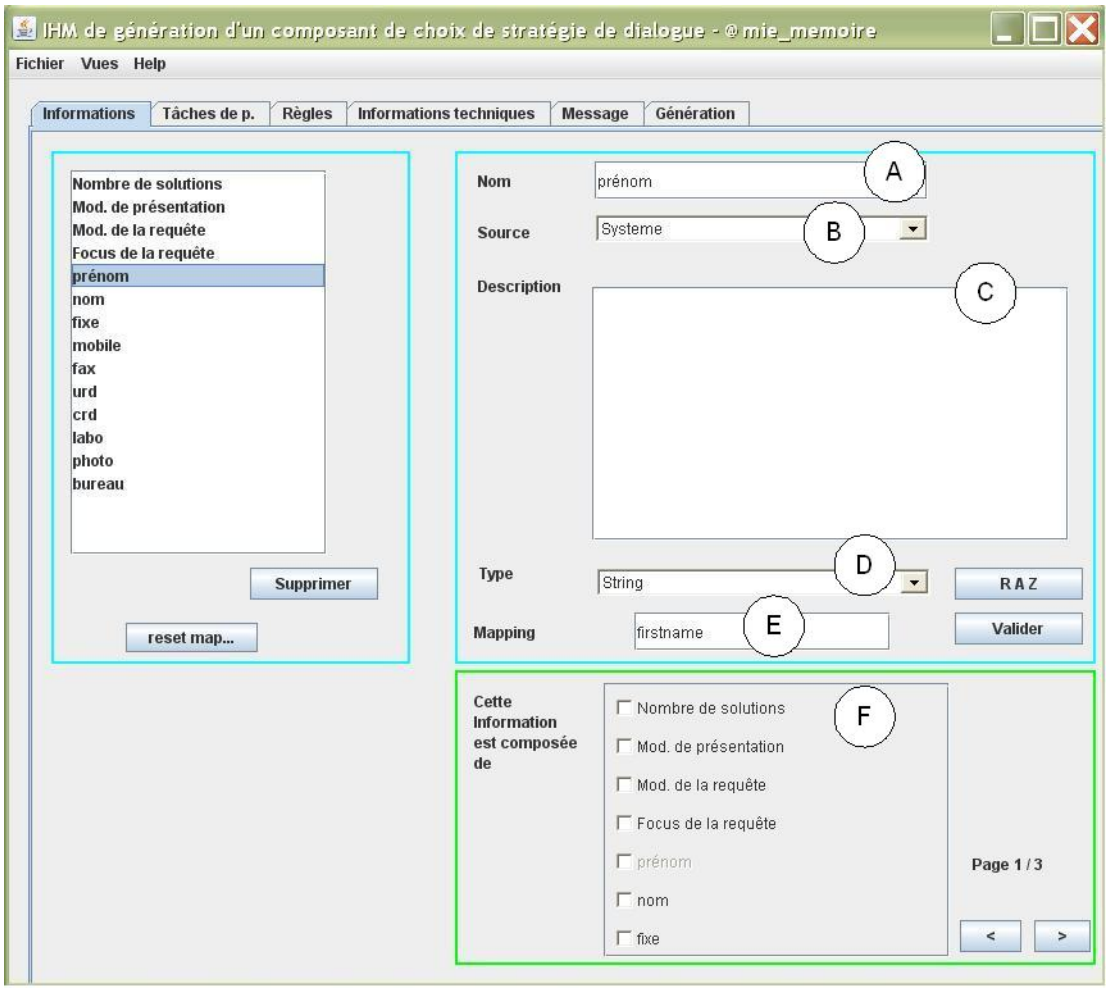

FIG. 7.3 – Définition des unités informationnelles

Pour que la correspondance entre les unités informationnelles et les données connues du système soit faite, l'indication du nom de la propriété correspondante pour chaque unité informationnelle ne suffit pas. Il faut également identifier la façon dont est récupérée cette unité informationnelle au niveau de la formule contenue dans l'acte communicatif que reçoit le composant de choix en entrée. Ceci est réalisé, généralement par l'informaticien, grâce à l'onglet "Message".

[1]Les cases de ces différentes unités informationnelles dans la zone F doivent être cochées.

7.2.3 Définition du message d'entrée

L'onglet "Message" est indispensable pour faire le lien entre les unités informationnelles déclarées dans l'onglet "Informations" et le message que reçoit en entrée le composant de choix. La structure de ce message en entrée, *i.e.* la formule FIPA-SL contenue dans l'acte communicatif transmis, est pré-formatée. Elle permet de considérer automatiquement le nombre de solutions comme une unité informationnelle susceptible d'être utilisée comme condition des règles de choix. La structure de ce message est la suivante :

```
"(content
:requestFocus ??requestFocus
:requestDescription ??requestDescription
:requestModality ??requestModality
:answerMode ??answerMode
:solutionCard ??nbSol
:solutionSet ??solutionSet
:restrictionSeq ??restrictionSeq
:relaxationSeq ??relaxationSeq
)"
```

En renvoyant aux unités informationnelles applicatives définies dans l'onglet "Informations", les paramètres suivants de la structure applicative du message d'entrée pré-formatée doivent être précisés grâce à l'onglet "Message" :

- les propriétés qui peuvent être utilisées comme centre d'attention de la requête (paramètre "requestFocus") ;
- l'ensemble des propriétés qui peuvent être utilisées pour décrire la requête de l'utilisateur (paramètre "requestDescriptionInput" qui est de la forme `"(set (description ... ) (description ... ) ...)"`) ;
- l'ensemble des propriétés qui définissent une solution (paramètre "solutionSet" qui est de la forme `"(set (solution ... ) (solution ... ) ...)"`) ;
- l'ensemble des propriétés qui définissent une solution approchée (paramètre "relaxationSeq" qui est de la forme `"(set (relaxation ... ) (relaxation ... ) ...)"`). Il s'agit du même ensemble que celui des propriétés qui définissent une solution ;
- l'ensemble ordonné des propriétés pouvant servir de critères de restrictions (paramètre "restrictionSeq" qui est de la forme `"(sequence ... )"`) ;

Le paramètre "requestModality" qui définit la modalité de la requête, le paramètre "answerMode" qui définit la contrainte de présentation de l'utilisateur et le paramètre "nbSol" qui définit le nombre de solutions ne sont pas à renseigner. Ils sont automatiquement inclus comme des unités informationnelles au niveau de l'onglet "Informations (*cf.* figure 7.2). Des valeurs par défaut sont associées dans le code aux deux premiers paramètres. En particulier, le paramètre "answerMode" peut prendre la valeur "null", qui indique une absence de contrainte de présentation, la valeur "HYPERTEXT", qui indique une contrainte de présentation visuelle et la valeur "ORAL" qui indique une

contrainte de présentation auditive. Ce message pré-formaté peut s'avérer insuffisant ou inadéquat avec le message reçu en entrée par le composant de choix. Dans ce cas, il doit être redéfini au niveau du code.

La correspondance entre les unités informationnelles et les données incluses dans la formule FIPA-SL de l'acte communicatif reçu par le composant de choix étant établie, ces unités informationnelles peuvent être utilisées pour la spécification des tâches de présentation abstraites grâce à l'onglet "Tâches de p.".

7.2.4 Définition des tâches de présentation abstraite

La caractérisation des tâches de présentation grâce à l'éditeur s'appuie sur la caractérisation des tâches de présentation abstraites présentée au chapitre 6. Comme le montre la figure 7.4, l'utilisateur-concepteur doit, pour chaque tâche de présentation abstraite, déterminer son nom (zone A), son type (à partir des types utilisées pour l'expérimentation du service Santiago présentées au début de ce chapitre, à savoir les types feedback, une réponse ou une relance, auxquels nous ajoutons l'aide), sa description optionnelle (zone C) et les modalités applicables (zone D). Les modalités applicables prédéfinies sont la modalité visuelle, la modalité auditive, une combinaison des modalités auditive et visuelle sans contrainte de coopération (VA Neutre), une combinaison complémentaire (VA Complémentaire) et une combinaison redondante (VA Redondance) de ces modalités. La définition de nouvelles modalités ou combinaison de modalités est possible, mais l'informaticien doit alors compléter le code de l'éditeur afin que ces modalités apparaissent dans l'onglet "Tâches de p.", qu'une couleur spécifique leur soit attribuée dans l'onglet "Règles" pour les tâches de présentation concrètes auxquelles elles sont allouées et qu'un nom leur soit attribué dans les tâches de présentation concrètes lors de la génération du composant de choix.

La sélection du type de la tâche de présentation abstraite nécessite de spécifier certaines caractéristiques dépendantes du type (zone F), par l'utilisateur-concepteur ou par l'informaticien. Les tâches de présentation de type "réponse" doivent être spécifiées en fonction de la stratégie de dialogue adoptée (relaxation, énumération ou restriction). Dans le cas d'une énumération (le terme anglais *statement* est utilisée dans l'interface, notamment à la figure 7.4), le centre d'attention de la tâche de présentation peut être spécifié. Ce centre d'attention est l'une des propriétés définies dans l'onglet "Message" pouvant caractériser une solution ou le nombre de solutions. Les tâches de présentation de type "relance" peuvent consister en une nouvelle requête ou en une demande de précision.

Du point de vue du système, les tâches de présentation concrètes (*i.e.* après allocation des modalités au niveau des règles) possibles sont définies comme suit :

- les tâches de présentation de type "feedback", dont les valeurs " ? ?requestFocus" et " ? ?requestDescription" sont récupérées de la formule FIPA-SL reçue en entrée du composant de choix :

```
(feedback
:requestFocus ??requestFocus
:description ??requestDescription
```

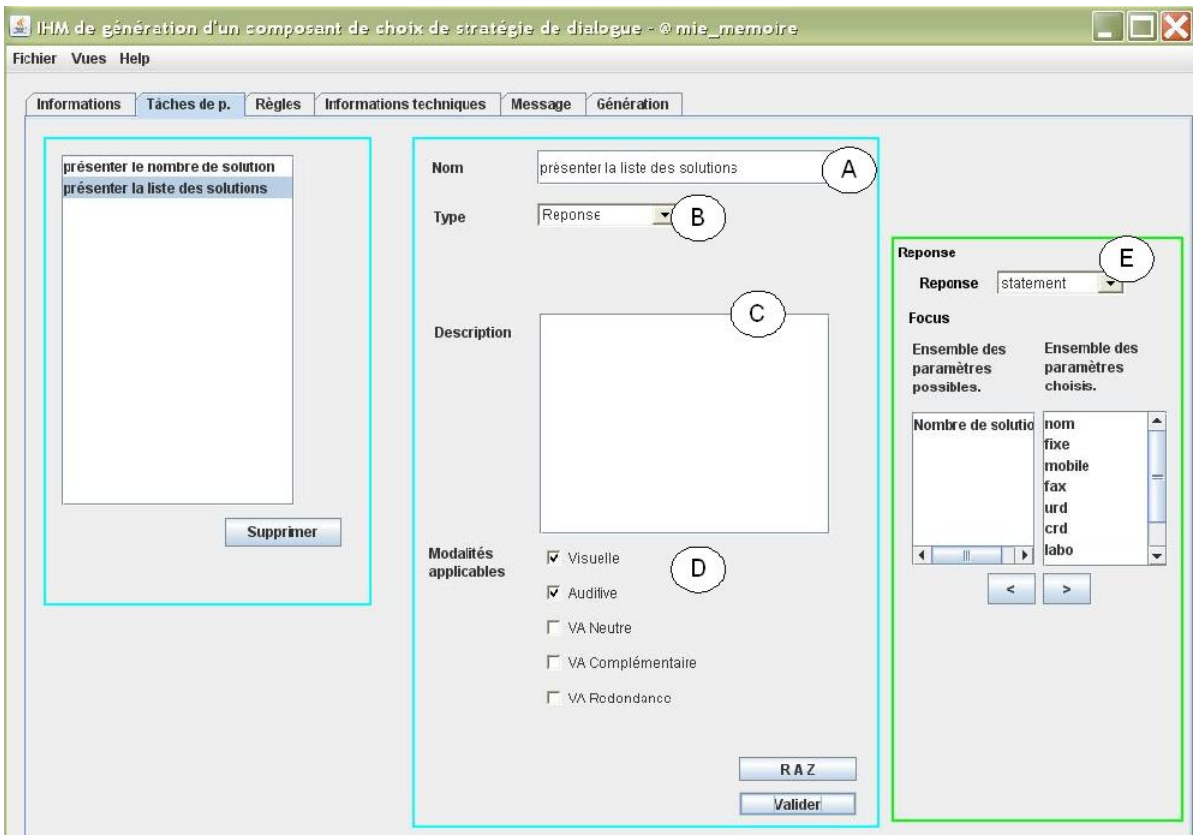

FIG. 7.4 – Définition des tâches de présentation abstraites

```
:modality ??modalityValue)
```

– les tâches de présentation de type "réponse-relaxation", dont la valeur "??relaxationSeq" est récupérée de la formule FIPA-SL reçue en entrée du composant de choix :

```
(reponse
:type relaxation
:description ??relaxationSeq )
:modality ??modalityValue
)
```

– les tâches de présentation de type "réponse-énumération", dont les valeurs "??solutionSet" et "??nbSol" sont récupérées de la formule FIPA-SL reçue en entrée du composant de choix, et la valeur "??focus" est un ensemble spécifié dans la définition de la tâche de présentation :

```
(reponse
:type statement
:focus ??focus
:nbSolutions ??nbSol
```

```
:description ??solutionSet )
:modality ??modalityValue
)
```

– les tâches de présentation de type "réponse-restriction", dont la valeur " ? ?restrictionSeq" est récupérée de la formule FIPA-SL reçue en entrée du composant de choix :

```
(reponse
:type restriction
:description ??restrictionSeq )
:modality ??modalityValue
)
```

– les tâches de présentation de type "relance-précision" :

```
(relance
:type precision
:modality ??modalityValue
)
```

– les tâches de présentation de type "relance-nouvelle requête" :

```
(relance
:type nvlRequete
:modality ??modalityValue
)
```

– les tâches de présentation de type "aide", dont la valeur " ? ?description" est un chaîne de caractères spécifiée dans la définition de la tâche de présentation : :

```
(aide
:description ??description
:modality ??modalityValue
)
```

Les valeurs " ? ?modalityValue", qui correspondent aux modalités allouées, sont déterminées pour chaque tâche de présentation lorsqu'elles sont utilisées, dans l'onglet "Règles", pour la définition de règles de choix.

7.2.5 Définition des règles de choix

Les règles sont définies selon les principes présentés au chapitre 6. La réalisation logicielle de cette partie de l'éditeur graphique repose sur l'utilisation de JGraph. Comme le montre la figure 7.5, l'utilisateur-concepteur peut créer des règles (zone C) à partir des unités informationnelles (zone A) et des tâches de présentation abstraites (zone B) précédemment définies.

Lorsqu'une unité informationnelle est sélectionnée pour être utilisée comme condition de règle, l'utilisateur-concepteur doit déterminer l'opérateur et la valeur de comparaison. Cette dernière peut être prédéfinie (c'est le cas de la constante "NBMAXSOLVISUAL" de la figure 7.5 qui correspond au nombre maximum de solutions présentables visuellement) ou saisie manuellement. Une condition est donc un tripet <UI, C, V>, où UI est l'unité informationnelle considérée et définie grâce à la vue "Informations",

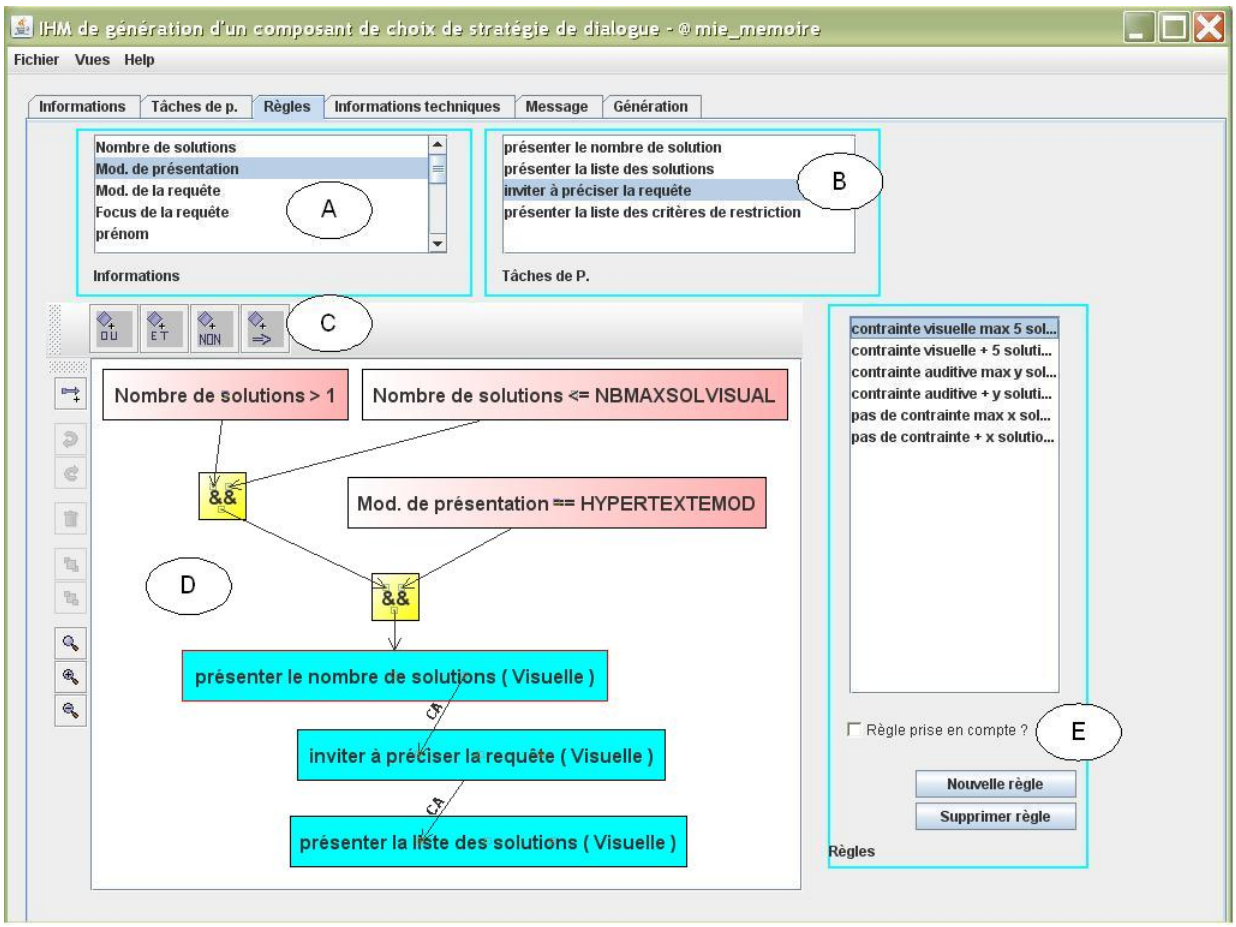

FIG. 7.5 – Définition des règles

C l'opérateur de comparaison et V la valeur de comparaison. Les conditions peuvent être combinées grâce aux opérateurs booléens classiques (zone C). Si la règle définie n'a qu'une condition, elle est reliée à la première tâche de présentation concrète de la spécification de présentation grâce à une flèche (boîte de droite de la zone C de la figure 7.5).

La ou les tâches de présentation concrètes qui constituent la spécification de présentation, *i.e.* la conclusion de la règle, sont définies à partir des tâches de présentation abstraites sélectionnées. Pour chacune de celles-ci, l'utilisateur concepteur doit déterminer la modalité appliquée parmi les modalités applicables indiquées lors de leurs définitions grâce à l'onglet "Tâches de p.". Les tâches de présentation concrètes doivent être reliées entre elles, et leurs relations temporelles doivent être spécifiées. Lorsqu'une règle est prise en compte lors de la génération du composant de choix, la spécification de présentation produite si la règle s'applique est formatée de la façon suivante :

```
(answer  (??operateurSynchro ??groupeTP ??groupeTP))
```

Cette formule FIPA-SL permet de prendre en compte les synchronisations possibles entre tâches de présentation. Contrairement à la réalisation logicielle pour le composant de choix présentée au début de ce chapitre, le symbole fonctionnel "answer" n'a pas pour valeur une séquence ordonnée de tâches de présentation. Il est constitué d'un terme fonctionnel constitué d'un symbole fonctionnel " ? ?operateurSynchro" et de deux termes " ? ?groupeTP". Le symbole fonctionnel renvoie à l'un des opérateurs de synchronisation suivants : IND (pour "indifférent"), CA (pour "commence avant"), CMT (pour "commence en même temps"), CP (pour "commence après"), TMT (pour "termine en même temps") et CTMT (pour "commence et termine en même temps"). Les deux termes peuvent être soit une tâche de présentation, soit un terme fonctionnel `(??operateurSynchro ??groupeTP ??groupeTP)`.

Lors de la génération du composant de choix , chaque règle peut être prise en compte, ou non (zone E). Ainsi plusieurs comportements peuvent-ils être testés pour un même système considéré. Les principes de génération sont les suivants.

7.2.6 Génération du composant de choix

Le composant de choix conçu avec l'interface graphique peut être généré grâce à l'onglet "Génération". Comme le montre la figure 7.6, cet onglet propose, outre un bouton de génération ("Générer"), un bouton de vérification de la validité des éléments définis ("Vérifier") et un bouton pour lancer une simulation avec le composant de choix généré ("Générer Sim."). La vérification consiste à vérifier que tous les éléments (unités informationnelles, tâches de présentation abstraites et règles) sont correctement renseignés. Si tous les éléments ne sont pas valides, le bouton "Vérifier" provoque l'affichage de feux rouges pour les aspects mal renseignés, (partie gauche de la figure 7.6). Plus précisément, il est vérifié (selon l'ordre d'affichage des feux rouges) que :

- les spécifications de présentation ont bien un point d'entrée, *i.e.* une tâche de présentation désignée comme étant le point d'entrée dans le graphe de la règle. Il s'agit d'une contrainte imposée par l'utilisation de JGraph ;
- les tâches de présentation possèdent bien des modalités applicables ;
- une modalité a bien été appliquée pour chaque tâche de présentation utilisée dans une règle ;
- un type a bien été appliqué à chaque tâche de présentation ;
- les conditions possèdent bien un opérateur et une valeur de comparaison ;
- un lien a bien été établi entre chaque unité informationnelle et une propriété connue du système.

Si une erreur est rencontrée, la génération ne peut pas se faire et sont affichées les règles, les conditions ou les tâches qui posent problème. Une fois la génération autorisée, l'action du bouton de simulation "Générer Sim." lance une plate-forme de simulation présentée dans la section suivante. Soulignons que l'ajout d'une règle nécessite une nouvelle génération. En effet, l'éditeur n'a pas été pensé pour être mis entre les mains d'utilisateurs souhaitant personnaliser le comportement d'un système. Il a été conçu pour faciliter la conception d'un composant de choix de stratégies de dialogue et de présentation dans le cadre d'expérimentations réalisées par des ergonomes/psychologues. Ces

expérimentations s'appuyant sur des hypothèses définies antérieurement et strictement cadrées, il n'est pas nécessaire de permettre une génération à la volée du composant de choix défini avec l'éditeur.

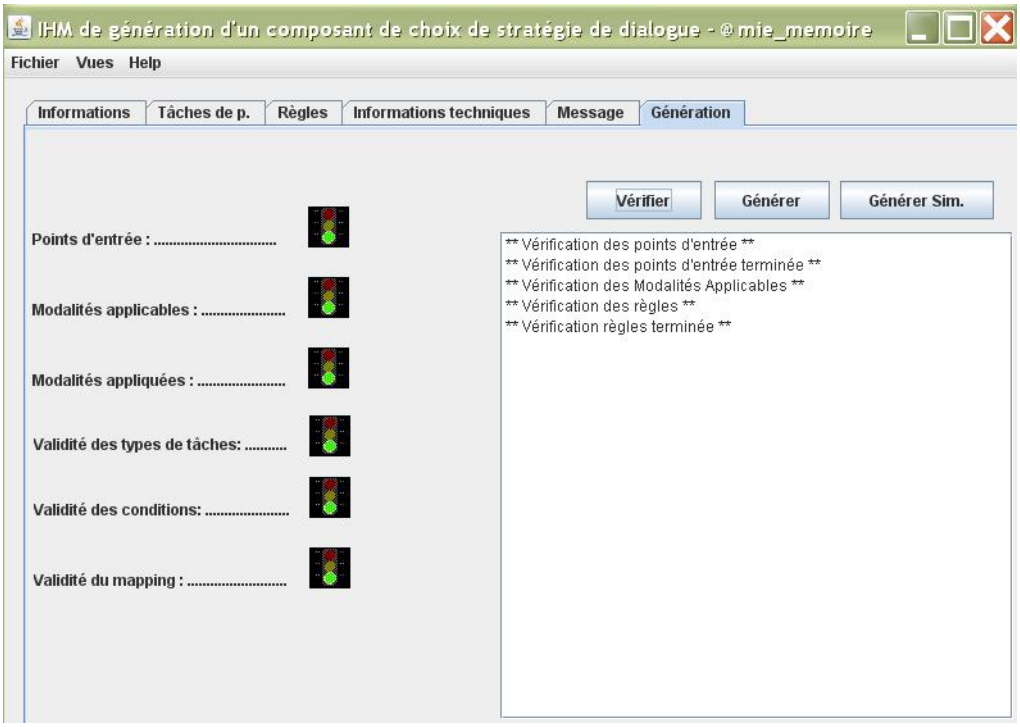

FIG. 7.6 – Génération du composant de choix de stratégies de dialogue et de présentation

La génération s'appuie directement sur l'analyse des règles pour produire un agent JSA. Conçues avec l'API JGraph, ces règles sont organisées sous forme de graphes qui peuvent être récupérés. Une classe a été conçue pour générer l'agent JSA correspondant à partir des graphes de règles actives. Elle permet l'écriture des chaînes de caractères qui correspondent au code de l'agent de choix. La génération utilise, outre le graphe des règles, un fichier XML qui recense les données nécessaires, telles que la liste des importations, le nom de la classe-mère, les constantes utilisées dans les règles (comme la constante NBMAXSOLVISUAL de la figure 7.5) et les noms des agents de simulation de l'entrée et de simulation de la sortie.

Les règles sont écrites sous forme de principes d'interprétation sémantique. Chaque règle correspond à un principe d'interprétation sémantique. Les conditions des règles sont traduites en conditions d'application interne au principe d'interprétation sémantique (*i.e.* portant sur les valeurs récupérées à partir du patron-type de la formule FIPA-SL définie dans l'onglet "Message"). Les tâches de présentation concrètes sont traduites en éléments de la formule FIPA-SL de sortie et leur synchronisation est directement récupérée du graphe de la règle considérée. Pour chaque type de tâche de présentation, une méthode spécifique a été implémentée pour permettre la récupération

des valeurs du message d'entrée prédéfini et la construction de la partie de la formule FIPA-SL à produire selon les principes du fonctionnement des agents JADE. Voici un exemple dans le cas d'une tâche de présentation de réponse-relaxation :

```
// Structure générale d'une tâche de présentation de réponse-relaxation
static final Term PATTERN_REPONSE_RELAXATION
= SL.fromTerm
("(reponse
:type ??type
:description ??description
:modalityReponse ??modality)");

// Construction des TP Reponse Relaxation
protected Term reponseRelaxationConstruction
(TermSequence description,String modality ){
 return PATTERN_REPONSE_RELAXATION
.instanciate("type",new WordConstantNode("relaxation"))
.instanciate("description",description)
.instanciate("modality",new WordConstantNode(modality));
```

Afin de pouvoir simuler le fonctionnement du composant de choix, qu'il soit développé ou généré, nous l'avons intégré dans une plate-forme de simulation selon l'architecture présentée dans le chapitre 5.

7.3 Plate-forme de simulation du composant de choix

Nous présentons une plate-forme de simulation qui respecte l'architecture proposée dans le chapitre 5 et permet de valider logiciellement le composant de choix en simulant les entrées envoyées par le composant de dialogue et la récupération des sorties envoyées au composant de présentation abstraite.

7.3.1 Architecture logicielle

L'architecture logicielle de la plate-forme de simulation, présentée dans la figure 7.7, comprend trois agents JSA et intègre une interface de simulation des entrées et des sorties du composant de choix.

Un premier agent se substitue aux étapes d'interprétation de la requête de l'utilisateur et du calcul de réaction réalisées, dans Arch, par les composants d'interaction et de présentation et par le composant de dialogue. Tout comme le ferait un composant de dialogue tel que défini dans le chapitre 5, cet agent de simulation de l'entrée envoie un message spécifiant les contenus possibles au deuxième agent, qui implémente le composant de choix proposé. La spécification de présentation produite en conséquence par le

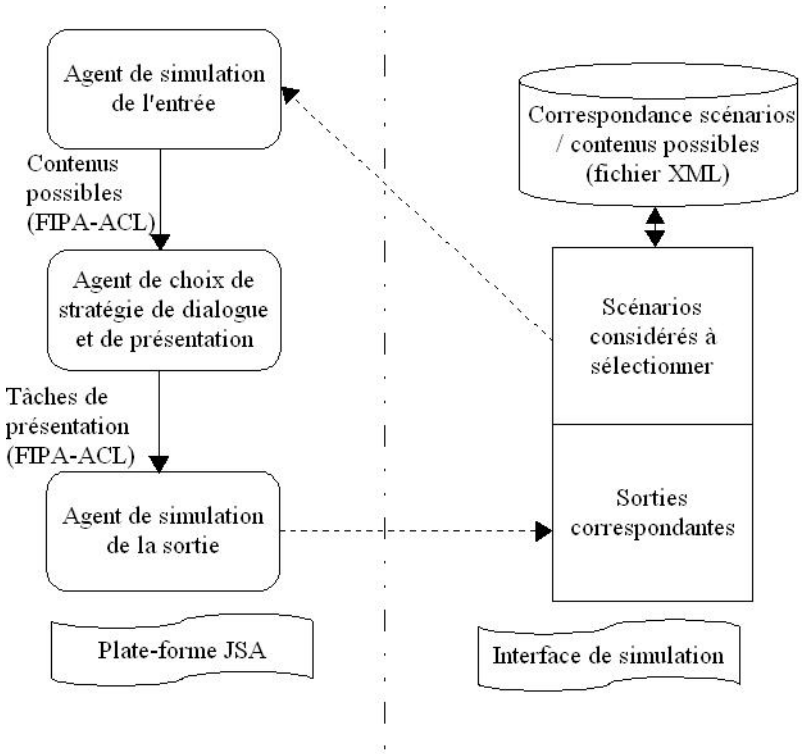

FIG. 7.7 – Architecture logicielle de la plate-forme de simulation

composant de choix est envoyée à un dernier agent qui se substitue aux composants de présentation. De façon à contrôler les entrées du composant de choix de stratégies de dialogue et de présentation d'une part et à observer ses sorties d'autre part, les agents d'entrée et de sortie partagent une interface graphique de simulation. Celle-ci permet, en entrée, de sélectionner un scénario-type dans une liste de scénarios prédéfinis. C'est cette sélection qui déclenche l'envoi au composant de choix des contenus possibles correspondants tels qu'il aurait été envoyé par le composant de dialogue. En sortie, l'interface de simulation permet d'observer le résultat des traitements du composant de choix sur les contenus possibles.

Le fonctionnement de cette architecture se fait de la façon suivante.

7.3.2 Fonctionnement

L'agent de simulation de l'entrée est un simple agent JADE, *i.e.* sans capacité sémantique particulière ni principes d'interprétation sémantique. Il est central dans l'architecture logicielle adoptée. De façon à simuler les entrées, l'interface permet de charger un ensemble de scénarios-types qui correspondent à des états de communication possibles dans le cas applicatif considéré. Ces états de communication précisent à la fois la requête de l'utilisateur et les éventuelles contraintes de présentation imposées par l'utilisateur. Ces scénarios-types sont sauvegardés dans un fichier XML où chaque scénario

est associé à une formule FIPA-SL qui correspond au contenu possible que produirait le composant de dialogue du système réel considéré dans l'état de communication indiqué. Cette formule a la structure suivante :

```
((choice (content
:answerMode _valeur_
:requestDescription (set
(description :_nom_propriete_ _valeur_ )
(description ... )
)
:requestFocus _nom_propriete_
:requestModality _valeur_
:restrictionSet (sequence _nom_propriete_ _nom_propriete_ ...)
:solutionSet (set (solution :_nom_propriete_ _valeur_ ...)(solution ... ))
:relaxationSet (set (relaxation :_nom_propriete_ _valeur_ ...)(restriction ...
))

)))
```

Le texte précédé de " :" correspond à des paramètres. Le texte qui suit immédiatement renvoie à la valeur de ce paramètre. Les paramètres "answerMode", "requestDescription", "requestFocus", "requestModality", "restrictionSet", "solutionSet" et "restrictionSet" sont génériques. "answerMode" renvoie à la contrainte de présentation de l'utilisateur ; "requestDescription" à la description de cette requête sous forme d'un ensemble de descriptions précisant chacune un ou plusieurs paramètres (*i.e.* propriétés caractérisant la ou les solutions cherchées), "requestFocus" au centre d'attention de l'utilisateur, *i.e.*, s'il y a lieu, à l'unité informationnelle explicitement demandée, "requestModality" à la modalité utilisée par l'utilisateur pour exprimer sa requête, "restrictionSet" à l'ensemble des critères permettant de restreindre la requête, "solutionSet" à l'ensemble des solutions à la requête (incluant tous les paramètres, *i.e.* les propriétés, renseignées), et "restrictionSet" à l'ensemble des solutions approchées à la requête de l'utilisateur lorsqu'aucune solution n'est trouvée à la requête stricte de l'utilisateur. L'expression "_nom_propriete_" fait référence à l'une des propriétés possibles susceptibles de caractériser une solution. Ces propriétés sont purement applicatives. Par exemple, dans le cas d'un annuaire, les prénoms, nom et numéros de téléphone sont des propriétés. L'expression "_valeur_" indique une valeur prise par le paramètre qui dépend directement de la requête de l'utilisateur, *i.e.*, dans notre cas, du scénario sélectionné.

Le fichier XML n'est donc pas générique et dépend du contexte applicatif considéré. La sélection d'un scénario-type au niveau de l'interface entraîne l'envoi d'un acte communicatif d'information de l'agent de simulation de l'entrée à l'agent de choix. Cet acte communicatif porte sur la formule associée au scénario-type en question dans le fichier XML.

L'agent de choix est un agent JSA qui remplit le rôle du composant de choix de stratégies de dialogue et de présentation décrit dans le chapitre 5. Grâce aux principes

d'interprétation sémantique génériques dont hérite cet agent, la formule FIPA-SL sur laquelle porte l'acte communicatif d'information reçu est admise comme une représentation sémantique qu'il croit vraie et qui déclenche l'application d'un principe d'interprétation sémantique applicatif. Les principes d'interprétation sémantique sont donc appliqués en fonction des unités informationnelles permettant de répondre à la requête de l'utilisateur et en fonction des contraintes de présentation de l'utilisateur. Toutes les conditions de choix de stratégies de dialogue et de présentation ne peuvent pas être prises en compte au niveau de la forme générale de représentation sémantique associé à un principe d'interprétation sémantique car toutes les distinctions n'ont pas une portée sémantique traitée par le langage FIPA-SL. Aussi, l'application d'un principe d'interprétation sémantique entraîne l'évaluation de conditions supplémentaires dans le cas de comparaison de valeurs, telles que le nombre de solutions. Un même principe d'interprétation sémantique peut donc conduire à des sorties différentes. En fonction du principe d'interprétation sémantique appliqué et des évaluations internes à ce principe d'interprétation sémantique, une spécification de présentation est produite, comprenant plusieurs tâches de présentation modalement allouées, est produit. Elle est ensuite envoyée aux autres agents, *i.e.* à l'agent de simulation de l'entrée (ce qui correspond au retour du composant de choix vers le composant de dialogue) et à l'agent de simulation de la sortie, sous la forme du contenu d'un acte communicatif de type "informer".

L'agent de simulation de la sortie est un agent JSA. Ses principes d'interprétation sémantique applicatifs analysent l'acte communicatif reçu de l'agent de choix. Si la représentation sémantique contenue correspond à la forme générale associée au principe d'interprétation sémantique considéré, celui-ci est appliqué, ce qui se manifeste par l'affichage dans l'interface de simulation des unités informationnelles contenues dans les tâches de présentation.

Comme le montre la figure 7.7, l'interface de simulation est découpée en deux parties. Une partie est destinée aux scénarios-types considérés et à leur sélection, et une autre partie permet de visualiser les unités informationnelles présentées pour le scénario sélectionné. Cette simulation de la sortie distingue les unités informationnelles présentées visuellement de celles présentées auditivement.

Nous avons utilisé cette plate-forme de simulation dans le cadre de deux systèmes-exemples. Nous décrivons les implémentations réalisées pour chacun d'eux dans la section suivante.

7.4 Exemples implémentés

Le premier système-exemple est le service Santiago, évoqué dans le chapitre précédent, initialement utilisé pour une expérimentation en magicien d'Oz permettant d'étudier l'impact des stratégies de présentation sur l'utilisateur et sur la communication. Cette expérimentation nous a servi de base pour faire la preuve que le composant de choix peut être utilisé durant le cycle de conception d'une application, dès les expérimen-

tations en Magicien d'Oz réalisées généralement en début de conception. Le deuxième système-exemple est le système @mie qui nous a servi d'exemple illustratif tout au long de ce mémoire. Pour chacun de ces systèmes, nous décrivons leur fonctionnement et leur utilisation pour la validation logicielle de nos travaux.

7.4.1 Santiago

Avant de proposer l'éditeur graphique de spécification du composant de choix, nous avons d'abord implémenté la plate-forme de simulation destinée à permettre des expérimentations en magicien d'Oz. Cette plate-forme a été motivée par le constat que, dans la plupart des cas, le matériel d'expérimentation développé par les expérimentateurs pour de tels tests les limitent dans leurs hypothèses sous peine d'être surchargés lors de l'expérience (*cf.* la section 6.2). De façon à donner une idée de la charge cognitive de l'expérimentateur, voici la façon dont se déroule généralement une expérimentation en Magicien d'Oz.

Les sujets sont placés face à une interface qu'ils utilisent pour réaliser un scénario prédéfini grâce à un système. Dans les faits, ce système est simulé par l'expérimentateur, dit magicien d'Oz. Il observe l'interaction du sujet et utilise les entrées produites par ce dernier pour déterminer l'état de la communication dans le scénario prédéfini. Le magicien indique alors cet état à un programme conçu pour l'expérimentation via une commande , commande qui déclenche la réalisation d'un pseudo-comportement au niveau de l'interface des sujets. Dans le cas de l'expérimentation réalisée pour Santiago, le programme de test consiste en des scripts de contrôle conçus avec Macromedia Director©. À titre d'exemple, le dispositif technique de cette expérimentation est présenté dans la figure 7.8.

Le magicien d'Oz remplit les fonctions des composants d'entrée et du composant de dialogue pour la partie interprétation du pseudo-système en test. La détermination de la réaction du pseudo-système, qui est généralement à la charge du composant de dialogue, ainsi que les fonctions des composants de présentation (concrète et abstraite) sont réalisées grâce à un programme de test créé pour l'application concernée. Cela signifie toutefois que le magicien doit (1) observer l'interaction du sujet, (2) l'interpréter et identifier l'état du dialogue (3) identifier la commande associée à cet état et (4) sélectionner cette commande au clavier.

De façon à faciliter la tâche du magicien d'Oz, nous avons décidé de lui proposer une interface de simulation de l'entrée qui lui permet d'identifier l'état de la communication pour les scénarios considérés plutôt que de devoir identifier la commande correspondant à l'état de la communication : ainsi, le magicien n'a-t-il plus à se soucier de retrouver la commande associée à l'état du dialogue qu'il a identifié. Cette interface simplifie donc la tâche du magicien d'Oz de façon à ce qu'il se concentre sur l'interprétation des interactions des sujets avec le pseudo-système. De plus, plutôt que cette interface génère des scripts programmés pour la circonstance, nous avons opté pour qu'elle envoie un message à un composant de choix de stratégies de dialogue et de présentation, comme le ferait un réel composant de dialogue. Le composant de choix a donc été associé à une interface de simulation et à des agents de simulation de l'entrée et de la sortie pour

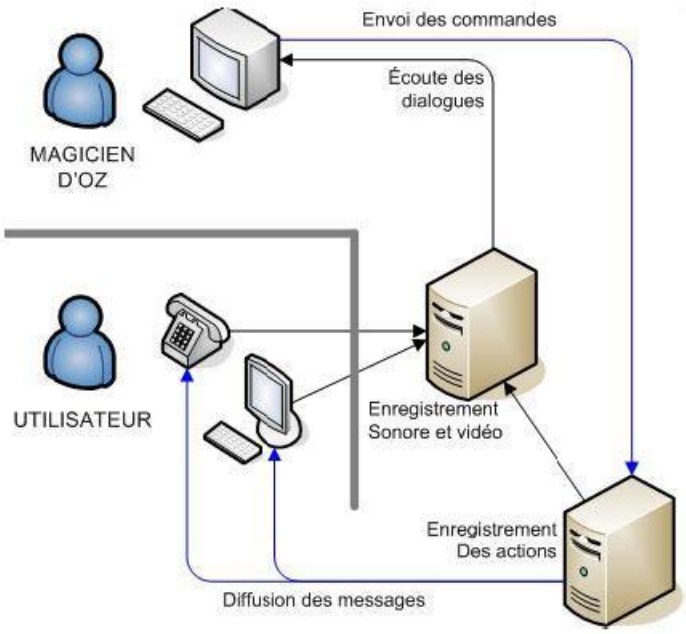

FIG. 7.8 – Un exemple de dispositif technique pour une expérimentation en magicien d'Oz (extrait de [Fréard, 2006])

former une plate-forme de simulation.

De façon à valider l'utilisation du composant de choix dans le cas d'une expérimentation en magicien d'Oz, nous avons décidé de reproduire le fonctionnement du programme de contrôle développé pour l'expérimentation du système Santiago (*cf.* la section 6.2), proposant en plus une interface de contrôle pour le magicien d'Oz. Nous avons donc développé une plate-forme de simulation destinée au magicien d'Oz qui comprend un composant de choix reprenant le fonctionnement des scripts dans l'expérimentation originelle.

Nous rappelons que ce système permet la prise de rendez-vous avec un médecin. La réponse du système comprend un feedback, une réponse et une relance (*cf.* la figure 6.1). Les modalités considérées étant le langage naturel écrit perceptible visuellement et le langage naturel oral perceptible auditivement, quatre configurations sont donc testées (*cf.* la figure 6.2) :

- deux configurations monomodales :
  - une configuration monomodale auditive (dite "AAA" car feedback, réponse et relance sont présentés auditivement) ;
  - une configuration monomodale visuelle (dite "VVV" car feedback, réponse et relance sont présentées visuellement) ;
- deux configurations multimodales complémentaires :
  - une configuration multimodale où les tâches de présentation à fonction interactive - *i.e.* feedback et relance - sont présentées visuellement et la tâche de présentation à fonction applicative qu'est la réponse est présentée auditivement

(configuration dite "VAV") ;

- la configuration multimodale inverse avec les tâches de présentation à fonction interactive présentées auditivement et la tâche de présentation à fonction applicative présentée visuellement (configuration dite "AVA").

Plusieurs scénarios ont été imaginés pour l'expérimentation initiale. Dans chacun de ces scénarios, trois étapes du dialogue sont distinguées, l'étape de la formulation de la requête, l'étape de présentation des solutions trouvées et l'étape de confirmation de la réservation. Dans le cadre de l'implémentation du composant de choix et de la plate-forme de simulation, nous nous sommes concentrés sur l'étape de la présentation des solutions, *i.e.* où le médecin et l'horaire de rendez-vous souhaités ont déjà été précisés mais où le sujet n'a pas encore fait son choix. Chaque scénario présenté au niveau de l'interface de simulation correspond donc à une requête complète identifiée par le magicien d'Oz. C'est ce dernier qui utilise l'interface pour sélectionner l'état de la communication, en l'occurrence le scénario que le sujet est en train de réaliser.

Comme le montre la figure 7.9 qui présente l'interface de simulation, sept scénarios sont distingués. La sélection d'un scénario entraîne l'apparition d'une boîte de dialogue, présenté à la figure 7.9. Celle-ci permet le choix de la configuration d'allocation des modalités appliquées. C'est la sélection d'une configuration qui provoque l'envoi d'un acte communicatif d'information par l'agent de simulation de l'entrée à l'agent de choix. Le contenu de cet acte est récupéré dans un fichier XML qui met en correspondance les scénarios considérés et les formules décrivant les contenus possibles correspondantes. Voici un exemple de formule dans le cas du système-exemple Santiago, qui correspond au scénario sélectionné dans la figure 7.9 :

```
"
(choice (content
:requestDescription (set (description :nomDocteur "dubois" :date "06/09"))
:relaxationSet (set
(relaxation :jour "samedi" :date "03/09" :heure "11h30")
(relaxation :jour "lundi" :date "05/09" :heure "9h30")
(relaxation :jour "lundi" :date "05/09" :heure "11h30")
(relaxation :jour "lundi" :date "05/09" :heure "18h00")
(relaxation :jour "mardi" :date "06/09" :heure "11h00")
)
))"
```

Le comportement de l'agent de choix s'appuie sur l'acte communicatif reçu. Les scénarions testés étant connus, les actes communicatifs susceptibles d'être reçus sont parfaitement définis. Nous avons donc implémenté un principe d'interprétation sémantique pour chaque acte communicatif possible. Le composant de choix fonctionne ici comme une table de correspondance, les stratégies de dialogue et de présentation n'ayant pas été choisies en fonction des contenus possibles ou d'éventuelles contraintes de présentation. Nous rappelons que nous ne faisons que reproduire l'expérimentation réalisée sur Santiago pour faire la preuve qu'il est possible d'utiliser la plate-forme de simula-

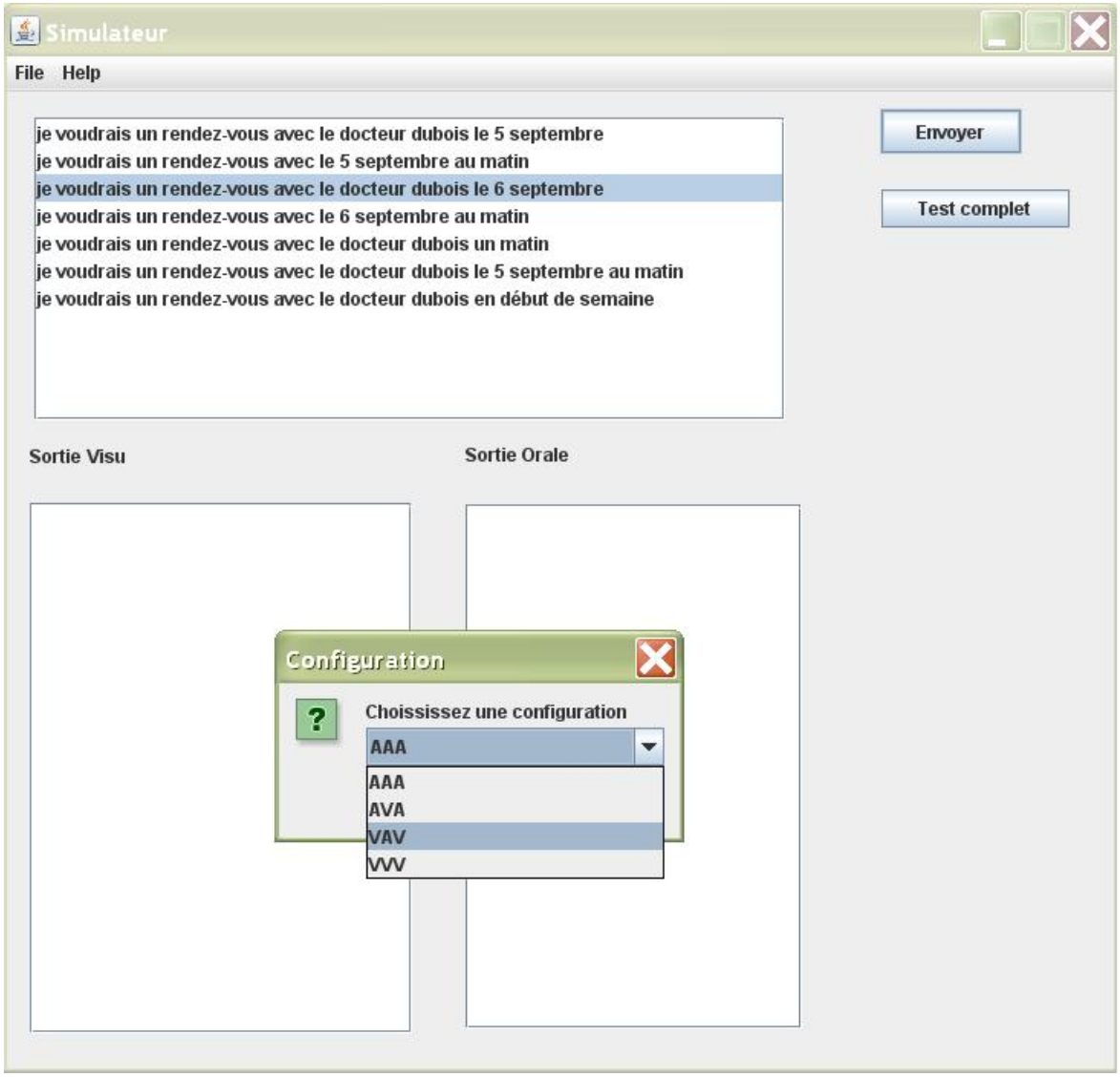

FIG. 7.9 – Exemple de simulation dans le cas de l'expérimentation pour le système-exemple Santiago

tion intégrant un composant de choix dès le début du cycle de développement pour des expérimentations en magicien d'Oz.

L'application d'un principe d'interprétation sémantique, *i.e.* l'identifiation d'une représentation sémantique qui correspond à un scénario donné, produit une spécification de présentation qui respecte la structure suivante :

```
"(answer (sequence ??feedback ??reponse ??relance))"
```

Le feedback (" ? ?feedback"), la réponse (" ? ?reponse") et la relance (" ? ?relance")

sont construits en fonction de la représentation sémantique récupérée dans le message d'entrée. Le problème du choix de la stratégie de dialogue ne se pose pas, étant donné que les scénarios possibles sont parfaitement définis pour les besoins de l'expérimentation. La principale contrainte, outre les solutions possibles à la requête de l'utilisateur, réside dans l'allocation des modalités. Celles-ci sont allouées en respectant la configuration testée. Pour cela, une constante pour la modalité de chaque tâche de présentation est créée et une valeur lui est attribuée en fonction de la configuration sélectionnée. Une fois la spécification de présentation construite, elle est envoyée à l'agent de simulation de la sortie. Dans l'exemple de la figure 7.9, la formule produite est la suivante :

```
"(answer (sequence
(requestReminder
:requestDescription (set (description :nomDocteur "dubois" :date "06/09"))
:modality HYPERTEXTEMOD
)
(relaxation :relaxationSet (set
(relaxation :jour "samedi" :date "03/09" :heure "11h30")
(relaxation :jour "lundi" :date "05/09" :heure "9h30")
(relaxation :jour "lundi" :date "05/09" :heure "11h30")
(relaxation :jour "lundi" :date "05/09" :heure "18h00")
(relaxation :jour "mardi" :date "06/09" :heure "11h00")
) :modality ORALMOD)
(invit :type precision :modality HYPERTEXTEMOD)
))"
```

En fonction de la modalité appliquée, l'agent de simulation de la sortie va afficher les unités informationnelles présentées dans la partie visuelle ou dans la partie auditive. Le magicien d'Oz a ainsi la spécification de présentation correspondant à la réponse à présenter au sujet (qui pourrait se faire grâce à un agent dédié), ce qui n'était pas le cas dans la version initiale de l'expérimentation du service Santiago. L'expérimentateur a donc accès au contexte du sujet et n'a pas à s'en souvenir, ni à se souvenir des commandes qui correspondent à l'état de la communication.

Cette implémentation avait pour but de permettre des expérimentations en magicien d'Oz avec une plate-forme intégrant un composant de choix. Ainsi celui-ci peut-il être utilisé dès le début du cycle de conception. Pour vérifier la validité de notre proposition, nous avons reproduit le matériel expérimental utilisé pour l'expérimentation sur le service Santiago décrite dans la section 6.2. Cette implémentation a rempli ses objectifs a été à l'origine de l'identification de l'intérêt de l'éditeur graphique introduit dans le chapitre 6. L'adaptation de cette plate-forme de simulation pour de potentiels systèmes à tester passe (1) par la définition du fichier XML utilisé pour le simulateur en entrée et (2) par l'implémentation du composant de choix pour le système considéré.

L'essentiel de nos réalisations logicielles a porté sur le système @mie. Nous les détaillons dans la section suivante.

7.4.2 @mie

7.4.2.1 Fonctionnement du système initial

Pour rappel, @mie est un annuaire multimodal intelligent d'entreprise, qui permet aux employés de chercher des informations sur leurs collègues (par exemple, leurs prénoms, noms, fonctions, adresses courriels, numéros de téléphone et de bureau, sites, équipes, etc.) mais aussi sur les équipes (notamment sur leurs numéros de fax, leurs acronymes, leurs noms, les descriptions de leurs activités, etc.) et sur les sites de l'entreprise (en particulier, les villes, pays, plans d'accès, etc.). Ce prototype est dit intelligent car il est développé selon les principes de la technologie ARTIMIS qui, s'appuyant sur la théorie de l'interaction rationnelle proposée par Sadek [Sadek, 1999], permet la conception de systèmes en tant qu'agents rationnels dialoguants qui communiquent naturellement en étant guidés par leurs croyances, leurs buts et leurs intentions.

En tant que système basé sur ARTIMIS, le prototype @mie déduit, à partir d'une interprétation souple du message d'entrée de l'utilisateur en langage naturel oral ou écrit, l'état mental de ce dernier et, en particulier, ses buts. Dans une démarche coopérative, le système va chercher à satisfaire les buts de l'utilisateur. La détermination de la réaction du système se fait par inférence sur l'état mental de l'utilisateur. L'axiomatique de ce mécanisme d'inférence est en partie générique et en partie propre à un système donné. Pour la partie générique, elle met en œuvre les caractéristiques de convivialité présentées dans le chapitre 3. Pour la partie applicative, elle détermine le bon déroulement de la communication, gère l'accès à l'équivalent des composants du domaine (appelé "unité de gestion des connaissances" dans la figure 4.5) et la sélection de la réaction à adopter et des contenus à présenter, *i.e.* de la stratégie de dialogue. L'inférence sur l'état mental de l'utilisateur conduit à la production d'actes communicatifs qui constituent la réaction du système. Ces actes communicatifs suivent le formalisme FIPA-ACL, le langage de communication entre agents (à l'origine du sigle ACL *Agent Communication Language*) standard adopté par le consortium de standardisation FIPA (pour *Foundation for Intelligent Physical Agents*) [FIPA, 2002a]. Ils sont indépendants des modalités.

La concrétisation de ces actes communicatifs en messages perceptibles, compréhensibles par l'utilisateur, se fait en deux temps : d'abord, la transformation des actes communicatifs produits par le cœur du système (*i.e.* l'équivalent du composant de dialogue appelé "unité rationnelle" dans figure 4.5, page 111) en messages compréhensibles par les dispositifs physiques, puis la concrétisation de ces messages par ces dispositifs. Dans les systèmes basés sur ARTIMIS en général et dans le prototype @mie en particulier, la stratégie de présentation est déterminée à la conception. Par conséquent, l'utilisateur ne peut pas l'influencer, ce qui va à l'encontre de l'accessibilité sensoriactionnelle. Par exemple, le comportement qui consiste à présenter le nombre de solutions via un message auditif oral et à afficher la liste des solutions sur l'écran (exemple de gauche de la figure 1, page 3) est la stratégie de présentation standard lorsque le système identifie plus d'une solution à la requête de l'utilisateur dans le cas où cette dernière porte sur une personne ou sur des propriétés de cette personne (comme son prénom, son nom, son numéro de téléphone, etc.). L'utilisateur qui ne peut pas regarder son écran n'a pas

accès aux informations présentées sur ce dispositif physique. Il doit réussir à préciser suffisamment sa requête, sans aide, sans coopération de la part du système, pour que les informations recherchées puissent être présentées de façon auditive. Il peut même échouer dans sa tâche si les critères de recherche qu'il connaît ne sont pas assez précis.

Alors que la réaction d'@mie est coopérative, la présentation de cette réaction est complètement figée. Notre proposition d'un composant de choix de stratégies de dialogue et de présentation vise justement à rendre la présentation plus souple, en la choisissant conjointement avec la réaction de façon à assurer les accessibilités sensoriactionnelle, cognitive et rhétorique des informations et des capacités d'action proposées à l'utilisateur.

Pour mettre au point le composant de choix de stratégies de dialogue et présentation dans @mie, nous avons utilisé notre plate-forme de simulation. Nous présentons l'implémentation de l'agent correspondant au composant de choix dans le paragraphe suivant. Le composant mis au point et testé, l'étape suivante à ce travail consisterait à intégrer directement le composant de choix avec ARTEMIS, afin d'intégrer le composant dans le système @mie. Cette intégration nécessiterait d'adapter les actes communicatifs fournis par ARTEMIS (le contrôleur de dialogue) au composant de choix afin qu'il fournisse non pas les contenus choisis, mais les contenus possibles.

7.4.2.2 Implémentation du composant de choix

Nous avons implémenté le composant de choix dans le cas des conditions prises en compte pour l'exemple illustratif du chapitre 5. Cet exemple tient compte du nombre de solutions et de la contrainte de présentation imposée par l'utilisateur dans le cas où il y plus d'une solution à la requête. Laissant de côté les éventuelles étapes de méta-dialogue (accueil et aide notamment), nous avons considéré 33 scénarios possibles, 11 avec une contrainte de présentation visuelle, 11 avec une contrainte de présentation auditive et 11 sans contrainte de présentation. Pour chaque contrainte de présentation possible, les scénarios possibles concernent une requête portant sur "Carole" ou sur une de ses propriétés. Nous nous sommes restreints aux propriétés sur lesquelles porte couramment la requête de l'utilisateur, à savoir : le numéro de téléphone, le numéro de mobile, la photo, la localisation (*i.e.* sur quel site), le bureau, l'adresse courriel, le numéro de fax, la fonction et l'équipe (à deux niveaux hiérarchiques différents que sont l'urd et le crd). Les autres propriétés prises en compte et utilisées dans le message d'entrée sont le prénom, le nom et le laboratoire (un troisième niveau hiérarchique pour situer l'équipe). Elles sont généralement utilisées pour décrire la requête de l'utilisateur, *i.e.* comme valeur des paramètres "requestDescription" et "requestFocus".

Comme le montre la figure 7.10, l'interface de simulation permet de charger tous les scénarios ou uniquement les scénarios incluant une contrainte de présentation. Ces scénarios sont listés dans un fichier XML. Celui-ci fait le lien entre un scénario et la formule FIPA-SL correspondant aux contenus possibles produits qu'aurait produit le composant de dialogue dans le système @mie initial. L'annexe 3 présente une partie de ce fichier XML.

La sélection et la validation d'un scénario entraîne l'envoi d'un acte communicatif

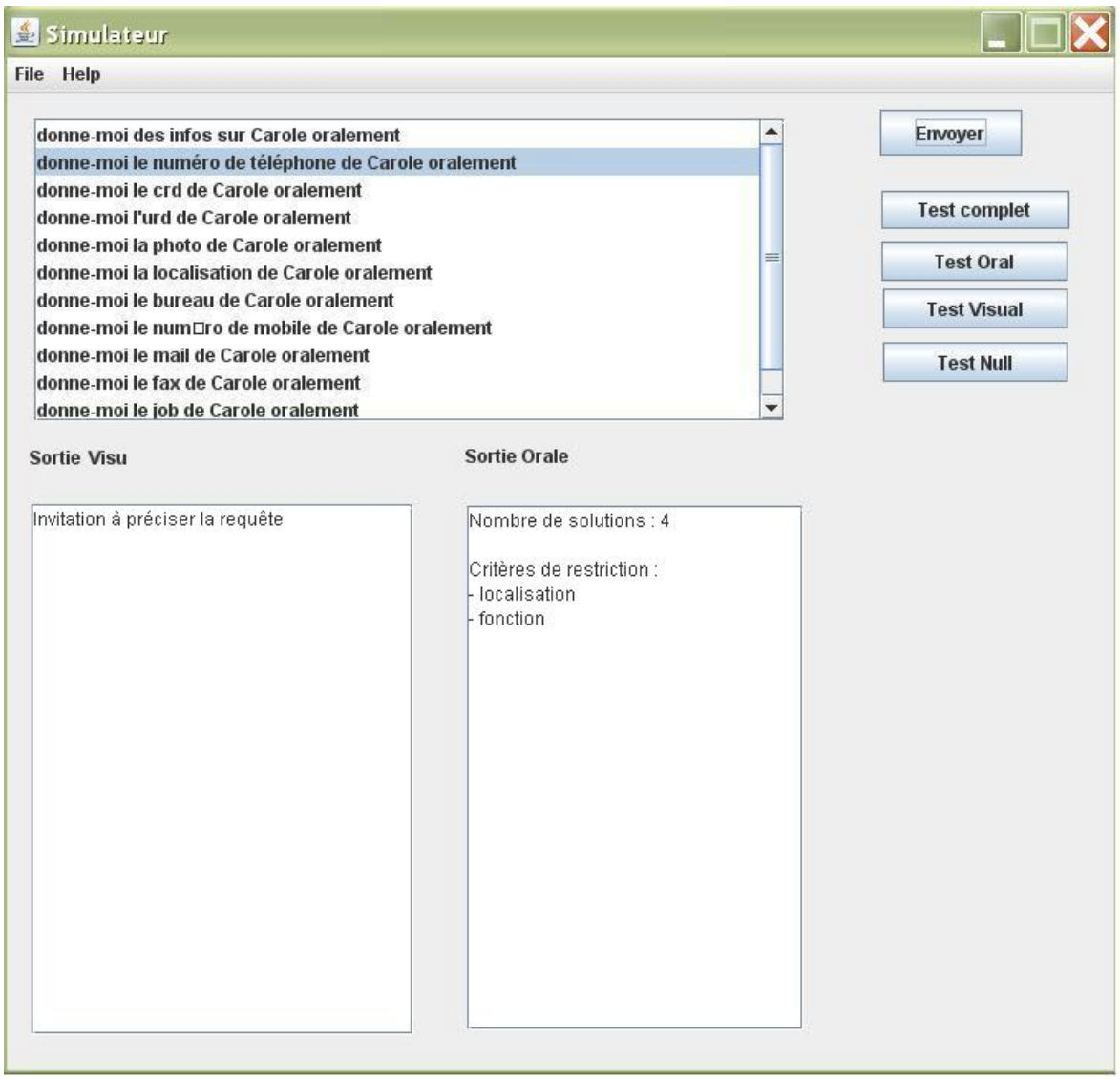

FIG. 7.10 – Exemple de l'interface simulation pour le système-exemple @mie

d'information de l'agent de simulation de l'entrée à l'agent de choix. Le contenu de cet acte communicatif est le message associé au scénario sélectionné dans le fichier XML. La réception de cet acte communicatif d'information entraîne, dans un premier temps, la transformation de son contenu en croyance de l'agent de choix par un principe d'interprétation sémantique générique ; dans un second temps, ce contenu est comparé, pour chaque principe d'interprétation sémantique applicatif, à la forme générale de représentation sémantique associée. Pour cette première version de la réalisation logicielle du composant de choix, nous avons choisi de faire un seul principe d'interprétation sémantique applicatif valable pour tout acte communicatif impliquant un choix et dont

la forme générale de représentation sémantique est la suivante :

```
(B ??myself (choice ??content))
```

Notre principe d'interprétation sémantique applicatif s'applique donc sur les représentations sémantiques qui implique une croyance ("B") de l'agent (identifié par "??myself") sur une formule qui est identifiée par un symbole fonctionnel de choix ("choice") et par une expression à identifier ("??content"). Les expressions précédées par un double point d'interrogation peuvent être récupérées et décomposées grâce à des patrons. Dans l'agent de choix, les règles de choix portent sur les valeurs de paramètres extraites grâce à ces patrons.

Le patron-type utilisé dans le cadre du système @mie est le suivant :

```
"(content
:requestFocus ??requestFocusInput
:requestDescription ??requestDescriptionInput
:requestModality ??requestModalityInput
:answerMode ??answerModeInput
:solutionSet ??solutionSetInput
:restrictionSet ??restrictionSequenceInput)"
```

La récupération des valeurs qui correspondent à la requête de l'utilisateur et aux contenus possibles de réponse se fait de la façon suivante :

```
// identification du patron de "content"
Term contentP = SL.fromTerm("(content
:requestFocus ??requestFocusInput
:requestDescription ??requestDescriptionInput
:requestModality ??requestModalityInput
:answerMode ??answerModeInput
:solutionSet ??solutionSetInput
:restrictionSet ??restrictionSequenceInput)").getSimplifiedTerm();

MatchResult contentMatch = contentP.match(applyResult.term("content"));

if ( contentMatch != null ) {
// récupération des éléments du pattern "content"
Constant answerModeInput =
(Constant) contentMatch.term("answerModeInput");
TermSet solutionSetInput =
(TermSet) contentMatch.term("solutionSetInput");
Constant requestFocusInput =
(Constant) contentMatch.term("requestFocusInput");
TermSet requestDescriptionInput =
(TermSet) contentMatch.term("requestDescriptionInput");
```

```
TermSequence restrictionSequenceInput =
(TermSequence) contentMatch.term("restrictionSequenceInput");

// règles de choix à compléter
}
```

Les expressions entre parenthèses qui précèdent l'application de la méthode "term" dépendent du type de données que l'on cherche à récupérer. Ce type a un effet direct sur les actions possibles sur les données. Par exemple, une méthode qui s'applique sur les ensembles (qui sont déclarés en "TermSet") permet de récupérer le nombre d'éléments dans cet ensemble : elle est utilisée sur la donnée "solutionSetInput" pour récupérer le nombre de solutions trouvées à la requête de l'utilisateur. C'est l'une des conditions de règles prises en compte dans l'exemple @mie. L'autre condition porte sur la contrainte de présentation imposée par l'utilisateur, *i.e.* sur la valeur de "answerModeInput". Trois valeurs sont possibles : "visual" dans le cas d'une contrainte de présentation visuelle, "aural" dans le cas d'une contrainte de présentation auditive ou "null" dans le cas d'absence de contrainte de présentation. L'évaluation de la valeur de la contrainte de comparaison est faite de la façon suivante :

```
// contrainte de présentation auditive ?
if ( answerModeInput.stringValue().equals("aural") ){
// construction du message de sortie correspondant
}
```

Lorsque les conditions d'application d'une règle sont réunies, *i.e.* que le nombre de solutions et que la contrainte de présentation sont identifiées, la spécification de présentation correspondante est construite. Par exemple, dans le cas du scénario sélectionné dans la figure 7.10, sachant qu'il y a quatre "Carole" et que le nombre Y limite pour énumérer les solutions oralement est fixé à 4, la règle implémentée est la suivante :

si le nombre de solutions est supérieur à 3 ET si la modalité de réponse est auditive, alors appliquer le comportement suivant (présenter le nombre de solutions auditivement, présenter la liste des critères de restriction auditivement, inviter à préciser la requête visuellement)

Comme présenté dans l'exemple du chapitre 5, une spécification de présentation est composée de trois tâches de présentation. Elle est donc une formule composée d'une séquence de trois tâches de présentation de la forme suivante :

```
"(answer (sequence ??answerPart1 ??answerPart2 ??answerPart3))"
```

Chaque tâche de présentation est complétée grâce à des méthodes spécifiques au type de tâche de présentation. Dans l'exemple de la figure 7.10, il y a une méthode pour la tâche de présentation du nombre de solutions, une pour la tâche de présentation de la liste des critères de restriction et une pour la tâche de présentation d'invitation à préciser la requête. Au final, la spécification de présentation obtenue et envoyée à l'agent de simulation de la sortie en tant qu'acte communicatif d'information est la suivante :

```
"(answer
(sequence
(info :nbSol "4" :modality oral)
(restriction :restrictionSet (sequence  localisation job) :modality oral)
(invit :type precision :modality hypertext)
)
)"
```

L'agent de simulation de la sortie est aussi un agent JSA. Il adopte donc la formule de l'acte communicatif reçu comme une croyance. Plusieurs principes d'interprétation sémantique applicatifs sont utilisés, un pour chaque type de tâche de présentation modalement allouée. Il y a donc deux principes d'interprétation sémantique différents pour la tâche de présentation "présenter la liste des solutions visuellement" et la tâche de présentation "présenter la liste des solutions auditivement". Ce choix a été fait pour anticiper le couplage à venir de l'agent de choix avec des composants de présentation effectifs qui permettent la concrétisation des tâches de présentation. Toutefois, tous les principes d'interprétation sémantique ont la même forme générale de représentation sémantique qui correspond à l'identification d'une réponse à concrétiser grâce au symbole fonctionnel "answer". L'application d'un principe d'interprétation sémantique entraîne l'affichage, dans la partie sortie, *i.e.* la partie basse de l'interface de simulation, des unités informationnelles en fonction de leur présentation visuelle (champ de texte "Sortie Visu" de la figure 7.7) ou auditive (champ de texte "Sortie Orale" de la figure 7.7).

L'interface de simulation a, dans les faits, été introduite après une première implémentation de l'agent de choix. Elle a été motivée, initialement, par la volonté de valider l'utilisation du composant de choix pour des expérimentations en magicien d'Oz. Pour cela, nous nous sommes appuyés sur des expérimentations faites dans le cadre d'un autre système, le service Santiago.

7.4.2.3 Composant de choix spécifié avec l'éditeur graphique : @mie

L'utilisation de l'interface graphique de génération par un ergonome dans le cadre d'une expérimentation sur le choix des stratégies de dialogue et de présentation est en cours. De façon à permettre l'expérimentation, l'interface de simulation est en train d'être complétée de façon à simuler la sortie sur un émulateur de téléphone portable. Avant d'entamer cette validation logicielle, nous avons effectué une validation dans le cadre du système-exemple @mie. Nous nous appuyons sur les règles de conception considérées dans l'exemple illustratif du chapitre 5.

L'utilisateur concepteur commence par définir les unités informationnelles nécessaires dans l'onglet "Informations". En plus des quatre unités informationnelles prédéfinies (les quatre premières de la liste dans la figure 7.3, il crée les unités informationnelles qui permettent de caractériser un employé, à savoir son prénom, son nom, son numéro de téléphone fixe, son numéro de téléphone mobile, son numéro de fax, son équipe (caractérisée par une "urd", un "crd" et un laboratoire), sa photo et le bureau. Pour chacune de ces informations, il détermine au minimum le nom et la source.

L'informaticien intervient alors pour déterminer le type de chacune de ces unités informationnelles du point de vue système, ainsi que le nom de la propriété à laquelle elle renvoie. Le prénom, le nom, l'équipe, la photo et le bureau sont des chaînes de caractère, alors que le numéro de téléphone fixe, le numéro de téléphone mobile et le numéro de fax sont numériques. De plus, le prénom est associée à la propriété "firstname", le nom à la propriété "lastname", le numéro de téléphone fixe à la propriété "phone", le numéro de téléphone mobile à la propriété "mobilePhone", etc.

L'informaticien peut à présent formater la formule FIPA-SL que pourra recevoir l'agent de choix dans l'onglet "Message". Comme le montre la figure 7.11, sont concernés par ce formatage les paramètres "requestFocus", "requestDescription", "solutionSet", "relaxationSeq" et "restrictionSeq". Par exemple, le paramètre "requestDescription" (qui correspond à la description de la requête de l'utilisateur) peut être défini par le prénom ("firstname"), le nom ("lastname"), l'équipe ("urd", "crd" et "labo") et le bureau ("bureau"). L'informaticien peut contrôler la structure générale du message d'entrée du composant de choix dans le même onglet.

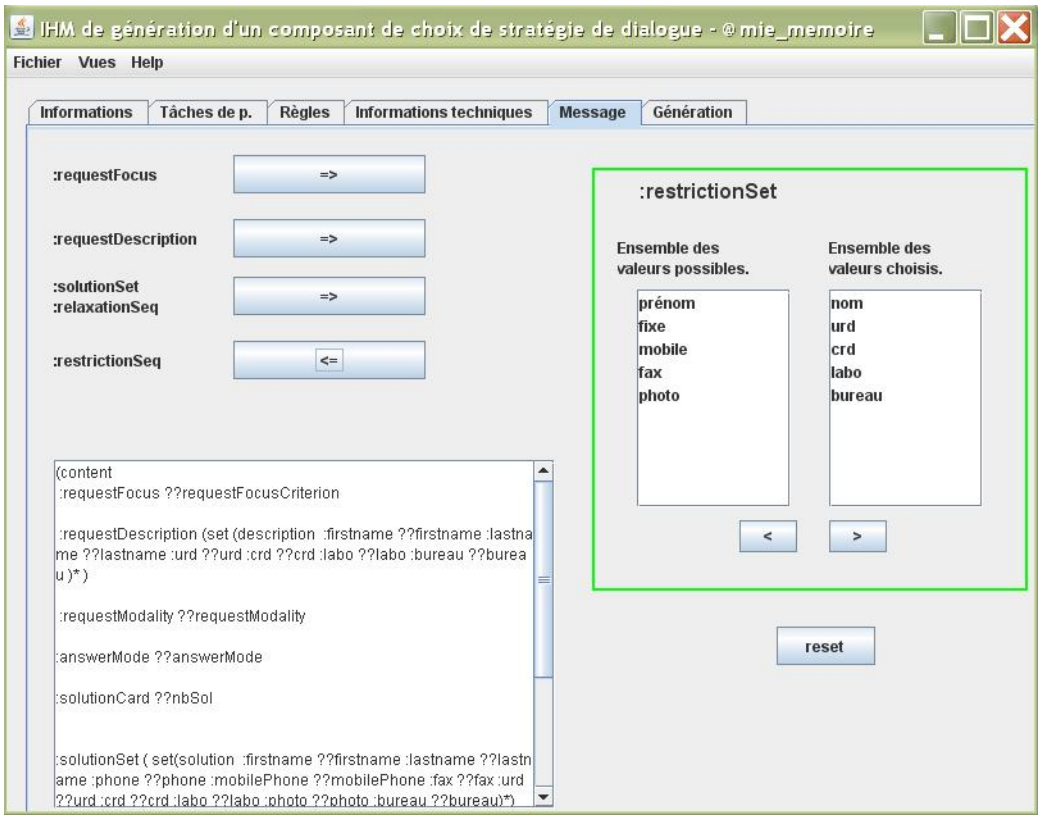

FIG. 7.11 – Formatage du message en entrée du composant de choix généré

L'utilisateur-concepteur peut alors définir les tâches de présentation abstraites dont il a besoin dans l'onglet "Tâches de p.". Pour l'exemple considéré, il définit quatre tâches de présentation dont les modalités applicables sont les modalités visuelle et auditive :

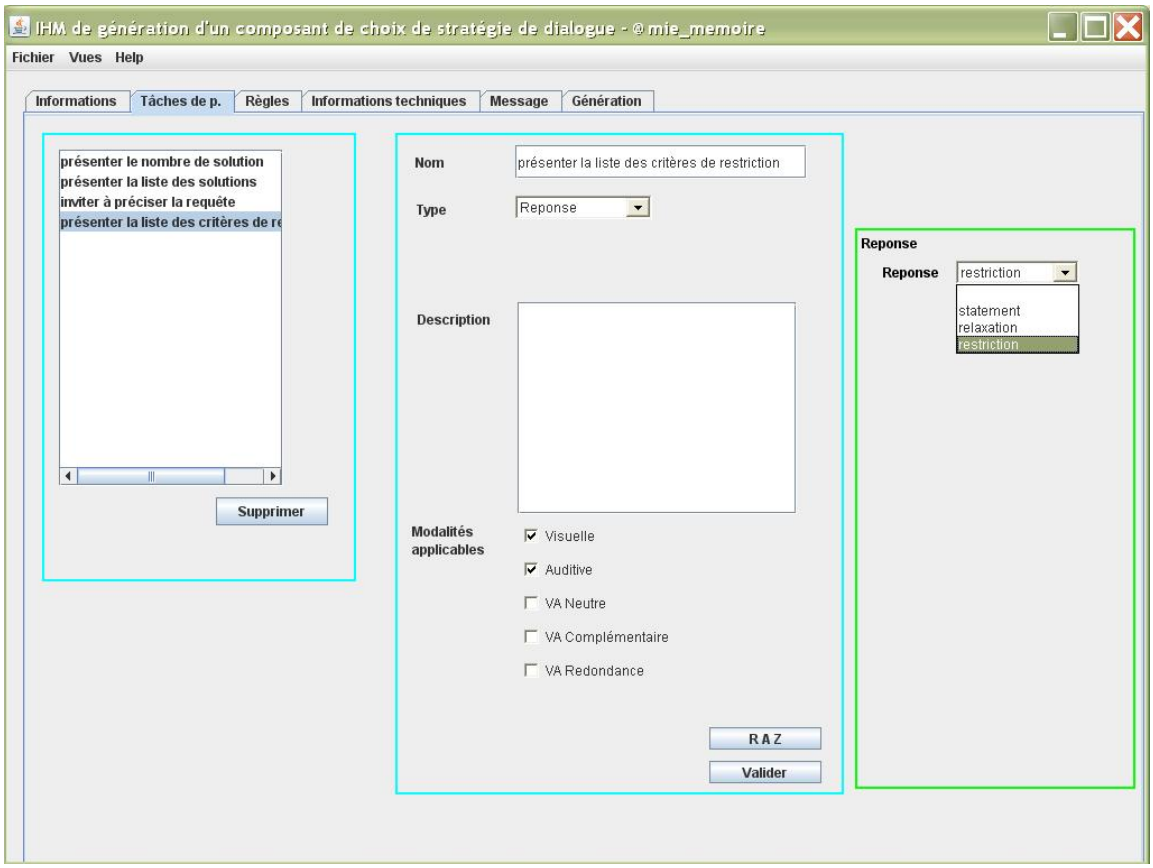

FIG. 7.12 – Définition de la tâche de présentation abstraite "présenter la liste des critères de restriction"

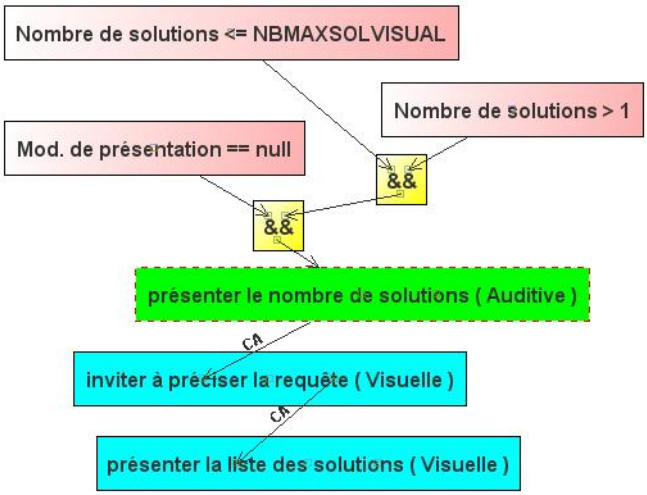

FIG. 7.13 – Règle dans le cas d'absence de contrainte de présentation et un nombre de solutions inférieur à un maximum

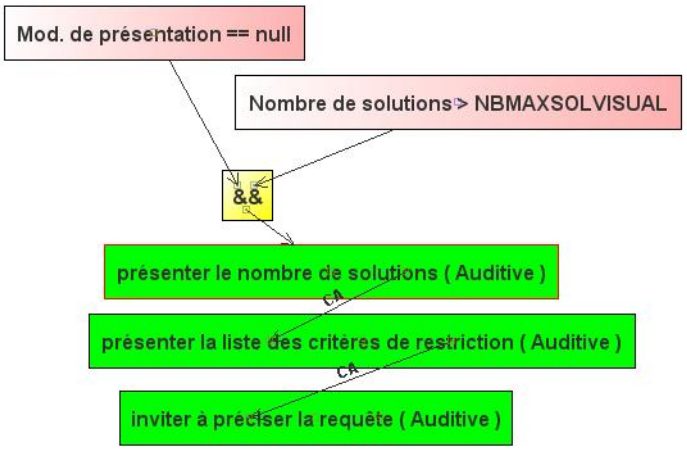

FIG. 7.14 – Règle dans le cas d'absence de contrainte de présentation et un nombre de solutions trop important

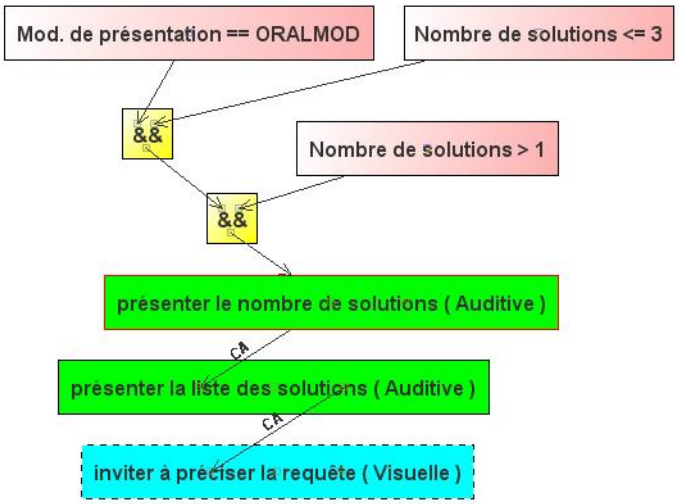

FIG. 7.15 – Règle dans le cas d'une contrainte de présentation auditive et un nombre de solutions inférieur à un maximum

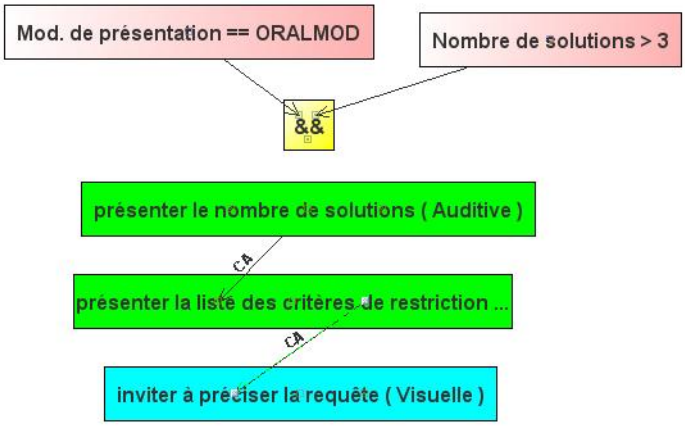

FIG. 7.16 – Règle dans le cas d'une contrainte de présentation auditive et un nombre de solutions trop important

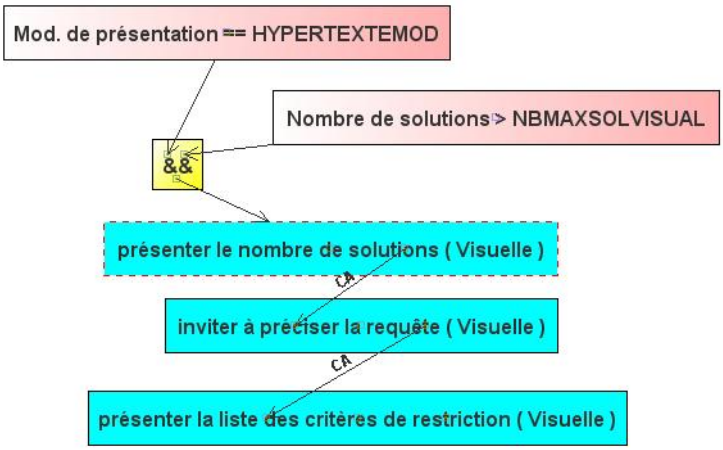

FIG. 7.17 – Règle dans le cas d'une contrainte de présentation visuelle et un nombre de solutions trop important

"présenter le nombre de solutions", "présenter la liste des solutions", "inviter à préciser la requête" et "présenter la liste des critères de restriction". Comme le montre les figures 7.4 et 7.12, la caractérisation des tâches de présentation dépend du type de tâche de présentation. La spécification du nom de la tâche de présentation, des modalités applicables et du type est à définir par l'utiliateur-concepteur. L'informaticien intervient éventuellement pour caractériser plus finement les tâches de présentation et vérifier qu'elles sont complètement définies.

L'utilisateur-concepteur peut alors spécifier les dix règles définies pour l'exemple. La figure 7.5 en présente une (nombre de solutions limité et contrainte de présentation visuelle) et les figures 7.13, 7.14, 7.15, 7.16 et 7.17 présentent les autres.

La génération du composant de choix est alors possible. Un fichier Java est alors obtenu, qui correspond à un composant de choix implémenté sous forme d'un agent JADE-JSA. Une partie de ce fichier est reproduite dans l'annexe 4. À n'importe quel moment, l'utilisateur-concepteur peut définir de nouvelles unités informationnelles, de nouvelles tâches de présentation ou de nouvelles règles et générer un nouveau composant de choix.

Nous avons présenté une implémentation de l'agent correspondant au composant de choix dans le système @mie et l'avons testé dans notre plate-forme de simulation. Nous avons également décrit la spécification du composant de choix dans le cas d'@mie avec l'éditeur graphique. Mais cette spécification a été réalisé par nos soins. Une validation de l'intérêt de l'éditeur graphique par un ergonome est en cours, dont nous évoquons les premiers éléments avant de conclure ce chapitre.

7.4.3 Premiers retours sur l'utilisation de l'éditeur graphique par un ergonome

La validation de l'éditeur graphique passe par son utilisation par un ergonome pour spécifier un composant de choix. Cette validation est en cours dans le cadre de l'élaboration d'une expérimentation en magicien d'Oz qui poursuit l'expérimentation réalisée sur Santiago (*cf.* la section 6.2). L'objectif de cette expérimentation est de préciser un peu plus l'impact des choix de stratégies de présentation sur l'utilisateur et sur la communication. Considérant cette fois la recherche de film, l'ergonome cherche à évaluer la combinaison du langage naturel écrit hypertextuel visuel et du langage naturel oral auditif en modifiant le style des présentations visuelles. Plus précisément, celles-ci peuvent être minimales ou plus développées. L'utilisation de l'éditeur graphique pour cette expérimentation montre que le composant de choix peut permettre un choix fin des stratégies de présentation même si le niveau d'abstraction de la production du comportement du système est limité.

Les premiers retours de la part de l'ergonome sont que la formalisation de ses choix de stratégie de présentation n'est pas toujours aisée. La première raison à cela est que l'ergonome a du mal à identifier la règle qui généralise un comportement particulier : il a tendance à vouloir faire une règle pour chaque information possible (par exemple, une règle qui affiche le résumé du film "La mort aux trousses" si le sujet a demandé

ce film, ainsi qu'une règle qui affiche le résumé du film "Fenêtre sur cour" si le sujet a demandé ce film, etc.). Ceci est dû au fait que l'ergonome n'ayant pas accès aux données manipulées par le système, il a du mal à adopter un raisonnement orienté système. Une deuxième raison à la difficulté de formalisation des règles de choix est que, dans l'état actuel de l'éditeur, l'ergonome ne peut pas observer immédiatement le comportement qui correspond à la spécification de présentation d'une règle. Pour pallier ce problème, un permier couplage entre le composant de choix généré et des composants de présentation a été réalisé. Ce couplage s'appuie sur un fichier XML qui permet de faire le lien entre une tâche de présentation définie et une action sur un émulateur de téléphone mobile ou la diffusion d'un message vocal. Voici un extrait de ce fichier XML.

```
<concreteRealisation>
<type>Reponse</type>
<modality>Auditive</modality>
<param>(sequence TitreFr)</param>
<realisation>play(C:/WINDOWS/Media/tada.wav)</realisation>
</concreteRealisation>

<concreteRealisation>
<type>Reponse</type>
<modality>Visuelle</modality>
<param>(sequence TitreFr)</param>
<realisation>Titre:_VALUE_</realisation>
</concreteRealisation>
```

La première tâche de présentation est réalisée par la diffusion d'un fichier son "tada.wav" via les hauts-parleurs de l'ordinateur sur lequel est émulé. Le résultat de la deuxième tâche de présentation sur l'émulateur est présenté dans la figure 7.18.

Une dernière raison à la difficulté de formalisation des règles de choix est que l'ergonome souhaiterait caractériser des tâches de présentation selon des attributs qui n'ont pas été envisagés. En particulier, la conception des règles seraient facilitées si l'ergonome pouvait spécifier qu'une tâche de présentation est plutôt applicative ou interactive (*cf.* la section 6.2). Cet attribut pourrait simplifier la conception des règles pour le test d'une configuration donnée en définissant des profils de stratégies de présentation qui appliquent une certaine modalité aux tâches de présentation applicative et une aux tâches de présentation interactive.

Comme tout logiciel, l'éditeur graphique proposé, s'il permet à un ergonome de simplifier la conception des comportements d'un système d'information en spécifiant un composant de choix de stratégies de dialogue et de présentation, doit encore être amélioré pour faciliter cette conception. Les premiers retours sur l'utilisation de cet éditeur par un ergonome ont déjà contribué à explorer une possibilité de couplage entre le composant de choix généré et des composants de présentation et à envisager des pistes d'amélioration.

FIG. 7.18 – Tâche de présentation visuelle d'un titre de film pour l'expérimentation en cours suite au couplage du composant de choix de choix avec un émulateur de mobile

7.5 Conclusion

Nous avons présenté une implémentation du composant de choix de stratégies de dialogue et de présentation du chapitre 5 sous la forme d'un agent JADE-JSA qui échange avec l'extérieur des actes communicatifs selon le standard FIPA-ACL. L'un des atouts de ce choix d'implémentation est de pouvoir connecter directement sans interface logicielle à écrire notre composant de choix à tout contrôleur de dialogue fournissant des actes communicatifs FIPA-ACL. Basé sur cette implémentation, nous avons ensuite présenté notre outil graphique de génération du composant de choix décrit au chapitre 6. Cet outil graphique est destiné à des ergonomes/psychologues, avec l'aide d'un informaticien pour le couplage avec le reste de l'application. Cet outil favorise l'exploration de plusieurs stratégies de dialogue et de présentation car il permet une spécification graphique à base de règles à partir de laquelle le composant de choix est généré.

Afin de valider l'implémentation du composant de choix de stratégies de dialogue et de présentation qu'il soit généré par notre outil graphique ou programmé, nous avons considéré deux exemples complémentaires. Le premier système est @mie, le système que nous avons utilisé comme exemple tout au long de l'exposé de nos travaux. Avec @mie, nous avons étudié 33 scénarios différents considérant l'usage des deux modalités, l'une visuelle et l'autre auditive. Le deuxième système considéré, Santiago, est complémentaire puisqu'il s'agit d'expérimentations magicien d'Oz. En considérant ce système, notre objectif n'était pas de valider le composant de choix au sein d'un deuxième système d'information, mais d'utiliser le composant de choix comme une aide au compère pour définir la présentation multimodale renvoyée au sujet. Une dernière validation est en cours : elle consiste en l'utilisation de l'éditeur graphique par un ergonome afin de concevoir un composant de choix qui sera utilisé, au sein de la plate-forme de simulation, pour une expérimentation en magicien d'Oz.

Pour les deux systèmes considérés, et dans un premier temps, le composant de choix n'a pas été intégré aux deux systèmes et nous avons développé un simulateur, simulant les actes communicatifs envoyés par le contrôleur de dialogue au composant de choix, et affichant sous la forme textuelle les résultats du composant de choix.

Nos perspectives immédiates pour ces réalisations sont d'intégrer le composant de choix au sein des systèmes directement sans passer par le simulateur. Ceci implique un couplage entre le composant de choix et des composants de présentation (abstraite et concrète) de façon à concrétiser la présentation multimodale spécifiée en sortie du composant de choix. Pour cela, nous envisageons d'étendre le couplage réalisé pour l'expérimentation en magicien d'Oz en cours utilisant la plate-forme de simulation. Une autre perspective est de faire évoluer l'éditeur graphique en fonction des retours de l'ergonome qui l'utilise actuellement. Bien que l'outil ait été conçu dans le cadre de collaborations avec des ergonomes ou psychologues, il nous semble important de valider les concepts manipulés dans l'outil et d'étudier plus finement la collaboration entre ergonomes et informaticiens que notre outil implique.

Conclusion

Contribution de la thèse

À l'heure où les systèmes informatiques sont plus courants que jamais, où les terminaux se multiplient et se miniaturisent et où les supports convergent pour proposer tous les services, il est temps de travailler sur la communication humain-machine naturelle de demain, qui devra sembler naturelle aux utilisateurs. Nous défendons l'idée que ce naturel ne transparaît pas seulement à travers les capacités d'action offertes à l'utilisateur et les capacités d'expression du système, mais aussi à travers le comportement des systèmes, que ce soit au niveau de leur réaction ou au niveau de la présentation de cette réaction. Dans des systèmes d'information coopératifs, ceci implique une prise en compte des contraintes de présentation lors du choix de la stratégie de présentation et lors du choix de la stratégie de dialogue. Ces deux choix de stratégies doivent être simultanés, à défaut d'être concertés, de façon à assurer aux utilisateurs l'accessibilité sensoriactionnelle, l'accessibilité cognitive et l'accessibilité rhétorique des informations et des capacités d'action. Ce choix conjoint des stratégies de dialogue et de présentation revient à plus rapprocher les deux paradigmes de communication humain-machine que sont le paradigme dialogique issu de l'Intelligence Artificielle (IA) et le paradigme actionnel issu de l'Interaction Homme-Machine (IHM).

C'est à ce rapprochement que nous avons consacré nos travaux et au choix conjoint des stratégies de dialogue et de présentation que nous avons contribué. En focalisant sur les systèmes d'information grand-public multimodaux en sortie, nous avons proposé une solution conceptuelle et un outil de conception. La solution conceptuelle a pour objectif de permettre la prise en compte des contraintes de présentation dans la détermination du comportement des systèmes. Au sein du modèle d'architecture de référence Arch, nous introduisons un composant de choix intermédiaire entre le composant de dialogue et le composant de présentation (abstraite) dédié au choix des stratégies de dialogue et de présentation. Remarquant le manque d'étude sur l'impact de ces stratégies sur l'utilisation et sur la poursuite de la communication d'une part et l'intérêt croissant pour cette problématique chez les ergonomes et en psychologie expérimentale d'autre part, nous proposons un outil destiné à faciliter de telles études dans le cadre du cycle de conception de systèmes d'information multimodaux. Plus précisément, nous présentons un éditeur graphique de génération de composants de choix de stratégies de dialogue et de présentation, destiné aux non-informaticiens. Cet éditeur permet la conception incrémentale d'un composant de choix pour un système applicatif donné. Une plate-

forme de simulation permet d'utiliser le composant de choix généré grâce à cet éditeur pour des expérimentations en magicien d'Oz. Chacune de ces deux contributions, le composant de choix de stratégies de dialogue et de présentation et l'éditeur graphique de génération, sont validés par une réalisation logicielle.

Perspectives de développement

Si nos contributions constituent un premier pas vers des systèmes d'information communiquant naturellement, la prise en compte dans ces derniers des contraintes de présentation pour le comportement du système, i.e. sa réaction et sa présentation, ne s'avère pas suffisante. Nos perspectives à court et à moyen terme sont donc les suivantes.

Extensions

Dans nos travaux, nous nous sommes concentrés sur le choix des stratégies de dialogue et de présentation à un haut niveau d'abstraction. Nous travaillons actuellement à un couplage du composant de choix avec des composants de présentation (abstraite et concrète) de façon à observer sur des exemples réels la prise en compte des contraintes de présentation sur le comportement du système. Si ce couplage est, pour l'instant, envisagé dans un cadre applicatif restreint, nous souhaitons l'étendre pour proposer une approche unifiée de la conception de la sortie de systèmes d'information multimodaux. Une solution serait le couplage du composant de choix avec la plateforme ICARE en sortie [Mansoux, 2005, Mansoux *et al.*, 2006].

L'éditeur graphique de génération est en cours de validation dans le cadre d'une utilisation réelle par un ergonome pour la conception d'un composant de choix de stratégies de dialogue et de présentation. Le composant de choix généré sera utilisé, au sein de la plate-forme de simulation, pour une expérimentation en magicien d'Oz. Les premiers retours de l'ergonome nous ouvrent des pistes d'affinement des concepts manipulés par cet éditeur et d'amélioration de fonctionnement de cet outil. Cette expérimentation a pour but d'étudier l'impact des stratégies de répartition multimodale (langage naturel oral auditif et langage naturel écrit / hypertexte visuels) des informations sur l'utilisateur, son intégration des informations et son interaction avec le système. Elle va également nous permettre de considérer toutes les étapes de la communication, en particulier les interventions relevant du méta-dialogue (messages de bienvenue, d'incompréhension, d'aide ...), et peut-être d'élargir la notion de "stratégie de dialogue" que nous proposons.

Nous avons, dans l'éditeur graphique proposé, commencé à introduire la problématique de la synchronisation de la présentation qui correspond à la prise en compte d'autres combinaisons de modalités que les combinaisons sémantique et modalitaire sur lesquelles nous nous étions concentrés. Nous envisageons d'affiner cette problématique en étudiant les synchronisations possibles, temporellement ainsi que spatialement, des informations présentées en tenant compte de la caractéristique des modalités utilisées. Les caractéristiques temporelles et spatiales des modalités sont des contraintes de présentation qui peuvent être problématiques si elles ne sont pas envisagées à la fois à un

haut et à un bas niveau d'abstraction. Si une synchronisation à un haut niveau d'abstraction peut sembler adéquate, elle peut être limitée par les capacités des dispositifs physiques disponibles. Dans ce cadre, nous pensons que l'introduction des modalités tactiles peut être intéressante, car elle pose plus directement le problème de la saillance des informations en tenant compte de la capacité exploratoire de l'utilisateur. L'étude de la synchronisation temporelle et spatiale des informations est indispensable dans la prise en compte de l'activité (par opposition à la passivité) de l'utilisateur et permet de ré-envisager la communication humain-machine naturelle comme une boucle et non simplement comme deux sens (sortie et entrée) dissociés.

Prolongements

Nos perspectives à moyen terme visent à poser des jalons supplémentaires pour une communication humain-machine plus naturelle. Il nous semble primordial de continuer à travailler sur les contraintes de présentation inhérentes aux modalités, en nous reposant sur leurs caractéristiques et leurs impacts sur les stratégies de dialogue et de présentation. En particulier, et dans l'appréhension de la communication en tant que boucle entrée-sortie, nous souhaitons concentrer nos efforts sur l'extraction de régularités d'usage pour anticiper les contraintes et les attentes de présentation et réaction de l'utilisateur. L'interprétation de ces régularités par rapport à un système donné permettrait d'anticiper et de mettre en avant certaines informations et/ou capacités d'action. Les stratégies de dialogue et de présentation ne se limiteraient alors pas à un haut niveau d'abstraction, mais s'étendraient à des choix de comportements plus fins, en agissant en faveur des différents types de saillance possibles [Landragin, 2004b].

Cette prise en compte des régularités d'usage rejoint une deuxième perspective à moyen terme. Pour des raisons de simplicité, nous avons choisi un mécanisme de choix de stratégies de dialogue et de présentation à base de règles. Or ce formalisme est trop figé, que ce soit au moment de la conception ou au moment de l'exécution. La prise en compte des régularités d'usage nécessite un assouplissement de ce formalisme et peut en même temps favoriser cette souplesse. De plus, de façon à ce que l'éditeur de génération ne soit pas trop contraignant, il serait souhaitable que l'utilisateur-concepteur ne soit pas obligé de concevoir un jeu de règles complet.

Pour permettre ces deux améliorations, les règles pourraient, dans un premier temps, être conçues selon les principes de la logique floue de façon à être plus souple - à la conception et à l'exécution. Par la suite et de façon complémentaire, nous envisageons d'attribuer des poids aux règles afin que le comportement du système soit déterminé par la règle qui s'applique le mieux à une situation d'utilisation donnée. Les poids pourraient alors été modifiés en fonction des régularités d'usage, i.e. de l'appréciation ou non du comportement du système par l'utilisateur - ce qui nécessite de pouvoir évaluer cette appréciation. Par ailleurs, les régularités d'usage pourraient également permettre la création de nouvelles règles, influençant alors directement la stratégie de dialogue du système (par exemple, quand, après une requête de l'utilisateur et un même type d'information présentée, l'utilisateur demande systématiquement un autre type d'information particulier). Nous pensons que l'assouplissement du fonctionnement du

composant de choix de stratégies de dialogue et de présentation permettra d'affiner le continuum de combinaison de modalités entre complémentarité et redondance et d'envisager des niveaux plus fins de choix de stratégies de présentation.

Enfin, de façon à tenir compte de l'évolution des terminaux et à anticiper les attentes des utilisateurs en ce qui concerne les dispositifs physiques multimodaux, nous envisageons également de travailler sur les modalités haptiques. De telles modalités peuvent porter des informations sémantiquement riches dans le cas du braille, et nous pensons qu'elles peuvent aussi intervenir en complément des modalités classiques visuelles et auditives pour rendre certaines informations plus saillantes. Leur prise en compte dans le choix des stratégies de dialogue et de présentation peut s'avérer particulièrement intéressante dans le cas des systèmes d'information en réalité augmentée ou encore en situation de mobilité où visuel et auditif sont plus difficilement exploitables.

Bibliographie

[Allen *et al.*, 2001] Allen, A., Ferguson, G., et A., S. (2001). An architectural for more realistic conversational systems. *In Proceedings of the Intelligent User Interfaces conference IUI'01.*
(cité aux pages 83, 91, 104, 105, 107, 108, 291)

[André, 2000] André, R. (2000). *A handbook of natural language processing : techniques and applications for the processing of language as text*, chapitre The generation of multimedia presentations. Marcel Dekker Inc.
(cité à la page 135)

[Bachimont, 2004] Bachimont, B. (2004). *Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* Habilitation à diriger des recherches, Université de Technologie de Compiègne.
(cité à la page 32)

[Battail, 1997] Battail, G. (1997). *Théorie de l'information – application aux techniques de communication.* Dunod.
(cité aux pages 10, 11, 55)

[Baus *et al.*, 2007] Baus, J., Wasinger, R., Krüger, A., Maier, A. et Schwartz, T. (2007). Auditory perceptible landmarks in mobile navigation. *In Proceedings of the Intelligent User Interfaces conference IUI'07.*
(cité à la page 88)

[Beaudoin-Lafon, 2004] Beaudoin-Lafon, M. (2004). Designing interaction, not interfaces. *In Proceedings of the Advanced Visual Interfaces conference AVI'04.* Invited paper.
(cité aux pages 2, 47, 83, 99, 101, 129, 130)

[Begault, 1994] Begault, D. (1994). *3-D sound for virtual reality and multimedia.* Academic Press Professional.
(cité à la page 88)

[Bell *et al.*, 2005] Bell, B., Feiner, S. et Höllerer, T. (2005). *Multimodal intelligent information presentation*, chapitre Maintaining visibility constraints for view management in 3D user interfaces, pages 255–277. Springer.
(cité à la page 88)

[Bellik, 1995] Bellik, Y. (1995). *Interfaces multimodales : concepts, modèles et architectures.* Thèse d'informatique, Paris XI.
(cité aux pages 19, 20, 29, 30, 67, 68, 152, 166, 202, 291, 295)

[Benali Khoudja *et al.*, 2005] Benali Khoudja, M., Sautour, A. et Hafez, M. (2005). Tactile feedback : towards a new tactile language to communicate emotions. *In Proceedings of the Virtual Concept 2005 conference.*
(cité à la page 88)

[Bernsen, 1994] Bernsen, N. O. (1994). Foundations of multimodal representations : a taxonomy of representational modalities. *Interacting with computers*, 6(4):347–371.
(cité aux pages 14, 15, 29, 30, 55, 124, 152, 166, 202)

[Bernsen, 1997] Bernsen, N. O. (1997). A reference model for output information in intelligent multimedia presentation systems. *Computer standards and interfaces, Special double issue*, 18(6-7):537–553.
(cité aux pages 14, 15, 16, 29, 30, 55, 124, 152, 166, 202, 291)

[Berthoz, 1997] Berthoz, A. (1997). *Le sens du mouvement.* Odile Jacob.
(cité aux pages 47, 75)

[Bodell *et al.*, 2003] Bodell, M., Johnsont, M., Kumar, S., Potter, S. et Waters, K. (2003). W3c multimodal interaction framework. http ://www.w3.org/TR/mmi-framework/.
(cité aux pages 120, 121, 291)

[Boehm, 1986] Boehm, B. (1986). A spiral model of software development and enhancement. *SIGSOFT Software Engineering Notes*, 11(4).
(cité à la page 174)

[Bolt, 1980] Bolt, R. A. (1980). "put-that-here" : voice and gesture at the graphics interface. *In Proceedings of the ACM SIGGRAPH conference*, pages 262–270. ACM.
(cité aux pages 87, 120)

[Bordegoni *et al.*, 1997] Bordegoni, M. G., Faconti, G., Feiner, S., Maybury, M. T., Rist, T., Ruggieri, S., Trahanias, P. et Wilson, M. (1997). A standard reference model for intelligent multimedia presentation systems. *Computer standards and interfaces : international journal on the development and application of standards for computers, data communications and interfaces.*, 18(6-7):477–496.
(cité aux pages 21, 30, 87, 135, 136, 150, 160, 161, 165, 291, 292)

[Bouchet, 2006] Bouchet, J. (2006). *Ingénierie de l'interaction multimodale en entrée : approche à composants ICARE.* Thèse de doctorat, Université Joseph Fourier - Grenoble 1.
(cité aux pages 162, 184)

[Breton et Proulx, 2002] Breton, P. et Proulx, S. (2002). *L'explosion de la communication à l'aube du XXIème siècle.* La Découverte.
(cité aux pages 34, 38, 71, 72)

[Brewster, 1997] Brewster, S. (1997). Using non-speech sound to overcome information overload. *Displays, special issue on multimedia displays,*, 17:179–189.
(cité à la page 88)

[Bui, 2006] Bui, T. H. (2006). Multimodal dialogue management - state of the art. Rapport technique 06-01, CTIT technical report.
(cité à la page 100)

[Bunt, 1994] Bunt, H. (1994). Context and dialogue control. *Think*, 3:19–31.
(cité à la page 194)

[Bunt *et al.*, 2005] Bunt, H., Kipp, M., Maybury, M. et Wahlster, W. (2005). *Multimodal intelligent information presentation*, chapitre Fusion and coordination for multimodal interactive information presentation, pages 325–339. Springer.
(cité aux pages 139, 292)

[CAC, 2007] CAC (2007). The context-aware computing group of the media lab (massachusetts institute of technology). http ://context.media.mit.edu. (accès le 16 octobre 2007).
(cité à la page 89)

[Caelen, 2003] Caelen, J. (2003). Stratégies de dialogue. *In Actes de la conférence Modèles Formels de l'Interaction MFI'03.*
(cité aux pages 101, 153, 298)

[Calvet, 1996] Calvet, L.-J. (1996). *Histoire de l'écriture.* Hachette.
(cité aux pages 31, 71)

[Carbonell, 2005] Carbonell, N. (2005). *Universal access in health telematics*, volume 3041/2005 de *Lecture notes in computer science*, chapitre Multimodal interfaces – a generic design approach, pages 209–223. Springer.
(cité aux pages 122, 123, 292)

[Cassell *et al.*, 2000] Cassell, J., Sullivan, J., Prevost, S. et Churchill, E., éditeurs (2000). *Embodied conversational agents.* MIT Press.
(cité aux pages 87, 103)

[Catizone *et al.*, 2003] Catizone, R., Setzer, A. et Wilks, Y. (2003). Multimodal dialogue management in the COMIC project. *In EACL 2003 - Workshop on Dialogue Systems : interaction, adaptation, and styles of management.*
(cité aux pages 137, 292)

[Clementz, 2003] Clementz, F. (2003). *Philosophies de la perception : phénoménologie, grammaire et sciences cognitives*, chapitre Le concept de propriété phénoménale, pages 133–155. Odile Jacob.
(cité à la page 47)

[Clémente, 2004] Clémente, P. (2004). *Vers la formalisation des capacités multimodales d'un agent rationnel dialoguant.* Thèse de doctorat, Université de Franche-Comté.
(cité aux pages 25, 29, 30, 64, 110, 152, 166, 166, 202, 202)

[Colineau et Paris, 2003] Colineau, N. et Paris, C. (2003). La génération de documents multimédia. *Cahiers romans de sciences cognitives Cognito*, 1(2):1–22.
(cité à la page 131)

[Coquery, 1994] Coquery, J.-M. (1994). *Traité de psychologie expérimentale*, volume 1, chapitre Processus attentionnels, pages 219–281. PUF.
(cité à la page 45)

[Corraze, 1980] CORRAZE, J. (1980). *Les communications non verbales.* PUF.
(cité aux pages 10, 36, 66, 71)

[Coutaz, 1987] COUTAZ, J. (1987). Pac : an implementation model for dialog design. *In Proc. Interact'87.*, pages 431–436.
(cité à la page 119)

[Coutaz *et al.*, 1993] COUTAZ, J., SALBER, D. et BALBO, S. (1993). Towards the automatic evaluation of multimodal user interfaces. *Knowledge-based systems, Special issue "“Intelligent user interfaces”"*, 6(4):267–274.
(cité aux pages 61, 124)

[Delorme, 1994] DELORME, A. (1994). *Traité de psychologie expérimentale 1*, chapitre Mécanismes généraux de la perception, pages 161–218. PUF.
(cité aux pages 43, 44, 46)

[Escarpit, 1991] ESCARPIT, R. (1991). *L'information et la communication : théorie générale.* Hachette.
(cité aux pages 10, 32, 33, 34, 71, 291)

[Exhalia, 2004] EXHALIA (2004). Solutions pour le multimédia olfactif. http ://www.exhalia.com. (accès le 5 novembre 2007).
(cité à la page 35)

[Feiner *et al.*, 1993] FEINER, S. K., LITMAN, D. T., MCKEOWN, R. et PASSONNEAU, R. J. (1993). *Intelligent multimedia interfaces*, chapitre Towards coordinated temporal multimedia presentations, pages 139–147. AAAI.
(cité à la page 203)

[Fernández *et al.*, 2007] FERNÁNDEZ, R., SCHLANGEN, D. et LUCHT, T. (2007). Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. *In The 2007 Workshop on the Semantics and Pragmatics of Dialogue DECALOG (SEMDIAL 2007).*
(cité à la page 83)

[FIPA, 2002a] FIPA (2002a). FIPA communicative act library specification. http ://www.fipa.org/specs/fipa00037/. (accès le 30 octobre2007).
(cité aux pages 205, 231)

[FIPA, 2002b] FIPA (2002b). FIPA SL content language specification. http ://www.fipa.org/specs/fipa00008/. (accès le 30 octobre 2007).
(cité à la page 206)

[Foster, 2002] FOSTER, M. E. (2002). State of the art review : multimodal fission. Rapport technique, COMIC project.
(cité aux pages 100, 135)

[Foster *et al.*, 2005] FOSTER, M. E., WHITE, M., SETZER, A. et CATIZONE, R. (2005). Multimodal generation in the comic dialogue system. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL 2005*, pages 45 – 48. Association for Computational Linguistics.
(cité à la page 137)

[Fréard, 2006] FRÉARD, D. (2006). Complémentarité et interférences dans l'interaction et le dialogue multimodal - rapport intermédiaire. Rapport technique, Convention de recherche France Télécom R&D - CRPSS.
(cité aux pages 227, 293)

[Fréard *et al.*, 2007] FRÉARD, D., JAMET, E., LE BOHEC, O., POULAIN, G. et BOTHEREL, V. (2007). Subjective measurement of workload related to a multimodal interaction task : Nasa-tlx versus workload profile. *In Proceedings of the HCI International 2007 HCII'07.*
(cité aux pages 46, 48, 74, 75, 91, 129, 150, 156, 181, 182, 186)

[Fraser et Gilbert, 1990] FRASER, N. M. et GILBERT, G. N. ., V. p. . (1990). Simulating speech systems. *Computer Speech and Language*, 5:81–99.
(cité à la page 189)

[Frohlich, 1996] FROHLICH, D. (1996). *Handbook of HCI : Second completely revised edition*, chapitre Chapter 22 : Direct manipulation and other lessons, pages 463–488. Elsevier Science.
(cité aux pages 12, 13, 56, 57, 83, 99, 117, 131, 291, 295)

[Frohlich, 1991] FROHLICH, D. M. (1991). The design space of interfaces, multimedia systems. *In* KJELLDAHL, L., éditeur : *Interaction and applications, 1er Eurographics Workshop*, pages 53–69. Springer-Verlag.
(cité aux pages 13, 30)

[Green, 1985] GREEN, M. (1985). *Seeheim Workshop on User Interface Management Systems*, chapitre Report on dialogue specification tools, pages 9–20. Pfaff, G.
(cité aux pages 116, 291)

[Grice, 1975] GRICE, H. P. (1975). *Syntax and semantics 3 : speech acts*, chapitre Logic and conversation, pages 41–58. Academic Press.
(cité aux pages 84, 263)

[GT ACA, 2007] GT ACA (2007). Groupe de travail sur les agents conversationnels animés. http ://www.limsi.fr/aca/. (accès le 16 octobre 2007).
(cité à la page 103)

[Gustafson, 2002] GUSTAFSON, J. (2002). *Developing multimodal spoken dialogue systems - empirical studies of spoken human-computer interactions.* Thèse de doctorat, KTH, Stockholm.
(cité aux pages 91, 99, 100, 103, 106, 107, 291)

[Hatwell, 1994] HATWELL, Y. (1994). *Traité de psychologie expérimentale*, chapitre Transferts intermodaux et intégration intermodale, pages 543–584. PUF.
(cité aux pages 73, 75)

[Herzog et Reithinger, 2006] HERZOG, G. et REITHINGER, N. (2006). *SmartKom : foundations of multimodal dialogue systems*, chapitre The SmartKom architecture : a framework for multimodal dialogue systems, pages 55–70. Springer.
(cité aux pages 140, 141, 142, 143, 292)

[Hoggan et Brewster, 2006] HOGGAN, E. E. et BREWSTER, S. A. (2006). Crossmodal icons for information display. *In Proceedings of the Computer-Human Interaction*

conference CHI 2006.
(cité à la page 88)

[Hone et Baber, 2001] HONE, K. S. et BABER, C. (2001). Designing habitable dialogues for speech-based interaction with computers. *International Journal of Human-Computer Studies*, 54(4):637–662.
(cité à la page 185)

[Horchani *et al.*, 2007a] HORCHANI, M., FRÉARD, D., CARON, B., JAMET, ., NIGAY, L. et PANAGET, F. (2007a). Stratégies de dialogue et de présentation multimodale : un composant logiciel dédié et son application à des expérimentations en magicien d'Oz. *In Actes de la 19ème conférence francophone sur l'Interaction Humain-Machine IHM'07.*
(cité aux pages 182, 187, 190, 193, 292)

[Horchani *et al.*, 2005] HORCHANI, M., NANARD, J. et NANARD, M. (2005). *Les hypermédias*, chapitre Les hypermédias comme paradigme d'interfaces adaptatives, pages 117–144. Hermès.
(cité à la page 82)

[Horchani *et al.*, 2007b] HORCHANI, M., NIGAY, L. et PANAGET, F. (2007b). A platform for output dialogic strategies in natural multimodal dialogue systems. *In Proceedings of the Intelligent User Interfaces conference IUI'07*, pages 206–215. ACM Press.
(cité aux pages 156, 177)

[Hutchins *et al.*, 1986] HUTCHINS, E. L., HOLLAND, J. D. et NORMAN, D. A. (1986). *User centered system design*, chapitre Direct manipulation interfaces, pages 87–124. Lawrence Erlbaum.
(cité aux pages 99, 114, 115)

[Imbert, 2006] IMBERT, M. (2006). *Traité du cerveau.* Odile Jacob.
(cité aux pages 40, 40, 40, 40, 40, 41, 41, 42, 45, 72, 73, 73)

[Jacob, 1995] JACOB, R. (1995). *Dialogue and instruction*, chapitre Natural dialogue in modes other than natural language, pages 289–301. Springer-Verlag.
(cité à la page 87)

[JADE,] JADE. http ://jade.tilab.com. (accès le 30 octobre 2007).
(cité à la page 205)

[Johnston *et al.*, 2002] JOHNSTON, M., BANGALORE, S., VASIREDDY, G., STENT, A., EHLEN, P., WALKER, M., WHITTAKER, S. et MALOOR, P. (2002). MATCH : An architecture for multimodal dialogue systems. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 376–383.
(cité aux pages 108, 109, 152, 291)

[Karsenty, 2000] KARSENTY, L. (2000). Shifting the design philosophy of spoken natural language dialogue : from invisible to transparent systems. *In CHI'2000, Workshop on natural language interfaces.*
(cité aux pages 46, 82, 85, 86)

[Karsenty, 2006] KARSENTY, L. (2006). Les déterminants du choix d'une modalité d'interaction avec une interface multimodale. *In Ergo-IA'06.*
(cité aux pages 46, 48, 71, 79, 82, 91, 129, 150, 182)

[Kopp *et al.*, 2004] KOPP, S., TEPPER, P. et CASSELL, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. *In Proceedings of the International Conference on Multimodal Interfaces (ICMI'04).*
(cité à la page 88)

[Krasner et Pope, 1988] KRASNER, G. et POPE, S. (1988). A cookbook for using the model-view-controller user interface paradigm in smalltalk-80. *Journal of Object Oriented Programming*, 1(3):26–49.
(cité à la page 119)

[LABORIUS, 2007] LABORIUS (2007). Laboratoire de robotique mobile et de systèmes intelligents (laborius) du l'université de sherbrooke. http ://www.gel.usherbrooke.ca/laborius/. (accès le 16 octobre 2007).
(cité à la page 103)

[Landragin, 2004a] LANDRAGIN, F. (2004a). *Dialogue homme-machine multimodal.* Lavoisier.
(cité aux pages 24, 30, 46, 66, 82, 83, 83, 87, 103)

[Landragin, 2004b] LANDRAGIN, F. (2004b). Saillance physique et saillance cognitive. *CORELA*, 2(2).
(cité aux pages 82, 203, 247)

[Le Bigot *et al.*, 2006] LE BIGOT, L., JAMET, E., ROUET, J.-F. et AMIEL, V. (2006). Mode and modal transfer effects on performance and discourse organization with an information retrieval dialogue system in natural language. *Computers in Human Behavior*, 22(3):467–500.
(cité aux pages 46, 48, 71, 74, 75, 79, 91, 129, 150, 181, 182, 185)

[Lehuen, 1997] LEHUEN, J. (1997). *Un modèle de dialogue dynamique et générique intégrant l'acquisition de sa compétence linguistique : le système COALA.* Thèse de doctorat, Université de Caen.
(cité aux pages 100, 101)

[Lieury, 2005] LIEURY, A. (2005). *Psychologie de la mémoire : histoire, théories, expériences.* Dunod.
(cité aux pages 71, 75, 79, 129, 298)

[Louis et Martinez, 2005] LOUIS, L. et MARTINEZ, T. (2005). Un cadre d'interprétation de la sémantique de FIPA-ACL dans JADE. *In Journées Francophones des Systèmes multi-agents.*
(cité à la page 206)

[Louis et Martinez, 2007] LOUIS, V. et MARTINEZ, T. (2007). *Developing multi-agent systems with JADE*, chapitre JADE semantics framework. Wiley.
(cité aux pages 206, 207, 292)

[Machrouh et Panaget, 2006] MACHROUH, J. et PANAGET, F. (2006). Visual interaction in naturel human-machine dialogue. *In* ANDRÉ, A., DYBKJAER, L., MINKER,

W., NEUMANN, H. et WEBER, M., éditeurs : *Perception and Interactive Technologies (PIT)*, volume 4021 de *Lecture Notes in Computer Sciences*, pages 152–163. Springer.
(cité à la page 87)

[Mansoux, 2005] MANSOUX, B. (2005). Distributed display environments in computer-assisted surgery systems. *In CHI'05 Workshop : The Future of User Interface Design Tools.*
(cité aux pages 17, 158, 162, 177, 184, 246)

[Mansoux *et al.*, 2006] MANSOUX, B., NIGAY, L. et TROCCAZ, J. (2006). Output multimodal interaction : the case of augmented surgery. *In The 20th British Human Computer Interaction Group conference HCI'06*, BCS Conference Series, pages 177–192. Springer-Verlag and ACM Press.
(cité aux pages 162, 184, 246)

[Martin, 1995] MARTIN, J.-C. (1995). *Coopérations entre modalités et liage par synchronie dans les interfaces multimodales.* Thèse d'informatique, Ecole Nationale Supérieur des Télécommunications.
(cité aux pages 9, 13, 18, 30, 57, 58, 73, 75, 148, 166, 175, 202)

[Moran et Dourish, 2001] MORAN, T. P. et DOURISH, P. (2001). Introduction to this special issue on context-aware computing. *Human-computer interaction*, 16(2-4):87–95.
(cité à la page 89)

[Nabaztag, 2006] NABAZTAG (2006). Le premier lapin communiquant. http ://www.nabaztag.com/fr/index.html. (accès le 16 octobre 2007).
(cité à la page 103)

[Nielsen, 1993] NIELSEN, J. (1993). Noncommand user interfaces. *Communications of the ACM*, 36(4):83–99.
(cité à la page 129)

[Nievergelt et Weydert, 1980] NIEVERGELT, J. et WEYDERT, J. (1980). *Methodology of interaction*, chapitre Sites, modes, and trails : Telling the user of an interactive system where he is, what he can do, and how to get places, pages 327–338.
(cité à la page 186)

[Nigay, 1994] NIGAY, L. (1994). *Conception et modélisation logicielle des systèmes interactifs : application aux interfaces multimodales.* Thèse d'informatique, Université Joseph Fourier - Grenoble 1.
(cité aux pages 17, 61, 62, 119, 124, 175)

[Nigay et Coutaz, 1996] NIGAY, L. et COUTAZ, J. (1996). Espaces conceptuels pour l'interaction multimédia et multimodale. *In TSI.*
(cité aux pages 9, 13, 17, 30, 57, 63)

[Norman, 1986] NORMAN, D. A. (1986). *User centered system design*, chapitre Cognitive engineering, pages 31–61. Lawrence Erlbaum.
(cité aux pages 115, 131)

[Novick et Lowe, 2005] Novick, D. G. et Lowe, B. (2005). Co-generation of text and graphics. *In SIGDOC'05.*
(cité à la page 88)

[Ochs *et al.*, 2007] Ochs, M., Pelachaud, C. et Sadek, D. (2007). An empathic rational dialog agent. *In Proc. of Affective Computing and Intelligent Interaction conference (ACII)*, pages 338–349.
(cité aux pages 88, 190)

[Oviatt et Lunsford, 2004] Oviatt, S., C. R. et Lunsford, R. (2004). When do we interact multimodally ? *In Proceedings of the International Conference on Multimodal Interfaces ICMI'04.*
(cité à la page 186)

[Oviatt, 1999] Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81.
(cité à la page 149)

[Paillard, 1994] Paillard, J. (1994). *Traité de psychologie expérimentale 1*, chapitre L'intégration sensori-motrice et idéo-motrice, pages 925–961. PUF.
(cité à la page 45)

[Pfaff, 1985] Pfaff, G., éditeur (1985). *Seeheim Workshop on User Interface Management Systems.* Springer-Verlag.
(cité à la page 116)

[Picard, 1995] Picard, R. W. (1995). Affective computing. Rapport technique 321, MIT - Media lab - Perceptual computing section.
(cité à la page 88)

[Poller et Tschernomas, 2006] Poller, P. et Tschernomas, V. (2006). *SmartKom : foundations of multimodal dialogue systems*, chapitre Multimodal fission and media design, pages 379–400. Springer.
(cité à la page 72)

[Prewett *et al.*, 2006] Prewett, M. S., Yang, L., Stilson, F. R. B., Gray, A. A., Coovert, M. D., Burke, J. L. Redden, E. et Elliot, L. R. (2006). The benefits of multimodal information : a meta-analysis comparing visual and visual-tactile feedback. *In Proceedings of the International Conference on Multimodal Interfaces ICMI 2006.*
(cité aux pages 88, 88)

[Purves *et al.*, 2003] Purves, D., Augustine, G. J., Fitzpatricj, D., L. C. Katz, L. C., LaMantia, A.-S., McNamara, J. O. et Williams, S. M. (2003). *Neurosciences.* de Boeck.
(cité aux pages 40, 41, 73)

[Ratzka, 2006] Ratzka, A. (2006). Combining modality theory and context models. *In* André, A., Dybkjaer, L., Minker, W., Neumann, H. et Weber, M., éditeurs : *Perception and intertactive technologies*, pages 141–151. Springer.
(cité aux pages 17, 29, 152, 166, 202)

[Reuchlin, 1977] Reuchlin, M. (1977). *Psychologie.* PUF.
(cité aux pages 43, 44, 45, 46, 47)

[Rey, 2005] Rey, G. (2005). *Contexte en interaction homme-machine : le contexteur.* Thèse de doctorat, Université Joseph Fourier - Grenoble 1.
(cité aux pages 94, 178, 190)

[Roth et Hefley, 1993] Roth, S. F. et Hefley, W. E. (1993). *Intelligent multimedia interfaces*, chapitre Intelligent multimedia presentation systems : research and principles, pages 13–58. AAAI.
(cité aux pages 132, 133, 145, 150, 160, 161, 292)

[Rousseau, 2006] Rousseau, C. (2006). *Présentation multimodale et contextuelle de l'information.* Thèse de doctorat, Université Paris-Sud 11.
(cité aux pages 27, 28, 30, 69, 89, 124, 125, 177, 185, 190, 291, 292)

[Sadek, 1999] Sadek, D. (1999). Design considerations on dialogue systems : from theory to technology - the case of artimis. *In IDS'99.* Invited talk.
(cité aux pages 2, 83, 87, 100, 110, 154, 205, 231, 263, 298)

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The bell system technical journal*, 27:375–457, 623–656.
(cité aux pages 10, 55, 70, 126, 291)

[Shneiderman, 1986] Shneiderman, B. (1986). *Designing the user interface : strategies for effective human-computer interaction.* Addison-Wesley Publishing Company.
(cité aux pages 87, 114, 115)

[Sinha et Landay, 2003] Sinha, A. et Landay, J. (2003). Capturing user tests in a multimodal, multidevice informal prototing tool. *In Proceedings of the International Conference of Multimodal Interfaces ICMI'03*, pages 117–124. ACM Press.
(cité à la page 184)

[Siroux *et al.*, 1989] Siroux, J., Gilloux, M., Guyomard, M. et Sorin, C. (1989). Le dialogue homme-machine en langue naturelle : un défi ? *Annales des télécommunications*, 44(1-2):53–76.
(cité aux pages 2, 84)

[Sottet *et al.*, 2007] Sottet, J.-S., Ganneau, V., Calvary, G., Coutaz, J., Favre, J.-M. et Demunieux, R. (2007). Mode-driven adaptation for plastic user interfaces. *In Proceedings of the 2007 Interact conference.*
(cité aux pages 89, 124, 152, 178, 190)

[Strothotte, 2007] Strothotte, T. (2007). Image-text interaction. *In Proc. of Intelligent user interfaces IUI'07*, page 3.
(cité à la page 88)

[Sweller, 1999] Sweller, J. (1999). *Instructional design in technical areas.* ACER Press.
(cité à la page 186)

[Tabbers *et al.*, 2001] Tabbers, H. K., Martens, R. L. et van Merriënboer, J. J. G. (2001). The modality effect in multimedia instructions. *In* Moore, J. et Stenning,

K., éditeurs : *The 23rd annual conference of the Cognitive Science Society*, pages 1024–1029. Lawrence Erlbaum.
(cité à la page 150)

[Thévenin, 2001] THÉVENIN, D. (2001). *Adaptation en interaction homme-machine : le cas de la Plasticité.* Thèse de doctorat, Université Joseph Fourier - Grenoble 1.
(cité aux pages 89, 124, 152, 178, 190)

[Turing, 1950] TURING, A. (1950). Computing machinery and intelligence. *Mind*, 59: 433–460.
(cité à la page 100)

[UBICOMP, 2007] UBICOMP (2007). International conferences on ubiquitous computing. http ://www.ubicomp.org. (dernier accès le 16 octobre 2007).
(cité à la page 89)

[UIMS, 1992] UIMS (1992). *SIGCHI bulletin*, volume 24, chapitre A metamodal for the runtime architecture of an interactive system, pages 32–37. SIGCHI.
(cité aux pages 117, 118, 135, 175, 291, 297)

[van Dam, 1997] VAN DAM, A. (1997). Post-wimp user interfaces. *Communications of the ACM*, 40(2):63–67.
(cité aux pages 113, 115, 130)

[Vernier, 2001] VERNIER, F. (2001). *La multimodalité en sortie et son application à la visualisation de grandes quantités d'information.* Thèse de doctorat, Université Joseph Fourier - Grenoble 1.
(cité aux pages 17, 23, 29, 30, 64, 65, 195, 202, 203, 291)

[Wahlster, 2006] WAHLSTER, W. (2006). *SmartKom : foundations of multimodal dialogue systems*, chapitre Dialogue systems go multimodal : the SmartKom experience, pages 3–27. Springer.
(cité aux pages 30, 140, 203)

[Wahlster *et al.*, 2001] WAHLSTER, W., REITHINGER, N. et BLOCHER, A. (2001). Smartkom : towards multimodal dialogues with anthropomorphic interface agents. *In* WOLF, G. et KLEIN, G., éditeurs : *Proceedings of International Status Conference "Human-Computer Interaction".*
(cité à la page 87)

[Weiser, 1994] WEISER, M. (1994). The world is not a deskop. *Interactions*, 1(1):7–8.
(cité à la page 89)

[Weiss, 2005] WEISS, P. L. (2005). *Multimodal intelligent information presentation*, chapitre Presentation technologies for people disabilities, pages 305–321. Springer.
(cité à la page 89)

[Weizenbaum, 1967] WEIZENBAUM, J. (1967). Eliza. http ://library.thinkquest.org/18242/eliza.shtml.
(cité à la page 100)

[Zhou et Aggarwal, 2004] ZHOU, M. X. et AGGARWAL, V. (2004). An optimization-based approach to dynamic data content selection in intelligent multimedia interfaces.

In Proceedings of UIST'04 conference, pages 227–236.
(cité à la page 152)

[Zhou *et al.*, 2005] Zhou, M. X., Wen, Z. et Aggarwal, V. (2005). A graph-matching approach to dynamic media allocation in iintelligent multimedia interfaces. *In Proceedings of the Intelligent User Interfaces conference IUI'05*, pages 114–121.
(cité aux pages 152, 164, 203)

Annexes

Annexe 1 - Caractérisation de la convivialité et de la coopération dans le dialogue

Sadek [Sadek, 1999] caractérise la convivialité de la façon suivante :

- capacité de négociation ;
- interprétation en contexte (cas le plus courant en communication inter-humaine) ; flexibilité du langage d'entrée – insuffisante à elle seule pour faire l'intelligence d'un système ;
- flexibilité de l'interaction : elle passe par une structure non figée de la dite interaction, qui s'avère nécessaire en cas de difficulté de communication. Il s'agit, en fait, de laisser à l'utilisateur la possibilité de s'écarter du cours « régulier » de l'interaction dès qu'il en ressent le besoin ;
- production de réponses coopératives : on parle de « réponse coopérative » lorsque la réponse s'étend de façon pertinente au-delà de la question posée explicitement. Dans cette catégorie, sont distinguées :
  - o les réponses complétives ou sur-informatives ;
  - o les réponses correctives, qui informent l'interlocuteur que certains de ses présupposés sont caducs ;
  - o les réponses suggestives, qui proposent une solution proche mais ne répondent pas directement à la requête formulée ;
  - o les réponses conditionnelles ;
  - o les réponses intentionnelles (dans le sens d'appartenance à une même classe, par opposition à « extension ») ;
- adéquation des formes/styles de réponses – tant au niveau du contenu que de la « mise en forme ».

Par ailleurs, Grice [Grice, 1975] définit les quatre principes suivants de coopération dans le dialogue :

- maxime de quantité : que la contribution comprenne ni plus ni plus que ce qui est nécessaire ;
- maxime de qualité : que la contribution soit vérifiée et non mensongère ;
- maxime de pertinence : que la contribution soit en adéquation avec les besoins immédiats ;

– maxime de manière : que la contribution soit claire, sans ambiguïté, brève et ordonnée.

Annexe 2 - Critères d'utilisabilité

Cette annexe consigne un ensemble de critères servant à caractériser la notion d'utilisabilité. Dans ce cadre, nous dirons que : Utilisabilité = Souplesse + Robustesse où :

- La Souplesse (de l'interaction) exprime l'éventail des choix (pour l'utilisateur et le système).
- La Robustesse (de l'interaction) vise la prévention des erreurs, l'augmentation des chances de succès de l'utilisateur.

Remarque : la facilité d'apprentissage n'est pas considérée.

Souplesse et robustesse sont déclinées en critères. Pour chacun d'eux, nous fournissons une définition complétée de remarques, d'exemples, de contre-exemples et, si possible, de métriques.

Souplesse

Atteignabilité

Définition : Capacité du système à permettre à l'utilisateur de naviguer dans l'ensemble des états observables du système. Un état q est atteignable à partir d'un état p s'il existe une suite de commandes ci qui font passer de l'état p à l'état q.

Remarque : Propriété non vérifiée si analyse de tâche ou analyse fonctionnelle défectueuse.

Métrique : Longueur de la trajectoire d'interaction entre p et q.

Non-préemption

Définition : Le prochain but souhaité par l'utilisateur est directement atteignable. Pour l'utilisateur, il n'y a pas de contrainte dans la trajectoire interactionnelle.

Contre-exemple : Boîte de dialogue modale : à n'utiliser qu'à bon escient (cas des commandes irrévocables). Un système interruptible est, du point de vue utilisateur, non préemptif. Mais trop de liberté peut nuire : L'utilisateur risque d'être perdu. Penser au concept de tableau de bord qui indique à l'utilisateur sa localisation dans l'espace des tâches.

Métrique : Si la longueur de la trajectoire d'interaction entre l'état actuel et l'état souhaité vaut 1, alors la propriété de non préemption est vérifiée.

Préemption globale

Définition : Interdiction, pour l'utilisateur, d'effectuer toute autre action que celle requise par le système (ici la station de travail).

Préemption locale

Définition : Ne bloque qu'un fil de dialogue. Laisse l'utilisateur continuer les autres fils. Préemption de ressources partagées par un utilisateur

Exemple : Collecticiels à "tour de parole" (turn-taking) : préemption du curseur partagé.

Interaction multifilaire

Définition : Capacité du système à permettre la réalisation "siumultanée" de plusieurs tâches.

Remarque : Analyse du caractère entrelacé ou parallèle à différents niveaux de granularité : niveau tâches au sein d'un logiciel donné, niveau actions (cf. les propriétés CARE).

Exemple : Edition de plusieurs documents à la fois.

Métrique : Nombre de tâches que l'on peut mener de manière entrelacée au regard de la charge cognitive.

Interaction multifilaire parallèle

Définition : Parallélisme vrai entre plusieurs tâches. Peu souhaitable si les compétences de l'utilisateur n'est pas du niveau expert ("skill level", au sens de Rasmussen).

Interaction multifilaire entrelacée

Définition : Les tâches peuvent être simultanées au sens de l'utilisateur, mais à un instant donné, l'interaction est restreinte à une seule tâche.

Multiplicité du rendu (ou représentation multiple d'un même concept)

Définition : Capacité du système à fournir plusieurs représentations pour un même concept.

Remarque : (1) En sortie : Formes différentes (par ex., pour le concept de température : un entier ou une représentation analogique comme un thermomètre. Contenus différents : le détail vs l'ensemble : Attention aux discontinuités visuelles. Penser aux techniques "fisheye" ou holophrastiques. (2) En entrée : Formes différentes : 6*4 et 24. (3) Principe d'égale opportunité : L'utilisateur choisit la nature de l'entrée et le système en déduit la sortie (principe des tableurs).

Réutilisabilité des données d'entrée et de sortie

Définition : Les sorties du système peuvent être utilisées comme des données d'entrée (couper-coller). Les entrées de l'utilisateur peuvent être réutilisées par le système en sortie (valeurs par défaut).

Remarque : Attention aux effets de bord dus aux conversions de types de données. Cas du couper-coller entre deux éditeurs graphiques, l'un vectoriel et l'autre du niveau pixel.

Adaptabilité

Définition : Personnalisation du système sur intervention explicite de l'utilisateur.

Remarque : Risque de perte de cohérence entre les systèmes d'une communauté d'utilisateurs sensés travailler ensemble. À considérer selon le degré de couplage des activités collectives. Les utilisateurs ne modifient pas les valeurs par défaut qui viennent à la livraison du collecticiel. Leçon à retenir : bien étudier les valeurs par défaut initiales en fonction des catégories/rôles des futurs utilisateurs. Exemple : Menus et formulaires d'options et de préférences. Macro d'encapsulation de commandes à caractère répétitif dont le niveau d'abstraction est trop bas. Toute donnée lexicale (ex. nom des commandes) doit être dans un fichier de ressources, donc pas dans le code source du logiciel. Mesure de vérification : produire le logiciel dans une autre langue sans le recompiler.

Adaptativité

Définition : Capacité du système à s'adapter à l'utilisateur sans intervention explicite de l'utilisateur.

Remarque : L'adaptativité s'appuie sur un modèle embarqué de l'utilisateur. Veiller à ce que le système ait un comportement prévisible. Ne pas surprendre l'utilisateur (caractère disruptif).

Plasticité

Définition : Capacité du système à s'adapter aux variations des ressources interactionnelles, computationnelles, comunicationnelles et environnementales tout en conservant la continuité ergonomique.

Exemple : IHM d'un agenda sur PalmPilot et sur PC.

Migrabilité de tâche

Définition : Capacité de délégation dynamique de tâches entre le système et l'utilisateur ou entre utilisateurs : changement dynamique de l'acteur(s) responsable(s) de l'accomplissement de la tâche.

Remarque : Manifestation à différents niveaux de granularité : (1) 6*4 et 24 (2) valeurs par défaut (3) détection de tâches répétitives puis prise en charge (Système Eager) (4) sauvegarde automatique de fichiers.

CARE (multimodalité)

Définition : Complémentarité, Assignation, Redondance, Equivalence. Caractérisation de la multimodalité offerte par un système. Modalité = <dispositif d'E/S, système représentationnel>. Complémentarité : plusieurs modalités distinctes sont nécessaires pour exprimer le but. Assignation : une seule modalité est disponible pour exprimer le but. Redondance : plusieurs modalités sont utilisables en "même temps" et expriment le même but. Equivalence : plusieurs modalités sont possibles pour exprimer le but. Une seule est utilisable à la fois. Notons que la redondance implique l'équivalence.

Exemple : Complémentarité : " mets ça là " vocal // geste. Redondance : " Montre-moi Grenoble " vocal // double-clic sur Grenoble affichée sur une carte.

CARE (collecticiel)

Définition : (1) Appliqué aux rôles des acteurs d'un collecticiel. : Complémentarité de rôle (cas des jeux). Assignation de tâche à un rôle. (2) Appliqué aux moyens technologiques pour collaborer : CARE entre Fax, email, Vphone etc.

Robustesse

Observabilité

Définition : Capacité du système à rendre perceptible l'état pertinent du système. Capacité pour l'utilisateur à évaluer l'état actuel du système. L'utilisateur PEUT PERCEVOIR.

Remarque : Inspectabilité (browsability) : capacité pour l'utilisateur d'explorer l'état interne du système au moyen de commandes articulatoires (ou passives) (c'est-à-dire qui ne modifient pas l'état du noyau fonctionnel) telles que zoom, défilement, etc.

Contre-exemple : "Defense in depth design" (Rasmussen).

Observabilité publiée

Définition : Capacité pour un utilisateur de rendre observables des variables d'état personnelles.

Remarque : Exemples de variables d'état personnelles : présence, niveau de disponibilité. Une variable publiée peut être filtrée. Filtrage de variable : opération de transformation de la valeur de la variable visant à protéger l'espace privé. Un filtre ne doit pas être réversible.

Réciprocité

Définition : Dans un collecticiel, capacité d'observation/inspection mutuelle des variables d'état personnelles.

Exemple : " Si je vous vois, vous me voyez. "

Réflexivité

Définition : Capacité d'inspecter ou d'observer les variables d'état personnelles publiées à autrui. Exemple : Vidéo miroir en vidéoconférence.

Insistance

Définition : Capacité du système à forcer la perception de l'état du système L'utilisateur DEVRA PERCEVOIR.

Remarque : Le retour d'information du système peut être pour un contexte donné : (1) Ephémère (ex. audio, vidéo) ou non (ex. écrit statique), (2) Evitable (retour visuel) ou inévitable (audio), (3) Entretenu par le système (ex. clignotement) ou par l'utilisateur (maintien d'un bouton enfoncé qui minimise les oublis). Attributs perceptuels additifs ; l'intensité sonore et l'intensité des couleurs ont un effet additif. Le ton sonore et la teinte ne sont pas additifs. Rétro-action de groupe. Awareness : compréhension des activités d'autrui, fournissant à l'action individuelle un contexte situationnel collectif

Honnêteté

Définition : Capacité du système à rendre observable l'état du système sous une forme conforme à cet état ET qui engendre une interprétation correcte de la part de l'utilisateur. L'utilisateur AURA UN MODELE CORRECT de l'état du système.

Remarque : WYSIWYG (What You See is What You Get), WYSIWIS (What You See Is What I See) et les versions relâchées. Distorsion des informations (ex. Les données linéaires ne doivent pas être présentées en deux dimensions). Conformité de l'état interne et de la présentation pas toujours compatibles avec les temps de réponse attendus : utiliser un indicateur pour exprimer que l'information a changé et qu'elle n'est pas encore réactualisée dans le rendu. Intégrité des messages entre l'émetteur et le récepteur. Dans les formulaires, veiller à la formulation des unités de mesure et du format des données à saisir. Veiller à une terminologie précise en accord avec le métier, l'utilisateur, etc.

Honnêteté sociale

Définition : Le système peut être honnête mais peut être détourné socialement.

Exemple : Enclencher son répondeur téléphonique pour simuler l'absence.

Curabilité

Définition : Capacité pour l'utilisateur de corriger une situation non désirée.

Remarque : (1) Curabilité arrière : capacité de défaire. Le défaire de profondeur 1 est facile à réaliser : on ne modifie le noyau fonctionnel qu'à l'interaction suivante. (2) Curabilité avant : reconnaissance de l'état actuel et capacité de négociation pour atteindre le but désiré = atteignabilité indispensable. Messages d'erreur explicatifs et correctifs. Principe de l'effort commensurable : ce qui est difficile à défaire doit être difficile à faire (exemple de la destruction d'un fichier).

Prévisibilité

Définition : Capacité pour l'utilisateur de prévoir, pour un état donné, l'effet d'une action.

Remarque : Cohérence : conformité aux règles/usages Les règles/usages de l'utilisateur ne sont pas nécessairement celles du concepteur. Cohérence interne : cohérence lexicale, syntaxique, sémantique. Cohérence externe : conformité à des normes d'IHM, conformité à l'expérience dans le monde réel (analogie, métaphore). Retour d'information proactif : principe du "do-nothing" ou de résistance passive : éléments interdits en grisé. Prévisibilité et stabilité des temps de réponse : rassurer si temps de réponse long. Attention aux solutions système de type ramasse-miette à la volée sur la stabilité dans le temps de réponse. Régularité des médias continus. Attention aux limites de tolérance

Tolérance du rythme

Définition : L'utilisateur plutôt que le système décide quand il peut agir.

Remarque : Attention aux temporisations qui font sens (ex. durée de pression sur un dispositif, tels les téléphones portables et les distributeurs de boissons).

Exemple : Dans le cas de la saisie anticipée : nombre d'actions anticipables (noter que sous Word, ce nombre est variable.).

Métrique : Durée de tolérance.

Viscosité

Définition : L'action de l'utilisateur à un effet sur son plan de tâches ou, pour un collecticiel, sur celui des autres.

Exemple : Ajout d'une ligne dans un texte qui provoque des orphelins dans la pagination.

Métrique : Longueur de la trajectoire d'interaction nécessaire à la correction de l'effet de viscosité.

Rejouabilité

Définition : Capacité pour l'utilisateur de rejouer des séquences informationnelles audio/vidéo.

Remarque : S'appuie sur l'existence d'un historique. En communication médiatisée, la rejouabilité d'un message audio permet de ré-écouter

Révisabilité

Définition : Capacité pour l'utilisateur de réviser un message avant de l'émettre.

Remarque : La révisabilité implique la rejouabilité. La révisabilité est une forme de curabilité.

Annexe 3 - Extraits du fichier XML pour la simulation d'entrée du composant de choix dans le cas du système-exemple @mie

```
<simulateur>

.../...

<mess>
<message>donne-moi le numéro de téléphone de Carole visuellement</message>
<enonce>
((choice (content
:answerMode visual
:requestDescription (set (description :firstName "Carole"))
:requestFocus phone
:requestModality null
:restrictionSet (sequence localisation job)
:solutionSet (set
(solution
:firstName "Josiane-Carole" :lastName "Cozanet" :crd "sirp"
:labo "cli" :urd "mix" :photo "isjo5626.jpg" :phone 299124491
:mobilePhone null :email "carole.cozanet@orange-ftgroup.com"
:fax 299123716 :job "chargée d'études clients et marchés"
)
(solution
:firstName "Carole" :lastName "Manquillet" :crd "tech"
:labo "easy" :urd "adn" :photo "00000505.jpg" :phone 296059464
:mobilePhone null :email "carole.manquillet@orange-ft.com"
:fax 299051129 :job "ing?nieur rd" :localisation "lannion"
:office "lb028"
)
```

```
(solution
:firstName "Carole" :lastName "Rivi?re" :crd "tech"
:labo "susi" :urd "inuit" :photo "ocra7421.jpg" :phone 296050000
:mobilePhone null :email "carole.rivi?re@orange-ft.com"
:localisation "lannion"
)
(solution
:firstName "Carole" :lastName "Paganus" :crd "resa"
:labo "net" :urd "nso" :photo "00001962.jpg" :phone 145296307
:mobilePhone null :email "carole.paganus@orange-ft.com"
:fax null :job "ing?nieur rd" :localisation "issy" :office "ie208"
)
)
)))
</enonce>
</mess>

.../...

<mess>
<message>donne-moi le numéro de téléphone de Carole oralement</message>
<enonce>
((choice (content
:answerMode aural
:requestDescription (set (description :firstName "Carole"))
:requestFocus phone
:requestModality null
:restrictionSet (sequence localisation job)
:solutionSet (set
(solution
:firstName "Josiane-Carole" :lastName "Cozanet" :crd "sirp"
:labo "cli" :urd "mix" :photo "isjo5626.jpg" :phone 299124491
:mobilePhone null :email "carole.cozanet@orange-ftgroup.com"
:fax 299123716 :job "chargée d'études clients et marchés"
)
(solution
:firstName "Carole" :lastName "Manquillet" :crd "tech"
:labo "easy" :urd "adn" :photo "00000505.jpg" :phone 296059464
:mobilePhone null :email "carole.manquillet@orange-ft.com"
:fax 299051129 :job "ing?nieur rd" :localisation "lannion"
:office "lb028"
)
(solution
:firstName "Carole" :lastName "Rivi?re" :crd "tech"
```

```
:labo "susi" :urd "inuit" :photo "ocra7421.jpg" :phone 296050000
:mobilePhone null :email "carole.rivi?re@orange-ft.com"
:localisation "lannion"
)
(solution
:firstName "Carole" :lastName "Paganus" :crd "resa"
:labo "net" :urd "nso" :photo "00001962.jpg" :phone 145296307
:mobilePhone null :email "carole.paganus@orange-ft.com"
:fax null :job "ing?nieur rd" :localisation "issy" :office "ie208"
)
)
)))
</enonce>
</mess>

<mess>
<message>donne-moi le num?ro de t?l?phone de Carole  </message>
<enonce>
((choice (content
:answerMode null
:requestDescription (set (description :firstName "Carole"))
:requestFocus phone
:requestModality null
:restrictionSet (sequence localisation job)
:solutionSet (set
(solution
:firstName "Josiane-Carole" :lastName "Cozanet" :crd "sirp"
:labo "cli" :urd "mix" :photo "isjo5626.jpg" :phone 299124491
:mobilePhone null :email "carole.cozanet@orange-ftgroup.com"
:fax 299123716 :job "chargée d'études clients et marchés"
)
(solution
:firstName "Carole" :lastName "Manquillet" :crd "tech"
:labo "easy" :urd "adn" :photo "00000505.jpg" :phone 296059464
:mobilePhone null :email "carole.manquillet@orange-ft.com"
:fax 299051129 :job "ing?nieur rd" :localisation "lannion"
:office "lb028"
)
(solution
:firstName "Carole" :lastName "Rivi?re" :crd "tech"
:labo "susi" :urd "inuit" :photo "ocra7421.jpg" :phone 296050000
:mobilePhone null :email "carole.rivi?re@orange-ft.com"
:localisation "lannion"
)
```

```
(solution
:firstName "Carole" :lastName "Paganus" :crd "resa"
:labo "net" :urd "nso" :photo "00001962.jpg" :phone 145296307
:mobilePhone null :email "carole.paganus@orange-ft.com"
:fax null :job "ing?nieur rd" :localisation "issy" :office "ie208"
)
)
)))
</enonce>
</mess>

.../...

</simulateur>
```

Annexe 4 - Extraits du composant de choix généré avec l'éditeur graphique dans le cas du système-exemple @mie

```
package @mie_memoire;
import jade.core.AID;
import jade.lang.acl.ACLMessage;
import jade.semantics.interpreter.SemanticAgentBase;
import jade.semantics.interpreter.SemanticCapabilities;
import jade.semantics.interpreter.SemanticInterpretationPrincipleTable;
import jade.semantics.interpreter.SemanticRepresentation;
import jade.semantics.interpreter.sips.adapters.ApplicationSpecificSIPAdapter;
import jade.semantics.lang.sl.grammar.Formula;
import jade.semantics.lang.sl.grammar.Term;
import jade.semantics.lang.sl.grammar.TermSet;
import jade.semantics.lang.sl.grammar.WordConstantNode;
import jade.semantics.lang.sl.tools.MatchResult;
import jade.semantics.lang.sl.tools.SL;
import jade.semantics.lang.sl.grammar.Constant;
import jade.util.leap.ArrayList;
import jade.semantics.lang.sl.grammar.FunctionalTermParamNode;
/**
* Classe générée automatiquement
* Composant de stratégie de choix de dialogue et de présentation
*/

public class Choix extends SemanticAgentBase {
static final String HYPERTEXTEMOD = "hypertext";
static final String ORALMOD = "oral";
static final int NBMAXSOLVISUAL = 5;
static final int NBCRITRESTRICTOUTPUT = 3;
```

```
static final Term GENERIQUE_ANSWER =
SL.fromTerm("(answer ??tache (synchronisation ??synch))");
static final Term GENERIQUE_SOLUTION =
SL.fromTerm("(solution firstname ??firstname lastname ??lastname phone ??phone
mobilePhone ??mobilePhone fax ??fax urd ??urd crd ??crd labo ??labo photo ??photo
bureau ??bureau ");
static final Term PATTERN_FEEDBACK =
SL.fromTerm("(feedback :focus  ??focus  :description ??description
:modalityFeedback ??modalityFeedback )");
static final Term PATTERN_AIDE =
SL.fromTerm("(aide :description ??description :modalityAide ??modalityAide)");
static final Term PATTERN_RELANCE =
SL.fromTerm("(relance :type ??type :modality ??modalityRelance)");
static final Term PATTERN_REPONSE_STATEMENT =
SL.fromTerm("(reponse :type ??type focus ??focus :description ??description
:modalityReponse ??modalityReponse)");
static final Term PATTERN_REPONSE_RESTRICTION =
SL.fromTerm("(reponse :type ??type :nbSolution ??nbSolution
:description ??description :modalityReponse ??modalityReponse)");
static final Term PATTERN_REPONSE_RELAXATION =
SL.fromTerm("(reponse :type ??type :description ??description
:modalityReponse ??modalityReponse)");


private String msg = new String();
class ChoixCapacites extends SemanticCapabilities {
/**---------------------------------**/
/** définition des SIP applicatifs **/
/**---------------------------------**/

class sipApp2 extends ApplicationSpecificSIPAdapter {

public sipApp2(){
super(ChoixCapacites.this,
"(B ??myself (choice ??content))");
}
// conséquence du SIP
protected ArrayList doApply(MatchResult applyResult,
ArrayList result, SemanticRepresentation sr) {
Term contentP = SL.fromTerm(
"(content :requestFocus ??requestFocus :requestDescription ??requestDescription
 :requestModality ??requestModality :answerMode ??answerMode :solutionCard ??nbSol
 :solutionSet ??solutionSet  :restrictionSeq ??restrictionSeq
 :relaxationSeq ??relaxationSeq )").getSimplifiedTerm();
```

```
MatchResult contentMatch = contentP.match(applyResult.term("content"));
if (contentMatch != null) {

Constant requestFocus = (Constant) contentMatch.term("requestFocus");
TermSet requestDescription = (TermSet) contentMatch.term("requestDescription");
TermSet solutionSet = (TermSet) contentMatch.term("solutionSet");
TermSequence restrictionSeq = (TermSequence) contentMatch.term("restrictionSeq");
TermSequence relaxationSeq = (TermSequence) contentMatch.term("relaxationSeq");
Constant answerMode = (Constant) contentMatch.term("answerMode");
Constant nbSolC = (Constant) contentMatch.term("nbSol");
int solutionCard =Integer.parseInt(nbSolC.stringValue());
if (answerMode.stringValue().equals(HYPERTEXTEMOD)
&&  solutionCard > NBMAXSOLVISUAL){
/* construction du message de sortie */
TermSet taches = new TermSetNode(new ListOfTerm());
if (solutionSet!=null)
taches.addTerm( reponseStatementConstruction( " solutionCard",
construireSolutionSet(solutionSet,solutionCard,
{"firstname", "lastname", "urd", "crd", "labo", "bureau"}),
 " Visuelle ") );
}
taches.addTerm(relanceConstruction( "null", " Visuelle ");
if (restrictionSeq!=null){
taches.addTerm( reponseRestrictionConstruction( "solutionCard",
construireRestrictionSeq(restrictionSeq), " Visuelle ") );
}

String synchro =" (  CA  0 (  CA  1 2  ) )";
result.clear();
result.add(new SemanticRepresentation(sr.getMessage(),
  answerConstruction(taches,synchro),
1));
try {
//nouveau message
String mess
ACLMessage messageACL = new ACLMessage(ACLMessage.INFORM);
//Recepteur Julie: simulateur de sortie
messageACL.addReceiver(new AID("Julie", AID.ISLOCALNAME));
//Emetteur Alice: composant de choix
messageACL.setSender(new AID("Alice", AID.ISLOCALNAME));
mess = result.get(0).toString();
messageACL.setContent(mess);
//on peut envoyer le message
```

```
send(messageACL);
} catch (Exception e) {
System.out.println(e);
}
}
  return result;
}else{
//le pattern ne s'applique pas
System.out.println("Le pattern sipApp2 ne s'applique pas.");
return null;
}
}
}
class sipApp3 extends ApplicationSpecificSIPAdapter {

public sipApp3(){
super(ChoixCapacites.this,
"(B ??myself (choice ??content))");
}
// conséquence du SIP
protected ArrayList doApply(MatchResult applyResult,
ArrayList result, SemanticRepresentation sr) {
Term contentP = SL.fromTerm(
"(content :requestFocus ??requestFocus :requestDescription ??requestDescription
:requestModality ??requestModality :answerMode ??answerMode :solutionCard ??nbSol
:solutionSet ??solutionSet  :restrictionSeq ??restrictionSeq
:relaxationSeq ??relaxationSeq )").getSimplifiedTerm();

MatchResult contentMatch = contentP.match(applyResult.term("content"));
if (contentMatch != null) {

Constant requestFocus = (Constant) contentMatch.term("requestFocus");
TermSet requestDescription = (TermSet) contentMatch.term("requestDescription");
TermSet solutionSet = (TermSet) contentMatch.term("solutionSet");
TermSequence restrictionSeq = (TermSequence) contentMatch.term("restrictionSeq");
TermSequence relaxationSeq = (TermSequence) contentMatch.term("relaxationSeq");
Constant answerMode = (Constant) contentMatch.term("answerMode");
Constant nbSolC = (Constant) contentMatch.term("nbSol");
int solutionCard =Integer.parseInt(nbSolC.stringValue());
if ((solutionCard > 1  &&  solutionCard <= NBMAXSOLVISUAL)
&&  answerMode.stringValue().equals(HYPERTEXTEMOD)){
/* construction du message de sortie */
TermSet taches = new TermSetNode(new ListOfTerm());
if (solutionSet!=null)
```

```
... // construction de la spécification de présentation correspondante

}
}
}

// Un principe d'interprétation sémantique appliqué par règle

// ajout des SIP applicatifs à la table des SIP
protected SemanticInterpretationPrincipleTable
setupSemanticInterpretationPrinciples() {
SemanticInterpretationPrincipleTable sipTab =
super.setupSemanticInterpretationPrinciples();
sipTab.addSemanticInterpretationPrinciple(new sipApp2());
sipTab.addSemanticInterpretationPrinciple(new sipApp3());
sipTab.addSemanticInterpretationPrinciple(new sipApp4());
return sipTab;
}

// Construction des TP Feedback
protected Term feedbackConstruction
(String focus,TermSet description,String modality ){
 return PATTERN_FEEDBACK
.instanciate("focus",new WordConstantNode(focus))
.instanciate("description",description)
.instanciate("modality",new WordConstantNode(modality));

// Construction des TP Aide
protected Term aideConstruction
(String description,String modality ){
 return PATTERN_AIDE
.instanciate("description",new WordConstantNode(description)))
.instanciate("modality",new WordConstantNode(modality));

// Construction des TP Relance
protected Term relanceConstruction
(String type,String modality ){
 return PATTERN_RELANCE
.instanciate("type",new WordConstantNode(type))
.instanciate("modality",new WordConstantNode(modality));

// Construction des TP Reponse Statement
protected Term reponseStatementConstruction
```

```
(String focus,TermSet description,String modality ){
 return PATTERN_REPONSE_STATEMENT
.instanciate("type",new WordConstantNode("statement"))
.instanciate("focus",new WordConstantNode(focus))
.instanciate("description",description)
.instanciate("modality",new WordConstantNode(modality));

// Construction des TP Reponse Restiction
protected Term reponseRestictionConstruction
(String nbSolution,TermSequence description,String modality ){
 return PATTERN_REPONSE_RESTRICTION
.instanciate("type",new WordConstantNode("restriction"))
.instanciate("nbSolution",new WordConstantNode("nbSolution"))
.instanciate("description",description)
.instanciate("modality",new WordConstantNode(modality));

// Construction des TP Reponse Relaxation
protected Term reponseRelaxationConstruction
(TermSequence description,String modality ){
 return PATTERN_REPONSE_RELAXATION
.instanciate("type",new WordConstantNode("relaxation"))
.instanciate("description",description)
.instanciate("modality",new WordConstantNode(modality));

// Construction de Answer
protected Term answerConstruction
(TermSet taches,String synch ){
 return GENERIQUE_ANSWER
.instanciate("taches",taches)
.instanciate("synch",new WordConstantNode(synch));

protected TermSequence construireRestrictionSeq
(TermSequence restrictionSeq){
TermSequence restrictionSequenceOutput = new TermSequenceNode(new ListOfTerm());
int cardRestrictionSequence = restrictionSequenceInput.size();
for(int i = 0; (i < NBCRITRESTRICTOUTPUT && i < cardRestrictionSequence); i++) {
restrictionSequenceOutput.addTerm(restrictionSeq.getTerm(i));
}
return restrictionSequenceOutput;
}

protected TermSequence construireRelaxationSeq
(TermSequence relaxationSeq){
TermSequence relaxationSequenceOutput = new TermSequenceNode(new ListOfTerm());
```

```
int cardRelaxationSeq = relaxationSeq.size();
for(int i = 0; i < cardRelaxationSeq); i++) {
relaxationSequenceOutput.addTerm(relaxationSeq.getTerm(i));
}
return relaxationSequenceOutput;
}

protected TermSet construireSolutionSet
(TermSet solutionSet,int nbSol,String tableau){
FunctionalTermParamNode solEncours;
TermSet solutionSetOutput = new TermSetNode(new ListOfTerm());
for(int i = 0; i < nbSol; i++) {
solEncours = (FunctionalTermParamNode)solutionSet.getTerm(i);
setInfoOutput.addTerm(solutionConstruction(solEncours,tableau));
}
return solutionSetOutput;
}

protected Term solutionConstruction
(FunctionalTermParamNode solEncours,String[] tableau){
 Term GENERIQUE_SOLUTION_TEMP =  initGeneriqueSolution();
 for (int i=0;i<tableau.length();i++){
 GENERIQUE_SOLUTION_TEMP
.instantiate(tableau[i],solEncours.getParameter(tableau[i]));
 }
 return GENERIQUE_SOLUTION_TEMP
}

protected Term initGeneriqueSolution
(){
 Term GENERIQUE_SOLUTION_TEMP =  GENERIQUE_SOLUTION;
 Term GENERIQUE_SOLUTION_TEMP
.instantiate(firstname,new WordConstantNode("non_instancie"))
.instantiate(lastname,new WordConstantNode("non_instancie"))
.instantiate(phone,new WordConstantNode("non_instancie"))
.instantiate(mobilePhone,new WordConstantNode("non_instancie"))
.instantiate(fax,new WordConstantNode("non_instancie"))
.instantiate(urd,new WordConstantNode("non_instancie"))
.instantiate(crd,new WordConstantNode("non_instancie"))
.instantiate(labo,new WordConstantNode("non_instancie"))
.instantiate(photo,new WordConstantNode("non_instancie"))
.instantiate(bureau,new WordConstantNode("non_instancie"))
;  return GENERIQUE_SOLUTION_TEMP
}
```

```
}

/******************/
/** CONSTRUCTEUR **/
/******************/
public Choix() {
semanticCapabilities = new ChoixCapacites();
}

/**************/
/** METHODES **/
/**************/
public void setup() {
super.setup();
}

}
```

Table des matières

Remerciements **i**

Sommaire **iii**

Introduction **1**

I Espace-problème : modalités, combinaison de modalités et communication naturelle **7**

1 Espace terminologique : modalité **9**
1.1 Théorie de la communication de Shannon : un canevas intégrateur . . . 10
1.2 Notion de "modalité" en informatique 12
1.2.1 Espace des interfaces selon Frohlich 12
Étude comparative et prise de position 13
1.2.2 Caractérisation des modalités représentationnelles selon Bernsen 14
Étude comparative et prise de position 15
1.2.3 Définitions pour les systèmes multi-sensori-moteurs selon Nigay et Coutaz 17
Étude comparative et prise de position 18
1.2.4 Communication humain-machine et notion de "modalité" selon Martin 18
Étude comparative et prise de position 19
1.2.5 Interfaces multimodales selon Bellik 19
Étude comparative et prise de position 21
1.2.6 Notions de base pour le modèle de référence des systèmes intelligents de présentation multimédia 21
Étude comparative et prise de position 22
1.2.7 Caractérisation des modalités de sortie selon Vernier 23
Étude comparative et prise de position 24
1.2.8 Dialogue humain-machine selon Landragin 24
Étude comparative et prise de position 25
1.2.9 Modèle de la communication selon Clémente 25

Étude comparative et prise de position 27
1.2.10 Composants d'interaction selon Rousseau 27
Étude comparative et prise de position 29
1.2.11 Synthèse : modalité en informatique 29
1.3 Notion de "modalité" dans d'autres domaines 30
1.3.1 Point de vue issu de l'histoire 31
Étude comparative et prise de position 32
1.3.2 Points de vue issus des sciences de l'information et de la communication 32
Étude comparative et prise de position 35
1.3.3 Point de vue issu de l'éthologie 36
Étude comparative et prise de position 38
1.3.4 Points de vue issus de la neurobiologie et des neurosciences . . . 40
Étude comparative et prise de position 43
1.3.5 Points de vue issus de la psychologie 43
Étude comparative et prise de position 46
1.3.6 Points de vue issus des sciences cognitives 47
Étude comparative et prise de position 48
1.3.7 Synthèse : modalité dans les autres domaines 50
1.4 Conclusion : notre terminologie 51

2 Combinaison de modalités 55
2.1 Combinaison des modalités en informatique 56
2.1.1 Combinaison des paradigmes d'action et de conversation selon Frohlich 56
Étude comparative et prise de position 56
2.1.2 Coopération entre modalités selon Martin 58
Étude comparative et prise de position 60
2.1.3 Classification des systèmes multimodaux : CASE, CARE et composition des modalités en sortie 61
2.1.3.1 Classification CASE 61
Étude comparative et prise de position 62
2.1.3.2 Propriétés CARE 62
Étude comparative et prise de position 64
2.1.3.3 Composition des modalités en sortie 64
Étude comparative et prise de position 65
2.1.4 Typologie des multimodalités selon Bellik et application à la multimodalité en sortie selon Rousseau 67
2.1.4.1 Types de multimodalités selon Bellik 67
Étude comparative et prise de position 68
2.1.4.2 Systèmes multimodaux et combinaison des modalités en sortie selon Rousseau 69
Étude comparative et prise de position 70
2.1.5 Synthèse : combinaison de modalités en informatique 70

2.2 Combinaison des modalités dans d'autres domaines 71
2.2.1 Point de vue issu des sciences de l'information et de la communication 71
Étude comparative et prise de position 72
2.2.2 Points de vue issus de la neurobiologie et des neurosciences . . . 72
Étude comparative et prise de position 73
2.2.3 Point de vue issu de la psychologie 73
Étude comparative et prise de position 75
2.2.4 Point de vue issu des sciences cognitives 75
Étude comparative et prise de position 76
2.2.5 Synthèse : combinaison de modalités dans d'autres domaines . . 77
2.3 Conclusion : notre approche de la combinaison de modalités 78

3 Notre approche : vers une communication multimodale naturelle 81
3.1 L'utilisateur et le contexte au centre de la communication naturelle . . . 81
3.2 Communication humain-machine naturelle 83
3.3 Communication naturelle : les approches existantes 87
3.4 Approche choisie et ses hypothèses 89
3.4.1 Caractéristiques et critères visés 89
3.4.2 Hypothèses de travail 90
3.4.2.1 Accessibilités et appropriation 91
3.4.2.2 Adaptation 91
3.4.2.3 Liens entre fond et forme 92
3.5 Limitations du cadre d'étude 93

II Espace-solution : vers le choix conjoint des stratégies de dialogue et de présentation 97

4 Dialogue et interaction multimodale : vers une approche intégrée 99
4.1 Dialogue humain-machine 100
4.1.1 Principes et modèles fondateurs du dialogue humain-machine . . 100
4.1.1.1 Approches du dialogue humain-machine 100
4.1.1.2 Architecture des systèmes de dialogue humain-machine 101
4.1.2 Architectures de systèmes de dialogue multimodaux en sortie . . 103
4.1.2.1 Architecture de TRIPS 104
4.1.2.2 Architecture proposée par Gustafson 106
4.1.2.3 Architecture de MATCH 108
4.1.2.4 Agents rationnels plurimodaux et multimodaux selon Clémente 110
4.1.3 Synthèse : systèmes de dialogue multimodaux 112
4.2 Interaction humain-machine 113
4.2.1 Principes fondateurs de l'interaction humain-machine 113
4.2.1.1 Interfaces WIMP et manipulation directe 114

4.2.1.2 Architecture des systèmes à manipulation directe 115
Modèle de Seeheim 116
Méta-modèle Arch 117
Modèles multi-agents 118
4.2.2 Architectures de systèmes multimodaux à manipulation directe . 120
4.2.2.1 Architecture du W3C pour l'interaction multimodale . . 120
4.2.2.2 Architecture générique des interfaces multimodales selon Carbonell 122
4.2.2.3 Modèle conceptuel WWHT selon Rousseau 124
4.2.3 Synthèse 126
4.3 Vers une approche intégrée de la communication humain-machine 127
4.3.1 Paradigmes dialogique et actionnel : des approches complémentaires 127
4.3.1.1 Du dialogue vers les interfaces à manipulation directe . 128
4.3.1.2 Des interfaces à manipulation directe vers le dialogue . 129
4.3.2 Des interfaces multimédias intelligentes à la présentation multimodale d'information 131
4.3.2.1 Architecture conceptuelle des interfaces intelligentes multimédias selon Roth et Hefley 132
4.3.2.2 Modèle de référence pour les systèmes intelligents de présentation multimédia 135
4.3.2.3 Architecture proposée dans le cadre du projet COMIC . 137
4.3.2.4 Architecture de référence pour les systèmes de présentation d'informations interactifs multimodaux 139
4.3.2.5 Architecture SmartKom 140
4.3.3 Synthèse et positionnement 143
4.4 Conclusion 145

5 Stratégies de dialogue et de présentation : un choix conjoint 147
5.1 Motivations et existant 148
5.1.1 Tenir compte des contraintes de présentation pour répondre aux situations particulières de communication 148
5.1.1.1 Pourquoi la présentation devrait influencer la réaction du système ? 149
5.1.1.2 Contrainte de présentation 151
5.1.2 L'existant 152
5.2 Stratégie de dialogue et stratégie de présentation 153
5.2.1 Stratégie de dialogue 153
5.2.2 Stratégie de présentation 156
5.2.3 Stratégies de dialogue et de présentation au sein du processus de génération multimodale 159
5.3 Composant de choix de stratégies de dialogue et de présentation 160
5.3.1 Pourquoi un composant dédié au choix conjoint de stratégies de dialogue et de présentation ? 161
5.3.2 Architecture globale : extension d'Arch 162

5.3.2.1 Modifications des composants de dialogue et de présentation . . . 164
5.3.2.2 Fonctionnement du composant de choix de stratégies de dialogue et de présentation . . . 165
5.3.2.3 Illustration des modifications des rôles fonctionnels des composants d'Arch . . . 167
5.3.3 Exemple : @mie . . . 169
5.4 Discussion et perspectives . . . 175
5.4.1 Synthèse de la contribution . . . 175
5.4.2 Limites et perspectives . . . 176
5.4.2.1 Couplage du composant de choix de stratégies de dialogue et de présentation à des composants de présentation 176
5.4.2.2 Récupération, sauvegarde et utilisation des contraintes de présentation . . . 177
5.4.2.3 Scalabilité, dynamicité et complétude du composant de choix de stratégies de dialogue et de présentation . . . 179

6 Spécifier les choix de stratégies de dialogue et de présentation 181
6.1 Motivations et existant . . . 182
6.1.1 Expertises en sciences humaines pour la conception . . . 182
6.1.2 Processus incrémental : affinement du choix des stratégies de dialogue et de présentation . . . 183
6.1.3 Outils existants . . . 184
6.2 Expérimentation de référence sur le service Santiago . . . 185
6.2.1 Caractérisation des informations présentées . . . 186
6.2.2 Expérience sur le service Santiago . . . 187
6.2.3 Résultats, discussion et besoins . . . 188
6.3 Éditeur de spécification du composant de choix . . . 189
6.3.1 Choix d'un éditeur graphique . . . 190
6.3.2 Concepts manipulés par l'éditeur . . . 191
6.3.2.1 Notion d'"unité informationnelle" . . . 191
6.3.2.2 Notion de "tâche de présentation" . . . 192
6.3.2.3 Notion de "règle" . . . 194
6.3.3 Processus de conception avec l'éditeur graphique . . . 196
6.3.4 Exemple : @mie . . . 197
6.4 Discussion et perspectives . . . 201
6.4.1 Synthèse de la contribution . . . 201
6.4.2 Limites de la contribution et perspectives . . . 201
6.4.2.1 Caractérisation de la notion d'"unité informationnelle" 202
6.4.2.2 Saillance comme moyen d'assouplissement des règles . . 202
6.4.2.3 Synchronisation des tâches de présentation . . . 203

7 Réalisations logicielles **205**
7.1 Composant de choix de stratégies de dialogue et de présentation 205
7.1.1 Agents JADE-JSA : principes 205
7.1.2 Intérêt du choix d'un agent JADE-JSA pour implémenter un composant de choix de stratégies de dialogue et de présentation . . . 207
7.1.3 Principes d'implémentation d'un composant de choix avec un agent JADE-JSA 208
7.2 Éditeur graphique de spécification du composant de choix 211
7.2.1 Organisation générale de l'éditeur 212
7.2.2 Définition des unités informationnelles 212
7.2.3 Définition du message d'entrée 215
7.2.4 Définition des tâches de présentation abstraite 216
7.2.5 Définition des règles de choix 218
7.2.6 Génération du composant de choix 220
7.3 Plate-forme de simulation du composant de choix 222
7.3.1 Architecture logicielle 222
7.3.2 Fonctionnement 223
7.4 Exemples implémentés 225
7.4.1 Santiago 226
7.4.2 @mie 231
7.4.2.1 Fonctionnement du système initial 231
7.4.2.2 Implémentation du composant de choix 232
7.4.2.3 Composant de choix spécifié avec l'éditeur graphique : @mie 236
7.4.3 Premiers retours sur l'utilisation de l'éditeur graphique par un ergonome 241
7.5 Conclusion 243

Conclusion **245**

Bibliographie **249**

Annexes **263**

Annexe 1 - Caractérisation de la convivialité et de la coopération dans le dialogue **263**

Annexe 2 - Critères d'utilisabilité **265**

Annexe 3 - Extraits du fichier XML pour la simulation d'entrée du composant de choix **271**

Annexe 4 - Extraits du composant de choix généré avec l'éditeur graphique dans le cas du système-exemple @mie **275**

Table des matières **283**

Table des figures **291**

Table des tables **295**

Glossaire **297**

Sigles et acronymes **301**

Table des figures

1 Exemples de sorties multimodales du système @mie 3
2 Schématisation de l'organisation du mémoire 4

1.1 Diagramme schématique d'un système de communication selon la théorie de la communication de Shannon (extrait de [Shannon, 1948]) 10
1.2 Espace des interfaces selon Frohlich (extrait de [Frohlich, 1996]) 12
1.3 Taxonomie des modalités représentationnelles de sortie de Bernsen (extrait de [Bernsen, 1997]) 16
1.4 Des contenus à transmettre aux informations perceptibles dans les systèmes intelligents de présentation multimédia (extrait de [Bordegoni *et al.*, 1997]) 21
1.5 Exemple de diagramme des composants d'interaction dans le cas d'un téléphone portable en réception d'appel (extrait de [Rousseau, 2006]) . . 28
1.6 Diagramme schématique d'un système de communication en sciences de l'information et de la communication (extrait de [Escarpit, 1991]) 33

2.1 Composition des modalités selon Vernier (extrait de [Vernier, 2001]) . . 65
2.2 Types de multimodalité selon Bellik (extrait de [Bellik, 1995]) 68

4.1 Architecture des systèmes de dialogue humain-machine en langage naturel oral 102
4.2 Architecture de TRIPS (extrait de [Allen *et al.*, 2001]) 105
4.3 Architecture des systèmes multimodaux proposées par Gustafson (extrait de [Gustafson, 2002]) 107
4.4 Architecture de l'application MATCH (extrait de [Johnston *et al.*, 2002]) 109
4.5 Architecture des systèmes développés en s'appuyant sur la technologie ARTIMIS 111
4.6 Modèle de Seeheim (extrait de [Green, 1985]) 116
4.7 Méta-modèle Slinky et les modèles Arch dérivés (extrait de [UIMS, 1992]) 118
4.8 Composants principaux d'un système multimodal selon le W3C (extrait de [Bodell *et al.*, 2003]) 120
4.9 Composants intervenant dans la sortie d'un système multimodal selon le W3C (extrait de [Bodell *et al.*, 2003]) 121

4.10 Adaptation d'Arch aux interfaces multimodales selon Carbonell (extrait de [Carbonell, 2005] - "Int." et "Gen." désignent respectivement les interpréteurs et les générateurs dédiés aux entrées monomodales (MIi) et aux sorties monomodales (MOi)) 123
4.11 Étapes de conception d'une présentation multimodale adaptée au contexte d'interaction selon Rousseau (extrait de [Rousseau, 2006]) 125
4.12 Architecture conceptuelle pour les systèmes intelligents de présentation multimédia (extrait de [Roth et Hefley, 1993]) 133
4.13 Modèle de référence pour les systèmes intelligents de présentation multimédia (extrait de [Bordegoni *et al.*, 1997]) 136
4.14 Architecture simplifiée de l'application COMIC (extrait de [Catizone *et al.*, 2003]) 137
4.15 Architecture de systèmes de présentation d'informations interactifs multimodaux (extrait de [Bunt *et al.*, 2005]) 139
4.16 Architecture SmartKom (extrait de [Herzog et Reithinger, 2006]) 141
4.17 Fission multimodale dans SmartKom (extrait de [Herzog et Reithinger, 2006]) 142

5.1 Stratégies de dialogue et de présentation par rapport aux principales étapes de détermination du comportement d'un système multimodal . . 159
5.2 Le composant de choix de stratégies de dialogue et de présentation au sein d'une architecture Arch 163
5.3 Multimodalité en sortie : détails des informations échangées 163
5.4 Exemple de règle par défaut (*i.e.* sans prise en compte de contraintes de présentation) du composant de choix dans le cas où il y a plusieurs solutions 170
5.5 Exemple de règles du composant de choix dans le cas où il y a plusieurs solutions et une prise en compte de la contrainte de présentation auditive ou visuelle de l'utilisateur 171
5.6 Exemple de règles du composant de choix dans le cas où il y a plusieurs solutions, une prise en compte de la contrainte de présentation auditive ou visuelle de l'utilisateur et une prise en compte du nombre d'informations visualisables et audibles 173

6.1 Structure d'une interaction (extrait de [Horchani *et al.*, 2007a]) 187
6.2 Configurations testées (A = auditif ; V = visuel) (extrait de [Horchani *et al.*, 2007a]) 187

7.1 Processus d'interprétation des agents JSA (extrait de [Louis et Martinez, 2007]) 207
7.2 Éditeur graphique pour spécifier le composant de choix 213
7.3 Définition des unités informationnelles 214
7.4 Définition des tâches de présentation abstraites 217
7.5 Définition des règles 219

7.6 Génération du composant de choix de stratégies de dialogue et de présentation . . . 221
7.7 Architecture logicielle de la plate-forme de simulation . . . 223
7.8 Un exemple de dispositif technique pour une expérimentation en magicien d'Oz (extrait de [Fréard, 2006]) . . . 227
7.9 Exemple de simulation dans le cas de l'expérimentation pour le système-exemple Santiago . . . 229
7.10 Exemple de l'interface simulation pour le système-exemple @mie . . . 233
7.11 Formatage du message en entrée du composant de choix généré . . . 237
7.12 Définition de la tâche de présentation abstraite "présenter la liste des critères de restriction" . . . 238
7.13 Règle dans le cas d'absence de contrainte de présentation et un nombre de solutions inférieur à un maximum . . . 238
7.14 Règle dans le cas d'absence de contrainte de présentation et un nombre de solutions trop important . . . 239
7.15 Règle dans le cas d'une contrainte de présentation auditive et un nombre de solutions inférieur à un maximum . . . 239
7.16 Règle dans le cas d'une contrainte de présentation auditive et un nombre de solutions trop important . . . 240
7.17 Règle dans le cas d'une contrainte de présentation visuelle et un nombre de solutions trop important . . . 240
7.18 Tâche de présentation visuelle d'un titre de film pour l'expérimentation en cours suite au couplage du composant de choix de choix avec un émulateur de mobile . . . 243

Liste des tableaux

1.1 Modes humains de sortie (haut du tableau) et d'entrée (bas du tableau) et critères pertinents selon Bellik (adapté de [Bellik, 1995]) 20
1.2 Synthèse et comparatif des termes utilisés pour définir la notion de "modalité" 30

2.1 Paramètres de sélection du mode selon Frohlich (extrait de [Frohlich, 1996] 57

Glossaire

Ce glossaire rassemble de façon classique les termes utilisés de façon récurrente dans ce document. Nous précisons que les définitions proposées sont discutables et résultent de notre appréhension de la communication humain-machine. D'autres définitions existent dans la littérature.

accessibilité sont distinguées (1) l'*accessibilité sensoriactionnelle* qui correspond l'accès sensoriel aux informations et l'accès actionnel sur le système et sur ses fonctions, l'*accessibilité cognitive*, couramment appelée "utilisabilité", dépend de la charge mentale induite par l'utilisation du système et l'*accessibilité rhétorique*, qui peut être rapprochée de la pertinence de la saillance, dépend de l'organisation des informations et de leur mise en avant différente dans la présentation de la réaction du système ainsi que de la suggestion des capacités d'action dont dispose l'utilisateur

adaptabilité/adaptativité adaptation du système initiée explicitement par l'utilisateur. Dans le cas contraire, nous parlons d'adaptativité[2]

Arch architecture de référence pour les systèmes interactifs (*cf.* la section 4.2.1.2) par rapport à laquelle le composant de choix de stratégies de dialogue et de présentation proposé est positionné.

communication (humain-machine) comprend le paradigme dialogique (prédominant dans les systèmes de dialogue), le paradigme actionnel (prédominant dans les interfaces) ainsi que leurs intermédiaires (commes les IMMPS). Nous considérons que, pour une communication humain-machine soit naturelle, les systèmes doivent combiner les paradigmes dialogique et actionnel.

comportement (du système) inclut la réaction du système (détermination du contenu, du fond) et la présentation de cette réaction (détermination de la forme)

composants de présentation incluent, dans Arch, le composant de présentation abstraite qui détermine la présentation abstraite correspondant à la tâche de présentation décidée par le composant de dialogue (cœur du système) et le composant de présentation concrète qui concrétise et rend la présentation abstraite perceptible [UIMS, 1992]

[2]Notons que, pour certains auteurs, la distinction entre adaptabilité et adaptativité renvoie au moment de l'adaptation, à savoir à la conception ou à l'exécution.

convivialité d'après Sadek [Sadek, 1999], la convivialité d'un système dépend de son interprétation en contexte, de sa flexibilité d'interaction, de sa capacité de négociation, de sa capacité à produire des réponses coopératives et de l'adéquation des formes et des styles de ses réponses (*cf.* l'annexe 1).

coopération d'après Caelen [Caelen, 2003], la coopération est une des stratégies de dialogue que peut adopter un système. D'après Sadek [Sadek, 1999], la capacité à produire des réponses coopératives contribue à caractériser la convivialité d'un système.

dialogue (humain-machine) système qui adopte un paradigme de communication dialogique (*cf.* la section 4.1).

effet de modalité en psychologie, terme qui désigne la différence d'intégration et la modification du comportement du sujet en fonction de la modalité de présentation d'une information

comportement inclut la réaction / le fond / le contenu et la présentation / la forme qui composent la sortie d'un système informatique

dispositif physique correspond aux capacités d'action et de perception des systèmes informatiques.

interaction (humain-machine) *cf.* communication humain-machine

interface (humain-machine) système qui adopte un paradigme de communication actionnel (*cf.* la section 4.2).

langage d'interaction correspond à la grammaire dont les symboles terminaux sont produits ou perçus par les dispositifs physiques. Pour les systèmes informatiques, il permet de caractériser le processus de production ou d'intégration du message présenté ou perçu grâce aux dispositifs physiques.

modalité renvoie tour à tour à : (1) les capacités de perception et les capacités d'action humaines ; (2) les couples <dispositif physique ; langage d'interaction> des systèmes informatiques.

modalité (effet de) en psychologie, terme qui désigne la différence d'intégration et la modification du comportement du sujet pour un même ensemble d'informations présentées en fonction de la modalité de présentation utilisée [Lieury, 2005]

multimodal (système) pour nous, c'est un système qui peut, en sortie, utiliser au moins deux langages d'interaction et deux dispositifs physiques et mobiliser deux sensibilités.

sémantique (fission/fusion) fusion ou fission d'unités informationnelles, *i.e.* à un niveau sémantique.

sensibilité inclut la perception sensorielle et l'intégration de cette perception grâce à des mécanismes nerveux et cérébraux.

stratégie de dialogue pour nous, tâche communicative privilégiée par le système. Dans le cas des systèmes d'information coopératifs et et en ne tenant pas compte du méta-dialogue (messages de bienvenue, d'incompréhension, d'aide ...), sont distinguées la stratégie de relaxation, la stratégie d'énumération et la stratégie de restriction (*cf.* la section 5.2.1).

stratégie de présentation pour nous, choix de la présentation (sélection et allocation des modalités) pour un contenu, une réaction donnée (*cf.* la section 5.2.2).

utilisabilité sert traditionnellement à caractériser un système à la fois robuste et souple (*cf.* l'annexe 2). Nous préférons parler d'accessibilité cognitive.

Sigles et acronymes

ACL *Agent Communication Language*

@MIE Annuaire Multimodal Intelligent d'Entreprise

API *Application programming Interface*

ARTIMIS Agent Rationnel à base d'une Théorie de l'Interaction mise en œuvre par un Moteur d'Inférence Syntaxique

BDI *Belief, Intention, Desire*

COMIC *COnversational Multimodal Interaction with Computers*

FIPA *Foundation for Intelligent Physical Agents*

IBM *International Business Machines*

JADE *Java Agent DEvelopment framework*

JSA *Jade Semantics Add-on*

IMMPS *Intelligent MultiMedia Presentation Systems*

MATCH *Multimodal Access To City Help*

MOSTe *Multimodal Output Specification Tool editors*

MOSTs *Multimodal Output Specification Tool simulator*

RIA *Responsive Information Architect*)

SL *Semantic Language*

W3C *World Wide Web*

WIMP *Windows, Icons, Menus, Pointers*

XML *eXtensible Markup Language*

Titre : Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales

Résumé : Cette thèse a pour thème la communication humain-machine multimodale pour des systèmes d'information grand-public. Dans ce contexte, la communication naturelle repose sur l'accessibilité sensoriactionnelle, cognitive et rhétorique aux informations et aux moyens d'action. Pour cela, nous identifions le rôle clef que jouent les stratégies de dialogue et de présentation : 1) La stratégie de dialogue pour des systèmes coopératifs définit la tâche dialogique qui oriente la suite du dialogue et conditionne, voire contraint, le choix du contenu, comme la relaxation, la présentation ou la restriction. 2) La stratégie de présentation définit une configuration multimodale des unités informationnelles à rendre perceptibles par l'utilisateur. Nous prônons le choix concerté de la stratégie de dialogue avec celle de présentation et nous proposons un composant logiciel dédié au choix conjoint des stratégies de dialogue et de présentation au sein de l'architecture logicielle de référence Arch. Ce nouveau composant, intermédiaire entre le contrôleur de dialogue et les composants de présentation concrète, prend en compte les contraintes de présentation, qu'elle soient définies par l'utilisateur ou issues du contexte d'utilisation et/ou d'études ergonomiques et cognitives, pour déterminer la réaction multimodale du système interactif. Outre la réalisation logicielle du composant au sein de deux systèmes, nous proposons un outil de conception destiné à des non-informaticiens qui permet la conception incrémentale et la génération d'un composant de choix pour un système donné grâce à une interface graphique. L'outil est associé à une plate-forme de simulation pour des expérimentations magicien d'Oz.

Mots-clefs : Communication Homme-Machine, Interaction Homme-Machine, Interaction Multimodale, Dialogue Naturel, Stratégie de Dialogue et de Présentation, Architecture Logicielle.

Title : Towards natural human-machine communication : multimodal dialogue and presentation strategies

Abstract : This thesis focuses on multimodal human-computer communication for public service systems. In this context, the naturalness of the communication relies on sensory-motor, cognitive and rhetorical accessibility of the interactive system by the user. Towards this goal of natural communication, we identify the key role of dialogic and presentation strategies : 1) A dialogic strategy of a cooperative information system defines the behavior of the system and the content to convey : Examples include relaxation, statement and restriction. 2) A presentation strategy defines a multimodal presentation specification, specifying the allocated modalities and how these modalities are combined. Highlighting the intertwined relation of content and presentation, we identify a new software component, namely the dialogic strategy component, as a mediator between the dialogue controller and the concrete presentation components within the reference Arch software architecture. The content selection and the presentation allocation managed by the dialogic strategy component are based on various constraints including the inherent characteristics of modalities, the availability of modalities as well as explicit choices or preferences of the user. In addition to the software development of this new component in two systems, we developed a design tool, which allows ergonomists and non-programmers to define and configure the dialogic strategies and to generate the dialogic strategy component of a particular system, as part of an iterative user-centered design process. Moreover this tool is integrated with a simulation framework for wizard of Oz experiments.

Keywords : Human-Machine Communication, Human-Machine Interaction, Multimodal Interaction, Natural Dialogue, Dialogic and Presentation Strategies, Software Architecture.