Université Paris 5 – René Descartes
Institut de Psychologie
Laboratoire d'Ergonomie Informatique

45 rue des Saints-Pères
F-75270 PARIS CEDEX 06

LIMSI-CNRS

BP 133
F-91403 ORSAY CEDEX

N° attribué par la bibliothèque

__ __ __ __ __ __ __ __ __ __

Thèse pour obtenir le grade de Docteur de l'Université Paris 5
Discipline : Psychologie Cognitive – Ergonomie

Présentée et soutenue publiquement par

Stéphanie BUISINE

Conception et Evaluation d'Agents Conversationnels Multimodaux Bidirectionnels

2005

Directeur de Thèse :

Jean-Claude SPERANDIO, Université de Paris 5

Tuteur :

Jean-Claude MARTIN, LIMSI-CNRS

Remerciements

Je remercie vivement tous les membres de mon jury, pour l'attention qu'ils ont portée à mon travail et pour leurs précieuses remarques grâce auxquelles j'ai pu améliorer ce rapport jusqu'au bout. Merci au Professeur Alain Lancry, de l'Université de Picardie, d'avoir présidé ce jury avec bienveillance. Merci à Françoise Détienne, Directrice de Recherche INRIA, et au Professeur Catherine Pelachaud, de l'Université Paris 8, d'avoir accepté de rapporter cette thèse et de m'avoir fait profiter de leurs conseils éclairés. Merci enfin à Marion Wolff, de l'Université de Paris 5, pour son aide et pour la précision de ses remarques.

Mes sincères remerciements au Professeur Jean-Claude Sperandio, pour l'intérêt qu'il a porté à ma thèse et pour avoir accepté d'en être le directeur. Merci également de m'avoir conviée aux séminaires mensuels du Laboratoire d'Ergonomie Informatique, qui ont été très riches et très bénéfiques à mon travail.

Mais c'est à Jean-Claude Martin que je dois cette thèse ! Un grand merci pour toute la confiance qu'il m'a accordée, pour tout le travail qu'il m'a confié, pour le projet NICE… C'était un défi, pour l'un comme pour l'autre, de se lancer dans cette collaboration, et je crois que nous nous en sommes bien sortis, dans la bonne humeur qui plus est (la plupart du temps en tout cas, car j'assume quelques sautes d'humeur…). L'expérience du travail et de la vie au LIMSI ont été extraordinaires pour moi, scientifiquement et humainement. Et puis c'est aussi grâce à Jean-Claude que j'ai connu l'ENSAM, où j'ai été accueillie ensuite.

J'en profite pour remercier Carole Bouchard et Améziane Aoussat, qui m'ont permis d'intégrer l'ENSAM et donc de terminer la rédaction de ma thèse sereinement – ce qui est inestimable ! Je les remercie aussi – ainsi que les autres membres du laboratoire Conception de Produits et Innovation – de m'avoir fait répéter et améliorer ma soutenance. Grâce à eux cela n'a pas été une expérience trop traumatisante.

Merci à Guillaume Pitel et surtout à Sarkis Abrilian, qui ont développé les applications que j'ai étudiées. Sans leur travail ma thèse n'avait aucun sens.

Merci aussi aux partenaires du projet NICE.

Merci à Marianne Najm, Fabien Bajeot et Clotilde Le Quiniou, pour tout le travail abattu et pour la jolie contribution qui en résulte. Merci à nouveau à Marion Wolff, car c'est grâce à elle que j'ai pu bénéficier du travail de ces étudiants et ainsi enrichir considérablement ma thèse.

Merci à Ouriel Grynszpan qui est allé présenter à ma place un exposé à CHI l'avant-veille de ma soutenance…

Merci à Janick Naveteur, de l'Université de Lille 1, qui m'a appris ce que c'était que la Recherche.

Merci aux collègues du LIMSI, à Jean-Paul Sansonnet et à tout le groupe AMI, à Amandine ma très chère « cupine », à Haïfa qui a été ma première amie du LIMSI, à Guillaume pour ses barbecues très populaires et pour toutes les personnes que j'ai rencontrées grâce à lui, à Nicolas pour les débats écologico-équitables dans le RER et pour les ballons éclatés (de rage ou de bonheur) dans le couloir (et dans l'indifférence générale), et à tous les autres…

Merci à mes amies et amis. Merci à Muriel et Sara qui sont un peu mes modèles dans la vie, à Barbara et Virginie qui ont été les amies qu'il me fallait au moment où il le fallait, à Boja pour toutes les nuits passées chez elle, pour la complicité et les habitudes de vieux couple qui se sont installées entre nous. Merci à mon grand ami Etienne.

Merci aussi à ma famille…

Mille mercis enfin à Fred, pour sa contribution multimodale (voire omnimodale). Je me demande bien à quoi ressemblerait cette thèse sans lui. Merci pour son intuition expérimentale et sa sagacité, qui m'ont aidée pour le contenu même de la thèse, merci de m'avoir nourrie et soutenue pendant la rédaction, jusqu'en « phase terminale », merci pour St Auban, merci pour le support logistique, merci pour tout… Et pour tout le reste aussi.

TABLE DES MATIERES

1. INTRODUCTION – CONTEXTE ET OBJECTIFS DE LA THESE 7

1.1. Thème de recherche 7

1.2. Psychologie, Ergonomie, Informatique 8

1.3. Historique du thème dans l'équipe 9

1.4. Cadre applicatif 9

1.5. Problématique de la thèse 10

1.6. Structure du rapport 12

2. LES AGENTS CONVERSATIONNELS ANIMES 15

2.1. Introduction 15

2.2. Agents Conversationnels et Multimodalité 23

2.2.1. Introduction 23

2.2.2. Multimodalité en communication Homme-Homme 25

2.2.3. Multimodalité en sortie chez les Agents 34

2.2.4. Multimodalité en entrée avec les Agents 40

2.3. Evaluation des Agents Conversationnels Animés 44

2.4. Conclusion 59

3. LES INTERFACES MULTIMODALES EN ENTREE 61

3.1. Introduction 61

3.2. Méthodes de conception 64

3.2.1. Principe général de conception 64

3.2.2. Magicien d'Oz 65

3.2.3. Annotation de corpus multimodaux 67

3.2.4. Taxonomies d'intégration des modalités 68

3.3. Comportement multimodal des utilisateurs 71

3.3.1. Interaction Homme-Homme 71

3.3.2. Interaction Homme-Machine 74

3.3.3. Interaction Homme-Agent 96

3.4. Conclusion 97

4. SYNTHESE DE L'ETAT DE L'ART ET PROBLEMATIQUE 99

4.1. Les Agents Conversationnels Multimodaux Bidirectionnels 99

4.2. Le comportement multimodal des Utilisateurs 99
4.3. Le comportement multimodal des Agents 101

5. LE COMPORTEMENT MULTIMODAL DES UTILISATEURS FACE AUX AGENTS . 105
5.1. Magicien d'Oz avec Agents 2D, scénario orienté tâche 105
5.2. Evaluation du prototype I avec Agent 3D, scénario conversationnel 136
5.3. Evaluation du prototype I avec Agent 3D, scénario orienté tâche 157
5.4. Conclusion 161

6. LE COMPORTEMENT MULTIMODAL DES AGENTS 167
6.1. Expérimentation préliminaire 168
6.2. Expérimentation répliquée 182
6.3. Conclusion 206

7. DISCUSSION SUR L'EVALUATION DU SYSTEME DEVELOPPE 209
7.1. Méthodes d'évaluation existantes 209
7.2. Vers de nouvelles méthodes 213
7.2.1. La multimodalité en entrée comme indice d'évaluation 213
7.2.2. L'expression multimodale des émotions 215
7.3. Conclusion 226

8. CONCLUSION GENERALE 229

REFERENCES BIBLIOGRAPHIQUES 233

ANNEXES 249

1. Introduction – contexte et objectifs de la thèse

1.1. Thème de recherche

Ce travail de thèse a été réalisé au LIMSI-CNRS[1] (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur), et financé dans le cadre du projet européen IST-NICE[2] (*Natural Interactive Communication for Edutainement*). La thématique générale concerne les **Agent Conversationnels**, qui sont des personnages virtuels interactifs (partenaires de jeu, tuteurs pédagogiques, assistants d'interface…). Ils sont dits **Multimodaux** lorsqu'ils peuvent utiliser plusieurs modalités de communication : dans le cas qui nous intéresse, ils parlent (synthèse vocale) et peuvent également s'exprimer de façon non verbale (gestes des mains, expressions faciales, postures…).

Pour permettre une meilleure symétrie de l'interaction entre Agents Conversationnels et utilisateurs humains, il semble intéressant d'offrir aux utilisateurs la possibilité de communiquer eux aussi par la parole et les gestes. Si les Agents multimodaux sont capables de traiter plusieurs modalités en entrée du système, la multimodalité devient alors **Bidirectionnelle**.

Le travail effectué au cours de cette thèse est une contribution à la conception et à l'évaluation de tels Agents. Pour cela, nous avons adopté principalement une méthodologie procédant de l'Ergonomie de l'Interaction Homme-Machine et de la Psychologie Expérimentale. Cette thèse s'inscrit donc dans un contexte fortement pluridisciplinaire : Psychologie, Ergonomie, Informatique.

[1] http://www.limsi.fr/

[2] http://www.niceproject.com/

1.2. Psychologie, Ergonomie, Informatique

L'histoire croisée de l'Informatique et de la Psychologie Cognitive témoigne des influences et des inspirations réciproques entre ces deux disciplines. Certains théoriciens tels Shannon & Weaver (théorie de l'information), Turing (machines programmables), von Neumann (systèmes experts), ou McCulloch & Pitts (réseaux de neurones) sont reconnus dans ces deux domaines de recherche. Ce sont en effet les mêmes modèles, mathématiques pour la plupart, qui ont donné naissance, en Informatique, à des architectures matérielles et logicielles performantes et, en Psychologie Cognitive, à des théories explicatives – et prédictives – du fonctionnement du cerveau.

Des liens étroits existent également entre l'Informatique et la Psychologie Ergonomique dans le domaine de l'Interaction Homme-Machine. L'objectif est ici d'optimiser la transmission d'informations entre le système informatique et l'utilisateur, et réciproquement. La Psychologie Ergonomique peut apporter deux types de contributions au domaine de l'Interaction Homme-Machine : une contribution théorique, par des connaissances sur le comportement humain qui peuvent être appliquées à la conception d'interfaces (ex : théorie de la Gestalt, loi de Fitts) ; et une contribution méthodologique, au cours des phases d'évaluation des interfaces (ex : paradigmes de mesure de l'attention, psychométrie).

Un aspect rarement mis en valeur est la contribution que l'Interaction Homme-Machine, en retour, peut apporter à la Psychologie. La confrontation entre utilisateurs et systèmes interactifs est, en effet, susceptible d'enrichir considérablement notre connaissance de l'être humain. Et ce potentiel est encore plus important lorsque c'est un Agent Conversationnel qui habite l'interface. L'Agent étant le miroir de l'humain, sa conception mobilise toutes les connaissances que la Psychologie a accumulées sur le comportement humain, en particulier sur les comportements de communication. Leur utilisation suscite alors de nombreuses questions exploratoires : Comment l'humain réagit-il face à cet Agent ? Comment communique-t-il spontanément ? Comment perçoit-il cet interlocuteur d'un nouveau genre ? En outre, les Agents étant avant tout des programmes informatiques, la possibilité de contrôler précisément leurs comportements offre un moyen intéressant de tester des hypothèses théoriques (ex : observer les effets d'un type de communication particulier), dans une démarche confirmatoire. Le thème de recherche des Agents Conversationnels semble donc constituer un lieu de rencontre extrêmement pertinent entre l'Informatique et la Psychologie.

1.3. Historique du thème dans l'équipe

Le LIMSI-CNRS, où ce travail a été réalisé, offre un cadre institutionnel particulièrement favorable à ce type de projet, puisqu'il n'a cessé, depuis sa création en 1972, de développer la pluridisciplinarité de ses membres et de ses thèmes de recherche. Plus précisément, cette thèse s'est déroulée au sein du groupe Architecture et Modèles pour l'Interaction (AMI), dans lequel collaborent Informatique, Sociologie et Ergonomie. Le thème des Agents Conversationnels est relativement récent dans cette équipe, mais a rapidement suscité un grand intérêt, impliquant aujourd'hui de manière plus ou moins étroite une dizaine de membres du groupe AMI. Le groupe a par ailleurs participé à la fondation d'un groupe de travail pluridisciplinaire regroupant une cinquantaine de personnes issues de la recherche académique et industrielle française (Groupe de Travail sur les Agents Conversationnels Animés[3]).

Cependant, au début de ce travail, le domaine des Agents Conversationnels Animés n'en était qu'à ses balbutiements au LIMSI et dans le groupe AMI, et était accueilli dans le thème « Interaction Multimodale ». Le thème « Agents Conversationnels » a été créé en 2003 comme thème autonome. Ce développement s'est accompli principalement sous l'impulsion de Jean-Claude Martin, Maître de Conférence en Informatique, qui a également co-dirigé l'ensemble de ce travail de thèse. A l'occasion de sa thèse en Informatique, Jean-Claude Martin avait été amené à explorer la littérature en Psychologie pour l'appliquer au développement d'interfaces multimodales. Il a adopté une démarche symétrique pour le présent projet, en proposant le sujet des Agents Conversationnels pour une thèse de Psychologie-Ergonomie.

1.4. Cadre applicatif

Cette thèse a été financée dans le cadre du projet européen IST-NICE, qui nous a fourni un contexte de recherche extrêmement riche et un terrain d'expérimentation stimulant tout au long de ce travail. Ce projet est dédié à l'écrivain danois Hans Christian Andersen (le fameux auteur des « Contes d'Andersen »), à l'occasion du bicentenaire de sa naissance (célébré en 2005). Il a pour objet de développer un prototype de système interactif mettant en scène, sous forme d'Agents Conversationnels Animés, Andersen lui-même et quelques-uns des personnages de ses contes. Ce prototype est destiné à l'interaction conversationnelle (ex : interroger Andersen sur sa vie, son époque) et ludique (ex : s'aventurer avec un des personnages dans l'univers des contes de fée).

[3] http://www.limsi.fr/aca/

En entrée du système, les utilisateurs disposent d'une interface multimodale : ils peuvent parler directement aux Agents, désigner des objets par le geste (à l'aide d'un dispositif du type écran tactile, par exemple), et également combiner ces deux modalités, par exemple en intégrant dans leur discours des références aux objets désignés. Le public visé est celui, notamment, des visiteurs du musée Andersen (à Odense, au Danemark). L'application doit donc convenir à une population large, initialement définie comme celle des enfants et adolescents de 9 à 18 ans.

D'un point de vue scientifique, un des enjeux du projet NICE est d'intégrer les technologies de traitement de la parole et des entrées multimodales au sein d'une application conversationnelle (domaine peu étudié) destinée à un public jeune (population également peu étudiée dans ce champ de recherche).

Les partenaires du projet NICE sont :

- Le NISLab (Natural Interactive Systems Laboratory, University of Southern Denmark) : le NISLab coordonne l'ensemble du projet, ainsi que la réalisation du prototype en langue anglaise. Le NISLab est spécialisé dans le traitement du langage naturel et la gestion de la conversation.
- TeliaSonera R&D (Suède) : TeliaSonera est une société de télécommunications, dont les travaux de recherche concernent notamment le traitement automatique de la parole et du dialogue. Dans le projet NICE, TeliaSonera coordonne le développement du prototype en suédois.
- Liquid Media (Suède), qui est une société de conception de jeux vidéo. Liquid Media réalise les graphismes 3D et leur animation dans projet NICE.
- ScanSoft (Allemagne), qui est responsable de la reconnaissance vocale en anglais et en suédois.
- Enfin, le LIMSI-CNRS, qui assure le développement des modules de reconnaissance et d'interprétation du geste, ainsi que le module de fusion multimodale pour les deux prototypes.

1.5. Problématique de la thèse

Dans le contexte du projet NICE, la thèse s'est orientée plus particulièrement vers l'étude du comportement multimodal (parole et gestes de désignation) des utilisateurs lorsqu'ils interagissent

avec des Agents. Cette problématique est en lien direct avec les développements informatiques confiés au LIMSI, et s'est également avérée très pertinente en termes de recherche :

- D'un point de vue applicatif, l'étude du comportement des utilisateurs permet d'optimiser le développement de l'interface multimodale en entrée, en fournissant les prémices d'un modèle d'utilisation et de combinaison des modalités. Les données comportementales sur les relations temporelles et sémantiques entre les modalités peuvent notamment guider le paramétrage du système et l'élaboration de l'algorithme de fusion multimodale.
- En outre, les études comportementales antérieures sur l'utilisation des systèmes multimodaux en entrée ont été menées avec des interfaces sans Agent, et avec des utilisateurs adultes. L'étude de l'interaction multimodale entre un Agent Conversationnel et un jeune utilisateur est donc susceptible d'apporter des connaissances nouvelles à ce domaine de recherche.

Parmi les variables étudiées dans le comportement spontané des utilisateurs, nous nous sommes intéressés notamment aux relations sémantiques entre les modalités : nous avons constaté que les utilisateurs combinent parole et gestes selon des types de coopération variés. Ce résultat nous a alors amenés à considérer, réciproquement, les relations entre parole et gestes chez les Agents : les types de coopération multimodale ont-ils une influence dans l'autre sens de communication, c'est-à-dire lorsqu'ils sont produits par les Agents ? Une étude de la littérature sur les coopérations sémantiques entre parole et gestes chez les humains, puis chez les Agents, montre que cette problématique a été très peu abordée antérieurement, et n'a jamais fait l'objet d'expérimentations contrôlées. Ce constat nous a poussés à réaliser deux études exploratoires sur ce thème, qui vont au-delà du projet NICE et au-delà de préoccupations applicatives strictes.

La thèse repose donc sur cette symétrie entre analyse du comportement multimodal en entrée (comment les utilisateurs combinent parole et gestes face aux Agents) et évaluation du comportement multimodal en sortie (quelle est l'influence des combinaisons entre parole et gestes des Agents). Les données d'observation sur la multimodalité en entrée nous ont permis de contribuer principalement au niveau de la **conception** des Agents (développement des modules traitant les entrées utilisateurs), et les résultats expérimentaux sur la multimodalité en sortie apportent principalement des connaissances au domaine de l'**évaluation** des Agents.

Enfin, nous discuterons de la réintégration de la multimodalité en entrée et en sortie : le fait d'étudier ces deux aspects séparément a fait naturellement émerger des axes de réunification, notamment sous la forme de méthodes d'évaluation.

1.6. Structure du rapport

Le chapitre 2 présente le domaine des Agents Conversationnels Animés. Dans un premier temps, nous dressons un aperçu des systèmes fonctionnels existants, puis, après avoir introduit la notion de communication multimodale, nous étudions les rapports entre les concepts d'Agents Conversationnels et de multimodalité. Nous présentons des travaux concernant la génération de la communication multimodale de la part des Agents, puis nous nous intéressons aux dispositifs d'entrée permettant aux utilisateurs de communiquer, eux aussi, par plusieurs modalités. Enfin, nous proposons un état de l'art sur l'évaluation expérimentale des Agents Conversationnels.

Le chapitre 3 est consacré au domaine des interfaces multimodales en entrée (sans Agent Conversationnel pour la plupart), qui permettent aux utilisateurs de communiquer par plusieurs modalités naturelles (en particulier la parole et le geste). Nous présentons des exemples de systèmes existants, et de méthodes de conception de ces interfaces. Nous nous intéressons ensuite plus particulièrement aux résultats d'observations sur le comportement multimodal humain : lors de la communication avec un partenaire humain, lors de l'interaction avec une interface multimodale, et enfin lors de l'interaction avec un Agent Conversationnel.

Dans le chapitre 4, nous dressons un bilan de ce double état de l'art : Que connaissons-nous du comportement multimodal des utilisateurs face aux Agents ? Que pouvons-nous anticiper de ce comportement dans une application telle que le projet NICE ? En sortie du système, quelles sont les capacités des Agents en termes de communication multimodale ? Que savons-nous de leur influence sur la qualité de la communication ? Les réponses – et parfois l'absence de réponses – apportées par la littérature à ces questions nous permettent de définir précisément nos axes de recherche.

Le chapitre 5 expose les études que nous avons menées sur la problématique du comportement multimodal des utilisateurs face aux Agents, dans le contexte du projet NICE et dans des contextes approchants. Plusieurs corpus d'interaction multimodale, recueillis auprès de différentes populations, ont été analysés dans le but d'en extraire des patterns comportementaux, qui ont ensuite servi de fondement au développement informatique du système NICE. Ces analyses ont également mis en éveidence des résultats de portée plus large, susceptibles d'être exploités au-delà du projet NICE, dans le domaine des Agents Conversationnels et des interfaces multimodales.

La seconde partie de notre problématique – le comportement multimodal des Agents – est traitée dans le chapitre 6. Plus précisément, nous avons testé l'effet de trois stratégies de coopération entre parole

et gestes chez les Agents – stratégie de redondance, de complémentarité et de spécialisation. Ces expériences ont été réalisées dans un contexte mimant une situation de pédagogie (les Agents présentaient de courts exposés techniques, que les utilisateurs avaient pour but de mémoriser). Les résultats sont largement en faveur de la stratégie redondante (qui va jusqu'à rendre l'Agent plus sympathique aux yeux des utilisateurs), mais font également apparaître des interactions fines avec certaines caractéristiques individuelles des utilisateurs (sexe et personnalité).

Dans le chapitre 7, nous revenons sur nos observations du comportement multimodal des utilisateurs, exposées au chapitre 5, afin d'avancer des hypothèses sur un nouveau mode d'évaluation des Agents. A la lumière des corpus que nous avons analysés, nous pensons en effet que les utilisateurs adaptent leur comportement multimodal aux capacités des Agents, ou à ce qu'ils en perçoivent. Il serait ainsi possible de considérer le comportement multimodal des utilisateurs comme un indice de la qualité du système ou de la perception que les utilisateurs en ont. Par ailleurs, nous proposons d'analyser également une modalité supplémentaire du comportement des utilisateurs : leurs expressions faciales, qui sont en effet susceptibles de porter des informations sur la qualité de l'interaction et de l'application. Nous discutons dans le chapitre 7 quelques éléments de validation de ces modes d'évaluation.

Ce rapport s'achève sur une conclusion générale, au chapitre 8.

2. Les Agents Conversationnels Animés

2.1. Introduction

Les Agents Conversationnels Animés sont des personnages virtuels interactifs : ils se définissent *a minima* par leur présence animée au sein d'un dispositif visuel (le plus souvent un écran), quelle que soit leur apparence (personnage humanoïde, fantaisiste…). Ils ont aussi un rôle communicatif, quel qu'en soit le degré d'élaboration et la modalité d'expression (parole orale, écrite, communication non verbale…). Il semble que le concept d'Agent Conversationnel soit apparu en 1987[4] dans une vidéo promotionnelle de la société Apple, où un utilisateur interagit en langage naturel avec l'Agent Phil (Figure 1), qui l'assiste dans son travail. Le système est fictif, mais cette vidéo anticipe de façon pertinente les réalisations technologiques qui ont vu le jour par la suite. Les Agents Conversationnels existants aujourd'hui sont des prototypes de recherche pour la plupart, mais quelques exemplaires avec des capacités communicatives simplifiées ont investi notre environnement quotidien : c'est le cas par exemple du fameux trombone de Microsoft, ou des assistants web tels que ceux que la société La Cantoche[5] développe pour ses clients.

L'Agent Conversationnel idéal, tel que les chercheurs du domaine l'idéalisent (voir par exemple Isbister & Doyle, 2004), devrait être intelligent, capable de comportements sociaux, et tirer parti de sa représentation visuelle pour renforcer sa crédibilité (notamment par des comportements non verbaux sophistiqués et pertinents, par l'expression d'émotions…). Selon la formule de Ball et Breese (2000) : pour être utiles, les Agents Conversationnels doivent être compétents et rendre des services ; pour être

[4] Si on exclut le domaine de la Science Fiction, où on pouvait croiser des Agents Conversationnels dès les années 60 (voir par exemple Heinlein, 1966).

[5] http://www.cantoche.com/

utilisables, ils doivent communiquer de manière efficace et robuste ; et pour être confortables, ils doivent satisfaire nos attentes sur la manière dont une conversation se déroule. La conception de tels Agents est par nature pluridisciplinaire : les domaines concernés incluent le graphisme, l'animation, l'intelligence artificielle, le traitement automatique de la langue, l'architecture logicielle, le design d'interface, la synthèse vocale, l'ergonomie, la psychologie, la sociologie, les arts du spectacle…

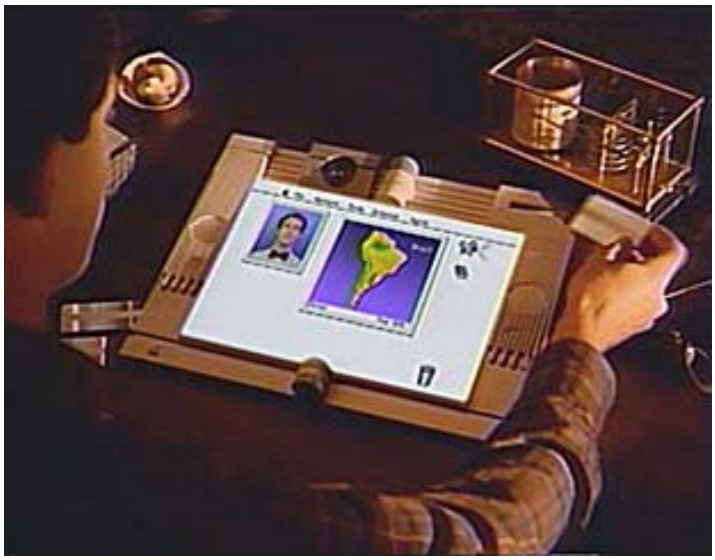

Figure 1 : Le *Knowledge Navigator*, imaginé par Apple en 1987, dont l'interface est munie d'un assistant intelligent, Phil.

Il semble évident que peu d'équipes regroupent une telle diversité de compétences. Les recherches sur les Agents Conversationnels se focalisent donc généralement sur un ou plusieurs points de leur chaîne de conception, faisant plus ou moins abstraction des autres. L'objectif de ce premier paragraphe est de présenter quelques unes des réalisations techniques les plus connues dans la communauté de recherche sur les Agents Conversationnels[6]. Nous recentrerons ensuite notre état de l'art sur la question spécifique de la multimodalité chez les Agents. Enfin, nous traiterons dans un troisième paragraphe de la question de l'évaluation des Agents Conversationnels.

[6] Le domaine des Agents Conversationnels rejoint par certains points celui des Avatars, qui ont cette spécificité de ne pas être autonomes, mais d'être au contraire contrôlés par des utilisateurs. Les avatars sont principalement utilisés dans les jeux vidéo (le joueur guide les actions du personnage grâce à une manette de jeu, par exemple), ainsi que dans les espaces de discussions virtuels, où, dans le but d'augmenter la convivialité, ils permettent de donner une représentation graphique à chacun des utilisateurs prenant part à la discussion. Dans ce chapitre et plus généralement dans l'ensemble de ce rapport, nous excluons les avatars de notre champ d'étude : par Agents Conversationnels, nous désignerons exclusivement des Agents autonomes, qu'ils soient fonctionnels ou simulés, interactifs ou pas, mais en tout cas non contrôlés directement par l'utilisateur.

Applications conversationnelles

En 1994 est présenté le projet *Animated Conversation* (Cassell *et al.*, 1994), dans lequel des Agents en trois dimensions sont capables de converser entre eux de manière autonome (Figure 2). Le système sous-jacent produit, à partir des buts de chaque Agent, le discours verbal de chacun d'eux et y associe automatiquement des intonations, des expressions faciales et des gestes des mains congruents avec le contenu du discours. Ce système permet de générer de courts dialogues entre plusieurs Agents sur des domaines très limités, et n'a pas été conçu pour l'interaction avec des utilisateurs humains.

Figure 2 : Une illustration du projet *Animated Conversation* (Cassell *et al.*, 1994).

Le système REA (*Real Estate Agent*) permet au contraire l'interaction en temps réel entre un Agent Conversationnel et un utilisateur humain (Bickmore & Cassell, 1999). Le contexte d'interaction est un entretien entre l'agent immobilier Rea et un client potentiel. Rea a une apparence humanoïde (Figure 3) lui permettant d'avoir des comportements verbaux et non verbaux (regard, expressions faciales, posture et gestes des mains). Pour renforcer le réalisme de l'interaction, l'Agent et son environnement 3D sont projetés sur grand écran, mettant ainsi Rea à taille humaine. Le domaine d'interaction de Rea (les transactions immobilières) étant fortement lié à une notion de confiance entre Agent et utilisateur, Cassell et Bickmore (2001) se sont interrogés sur la manière d'instaurer un climat de confiance avec l'utilisateur. Ils ont ainsi donné à Rea la capacité, avant d'aborder les questions principales relatives au domaine immobilier, d'engager la conversation sur des sujets légers, de parler de la pluie et du beau temps (*small talks*).

Figure 3 : L'Agent Rea (Bickmore & Cassell, 1999).

Applications pédagogiques

L'enseignement est un champ d'application intéressant pour les Agents Conversationnels, qui sont susceptibles d'introduire un aspect ludique dans l'apprentissage. Stone et Lester (1996) ont été parmi les premiers à développer une application pédagogique fonctionnelle avec un Agent interactif. Le domaine de connaissance est celui de la botanique : le logiciel *Design-A-Plant* enseigne à des élèves (de niveau collège) le développement des plantes, et en particulier les mécanismes de la photosynthèse, dans un contexte constructiviste (les élèves doivent choisir un type de racines, de tronc et de feuilles approprié à l'environnement dans lequel la plante doit se développer). Un personnage 2D (*Herman the Bug*, Figure 4) est constamment présent visuellement : il suit la progression de l'utilisateur et interagit avec les objets 3D (les plantes).

Figure 4 : Deux vues de l'application *Design-A-Plant* et de l'Agent *Herman the Bug* (Stone & Lester, 1996).

Les animations de l'Agent consistent en une série de clips et de répliques verbales qui sont activés soit par une demande d'aide de l'utilisateur, soit par une de ses actions (ex : solution incorrecte), soit par l'absence d'action de l'utilisateur au bout d'une certaine période de temps. Les interventions de l'Agent visent à fournir des conseils appropriés au contexte, en rappelant des principes biologiques (ex : des feuilles plus larges permettent de capter plus de lumière) ou en donnant des recommandations concrètes (ex : dans un environnement sombre, il faut que la plante ait de larges feuilles). Il donne aussi des feedbacks réguliers sur la progression de l'utilisateur. Enfin, sa crédibilité (*believability*[7]) est renforcée par des comportements annexes à sa mission pédagogique (Lester & Stone, 1997) : par exemple des changements de posture (ex : debout, assis, couché), des déplacements à l'écran (ex : en marchant, sautant, volant, en faisant des acrobaties), des mouvements auto-centrés (ex : se gratter la tête ou le dos, nettoyer ses lunettes), etc. A plusieurs reprises, cette application a été testée avec succès par des élèves d'une douzaine d'années (30 utilisateurs mentionnés par Stone & Lester, 1996 ; 100 utilisateurs dans Lester *et al.*, 1997a).

Figure 5 : L'Agent Cosmo et son environnement pédagogique (Lester *et al.*, 1997b).

Lester *et al.* (1997b ; 1999c) proposent une seconde application pédagogique sur l'aiguillage des données sur Internet, où un Agent, Cosmo, est intégré à l'environnement 3D (Figure 5). Cosmo donne des explications techniques sur la transmission de données numériques, propose des exercices aux utilisateurs, contrôle leurs actions et leur donne des conseils. Il est également capable de montrer des réactions émotionnelles, du type « interrogation » quand l'utilisateur sollicite de l'aide, « concentration » quand Cosmo donne une explication, « félicitation » quand l'utilisateur fait un choix

[7] Se définit par le fait que les utilisateurs pensent que l'Agent est doué de sensation, a ses propres croyances, ses propres désirs et sa propre personnalité (Lester & Stone, 1997). Selon Johnson *et al.* (2000), la crédibilité dépend des propriétés visuelles de l'Agent et de la création des comportements en réponse à l'interaction avec l'utilisateur.

correct ou « déception » quand il fait une erreur (Towns *et al.*, 1998). Ces réactions sont produites simultanément sur différentes parties du corps de l'Agent : expression faciale (en particulier par la forme de la bouche et des sourcils), position de la tête, gestes des mains, posture. Cosmo dispose également d'une modalité supplémentaire pour exprimer ses affects : la position de ses antennes (Lester *et al.*, 2000).

Steve (*Soar Training Expert for Virtual Environments*) est un autre Agent pédagogique, techniquement sophistiqué (Rickel & Johnson, 1999) : il habite un environnement 3D dans lequel des étudiants sont immergés (par l'intermédiaire d'un casque de réalité virtuelle). L'objectif de Steve (Figure 6) est de leur apporter des connaissances procédurales, par exemple comment faire fonctionner ou comment réparer des équipements complexes (il a notamment été utilisé dans le cadre d'entraînements pour la marine militaire). Steve est capable de réaliser ces actions lui-même, de répondre aux questions des étudiants et de contrôler leur exécution de ces tâches (les étudiants peuvent agir grâce à une souris 3D et des *datagloves*). Cet Agent est composé de trois modules principaux : un module de perception, un module de cognition et un module de contrôle moteur. Steve peut aussi exprimer des réactions émotionnelles, qui sont liées aux buts qui lui ont été définis (Elliott *et al.*, 1999). Par exemple, un des buts de Steve est de donner des explications sur le domaine considéré. Il manifeste donc un certain enthousiasme lorsqu'il a l'opportunité de discuter un détail intéressant avec un étudiant. Un autre de ses buts est de maintenir l'attention de l'étudiant : il est donc parfois amené à rappeler à l'ordre de manière ferme un étudiant qui ne regarde pas ce que Steve est en train de montrer. Enfin, un autre but de Steve est que les étudiants retiennent ses leçons : il manifeste donc de la joie s'il constate que l'étudiant a retenu un point important ; ou au contraire il apparaît préoccupé s'il constate l'inverse.

Figure 6 : Deux vues de l'environnement de Steve (Johnson *et al.*, 2000).

Steve peut également accueillir des équipes d'instructeurs/élèves : il est capable de coopérer avec d'autres agents virtuels et de gérer le travail d'équipe. Dans l'environnement virtuel, les capacités et l'apparence de tous les Agents sont similaires à celles de Steve (avec des coiffure et couleur de peau éventuellement différentes, des vêtements différents et des voix distinctes ; Figure 6). Les étudiants sont quant à eux représentés par une tête (éventuellement avec leur propre visage) et deux mains.

L'utilisation de Steve, en réalité virtuelle immersive, nécessite un équipement lourd. A l'inverse, l'Agent féminin Adele (Agent for Distance Learning – light edition) ne requiert qu'un navigateur web, et peut donc être utilisée pour l'enseignement à distance, par exemple (Shaw *et al.*, 1999 ; Johnson *et al.*, 2000). Adele enseigne la médecine, en particulier l'analyse de cas cliniques (Figure 7), et propose des exercices aux étudiants, dont elle contrôle et évalue la réalisation. Elle suit du regard le pointeur de la souris, fournit des feedbacks verbaux et non verbaux (hoche la tête ou sourit pour les bonnes réponses, montre de la perplexité quand l'étudiant fait une erreur), commente et conseille. Ce système a été testé avec des étudiants en médecine et dentaire, soit 240 utilisateurs en tout (Johnson *et al.*, 2003).

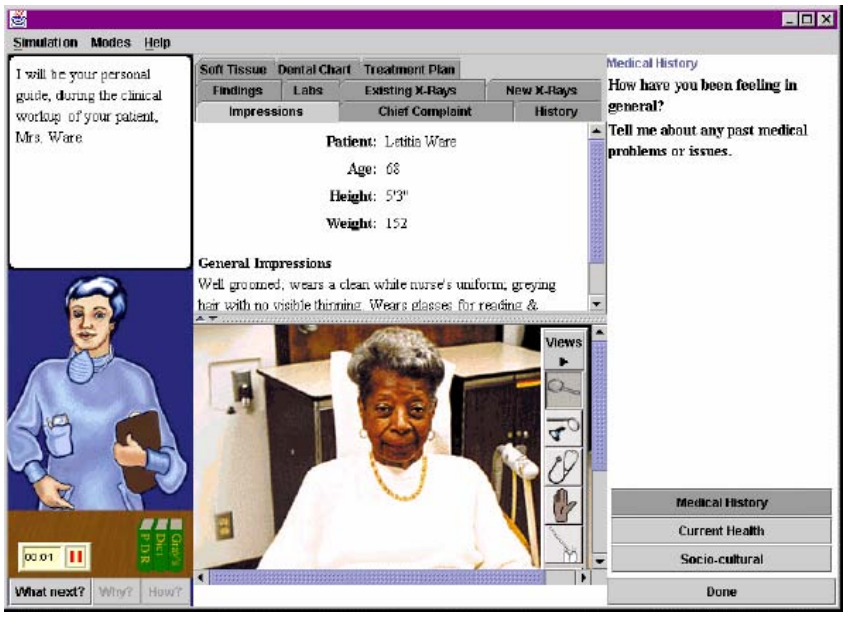

Figure 7 : A gauche l'Agent Adele, commentant une étude de cas (Johnson *et al.*, 2003).

Présentation d'informations

PPP (*Personalized Plan-based Presenter*) Persona (André *et al.*, 1999 ; 2000) est un Agent « de présentation », c'est-à-dire capable de faire des exposés générés dynamiquement. PPP Persona a été appliqué à des situations pédagogiques, mais c'est un Agent plus générique que les précédents. Il a en

effet également été appliqué à des contextes d'assistance technique, ou encore au domaine touristique (présentation d'informations sur Internet). PPP Persona présente verbalement des informations et désigne des éléments graphiques à l'aide d'un pointeur (baguette virtuelle dont la longueur et l'orientation s'adaptent aux éléments devant être désignés, Figure 8). L'apparence de cet Agent n'est pas limitée au personnage 2D présenté dans la Figure 8 : les poses de l'Agent 2D peuvent être remplacées par des photos d'une personne humaine et être animées de la même manière. Le comportement de l'Agent est en partie déterminé par les objectifs de l'application (un certain nombre d'informations doivent être présentées), et en partie par sa propre initiative (ex : déplacements, changement de posture pendant les temps morts). Cette dernière catégorie de comportements est susceptible d'être influencée par la personnalité[8] de l'Agent (à cet égard, André *et al.*, 2000, ont étudié les traits d'extraversion et d'amicalité). Cependant, les capacités d'interaction de PPP Persona sont plus limitées que celles des précédents Agents pédagogiques : les utilisateurs peuvent influencer les sujets abordés (ex : s'ils cliquent sur un élément précis de l'image, PPP s'interrompt pour parler de cet élément), mais ne peuvent pas dialoguer avec l'Agent.

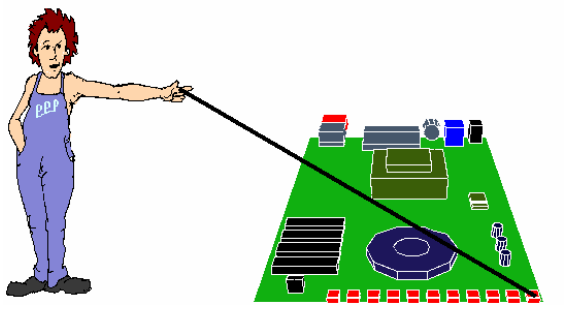

Figure 8 : PPP Persona expliquant le fonctionnement d'un modem (André *et al.*, 1999).

André et Rist (2000) explorent également un autre mode de présentation, dans lequel plusieurs Agents sont simultanément présents dans l'interface et dialoguent entre eux. L'utilisateur assiste à cette interaction, sans y prendre part. Ce type de présentation constitue une alternative aux systèmes interactifs, qui sont complexes à mettre en œuvre et dont la qualité n'est pas toujours garantie. De plus, une équipe d'Agents peut présenter de manière efficace différents points de vue sur une même question, discuter des points que l'utilisateur n'aurait peut-être pas spontanément abordés, etc. André et Rist (2000) appliquent ce concept à une situation de vente de voiture (Figure 9 Gauche) : l'utilisateur peut choisir l'apparence des Agents, leur rôle (vendeur, acheteur) et leur personnalité

[8] La personnalité des Agents est un élément de crédibilité. Les traits de personnalité les plus populaires dans le domaine des Agents sont : l'introversion/extraversion, la dominance/soumission, l'amicalité/hostilité (voir par exemple Ball & Breese, 2000 ; Barker, 2003), ou plus généralement les traits du Big Five (Digman, 1990), également connu sous le terme de modèle OCEAN (*Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*).

(degrés d'amabilité et d'extraversion). La présentation du produit est ensuite générée automatiquement à partir de ces paramètres.

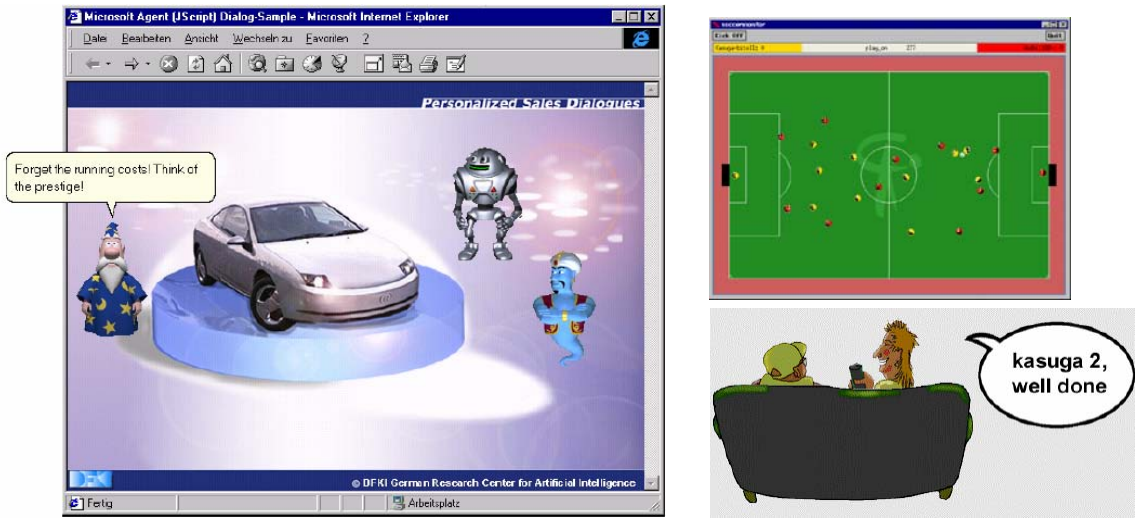

Figure 9 : Présentations sous forme de dialogues entre Agents (André & Rist, 2000).

Dans une autre application, deux Agents commentent un match de football (Figure 9 Droite). Dans ce cas également, l'utilisateur peut définir le rôle de chaque Agent (supporter d'une des deux équipes, neutre) et ajuster sa personnalité (sur les dimensions d'extraversion et d'ouverture d'esprit), ce qui aura pour conséquence d'influencer son discours, la manière dont il s'exprime, et ses réactions émotionnelles au cours du match.

Les quelques Agents présentés dans ce paragraphe ont des caractéristiques très variées, dans leur apparence, leurs fonctions et leurs compétences. Cependant, un trait commun à tous ces exemples est qu'ils peuvent tous utiliser plusieurs modalités de communication : ces Agent parlent et, grâce à leur apparence corporelle, ils peuvent également utiliser la communication non verbale (gestes des mains, expressions faciales, postures…). Dans le paragraphe suivant, nous concentrerons notre exposé autour de cette notion de **communication multimodale** avec les Agents Conversationnels Animés.

2.2. Agents Conversationnels et Multimodalité

2.2.1. Introduction

Dans le cadre de l'Interaction Homme-Machine, la multimodalité caractérise la capacité d'un système à exploiter plusieurs modalités de communication en entrée (c'est-à-dire de l'utilisateur vers le système) et/ou en sortie (du système vers l'utilisateur). Les modalités désignent ici des canaux de

communication utilisés par l'être humain pour transmettre des informations (modalités motrices) et les recevoir (modalités perceptives).

Multimodalité en entrée

Un système multimodal en entrée est donc capable de traiter des informations provenant de plusieurs modalités : par exemple l'utilisateur peut communiquer par la modalité orale (parole), la modalité gestuelle (gestes des mains sur un plan en deux dimensions ou dans un espace à trois dimensions), la modalité oculaire (direction du regard) ou encore corporelle (déplacements, changements de posture).

La multimodalité en entrée présente un intérêt particulier pour les systèmes dotés d'Agents Conversationnels, puisqu'elle assure une certaine **symétrie dans l'interaction**. Nous présenterons dans ce chapitre quelques systèmes associant la multimodalité en entrée et les Agents Conversationnels Animés. Nous nous intéresserons en particulier aux systèmes permettant aux utilisateurs de communiquer par la **parole** et les **gestes des mains**, parce que ces modalités sont les plus courantes dans le domaine des interfaces multimodales en entrée, et parce que ce sont ces modalités qui seront intégrées en entrée du système NICE.

Multimodalité en sortie

Réciproquement, un système multimodal en sortie utilise plusieurs modalités de communication pour transmettre l'information à l'utilisateur : il donne par exemple des informations visuelles (ex : éléments présentés à l'écran, animations) et auditives (ex : sons, synthèse vocale). Certains systèmes donnent également des informations tactiles ou proprioceptives (ex : dispositifs haptiques, retour d'effort).

Dans le cas des Agents Conversationnels Animés, la multimodalité en sortie désigne principalement la production de **comportements verbaux** (synthèse vocale, prosodie) et de **comportements non verbaux**. Ces derniers incluent l'animation de nombreuses parties du corps de l'Agent : expressions faciales[9] (animation de la tête, des yeux, des sourcils, de la bouche, rides…), gestes des bras et des

[9] La génération d'expressions faciales chez les Agents Conversationnels est un domaine de recherche à part entière, qui a suscité de nombreux travaux en graphisme, en animation et en architecture logicielle. Pour un bref état de l'art, voir par exemple Gratch *et al*. (2002) ; pour la génération automatique d'expressions faciales, voir Pelachaud *et al*. (2002) ; pour les comportement oculaires, voir Poggi *et al*. (2000) et Pelachaud et Bilvi (2003) ; pour la synchronisation entre animation des sourcils et intonation de la voix, voir par exemple Krahmer et Swerts (2004) ; pour des travaux détaillés sur les mouvements articulatoires, voir Massaro (2003).

mains, mouvements posturaux[10], déplacements. Traiter du comportement multimodal des Agents dans son ensemble est une tâche considérable. De manière concordante avec le paragraphe précédent (multimodalité en entrée) et avec la notion de symétrie de l'interaction, nous nous intéresserons ici plus particulièrement aux **gestes des mains** et à leurs relations avec la **parole** dans le comportement des Agents.

L'étude de la multimodalité entre parole et gestes chez les Agents ou chez leurs interlocuteurs humains requiert cependant dans un premier temps de poser quelques bases de connaissances sur les relations entre parole et gestes dans la communication Homme-Homme.

2.2.2. Multimodalité en communication Homme-Homme

La littérature en communication Homme-Homme non verbale est extrêmement riche et étendue ; dans ce paragraphe, nous nous limitons à l'exposé de travaux qui sont en lien direct avec la thèse.

Analyse structurelle des gestes

Avant de développer les différentes catégories de gestes, leurs rôles et leurs relations avec la parole, il est important d'introduire les termes classiquement utilisés dans l'analyse des gestes. La réalisation des gestes se décompose en effet en plusieurs phases (McNeill, 1992 ; Kipp, 2004) : la phase dite du ***stroke*** est le moment principal du geste, la phase la plus énergique. La phase de **préparation** précède le *stroke* : elle correspond au moment ou les bras et les mains sont amenés jusqu'à la position de *stroke*. Le geste intègre parfois une phase de **maintien**, qui marque un arrêt juste avant ou juste après le *stroke*. Enfin, le geste s'achève généralement avec la phase de **rétraction**, lorsque bras et mains retournent à leur position de repos. La rétraction d'un geste est parfois confondue avec la phase de préparation du geste suivant. Il arrive également que plusieurs gestes se superposent ou s'enchaînent de telle sorte qu'on parle de **co-articulation** de gestes.

Catégories de gestes

Un des éléments principaux de compréhension de l'articulation entre parole et gestes chez l'humain consiste à observer et à catégoriser tous les types de gestes des mains concomitants à la parole. Ce travail correspond à une analyse fonctionnelle des gestes, par opposition à l'analyse structurelle mentionnée ci-dessus.

[10] Voir par exemple Gillies et Ballin (2003).

Cosnier *et al.,* (1982) proposent en premier lieu de distinguer les gestes **communicatifs** des gestes **extra-communicatifs**. Ces derniers surviennent pendant l'interaction mais ne portent pas de contenu sémantique. Ils peuvent être autocentrés (stéréotypie motrice, etc.), ludiques (manipulation d'objets, etc.) ou être liés au confort (changements de position, etc.). Les gestes **communicatifs** sont quant à eux subdivisés ainsi :

- Les gestes quasi linguistiques : ils portent une connotation affective ou des informations opératoires. Les gestes quasi linguistiques peuvent être conatifs (pour influencer autrui, par exemple en signifiant « stop », « viens »), phatiques (ex : dans les rituels de contact), opératoires, ou bien tenir lieu d'injures.
- Les gestes syllinguistiques, qui comprennent notamment :
    - Les gestes coverbaux :
        - Paraverbaux, liés aux traits phonétiques et syntaxiques (mouvements de la tête et des mains qui soulignent ou scandent le discours),
        - Illustratifs : déictiques, spatiographiques, kinémimiques, pictomimiques.
    - Les synchronisateurs, utilisés par exemple pour la passation de parole :
        - Phatiques (assurent le contact visuel ou corporel),
        - Régulateurs (par exemple un hochement de tête).

La typologie fournie par Scherer (1984) se situe sur un versant plus fonctionnel :

- Les signes non verbaux peuvent avoir une **fonction sémantique** (lorsqu'ils signifient un référent par eux-mêmes ou affectent la signification des signes verbaux concomitants). Cette fonction regroupe les cas suivants :
    - Signification indépendante ou substitution : les signes gestuels et faciaux sont codés de façon univoque et discrète (par exemple quand la communication verbale est impossible).
    - Amplification, illustration ou clarification : parmi les illustrateurs gestuels, on trouve : les bâtons (rythment, accentuent ou soulignent), les idéographes (dessinent le mouvement de la pensée), les mouvements déictiques (désignent un objet présent), les mouvements spatiaux (dépeignent une relation spatiale), les kinétographes (miment l'action corporelle), les pictographes (dessinent un tableau du référent).
    - Contradiction : par exemple dans l'ironie, ou dans les propositions discordantes.
    - Modification de la signification : lorsqu'il n'y a ni contradiction directe ni renforcement de la signification verbale, par exemple dans l'expression d'une certitude ou d'un doute.
- Les signes non verbaux remplissent parfois une **fonction syntaxique** (lorsqu'ils régulent l'organisation des signes verbaux simultanés ou suivants et des autres signes non verbaux) :

    - Segmentation de la chaîne parlée en unités hiérarchiquement organisées (tour de parole, segments de phrases, etc.).
    - Synchronisation des signes verbaux et non verbaux (ex : contour intonatif et longueur de phrase).
- Il existe aussi une **fonction pragmatique** (lorsque la communication non verbale renseigne sur des caractères ou des états des utilisateurs, émetteur ou récepteur) :
    - Fonction expressive : information sur des caractéristiques permanentes (individualité, sexe, âge, appartenance à un groupe social, personnalité) ou des états transitoires (émotions, attitudes, intentions).
    - Fonction réactive : manifestation de réactions très brèves (signaux d'attention, de compréhension, d'évaluation).
- Enfin, les signes non verbaux peuvent remplir une **fonction dialogique** (indication de la relation entre les partenaires dans une conversation) :
    - Relations structurales entre les partenaires (ex : posture corporelle, distance interpersonnelle, contact oculaire, comportement paralinguistique, sont importants pour les dimensions du statut et de la sympathie).
    - Régulation des états du système (début et fin d'une conversation, changement de thème, prises de tour de parole, etc.)

McNeill (1992), quant à lui, fonde son travail sur le continuum de Kendon (1988 ; cité par McNeill, 1992) qui distingue cinq catégories de gestes :

La gesticulation → les gestes langagiers → le mime → les emblèmes → les langues des signes.

A mesure que l'on progresse vers la droite de ce continuum, la présence de la parole est de moins en moins obligatoire, les propriétés langagières des gestes augmentent et leur caractère idiosyncrasique diminue au profit de règles conventionnelles. McNeill (1992) situe son travail à l'extrémité gauche, c'est-à-dire au niveau de la gesticulation : ce terme désigne les mouvements spontanés des mains et des bras accompagnant la parole (ex : la main qui fait un mouvement d'élévation lorsque le locuteur dit « il a escaladé le mur »). Au sein de cette catégorie de gesticulations, McNeill (1992) décrit cinq types de gestes :

- Les gestes déictiques : sont des gestes de pointage vers une entité concrète ou abstraite dans l'espace conversationnel.
- Les gestes iconiques : miment des caractéristiques concrètes du contenu sémantique (ex : forme, taille de l'objet ou de l'action évoquée).
- Les gestes métaphoriques : symbolisent un contenu abstrait du discours (ex : haussement d'épaules pour exprimer l'ignorance) ou illustrent la parole indirectement par un objet imaginaire jouant le rôle de métaphore (voir Kipp, 2004).

- Les gestes de battement : rythment le discours, apportent une emphase sur certains mots.
- Les gestes dits Butterworths, du nom d'un chercheur les ayant étudiés : ils apparaissent en cas de difficulté du canal verbal, par exemple lorsque le locuteur cherche un mot. Cette catégorie de gestes est légèrement différente des quatre précédentes puisqu'elle est produite en l'absence de parole simultanée.

D'autres typologies de gestes existent et ont parfois été proposées en lien direct avec la problématique de la génération de gestes chez les Agents. Par exemple, nous développerons plus loin en 2.2.3 la typologie de Poggi (2002).

Dans la communication Homme-Homme, le geste accompagnant le discours a un rôle facilitateur pour les deux interlocuteurs (Goldin-Meadow, 1999) : il fournit un format de représentation supplémentaire à l'auditeur, lui offrant ainsi un accès aux pensées non verbalisées du locuteur ; mais il fournit également au locuteur lui-même un format de représentation supplémentaire, réduisant son effort cognitif et servant d'aide à la verbalisation des pensées. Par ailleurs, la discordance entre discours verbal et gestuel (en particulier chez les enfants) révèle parfois une instabilité cognitive (un geste illustratif inapproprié peut montrer que l'individu ne maîtrise pas totalement ce dont il parle).

A propos du rôle des gestes pour le locuteur lui-même, nous pouvons également rapporter les observations de Rimé et Schiaratura (1991), qui ont montré que la quantité de comportements non verbaux ne diminue pas lorsque les interlocuteurs ne se voient pas l'un l'autre (comme c'est le cas dans une conversation téléphonique). Ces auteurs ont également constaté que si l'on immobilise la tête, les bras, les mains, les jambes et les pieds du locuteurs, on observe une augmentation de l'activité des sourcils, des yeux, de la bouche et des doigts. De plus, une analyse du contenu du discours montre une diminution de la vivacité de l'imagerie dans une telle situation. Rimé et Schiaratura (1991) parlent de pensée incarnée (*embodied thinking*). Par ailleurs, le taux de gestes des mains semble lié à la complexité du discours verbal : il est plus élevé lorsque les locuteurs doivent décrire un itinéraire complexe, parler dans une langue étrangère, ou sont atteints de pathologies du langage parlé (aphasie).

Par ailleurs, les mouvements des mains font partie intégrante du style communicatif d'un individu : l'utilisation des gestes dépend en partie de la tactique de communication adoptée. Rimé (1985) oppose en cela deux styles de discours, qui sont deux extrêmes d'un continuum : le discours peu éloquent et peu codifié, concret, subjectif, qui est accompagné de gestes iconiques (style direct verbal/gestuel) ; et le discours hautement éloquent, abstrait, objectif, conventionnel, accompagné principalement de gestes de battement (style élaboré). Rimé (1985) prévoit que le style a tendance à être plus direct quand le locuteur est personnellement investi dans le discours, quand ses compétences d'expression sont limitées, quand son interlocuteur lui est très intime, ou quand il est dans un environnement très

familier. Ajoutons qu'il existe des différences d'utilisation des gestes selon les cultures et les langues.

Relations temporelles entre paroles et gestes

Etant donnée la fonction des gestes de battement, qui rythment le discours oral ou mettent l'emphase sur certains éléments, leurs relations temporelles avec le discours ont été étudiées de manière approfondie. Il s'avère en effet qu'une synchronisation s'opère naturellement entre des unités de discours de longueurs variables et les gestes du locuteur. A cela vient s'ajouter, dans le cas du dialogue, une synchronisation de type social dans le rythme du discours verbal et non verbal des interlocuteurs (pour une revue de ces deux types de synchronisation, voir Knapp, 2002).

Les relations temporelles entre modalités ont ainsi été décrites en termes de synchronisation, voire de simultanéité. McNeill (1992) énonce et illustre trois règles à ce propos :

- La synchronie phonologique : dans le cas particulier de l'emphase, il a été observé que le *stroke* du geste (la phase la plus énergique, l'aboutissement) précède ou s'achève sur la syllabe accentuée phonologiquement.
- La synchronie sémantique : si le geste et la parole sont simultanés, ils doivent traiter la même unité de sens (McNeill, 1992, précise en effet n'avoir jamais observé, dans ses expériences, de cas de conflit sémantique entre les deux modalités). Il arrive parfois que le geste anticipe certains éléments du discours : cela peut être le cas lorsqu'un geste unique combine le sens de plusieurs éléments verbaux (par exemple pour la phrase « le bois flambe et la fumée s'échappe », les mains peuvent symboliser le foyer et le flamboiement puis monter et s'écarter pour mimer le mouvement de la fumée ; si ce geste est produit simultanément au segment « le bois flambe », il anticipe le sens du segment « la fumée s'échappe »).
- La synchronie pragmatique : si le geste et la parole sont simultanés, ils remplissent la même fonction pragmatique.

McNeill (1995) décrit une analyse plus détaillée de synchronisation entre les gestes illustrateurs et le discours verbal. Dans le corpus recueilli, des locuteurs devaient réaliser une action (ou imaginer qu'ils la réalisaient) et expliquer en temps réel ce qu'ils étaient en train de faire : l'analyse montre que parfois le locuteur essaie de maintenir synchronisés la parole et le geste, et modifie la parole et/ou le geste en ce sens (ex : suspend son geste le temps de finir sa phrase). D'autres fois, la parole et le geste sont alternés (ex : le sujet prend un tube de la main gauche et le soulève, suspend cette action puis dit « Je prends ceci avec ma main gauche »). Dans ce cas également, il y a modification du geste et/ou de la parole pour respecter ce second pattern de coordination temporelle. McNeill (1995) aboutit donc à un résultat plus nuancé : paroles et gestes s'ajustent l'un à l'autre, mais de manière tantôt simultanée,

tantôt séquentielle. Dans tous les cas, les deux modalités sont intimement liées l'une à l'autre et ne procèdent pas de flux séparés. Le geste n'est pas un accompagnement externe au discours verbal mais l'habite littéralement.

Gogate *et al.* (2000) ont examiné la manière dont les mères, pour parler à leur enfant, synchronisent naturellement leurs paroles et leurs gestes. Les auteurs ont observé vingt-quatre couples mère/enfant (enfants âgés de cinq à trente mois). Les mères avaient pour tâche d'apprendre quatre mots à leur enfant : deux noms de peluches (Gow et Chi) et deux noms d'action n'existant pas dans la langue (« flo » et « pru »). Une analyse fine du comportement des mères montre que :

- Le pattern le plus fréquent d'intégration de la parole avec le geste consiste à nommer (l'objet ou l'action) et à faire un geste (montrer l'objet ou mimer l'action) **simultanément**. Dans cette étude, le critère de simultanéité correspond à un décalage inférieur à 0,5 secondes entre les modalités (décalage moyen de 0,1 secondes).
- La fréquence d'utilisation de ce pattern simultané est d'autant plus élevée que l'enfant est jeune.
- Interrogées en fin de passation, les mères rapportent ne pas être conscientes de ce comportement.

La communication multimodale de la mère semble donc suivre naturellement une stratégie destinée à transmettre les relations mots/référents à l'enfant. La simultanéité temporelle semble être un facteur clé dans l'apprentissage de cette relation. Ce comportement s'atténue avec le temps pour s'adapter à la capacité de l'enfant à saisir cette relation mots/référents.

Relations sémantiques entre parole et gestes

Une partie des relations sémantiques entre parole et gestes en communication Homme-Homme a déjà été abordée dans le paragraphe traitant des catégorisations possibles des gestes. Nous y avons en effet exposé les quatre fonctions sémantiques des gestes vues par Scherer (1984) :

- Signification indépendante ou substitution à la parole,
- Amplification, illustration ou clarification du discours,
- Contradiction,
- Modification de la signification de la parole.

Knapp (2002) développe et illustre quant à lui six types de relations entre parole et gestes :

- Répétition,
- Conflit,

- Complémentarité,
- Substitution,
- Accentuation/modération,
- Régulation.

Les études illustrant l'utilisation de ces relations sémantiques semblent être assez peu nombreuses. Gogate *et al.* (2000), dans l'article précédemment cité sur la simultanéité de la parole et du geste chez les mères, rapportent que plusieurs études ont constaté que la communication de la mère vers l'enfant était typiquement **redondante**. Par exemple, les mères nomment un nouvel objet uniquement en présence de cet objet. Les résultats de Gogate *et al.* sur la simultanéité recouvrent également une certaine notion de redondance : l'apprentissage de la relation mot/référent se fait en nommant l'objet et en utilisant simultanément un geste pour désigner cet objet ou pour l'agiter face à l'enfant. Une troisième modalité (tactile) vient parfois renforcer encore la redondance : les auteurs ont observé que dans 17% des cas les mères touchent les enfants avec l'objet tout en le nommant.

La pédagogie est un domaine dans lequel les caractéristiques de la communication multimodale peuvent aussi être intéressantes. Goldin-Meadow *et al.* (1999) ont ainsi étudié le rôle des gestes des mains dans l'enseignement des mathématiques. Le comportement de huit enseignants donnant des cours particuliers à 49 enfants de 8 à 11 ans a été analysé. Les éléments liés à la leçon étaient relevés et codés en fonction de la modalité dans laquelle ils avaient été exprimés. Les cas de simultanéité entre éléments verbaux et gestuels étaient ensuite classifiés selon deux catégories : **concordance** et **discordance** entre les modalités. L'objet de la leçon était l'équivalence mathématique. Les enseignants devaient donc expliquer aux enfants comment aboutir à une égalité entre les deux membres d'une équation : à cet égard, deux stratégies sont possibles, soit considérer les deux membres de l'équation séparément, soit les considérer d'un seul bloc. Il pouvait y avoir discordance entre paroles et gestes lorsque les enseignants évoquaient l'une des stratégies par la parole et symbolisaient l'autre par leurs gestes. Ainsi, 36% des éléments de la leçon ont été produits par la parole seule ; 7% par le geste seul ; 39% de manière concordante entre parole et gestes et 19% de manière discordante entre les deux modalités. L'effet de ces stratégies (inconscientes) des enseignants sur l'apprentissage a également été évalué : Goldin-Meadow *et al.* rapportent que les enfants saisissent mieux la stratégie expliquée verbalement lorsqu'elle est accompagnée d'un geste concordant que lorsqu'elle n'est accompagnée d'aucun geste. Lorsque le geste est discordant, la compréhension est détériorée par rapport à la condition sans geste. Par ailleurs, les enfants reproduisent la stratégie exprimée par le geste seul dans 20% des cas, ce qui correspond au taux de reproduction des stratégies exprimées par la parole et par un geste discordant. Les gestes en eux-mêmes portent donc une information substantielle que les enfants sont à mêmes de décoder.

Ces résultats montrent que les combinaisons entre parole et gestes peuvent influencer l'efficacité de la communication, dans un sens comme dans l'autre. Le geste a cette particularité de permettre de cristalliser certaines propriétés visuelles ou imagées des concepts. Sa composante visuelle a un fort pouvoir d'illustration du discours et, non seulement elle est perçue par l'auditeur, mais elle doit jouer un rôle non négligeable dans la compréhension et la mémorisation des idées. Goldin-Meadow *et al.* (1999) citent un exemple très éclairant en introduction de leur article : si une phrase du type « la population a augmenté régulièrement » est accompagnée par un mouvement de la main décrivant une augmentation par paliers, c'est le geste et le geste seul qui révèle que le locuteur conçoit l'augmentation comme une fonction discrète et non continue. Il est possible que, dans le cadre d'une leçon, les élèves fixent ce mouvement de la main et l'associent durablement au contenu verbal.

Cassell *et al.* (1999a) ont également testé l'effet des gestes accompagnant le discours, mais avec une procédure différente : ils n'ont pas analysé le comportement spontané du locuteur, mais l'ont au contraire entraîné à le contrôler précisément et à le manipuler. Ce narrateur a en effet été filmé racontant une même histoire dans différentes conditions : certaines parties du discours étaient parfois accompagnées de gestes concordants, parfois de gestes discordants. Par exemple la phrase « Elle lui donne un penny » pouvait être accompagnée d'un geste « donner », où la main du narrateur avance en direction de l'auditeur (geste concordant) ou du geste « prendre », dirigé vers le narrateur lui-même (geste discordant). Huit sujets devaient regarder une de ces vidéos, puis raconter à leur tour l'histoire à un autre sujet. Les résultats montrent que le contenu verbal contient davantage d'inexactitudes sur les éléments qui étaient accompagnés de gestes discordants dans la vidéo initiale. Un troisième type de gestes (nommés gestes de manière par Cassell *et al.*, 1999a) était également testé : ces gestes apportaient une information supplémentaire par rapport au contenu verbal, qui n'était ni concordante ni discordante. Par exemple, la phrase « Il le fait venir auprès de lui » pouvait être accompagnée soit d'un appel de la main, soit d'un geste de saisie. Ce type de geste a également introduit des modifications dans le contenu verbal rapporté par les sujets. Cette étude confirme donc que les auditeurs traitent effectivement l'information portée par le geste du locuteur et modifient leur représentation du discours verbal en fonction de ces gestes.

Cassell *et al.* (2000) ont également étudié les coopérations sémantiques apparaissant naturellement dans le discours. Dans le cadre du développement de l'Agent Conversationnel Rea, qui joue le rôle d'un agent immobilier, ces auteurs ont en effet collecté un corpus d'interaction Homme-Homme dans lequel des sujets devaient visionner une vidéo et un plan d'une maison, puis décrire celle-ci à un autre sujet. La description se faisait sans support visuel, c'est-à-dire qu'il ne pouvait y avoir de gestes déictiques en direction d'éléments graphiques ; les gestes étaient essentiellement iconiques. L'analyse de ce corpus a montré que 50% des constructions multimodales étaient redondantes, et 50% complémentaires (ex : « Il y a un grand jardin » avec un geste circulaire de la main, suggérant que le

jardin entoure la maison).

Piwek et Beun (2002) ont également mené une expérience de communication Homme-Homme avec l'objectif d'obtenir des données pour la spécification d'Agents Conversationnels. Le contexte d'interaction était un jeu de Lego en binôme : un des sujets était l'instructeur, avec un modèle de construction sous les yeux, l'autre était le constructeur devant aboutir au même objet. Les deux acteurs étaient des sujets non informés du projet. Le dispositif était le suivant : l'instructeur et le constructeur pouvaient communiquer par la parole et par des gestes de désignation sur une plateforme. Les deux acteurs n'étaient pas en face-à-face, mais séparés par une petite cloison : les gestes sur la plateforme pouvaient être partagés mais aucun autre élément de communication non verbale (expression faciale, gestes métaphoriques ou iconique, etc.) n'était possible. L'instructeur avait uniquement le droit de parler et de désigner des éléments sur la plateforme ; le constructeur pouvait agir sur cette plateforme. Dix binômes ont participé à cette expérience. Les résultats de Piwek et Beun montrent que 44% des références aux objets ont été accompagnées de gestes de désignation. Les auteurs ont étudié différents paramètres en détail (ex : utilisation d'adjectifs démonstratifs proximaux *vs.* distaux comme *this* et *that* en anglais) mais n'ont pas distingué les références multimodales complémentaires et redondantes. Par ailleurs, ils ont observé que ce sont plutôt les constructeurs qui employaient des gestes de désignation ; au contraire, les instructeurs utilisaient davantage de descriptions verbales (en particulier sur la position des éléments). Piwek et Beun pensent que cet effet est lié à la proximité de chacun des acteurs avec les objets : les gestes de désignation sont plus appropriés pour le constructeur qui est au contact direct des éléments. Ces résultats ont été traduits en algorithmes pour la génération du comportement multimodal (utilisation de déictiques) des Agents Conversationnels.

Conclusion

Dans la mesure où les recherches sur les Agents Conversationnels visent à leur donner un comportement humain, ou tout au moins humanoïde, toutes les données précédentes semblent pertinentes et peuvent leur être appliquées de manière assez directe. Du moins, il semble généralement admis qu'on optimise l'efficacité de ces Agents en leur donnant des caractéristiques conversationnelles observées chez l'Homme réel. En cela, les études relatives à la classification des gestes accompagnant la parole nous fournissent un éventail de tous les comportements gestuels que les Agents Conversationnels devraient, idéalement, être capables de produire. Les études sur les relations temporelles entre parole et gestes suggèrent que la meilleure façon de combiner ces deux modalités est de les rendre simultanées. Enfin, les résultats sur les relations sémantiques donnent des indications sur la fréquence d'occurrence spontanée de certains gestes et de certaines combinaisons parole/geste. Ces travaux soulignent également l'importance des gestes, étant donné les conséquences délétères que peuvent avoir des gestes discordants au contenu du discours.

2.2.3. Multimodalité en sortie chez les Agents

Dans ce paragraphe, nous présentons un état de l'art de l'application technique de certaines des connaissances exposées dans le paragraphe précédent. Nous allons ainsi exposer de manière simplifiée comment certains systèmes génèrent automatiquement des gestes des mains de l'Agent, et comment ils coordonnent ceux-ci avec la parole.

Techniques d'animation 3D. Avant toute chose, précisons qu'il existe trois grandes stratégies d'animation pour les Agents 3D (Hartmann, 2002) :

- La capture de mouvement, qui permet d'animer un Agent en lui transférant les mouvements d'un acteur humain : l'inconvénient de ce mode d'animation est bien entendu l'absence de contrôle sur la génération des gestes de l'Agent.
- L'animation par *keyframes*, dans laquelle un artiste définit manuellement les poses décisives des comportements, et le système fait les interpolations entre ces poses. Cette méthode est plus généralisable que la précédente, mais nécessite tout de même une intervention humaine, avec des compétences artistiques de surcroît.
- Enfin, l'animation procédurale permet de générer des mouvements à partir d'algorithmes. Cette méthode est la plus complexe à mettre en œuvre, mais c'est la plus prometteuse en termes d'interactivité.

Les travaux rapportés dans ce paragraphe concernent cette troisième catégorie d'animation.

Synchronisation des modalités. La recherche sur la génération de comportements non verbaux (faciaux, posturaux, gestuels…) implique généralement de travailler aussi sur leur ajustement au contenu verbal. Le défi que ces systèmes doivent relever est de réussir à synchroniser très précisément les différents canaux de communication : par exemple, selon Cassell (2000), si un hochement de tête est décalé ne serait-ce que de 200 ms par rapport au contenu du discours, sa signification peut en être profondément modifiée[11]. Au niveau architectural, ceci implique que les différents canaux de communication soient générés en même temps, et procèdent donc d'une représentation unique (par exemple, les gestes ne doivent pas résulter du contenu verbal mais doivent être générés conjointement, en amont de la réalisation comportementale).

[11] A l'inverse, Krämer *et al.* (2003) rapportent les résultats d'un test dans lequel il s'est avéré qu'une désynchronisation d'une seconde entre parole et geste n'a pas influencé les évaluations des utilisateurs. Cette étude sera brièvement décrite dans le paragraphe 2.3.

Gestes déictiques. Cosmo, l'Agent pédagogique spécialiste d'Internet (Lester *et al.*, 1997b, 1999b), peut se déplacer dans son environnement 3D et faire des références multimodales (par la parole et par des gestes déictiques) aux objets. Le système dispose d'un modèle de l'environnement qui met à jour en temps réel la représentation de la position de tous les éléments graphiques (y compris l'Agent), et qui connaît aussi leur taille : cette architecture permet à Cosmo de faire des gestes de pointage non ambigus, en se déplaçant jusqu'à l'objet en question si nécessaire. L'Agent est également capable d'inclure dans son discours le démonstratif approprié en fonction de la proximité et du nombre d'objets désignés (*this, these, that, those*). La coordination des modalités se fait ainsi : la préparation du geste de pointage est réalisée pendant le déplacement de l'Agent et le *stroke* du geste coïncide avec la fin du déplacement ; le changement de direction du regard est également synchronisé au geste de pointage, de sorte que l'Agent regarde l'objet qu'il pointe lorsque le *stroke* se produit ; la référence verbale est simultanée au *stroke* du geste de pointage (Lester *et al.*, 1999b).

Steve, l'Agent évoluant en réalité virtuelle immersive, guide l'attention des étudiants et montre la sienne propre par des comportements déictiques utilisant son regard et ses mains (Johnson *et al.*, 2000) : il fait un geste déictique en direction des objets dont il parle, il regarde un objet avant de le manipuler ou de le pointer, il regarde les objets qui sont manipulés par les étudiants, il regarde les objets quand il vérifie leur état, il regarde les étudiants quand il les écoute ou leur parle, il peut aussi suivre du regard un objet en mouvement. Steve se déplace si nécessaire, et est capable d'emprunter le chemin le plus court en contournant les objets sur son passage. Les comportements non verbaux de Steve sont produits de différentes façons : certains correspondent à des actions volontaires générées par son module de raisonnement cognitif (ex : regarder un utilisateur quand il parle ou donner le tour de parole à quelqu'un) ; d'autres sont générés automatiquement par son module de contrôle moteur (ex : regarder un objet avant de le manipuler) ; enfin, certains sont contrôlés de manière externe par le moteur de réalité virtuelle, lorsqu'ils nécessitent la prise en compte des objets de l'environnement (ex : la locomotion et la course des gestes des mains).

Gestes iconiques. Pour générer des gestes iconiques, la plupart des systèmes ont recours à un dictionnaire de gestes (Cassell *et al.*, 1994 ; 2001b ; Hartmann *et al.*, 2002). Pour davantage de souplesse, Kopp *et al*. (2004b) proposent de s'en affranchir : les auteurs observent que dans les dialogues Homme-Homme, les gestes iconiques consistent généralement à donner des caractéristiques visuelles et/ou spatiales simplifiées de certains éléments du discours. De plus, ils identifient un certain nombre de patterns reliant ces caractéristiques visuelles et/ou spatiales avec des formes de gestes (ex : lorsque la main est plate et l'avant-bras vertical, cela réfère dans 80% des cas à un objet ayant un plan vertical ; au contraire, la main plate et l'avant-bras horizontal correspondent dans 85% des cas à une indication de direction). L'architecture proposée par Kopp *et al*. (2004) intègre donc un niveau

intermédiaire d'abstraction permettant d'extraire du processus de génération du langage les propriétés spatiales qui peuvent être représentées par un geste iconique, et d'y associer une forme de geste.

Gestes manipulatoires. Le système PAR (*Parameterized Action Representation*) permet de générer automatiquement les gestes de manipulation des objets de l'environnement (Allbeck & Badler, 2002). Une base de données contient les propriétés de chacun des objets présents (ex : la position, le volume, l'existence et la position de poignées sur cet objet, les actions pouvant être réalisées sur l'objet, les changements résultant de ces actions). Ces connaissances permettent ensuite de guider les comportements des Agents (ex : déplacement, préhension).

Une partie de ces fonctionnalités avaient déjà été implémentées dans le système Steve (Rickel & Johnson, 1999) : les déplacements et les gestes de l'Agent sont en effet calculés en fonction de la position des objets et de leurs propriétés (existence d'une face avant, d'éléments facilitant la préhension, de boutons, d'interrupteurs, de valves, etc.). L'articulation entre explications verbales / gestes déictiques / gestes manipulatoires se fait ainsi : (1) Steve se déplace jusqu'à l'objet dont il doit expliquer le fonctionnement ; (2) il explique verbalement ce qu'il va faire, tout en désignant l'objet de la main et du regard (regard bref en direction de l'objet à pointer, en alternance avec les regards vers la personne à qui Steve parle) ; (3) une fois l'explication verbale achevée, il réalise l'action ; (4) enfin, éventuellement, il explique le résultat de l'action.

Synchronisateurs dialogiques.

Fonction communicative	**Comportement**
Initiation – clôture	
Réaction face à une nouvelle personne	Bref regard à l'autre
Mettre fin à la conversation	Bref regard ailleurs
Dire au revoir	Regard vers l'autre, signe de la tête et de la main
Tours de parole	
Donner la parole	Regard vers l'autre, haussement de sourcils (puis silence)
Demander la parole	Lever les mains
Prendre la parole	Bref regard ailleurs, commencer à parler
Feedback	
Demander un feedback	Regard vers l'autre, haussement de sourcils
Donner un feedback	Regard vers l'autre, signe de la tête

Tableau 1 : Exemples de fonctions communicatives produites par Rea.

Rea (Bickmore & Cassell, 1999 ; Cassell & Bickmore, 2001), l'Agent féminin permettant de converser en temps réel sur le domaine de l'immobilier, a cette particularité de pouvoir gérer la régulation des tours de parole : elle est en effet capable de détecter chez l'utilisateur des signes non verbaux de prise de tour de parole, et de produire elle-même des gestes de synchronisation dialogique (Tableau 1). A noter que Rea a été conçue pour prendre l'initiative dans les dialogues et gère mal les situations où ce sont les utilisateurs qui initient la conversation (Cassell & Bickmore, 2001).

Gestes multiples. Dans le projet *Animated Conversation* (Cassell *et al.*, 1994), où deux Agents dialoguent de manière autonome, les gestes des mains des deux Agents sont générés automatiquement à partir du contenu de leur discours. Plus précisément, ce système permet de produire des gestes **iconiques**, **métaphoriques**, **déictiques**, ou des gestes de **battement**. Ces gestes sont générés très en amont, au niveau du planificateur de dialogue (à partir d'un dictionnaire de gestes), et sont ensuite synchronisés au discours verbal : les battements sont synchronisés aux intonations, les autres gestes sont synchronisés au contenu sémantique. A noter que les gestes de battements peuvent se superposer aux autres, par exemple lorsqu'un geste de battement marquant une emphase verbale montre des propriétés iconiques (forme adoptée par les mains) au moment du *stroke*.

Dans un souci de réutilisation logicielle, Cassell *et al*. (2001b) proposent la boîte à outils BEAT (*Behavior Expression Animation Toolkit*), adaptable à différents Agents, qui permet de générer automatiquement leur comportement non verbal à partir du contenu des répliques verbales. BEAT est fondé sur des règles linguistiques issues de la communication humaine et dispose d'une base de connaissance sur les comportements non verbaux (ex : quels gestes peuvent être associés avec quelles actions). Ce système permet également d'inclure des informations contextuelles (ex : l'historique de la conversation) de manière à affiner le comportement de l'Agent. L'animation qui en résulte combine de manière synchronisée les modalités suivantes : synthèse vocale, **gestes de battement**, **gestes iconiques**, mouvements des sourcils, direction du regard, et intonation de la voix.

Hartmann *et al*. (2002) proposent un moteur gestuel permettant de synthétiser des gestes (**déictiques**, **iconiques**, **métaphoriques**…) à partir de spécifications de haut niveau, afin d'alimenter un dictionnaire de gestes. Par exemple, un geste iconique pour désigner un objet de petite taille peut être généré à partir de l'annotation *small*, sans préciser la forme que la main doit adopter. Ceci permet au système d'animation de créer des transitions fluides entre les gestes consécutifs. Une légère déformation aléatoire (*addnoise*) peut être ajoutée afin d'éviter que les gestes semblent mécaniques et artificiels. Lorsque les spécifications contextuelles sont faibles, le système choisit un simple geste de **battement**. Un même geste peut combiner plusieurs fonctions (ex : celle d'un geste iconique et d'un geste de battement, comme dans le système de Cassell *et al*., 1994). Dans tous les cas, la vitesse des

phases de préparation et de rétraction est ajustée de manière à ce que le *stroke* du geste soit synchronisé avec le flux de la synthèse vocale et les emphases verbales.

Poggi (2002) explicite une procédure permettant au système de déterminer en temps réel quel type de geste générer en relation avec le discours verbal. Cette procédure repose sur une typologie des gestes croisant les dimensions suivantes :

- La relation que le geste a avec les autres signaux : gestes autonomes *vs*. co-verbaux.
- La construction cognitive du geste : gestes codifiés (appartenant à un lexique gestuel, comme c'est le cas pour les **emblèmes**) *vs*. créés en temps réel (comme les **iconiques** et **déictiques**).
- La relation entre le geste et son signifié : relation motivée (la signification du geste peut être immédiatement inférée par l'auditeur) *vs*. arbitraire.
- Le type de signification véhiculée : qui varie en fonction des croyances, des buts et des émotions du locuteur.

La typologie résultante est traduite en un algorithme que le système parcourt pour décider quel geste produire. Par exemple, s'il n'existe pas de geste codifié pour un signifié donné, que ce signifié n'est pas présent dans le contexte mais qu'il existe un objet O lié à ce signifié, alors le geste produit sera un déictique dirigé vers l'objet O (Poggi, 2002).

Expressivité des gestes. Comme nous l'avons vu plus haut, le système PAR (Allbeck & Badler, 2002) génère les gestes manipulatoires à partir d'une base de données des propriétés des objets. Les Agents étant eux-mêmes des objets possédant des propriétés particulières, les auteurs proposent de généraliser le système PAR à la modélisation des émotions et de la personnalité des Agents, de sorte que celles-ci transparaissent dans leurs comportements et dans la manière dont ils réalisent les actions.

De même, Ruttkay et Pelachaud (2002) proposent un modèle permettant de faire varier le **style des gestes** produits par un Agent, en fonction de facteurs tels que : la culture assignée à l'Agent, son âge, sa personnalité, sa catégorie socioprofessionnelle, le contexte situationnel (ex : émotion), etc. Le fait d'attribuer des valeurs à ces paramètres détermine l'intensité des gestes, leur vitesse, leur enchaînement (ex : fluide, fébrile, anguleux).

Le système EMOTE (*Expressive MOTion Engine*) proposé par Chi *et al*. (2000), vise à contrôler l'expressivité des gestes. EMOTE repose sur le modèle LMA (*Laban Movement Analysis*) qui permet de décrire les mouvements selon cinq éléments : le corps (parties du corps mises en jeu), l'espace (position, direction, trajectoire), la forme du mouvement, l'effort déployé pour son exécution, et la relation (définie par les modalités utilisées). Parmi ces notions, Chi *et al*. (2000) ont retenu celles de

forme et d'effort pour les implémenter dans EMOTE. Les paramètres définissant la forme sont appliqués sur les *keyframes*, c'est-à-dire les poses décisives des gestes. Les paramètres d'effort sont au contraire appliqués au niveau des interpolations entre *keyframes* successives. L'effort est défini sur quatre dimensions : l'espace (mouvement direct / indirect), le temps (soudaineté), le poids (force, rendu par l'accélération du mouvement) et la fluidité (degré de contrôle exercé sur le mouvement).

Hartmann *et al*. (soumis) contrôlent l'expressivité des gestes par six échelles de valeurs :

- Le niveau d'activation général (quantité de mouvements associée à un tour de parole),
- L'amplitude spatiale des gestes,
- Leur amplitude temporelle (durée),
- La fluidité de leur enchaînement,
- Leur intensité (douceur *vs*. puissance, tension),
- La répétition d'un même mouvement.

Le système traduit ces dimensions en paramètres d'animations. Le geste consécutif respectera les contraintes biomécaniques et articulatoires du corps humain, ainsi que la loi de Fitts (1954), qui lie l'amplitude spatiale et temporelle (ainsi que la précision du geste) dans les mouvements humains. Lorsque aucune contrainte n'est posée sur la forme de la main (par exemple dans les gestes de battement), celle-ci participe au rendu de l'intensité du geste (main relaxée pour intensité faible, main tendue pour intensité forte). Deux études perceptives (Hartmann *et al.*, soumis) ont été menées avec un total de 106 sujets, dans le but (1) de valider ces dimensions d'expressivité une à une et (2) de tester la capacité du modèle à produire une impression cohérente à partir de la combinaison des paramètres. Les résultats sont globalement positifs, et permettront en outre de continuer à raffiner ce modèle, puisqu'il est apparu que certaines dimensions sont moins bien reconnues que d'autres, et engendrent parfois quelques confusions.

Coopération sémantique. La question des relations sémantiques entre parole et gestes des mains n'étant pas gérée par le système *Animated Conversation* (Cassell *et al.*, 1994), Cassell et Prevost (1996) proposent de l'améliorer en déterminant, au niveau de la planification de la phrase, la répartition des éléments sémantiques à transmettre par chaque modalité : dans le cas d'une emphase verbale (marquée par un changement d'intonation), celle-ci est renforcée par un geste **redondant**. Pour les phrases sans emphase verbale, les éléments sémantiques sont répartis entre parole et geste de manière **non redondante**.

Le système Rea (Bickmore & Cassell, 1999 ; Cassell & Bickmore, 2001) repose sur une modélisation des comportements humains dans un contexte similaire à l'application (transactions immobilières ; voir l'étude de Cassell *et al*., 2000 exposée au paragraphe 2.2.2). Au niveau des combinaisons entre

parole et gestes des mains, Rea est capable de produire de la **redondance** et de la **complémentarité**. Elle sait aussi répartir de manière appropriée les informations à transmettre sur ces deux modalités (ex : certaines informations spatiales sont plus aisées à transmettre par un geste que par une explication verbale). Cassell (2001) indique que la possibilité de contrôler la redondance et la complémentarité constitue un progrès significatif par rapport aux Agents de Lester *et al*. (1999b), Rickel et Johnson (1999), André *et al*. (1999). Ce raffinement est en effet intéressant en soi et répond à un besoin de réalisme (puisqu'il repose sur un modèle de communication humaine), mais le gain en réalisme est-il perçu par les utilisateurs ? La communication est-elle améliorée de ce fait ? Etc. Malheureusement, ces questions n'ont pas fait l'objet d'évaluations du comportement multimodal de Rea.

Conclusion. Les systèmes existants ont atteint un niveau de sophistication remarquable dans la génération des gestes associés au discours des Agents. La problématique de la coordination temporelle et, dans une moindre mesure, celle de la coordination sémantique avec la parole, ont également fait l'objet de progrès importants. Cependant, ces systèmes ont trop peu été mis à l'épreuve de l'interaction et de l'évaluation avec des utilisateurs. Dans la troisième partie de ce chapitre, consacrée à l'évaluation des Agents Conversationnels, nous montrerons qu'il subsiste un manque de données expérimentales sur la multimodalité parole/geste produite par les Agents Conversationnels.

2.2.4. Multimodalité en entrée avec les Agents

Comme nous l'avons déjà souligné, la multimodalité en entrée présente un intérêt particulier pour les Agents Conversationnels : elle assure en effet une symétrie au moins partielle dans l'interaction, puisque l'Agent capable de comportements verbaux et non verbaux peut aussi recevoir les comportements verbaux et non verbaux des utilisateurs. Cassell *et al*. (1999b) incluent même la multimodalité en entrée dans leur définition des Agents Conversationnels Animés : selon eux, les Agents doivent avoir les mêmes propriétés que les humains lors d'une conversation en face-à-face, et notamment la capacité de reconnaître et de répondre aux entrées verbales et non verbales, la capacité de générer des sorties verbales et non verbales, l'utilisation de fonctions conversationnelles comme les tours de parole et les feedbacks, et la capacité de contribuer au discours.

Le système de Nagao et Takeuchi (1994a) est l'un des tous premiers à déployer la multimodalité en entrée. Il met en scène une tête parlante 3D pour l'interaction sociale avec un ou plusieurs utilisateurs (Figure 10). Le système est capable de détecter la position de chacun des interlocuteurs (reconnaissance par caméra) et de traiter leurs entrées verbales (micro casque). L'Agent s'oriente en

conséquence (regarde la personne en train de parler), et participe à la conversation par la parole et les expressions faciales.

Figure 10 : Deux utilisateurs en interaction sociale avec une tête parlante 3D (Nagao & Takeuchi, 1994a).

Gandalf (Cassell & Thorisson, 1999) est une tête parlante munie d'une main 2D (Figure 11). Comme le système précédent (Nagao & Takeuchi, 1994a), il est relativement limité comportements communicatifs en sortie (en particulier sur les aspects prosodiques et non verbaux). En revanche, il est capable de reconnaître et de traiter de nombreux signaux multimodaux de la part de l'utilisateur (paroles, mouvements de la tête, direction du regard, gestes de pointage). Un inconvénient de ce système est l'équipement que les utilisateurs doivent revêtir (capteurs sur la tête, sur les mains, devant les yeux, micro), et dont on peut supposer qu'il perturbe le naturel de l'interaction.

Figure 11 : Un utilisateur face à l'Agent Gandalph (Cassell & Thorisson, 1999).

Rea est également capable de comprendre et d'interpréter le comportement multimodal des utilisateurs, mais avec un équipement beaucoup moins contraignant (Figure 12). Les utilisateurs portent simplement un micro casque (détection et interprétation de la parole et de la prosodie) ; les autres modalités (gestes 3D, position et orientation de la tête) sont détectées par deux caméras placées au-dessus de l'écran de projection. Par exemple, Rea peut détecter chez l'utilisateur des signes non verbaux de prise de tour de parole (gestes faits par l'utilisateur pendant que Rea parle) et ainsi s'interrompre pour laisser l'utilisateur intervenir. Rea étant un système spécifiquement conversationnel, son architecture lui permet en priorité de traiter les comportements visant à la régulation des tours de parole, ainsi que les gestes de battement marquant des emphases (Cassell *et al.*, 1999b ; 2001a). Certains gestes déictiques sont également détectés et, s'ils sont produits simultanément à la parole, leur traitement peut être fusionné avec celui du contenu verbal (Cassell, 2001).

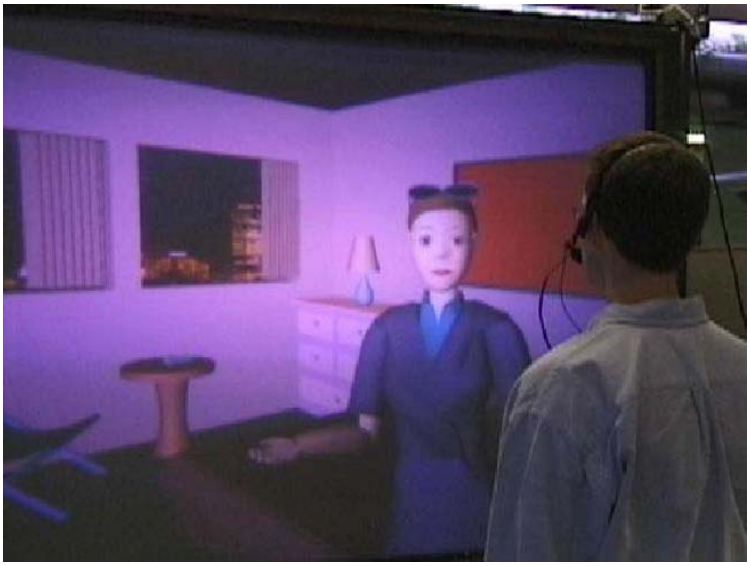

Figure 12 : Un utilisateur face à l'Agent Rea (Bickmore & Cassell, 1999 ; Cassell & Bickmore, 2001).

L'immersion dans le système Steve est assurée grâce à un casque de réalité virtuelle (Figure 13). L'utilisateur est muni de capteurs sur la tête et sur les mains, peut parler pour interagir avec les Agents, se déplacer dans l'environnement et manipuler des objets grâce à une souris 3D et des *datagloves*.

Max – *Multimodal Assembly eXpert* (Kopp *et al.*, 2004a) – est également un Agent d'environnement immersif (Figure 14). Les utilisateurs sont munis de lunettes stéréoscopiques, et leurs gestes et postures sont détectés à l'aide de nombreux capteurs (disposés sur le torse et les bras) ainsi que de *datagloves*. Max réalise des tâches collaboratives (de type construction d'objets) avec les utilisateurs :

il est capable non seulement de détecter et d'interpréter les gestes de son partenaire humain, mais encore d'imiter ces gestes (Kopp *et al.*, 2004a).

Figure 13 : Une utilisatrice du système Steve (Rickel & Johnson, 1999).

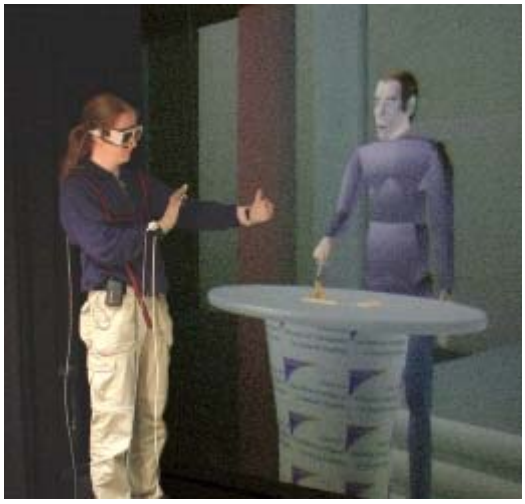

Figure 14 : L'Agent Max face à un utilisateur (Kopp *et al.*, 2004a).

Enfin, SenToy (Prada *et al.*, 2003) n'est pas un système multimodal, mais il offre un moyen d'interaction innovant (une interface dite tangible) avec les Agents : les utilisateurs manipulent une poupée (Figure 15) pour contrôler un Avatar présent à l'écran, qui doit affronter un Agent contrôlé par le système. Les gestes de la poupée induisent des réactions émotionnelles chez l'avatar présent dans le jeu (ex : placer les bras de la poupée devant son visage provoque une réaction de peur). Il existe aussi d'autres systèmes (Höysniemi *et al.*, 2005), dans lesquels un avatar reproduit les mouvements faits par l'utilisateur lui-même, avec son corps tout entier (les mouvements de l'utilisateur étant capturés par caméra) : il est ainsi possible par exemple de faire courir, sauter, ou nager l'avatar.

Figure 15 : La poupée de SenToy (à gauche) et deux utilisateurs de ce système (à droite).

Conclusion. Il existe aujourd'hui des systèmes avec Agents Conversationnels extrêmement sophistiqués. Nous avons vu que la génération de comportements multimodaux chez les Agents avait fait l'objet de nombreux travaux, couvrant une partie importante de l'éventail des comportements communicatifs humains. Il apparaît également que certains systèmes réalisent la bi-directionnalité de la communication multimodale, en autorisant des entrées verbales et gestuelles de la part des utilisateurs. Il semble cependant exister un déséquilibre dans la manière dont les problématiques de la sortie et de l'entrée multimodale sont traitées : alors que la multimodalité en sortie fait l'objet de développements approfondis, le fonctionnement du système en entrée est très rarement décrit, outre l'énumération des dispositifs physiques à disposition de l'utilisateur. A l'issue de ce paragraphe, il nous semble donc que les recherches sur la conception d'Agents Conversationnels Animés se soient davantage focalisées sur les comportements en sortie. Dans le paragraphe suivant, nous verrons si ce déséquilibre se retrouve dans les recherches sur l'évaluation des Agents Conversationnels.

2.3. Evaluation des Agents Conversationnels Animés

Le développement et l'évaluation des Agents Conversationnels Animés sont deux problématiques relativement découplées. La plupart des systèmes précédemment exposés dans ce chapitre ont été développés sur la base de principes généraux, souvent transposés directement de la communication Homme-Homme. Certains d'entre eux ont été évalués à titre de validation, c'est-à-dire mis à l'épreuve de l'utilisation par des individus naïfs, en fin de chaîne de conception. Encore plus rares sont les prototypes qui sont ensuite devenus des systèmes réellement opérationnels, exploités régulièrement et par des utilisateurs variés. Ces évaluations et données d'utilisation permettent de mettre en évidence

(et par conséquent de corriger) des erreurs de conception, des dysfonctionnements techniques ; elles permettent également de recueillir des données de performance, d'efficacité du système, ainsi que des données subjectives émanant des utilisateurs (ex : avis sur l'apparence de l'Agent, sur le mode d'interaction). Cependant, ces études renseignent davantage l'évaluation de chaque système singulier (avec ses composants sous-jacents propres) que la problématique générale de l'évaluation des Agents.

En effet, il va plutôt être question dans ce paragraphe d'études expérimentales cherchant à mettre en évidence les effets des Agents (effets de leur présence, de leurs caractéristiques graphiques, de leur comportement…) sur l'efficacité d'un système, et/ou sur les impressions subjectives des utilisateurs. De telles questions requièrent toujours une comparaison entre différentes conditions (ex : un même système avec ou sans Agent ; un même système avec des Agents d'apparences différentes, de comportements différents). En revanche, ces questions ne nécessitent pas forcément un système fonctionnel sur le plan technique : des expérimentations très pertinentes peuvent être menées avec des Agents ayant des capacités d'interaction restreintes, ou bien encore entièrement simulés. Les résultats de ces expérimentations fournissent des indications qui sont ensuite exploitables en amont de la conception de nouveaux Agents.

Dans la suite de ce chapitre, nous exposons de manière synthétique des résultats expérimentaux qui ont été obtenus en confrontant des utilisateurs humains avec des Agents Conversationnels. Le long Tableau 2 (réparti sur les pages suivantes) est structuré en cinq parties, qui distinguent les études menées sur différentes problématiques : les effets de la présence d'un Agent, les effets de l'apparence des Agents, les effets des modalités utilisées, les effets des comportements non verbaux, et enfin les effets de la personnalité ou du rôle des Agents. En complément du présent paragraphe, le lecteur pourra se reporter à la revue de question menée par Dehn et van Mulken (2000) sur l'évaluation des Agents Conversationnels.

Effets de la présence d'un Agent				
Référence	**Système**	**Objet du test**	**Protocole**	**Principaux résultats**
(Van Mulken *et al.*, 1998)	**Agent** 2D cartoon PPP Persona. Pas d'**entrée** (tâche passive). Système **fonctionnel**.	- Avec Agent *vs*. sans Agent. - Présentation technique *vs*. non techniques.	**Application** : Pédagogie. **30 utilisateurs** **Variables** : Performance de rappel et de compréhension ; Questionnaire.	L'Agent améliore les évaluations subjectives pour la présentation technique (facilité, amusement).
(Walker *et al.*, 1994)	**Agent** : Tête parlante 3D réaliste. **Entrée** : Clavier, souris. Mais **pas d'interaction** avec l'Agent.	- Texte écrit ; - Agent (expression neutre) + audio ; - Agent (expression sévère) + audio.	**Application** : Questionnaire d'enquête. **49 utilisateurs** **Variables** : Questionnaire.	La situation de réponse aux questions posées par les Agents est évaluée moins positivement que celle de répondre à la version textuelle.
(Lee & Nass, 1999)	**Agents** 2D, animation pauvre. **Entrée** : clavier, souris. Mais système en réalité **non fonctionnel**.	- Boîte de dialogue *vs*. Agent. - 1 *vs*. 4 interlocuteurs.	**Application** : Résolution de problème, stratégie. **48 utilisateurs** **Variables** : Prise de décision ; Questionnaire.	Les boîtes de dialogue induisent plus de comportements de conformisme (l'utilisateur se laisse influencer par son (ses) interlocuteur(s)) que les Agents.

Référence	Système	Objet du test	Protocole	Principaux résultats
(Rickenberg & Reeves, 2000)	**Agent** 3D cartoon. **Entrée** : clavier, souris. Mais **pas d'interaction** avec les Agents.	- Agent inquisiteur ; Agent indifférent ; pas d'Agent. - Utilisateurs avec *locus of control* interne (croyance que les événements de la vie peuvent être maîtrisés) *vs*. externe (croyance qu'ils échappent à tout contrôle).	**Application** : Navigation E-commerce. **84 utilisateurs** **Variables** : Performance ; Questionnaires (anxiété et évaluation des sites).	Présence d'un Agent est anxiogène. Agent inquisiteur plus anxiogène que Agent indifférent. Sujets externes sont plus sensibles à ces effets. Agent inquisiteur fait diminuer la performance mais augmente la confiance dans les sites web.
(Oviatt & Adams, 2000)	**Agents** 2D, animaux. **Entrée** : Parole + gestes 2D (stylo). Système *I SEE!* (**simulé***).	Conversation avec un Agent *vs*. avec un adulte.	**Application** : Pédagogie (biologie marine). **10 utilisateurs** (6 à 10 ans). **Variables** : Analyse des incorrections langagières.	Face aux Agents, le langage est plus répétitif mais les incorrections sont moins fréquentes.
(Moreno *et al.*, 2000)	**Agent** 2D fantaisiste (Herman the Bug). **Entrée** : souris. Système **fonctionnel**.	Avec Agent *vs*. sans Agent (texte).	**Application** : Pédagogie (botanique). **44 utilisateurs** (étudiants à l'université). **Variables** : Rappel, Transfert d'apprentissage, Questionnaire (motivation, facilité, sympathie…).	La présence de l'Agent n'a pas d'effet sur le rappel mais augmente le transfert d'apprentissage. L'Agent augmente aussi la motivation et l'intérêt des étudiants.
(Moreno *et al.*, 2001)	L'expérience précédente (Moreno *et al.*, 2000) est répliquée (même application, même protocole) avec de jeunes élèves (**48 utilisateurs** d'environ 12 ans). Les résultats sont eux aussi exactement répliqués.			

Référence	Système	Objet du test	Protocole	Principaux résultats
(Moundridou & Virvou, 2001)	**Agent** 3D animal (perroquet). **Entrée** : clavier, souris. Système **fonctionnel**.	Agent instructeur, Agent étudiant, pas d'Agent.	**Application** : Exercices de maths. **28 utilisateurs** (étudiants) **Variables** : log files, questionnaire, résolution des exercices.	La version Agent étudiant augmente l'amusement et la facilité perçue par rapport à la version sans Agent. Pas de différence entre les versions Agent étudiant / Agent instructeur.
(Ryokai *et al.*, 2003)	**Agent** 3D cartoon, enfant. **Entrée** : Parole, gestes sur un jouet (figurines dans un château). Système **simulé***.	- Jeu avec *vs.* sans Agent ; - Jeu avec *vs.* sans partenaire (enfant de même âge).	**Application** : Jeu et interaction verbale (narration d'histoires courtes). **28 utilisatrices** (filles de 5 ans). **Variables** : Complexité des histoires des enfants (compléments circonstanciels de temps et de lieu, discours direct des protagonistes).	La présence de l'Agent comme partenaire de narration augmente significativement les indices de complexité des histoires racontées par les enfants. Le fait que l'Agent lui-même raconte les histoires de façon décontextualisée permet aux enfants de progresser. La présence d'un autre enfant n'a pas d'effet.
(Beun *et al.*, 2003)	**Agent** d'apparences variées. Système **non interactif** ; pas d'**entrée**.	Tête parlante 3D réaliste ; Agent 3D fantaisiste ; Pas d'Agent. Animations faciales de base chez les Agents.	**Application** : Narration d'histoires. **61 utilisateurs** **Variables** : Rappel ; Questionnaire d'anthropomorphisation.	La présence d'un Agent (quel qu'il soit) améliore le rappel. Pas de corrélation entre rappel et anthropomorphisation.

Effets de l'apparence des Agents				
Référence	**Système**	**Objet du test**	**Protocole**	**Principaux résultats**
(Koda & Maes, 1996)	**Agents** : têtes expressives. **Entrée** : Clavier et souris. Système **fonctionnel**.	Pas d'Agent ; Agent 3D réaliste féminin ; Agent 2D cartoon féminin ; Agent 2D cartoon chien ; Agent 2D stylisé (emoticon).	**Application** : Jeu de poker. **10 utilisateurs** **Variables** : Questionnaire.	Les Agents augmentent l'engagement et la satisfaction. L'Agent réaliste est jugé le plus intelligent ; le chien jugé le plus agréable.
(McBreen & Jack, 2000)	**Agents** d'apparences variées. Système **non interactif** : évaluation d'extraits vidéos d'interaction où des utilisateurs interagissent par la parole avec des Agents.	Version homme et femme de : A1 : Voix seule ; A2 : Tête parlante 2D ; A3 : Tête parlante 3D ; A4 : Corps entier 2D ; A5 : Corps entier 3D ; A6 : A5 intégré à l'environnement (évolue dans la pièce à décorer).	**Application** : Décoration d'intérieur. **36 participants** **Variables** : Questionnaire.	1/3 des participants préfèrent la condition Voix seule. Pour les autres : Préférence pour corps entier. Préférence pour 3D. Préférence pour version féminine (résultat dû au domaine d'application ?). Pas d'effet de la position (intégré à l'environnement ou à l'extérieur).

Référence	Système	Objet du test	Protocole	Principaux résultats
(McBreen & Jack, 2001)	**Agents** d'apparences variées. Même système que précédemment (McBreen & Jack, 2000).	Version homme et femme de : A1 : Vidéo d'un humain (tête seule) ; A2 : Tête parlante 3D réaliste ; A3 : Tête parlante photo-réaliste ; A4 : Images fixes photo-réalistes ; A5 : Voix seule.	**Application** : Décoration d'intérieur et service de musique en ligne. **64 participants** **Variables** Questionnaires	Pas d'effet du domaine d'application. Pas d'effet du sexe ou de l'âge des participants. A1 est le mieux jugé, puis A5. A2 est le plus mal jugé. Préférence pour version masculine.
(McBreen *et al.*, 2001)	**Agents** 3D réalistes, moitié supérieure du corps. **Entrée** : Parole. Système **fonctionnel**.	Comparaison entre Agents de sexe et de styles vestimentaires différents (décontracté, formel).	**Application** : E-commerce (cinéma, agence de voyage, banque en ligne). **36 utilisateurs** **Variables** : Questionnaire.	Agents habillés décontractés sont plus appropriés pour la vente de places de cinéma ; Agents habillés de façon plus formelle pour une banque en ligne.
(Wonish & Cooper, 2002)	**Agents** inanimés (présentés sur des vignettes), d'apparences variées. Pas d'**entrée**. Pas de **système**.	- Agent contextuel (apparence liée à la fonction, ex : Agent médecin pour tutorial médical) *vs*. non contextuel. - Agent réaliste *vs*. fantaisiste.	**Applications** variées. **41 participants** **Variables** : Préférence.	Préférence pour les Agents contextuels. Le réalisme n'a pas d'effet sur la préférence.

Référence	Système	Objet du test	Protocole	Principaux résultats
(Moreno *et al.*, 2002)	**9 Agents** d'apparences variées. Pas d'**entrée**. Système **non interactif**.	Comparer l'efficacité et les évaluations subjectives d'Agents d'apparences variées.	**Application** : Pédagogie (informatique, physique, psychologie). La leçon est une conversation entre 3 Agents ; l'utilisateur ne participe pas. **36 utilisateurs** **Variables** : Rappel ; Questionnaire.	Les Agents diffèrent sur l'efficacité de l'apprentissage, ainsi que sur les dimensions subjectives, mais ces 2 types de variables ne sont pas liés.
(Woods *et al.*, 2003)	**Agents** 3D d'apparences variées. **Entrée** : clavier, souris. Système **fonctionnel**.	Comparaison des différents Agents impliqués dans les sketches.	**Application** : Sketches pour la prévention de la violence. **76 utilisateurs** (10 à 55 ans) **Variables** : Questionnaire.	Les utilisateurs de 10-18 ans préfèrent les Agents cartoon, ceux de 19-40 ans ont une légère préférence pour les Agents cartoon, ceux de 41-55 ans préfèrent les Agents réalistes. Aucun effet du sexe.
(Cowell & Stanney, 2003a)	**Agents** inanimés (présentés sur des vignettes), d'apparences variées. Pas d'**entrée**. Pas de **système**.	18 Agents croisant l'âge (jeune, intermédiaire, vieux), le sexe et l'origine ethnique (européenne, africaine, asiatique).	**Application** **45 utilisateurs** (20 originaire d'Europe, 9 d'Asie, 16 d'Afrique) **Variables** : Préférence des utilisateurs sur les Agents avec qui ils aimeraient travailler.	Tendance à préférer les Agents jeunes, de même origine ethnique et de sexe opposé à soi.

Effets des modalités utilisées				
Référence	**Système**	**Objet du test**	**Protocole**	**Principaux résultats**
(Seto *et al.*, 1994)	**Agent** : tête parlante 2D. **Entrée** : Parole. Système **fonctionnel**.	1. Agent, audio, texte, images 2. Agent, audio, texte 3. Agent, audio	**Application** : Commande de repas. **4 utilisateurs** **Variables** : Temps de dialogue.	Le temps de réalisation de la tâche est supérieur en conditions 2 et 3.
(Moreno *et al.*, 2000)	**Agent** 2D fantaisiste (Herman the Bug). **Entrée** : souris. Système **fonctionnel**. **Application** : Pédagogie (botanique).	1. Agent, parole ; 2. Parole seule ; 3. Agent, texte ; 4. Texte seul	**64 utilisateurs** **Variables** : Rappel, Transfert d'apprentissage, Questionnaire.	Pas d'effet de l'Agent. La parole améliore le rappel, le transfert et les évaluations subjectives.
		1. Vidéo humain, parole ; 2. Parole seule ; 3. Vidéo humain, texte ; 4. Texte seul.	**79 utilisateurs** **Mêmes variables.**	
		Synthèse vocale (voix humaine) sans Agent : Avec *vs.* sans interaction.	**39 utilisateurs** **Mêmes variables.**	La condition interactive augmente le rappel et le transfert d'apprentissage. Dans le cas du texte, l'interaction augmente aussi les évaluations subjectives.
		Texte écrit sans Agent : Avec *vs.* sans interaction.	**42 utilisateurs** **Mêmes variables.**	

Référence	Système	Objet du test	Protocole	Principaux résultats
(Craig *et al.*, 2002)	**Agent** 2D, moitié supérieure du corps.	9 conditions par croisement de : - Agent, Agent avec gestes (pointages imprécis), voix seule ; - Image statique, surbrillance, animation.	**Application** : Pédagogie (phénomène des éclairs). **135 utilisateurs** **Variables** : Performance de rappel, transfert d'apprentissage ; Questionnaire.	Pas d'effet de l'Agent ni de ses propriétés. La surbrillance et l'animation améliorent le rappel et le transfert.
		Comportement verbal de l'Agent : Texte (bulles) *vs*. synthèse vocale *vs*. les deux.	**71 utilisateurs** **Même application, mêmes variables**.	Synthèse vocale seule améliore les performances.
(Krämer *et al.*, 2003)	**Agent** 3D réaliste. Pas d'**entrée**, système **non interactif**.	Agent sans geste ; Agent avec gestes ; Agent avec gestes désynchronisés (1 sec) ; Texte ; Audio seul.	**Application** : Présentation d'informations techniques (magnétoscope). **106 utilisateurs** **Variables** : Rappel ; Questionnaire.	Agents jugés plus amusants que Texte et Audio mais font diminuer la performance de rappel. Agent sans geste est jugé plus naturel, son comportement ressemble plus à celui d'un humain. Aucune différence n'apparaît entre les Agents synchronisé et désynchronisé (une synchronisation parfaite des gestes des mains ne semble pas indispensable).

Référence	Système	Objet du test	Protocole	Principaux résultats
(Graesser *et al.*, 2003)	**Agent** : Tête parlante 3D. **Entrée** : Clavier. Système **fonctionnel** (AutoTutor).	AutoTutor *vs*. livre *vs*. rien.	**Application** : Pédagogie (informatique). **81 utilisateurs** (étudiants). **Variables** : Connaissances en informatique (pré- et post-test).	L'acquisition de connaissances est supérieure avec AutoTutor.
		Ecrit ; audio ; Agent + audio ; Agent + audio + écrit.		Pas d'effet des modalités de AutoTutor.
(Bickmore & Cassell, 2004)	**Agent** 3D réaliste (Rea). **Entrée** : Parole, gestes 3D Système **simulé***.	Agent en face-à-face *vs.* au téléphone. Avec *vs.* sans *small talk.* Utilisateurs actifs *vs.* passifs au cours de l'interaction (estimation *a posteriori*).	**Application** : Transactions immobilières. **58 utilisateurs** **Variables** : Confiance et qualité de l'interaction (questionnaire) ; Montant du loyer accepté par les utilisateurs.	Les utilisateurs préfèrent l'interaction sociale (avec *small talks*) au téléphone et l'interaction orientée tâche (sans *small talks*) en face-à-face. Les limitations de Rea pour l'interaction sociale en face-à-face sont attribuées à ses comportements non verbaux introvertis. L'interaction et Rea sont mieux jugés par les utilisateurs passifs (Rea est conçue pour mener elle-même le dialogue).

Effets des comportements non verbaux				
Référence	**Système**	**Objet du test**	**Protocole**	**Principaux résultats**
(Nagao & Takeuchi, 1994b)	**Agent** : Tête parlante 3D réaliste. **Entrée** : Parole, position de la tête. Système **fonctionnel**.	Agent avec expressions faciales *vs.* sans (expressions faciales remplacées par des précisions verbales, ex : « Je ne suis pas sûr » remplace l'expression faciale de doute).	**Application** : E-commerce. **32 utilisateurs** **Variables** : Qualité de l'interaction (ex : nombre de sujets abordés en 10 minutes).	Les expressions faciales améliorent l'interaction en début de conversation ; ce bénéfice initial est ensuite compensé par un effet d'entraînement.
(Granström *et al.*, 2002)	**Agent** : tête parlante 3D. Système **non interactif** : évaluation d'extraits vidéos d'interaction où des utilisateurs interagissent par la parole avec des Agents.	Feedbacks par combinaisons des paramètres : 1. Sourire ; 2. Mouvement de la tête ; 3. Mouvement des sourcils ; 4. Ouverture des yeux ; 5. Prosodie ; 6. Délai de réponse.	**Application** : Agence de voyage. **17 utilisateurs** **Variables** : Performance des utilisateurs à décoder les feedbacks de l'Agent (positif, négatif).	Le sourire est le feedback le mieux reconnu (puis la prosodie, puis le mouvement des sourcils). La combinaison de plusieurs indices a un effet additif sur la qualité de décodage du feedback.

Référence	Système	Objet du test	Protocole	Principaux résultats
(Cowell & Stanney, 2003b)	**Agent** 3D réaliste. **Entrée** : Clavier, souris. Système **fonctionnel**.	1. Pas de comportements non verbaux (sauf mouvements des lèvres) ; 2. Comportements faciaux de confiance ; 3. Comportements faciaux + corporels de confiance ; 4. Comportements faciaux + corporels inverses à la confiance.	**Application** : Assistance logiciel de tri de photos. **48 utilisatrices** **Variables** : Questionnaire, Réalisation du scénario.	Pas de différences entre versions 2 et 3 : inspirent effectivement la confiance. Version 1 jugée intermédiaire. Version 4 jugée désagréable ; les sujets perçoivent la tâche plus complexe et font plus d'erreurs. *Voir l'article pour les recommandations concrètes pour la création de comportements de confiance.*
Effets de la personnalité ou du rôle des Agents				
Référence	**Système**	**Objet du test**	**Protocole**	**Principaux résultats**
(Cassell & Bickmore, 2001)	**Agent** 3D réaliste (Rea). **Entrée** : Parole, gestes 3D de tours de parole. Système **simulé***.	Dialogue orienté tâche *vs.* introduction par *small talks* (parler de la pluie et du beau temps). Utilisateurs introvertis *vs.* extravertis.	**Application** : Transactions immobilières. **31 utilisateurs**. **Variables** : Confiance et qualité de l'interaction (questionnaire) ; Montant du loyer accepté par les utilisateurs.	Les *small talks* rendent Rea plus sympathique. Ils augmentent la confiance et la satisfaction des extravertis mais n'ont pas d'effet sur les introvertis (ou un effet inverse).

Référence	Système	Objet du test	Protocole	Principaux résultats
(Baylor, 2003)	**Agents** : têtes parlantes 3D réalistes. **Entrée** : clavier, souris. Système **fonctionnel**.	2 Agents (1 motivateur + 1 expert) *vs*. 1 seul Agent (mentor combinant les rôles de motivateur et d'expert).	**Application** : Planification d'un programme éducatif. **48 utilisateurs** **Variables** : Performance de rappel, de transfert d'apprentissage ; Questionnaire.	La division des rôles entre 2 Agents améliore le rappel et la facilité perçue. Il n'y a pas d'effet sur la motivation ressentie.
(Mori *et al.*, 2003 ; Prendinger *et al.*, 2003)	**Agent** : 2D réaliste. **Entrée** : clavier, souris. Système **fonctionnel**.	Agent avec comportement émotionnel (joyeux pour bonnes réponses, triste pour mauvaises, désolé pour bugs du système) *vs*. sans (donne juste les résultats).	**Application** : Quizz mathématique. **20 utilisateurs** (hommes, Japonais) **Variables** : Activité électrodermale et fréquence cardiaque (indicateurs de stress) ; Score ; Questionnaire.	Les excuses de l'Agent en cas de bug diminuent le stress. Les autres réactions émotionnelles n'ont pas d'effet.
(Darves & Oviatt, 2004)	**Agents** 2D, animaux. **Entrée** : Parole + gestes 2D (stylo). Système *I SEE!* (**simulé***).	Voix de l'Agent introvertie *vs*. extravertie (manipulation du volume et de la hauteur de la voix).	**Application** : Pédagogie (biologie marine). **24 utilisateurs** (7 à 10 ans). **Variables** : Analyse des incorrections langagières.	Les enfants posent davantage de questions pertinentes (sur la biologie marine) lorsque l'Agent a une voix extravertie.

Tableau 2 : Récapitulatif de résultats expérimentaux sur les Agents Conversationnels (classement chronologique au sein de chaque problématique).
*** Dans le cas des systèmes simulés, la méthode employée est le Magicien d'Oz : les principes en seront exposés au paragraphe 3.2.2.**

Discussion. La première partie de ce tableau synoptique a apporté des éléments de réponse à la problématique de l'**utilité** des Agents Conversationnels (comparaison d'un même système avec et sans Agent) : à cet égard, plusieurs études tendent à montrer que la présence d'un Agent permet d'améliorer diverses dimensions subjectives comme la facilité perçue d'une application, ou la motivation à l'utiliser (Koda & Maes, 1996 ; Van Mulken *et al.*, 1998 ; Moreno *et al.*, 2000 ; 2001 ; Moundridou & Virvou, 2001). D'autres résultats suggèrent en outre que les Agents seraient susceptibles d'améliorer l'efficacité d'une application (Moreno *et al.*, 2000 ; 2001 ; Beun *et al.*, 2003 ; Ryokai *et al.*, 2003). En général, il semble donc que l'insertion d'un Agent dans l'interface ait des effets plutôt positifs (avec un bémol apporté par les résultats de Walker *et al.*, 1994 ; Rickenberg & Reeves, 2000 ; Krämer *et al.*, 2003).

Y a-t-il des **caractéristiques graphiques** qui soient préférées de manière stable par les utilisateurs ? Il semble difficile de formuler une réponse univoque à cette question : la seconde partie du Tableau 2 montre en effet qu'il existe des résultats en faveur des Agents réalistes (Koda & Maes, 1996 ; McBreen & Jack, 2000, 2001 ; Woods *et al.*, 2003 avec les utilisateurs de 41 à 55 ans), et d'autres en faveur des Agents de style cartoon (Koda & Maes, 1996 ; Woods *et al.*, 2003 avec les utilisateurs de 10 à 18 ans)… Au niveau du look des Agents, il semble préférable de l'accorder avec le contexte de l'application (McBreen *et al.*, 2001 ; Wonish & Cooper, 2002).

Une autre série d'études recensées dans ce paragraphe concernent l'effet des **modalités de sortie** utilisées par les Agents : plusieurs résultats soulignent l'importance de la synthèse vocale (Moreno *et al.*, 2000 ; Craig *et al.*, 2002), ainsi que des expressions faciales (Nagao & Takeuchi, 1994b ; Granström *et al.*, 2002 ; Cowell & Stanney, 2003b). A l'inverse, les gestes semblent avoir un impact faible (Craig *et al.*, 2002 ; Cowell & Stanney, 2003b ; Krämer *et al.*, 2003). Deux études ont abordé la **coordination entre parole et gestes des mains** : Craig *et al.* (2002) ont montré que les gestes, sous forme de pointages imprécis sur une image pendant une explication, n'avaient aucun effet. Krämer *et al.* (2003) ont quant à eux montré qu'une désynchronisation d'une seconde entre parole et gestes n'était pas perçue des utilisateurs.

Au-delà des résultats obtenus, le choix des variables et l'interprétation qui en est faite sont parfois discutables dans certaines des études citées – par exemple lorsque la qualité de l'interaction avec les Agents est estimée par le nombre de sujets abordés. Nous reviendrons sur certains aspects méthodologiques de l'évaluation des Agents dans le chapitre 7.

2.4. Conclusion

L'objectif de ce chapitre était de dresser un état des lieux sur le domaine des Agents Conversationnels Animés : dans un premier temps, nous nous sommes attachés à donner un rapide aperçu des systèmes fonctionnels existants. Puis, après avoir posé quelques définitions et concepts de base sur la multimodalité, nous avons resserré cet état de l'art autour des questions spécifiques de :

- La multimodalité entre la parole et les gestes des mains chez les Agents.
- La multimodalité entre la parole et les gestes des mains chez les utilisateurs interagissant avec un (des) Agent(s).

La suite du chapitre a montré que :

- Les travaux sur la génération de comportements multimodaux en sortie, de la part des Agents, ont atteint un niveau d'élaboration important, et sont en constant progrès ; en revanche les données expérimentales sur cette problématique sont peu nombreuses.
- Certains systèmes proposent aux utilisateurs d'interagir avec plusieurs modalités en entrée face aux Agents, mais cet aspect du système reste flou : le fonctionnement technique et l'utilisation effective de l'entrée multimodale sont peu abordés dans la littérature sur les Agents Conversationnels. La multimodalité en entrée semble davantage considérée comme un outil que comme un objet d'étude à part entière.

Il semble ainsi que le champ de recherche sur les Agents Conversationnels (plus précisément leur conception et leur évaluation) se soit focalisé sur la communication en sortie, oubliant bien souvent la communication en entrée du système, ou en tenant peu compte. Pour mieux appréhender la problématique du comportement multimodal des utilisateurs face aux Agents, nous proposons donc d'explorer dans le chapitre suivant le domaine des interfaces multimodales en entrée, sans Agent Conversationnel.

3. Les interfaces multimodales en entrée

3.1. Introduction

Comme nous l'avons vu dans le chapitre précédent, la multimodalité en Interaction Homme-Machine caractérise la capacité d'un système à exploiter plusieurs modalités de communication en entrée (de l'utilisateur vers le système) et/ou en sortie (du système vers l'utilisateur). Dans ce chapitre, nous traiterons exclusivement de la multimodalité en **entrée**, c'est-à-dire de celle qui concerne les informations transmises par l'utilisateur à destination du système.

Les Interfaces Multimodales en entrée permettent d'utiliser, simultanément ou pas, de manière combinée ou pas, plusieurs modalités[12] de communication comme la parole et le geste. Ces systèmes visent notamment à se rapprocher de la communication Homme-Homme, plus intuitive pour les utilisateurs et plus flexible. Alors qu'avec les systèmes WIMP (Windows – Icons – Menus – Pointing) que nous utilisons tous les jours, la communication en entrée se fait par l'intermédiaire du clavier et de la souris, les Interfaces Multimodales sont généralement enrichies d'un module de reconnaissance vocale. Le média[13] gestuel est également souvent modifié, en substituant par exemple une tablette graphique ou un écran tactile à la souris.

Historiquement, le premier système multimodal a été conçu par Bolt (1980). Cette application

[12] Les modalités correspondent ici aux canaux de communication utilisés par l'être humain pour transmettre des informations. Dans la suite du texte, nous en considèrerons essentiellement deux : la modalité verbale (parole orale) et la modalité gestuelle (par l'intermédiaire des mains).

[13] Par média, nous désignons le dispositif physique permettant l'expression ou la capture d'une modalité d'entrée. Par exemple, la souris est un des médias supports de la modalité gestuelle, le micro est le média permettant d'utiliser la modalité verbale.

permettait de créer et manipuler des figures géométriques à l'aide de la parole (vocabulaire de 120 mots) et du geste en trois dimensions (pointage à distance). Dans ce système avant-gardiste, la parole et le geste n'étaient pas seulement utilisés de manière juxtaposée, mais pouvaient être combinés en expressions multimodales. Le titre de cet article, *« Put-That-There »*, illustre cette notion et est resté l'expression désignant par excellence le paradigme de Bolt. La possibilité d'augmenter la parole avec un ou plusieurs gestes de désignation permet l'utilisation de pronoms et adverbes, ce qui peut apporter un gain en naturel et en économie d'expression. Réciproquement, la possibilité d'augmenter le geste par la parole peut augmenter la précision et la puissance de référence. Bolt (1980) ne fournit cependant pas de données d'observation ou d'évaluation pour appuyer ces hypothèses.

Cohen *et al.* (1989) ont également souligné très tôt les bénéfices de l'utilisation synergique des modalités par rapport à leur utilisation séparée. Mais il ne s'agissait pas ici seulement d'un bénéfice du point de vue de l'utilisateur, mais aussi du point de vue du système. En effet, en intégrant les informations provenant de la parole et du geste, le traitement nécessaire au système pour interpréter le message s'en trouve simplifié. C'est ce qu'Oviatt (1999a) désignera plus tard sous le concept de « désambiguïsation mutuelle ».

A partir des années 90, quelques prototypes multimodaux ont été développés, pour des domaines d'application variés. Le Tableau 3, sans rechercher l'exhaustivité, donne un aperçu de ces systèmes en précisant à chaque fois les dispositifs utilisés en entrée et le contexte d'application. Pour un inventaire plus détaillé, ainsi qu'une taxonomie des tâches dans lesquelles la multimodalité peut être bénéfique, le lecteur peut se reporter à la revue de question menée par Benoit *et al.* (2000).

Référence	**Dispositif d'entrée**	**Application**
(Cohen *et al.*, 1989)	Micro, souris, clavier	ShopTalk : Monitoring industriel, surveillance et gestion d'une ligne de production en usine.
(Thorisson *et al.*, 1992)	Micro, *dataglove* (pour gestes 3D), *eye-tracker* (pour la direction du regard)	Manipulation d'éléments sur un plan.
(Nigay & Coutaz, 1993)	Micro, souris	VoicePaint : Dessin.
(Nigay & Coutaz, 1993)	Micro, souris, clavier	Notebook : Carnet mémo.
(Vo & Waibel, 1993)	Micro, stylo sur tablette-écran	Editeur de texte.
(Ando *et al.*, 1994)	Micro, écran tactile	Aménagement d'intérieur.

(Koons & Sparrell, 1994)	Micro, *datagloves* (pour gestes 3D)	Iconic : manipulations graphiques et spatiales, autorise les gestes iconiques.
(Bellik, 1995)	Micro, écran tactile, souris	LimsiDraw : Dessin.
(Cheyer & Julia, 1995)	Micro, stylo sur tablette-écran	Cartes touristiques.
(Coutaz *et al.*, 1995)	Micro, souris, clavier	MATIS (Multimodal Airline Travel Information System) : horaires d'avions.
(Denda *et al.*, 1997)	Micro, écran tactile	Cartes touristiques.
(Cohen *et al.*, 1997)	Micro, stylo sur tablette-écran portable	QuickSet : simulation militaire.
(Duncan *et al.*, 1999)	Micro, *cybergloves* (pour gestes 3D)	Entraînement à la maintenance d'avions en réalité virtuelle.
(Holzman, 1999)	Micro, stylo et clavier virtuel sur tablette-écran portable	NCR Field Medic Information System : médecine d'urgence sur le terrain, description d'informations médicales.
(Caelen & Bruandet, 2001)	Micro, clavier, souris	Navigateur Internet.
(Wahlster *et al.*, 2001)	Micro, écran tactile PDA ou reconnaissance gestes 3D et expressions faciales	SmartKom : interfaces de loisirs, chez soi (TV) ou à l'extérieur (kiosques, bornes, mobile). Assistant virtuel intégré à l'interface.

Tableau 3 : Exemples d'Interfaces Multimodales en entrée (classement chronologique).

Le champ de recherche sur les Interfaces Multimodales repose sur l'hypothèse que celles-ci seraient plus faciles à utiliser et à apprendre, et seraient préférées par les utilisateurs. Les Interfaces Multimodales pourraient également permettre le développement de systèmes interactifs plus sophistiqués, l'élargissement de leur utilisation à une frange plus large de la population, et l'adaptation à des conditions d'utilisation plus variées par rapport aux interfaces classiques (Oviatt *et al.*, 2000).

Dans le cadre de l'interaction avec des Agents Conversationnels tels que nous les avons décrits dans le chapitre précédent, la multimodalité en entrée se justifie d'autant plus qu'elle permet une certaine symétrie dans l'interaction (parole et gestes des utilisateurs en entrée, parole et gestes des Agents en sortie). Dans ce chapitre, nous allons donc exposer brièvement quelques méthodes de conception et d'évaluation pour les Interfaces Multimodales en entrée, puis nous rendrons compte des connaissances actuelles sur le comportement multimodal des utilisateurs face à de telles interfaces.

3.2. Méthodes de conception

3.2.1. Principe général de conception

Une des problématiques majeures dans le développement d'interfaces multimodales en entrée est la **fusion des modalités**. En effet, le traitement et l'interprétation des constructions multimodales (où un message unique est composé d'éléments verbaux et gestuels, par exemple la phrase « Mets ça ici » accompagnée de la désignation d'un objet et d'un emplacement) sont un véritable défi pour les systèmes multimodaux. Deux questions principales émergent à ce propos :

- Dans les constructions multimodales, comment les modalités sont-elles intégrées au niveau **temporel** ? Il faut en effet que le système soit capable de résoudre les références verbales (ex : « ça » et « ici ») avec les éléments gestuels pertinents (la sélection de l'objet, et la localisation spatiale, respectivement).
- Comment les modalités coopèrent-elles au niveau **sémantique** ? Par exemple, le système doit pouvoir détecter et traiter correctement les constructions redondantes (ex : « déplacer le cube bleu sur la droite » accompagné d'un déplacement du cube bleu par le geste) afin d'éviter de réaliser deux fois la même commande.

A l'instar d'Oviatt (1999b), nous pensons que le développement d'Interfaces Multimodales fonctionnelles et robustes ne peut être fondé sur la seule intuition. Dans un article intitulé *« Ten Myths of Multimodal Interaction »*, Oviatt s'est attachée à recenser dix idées préconçues sur le comportement multimodal en Interaction Homme-Machine, qui semblent largement répandues mais qui ont été contredites par des résultats empiriques. Par exemple, on a tendance à penser que les entrées gestuelles vont être simultanées aux références verbales. Oviatt (1999b) révèle qu'en réalité, les modalités ne se chevauchent que dans la moitié des constructions multimodales. Ce type de résultat suggère à quel point les données comportementales sont utiles au développement de systèmes informatiques en général, et d'Interfaces Multimodales en particulier. Seule une bonne anticipation de la manière dont les utilisateurs vont communiquer permet d'implémenter un système efficace : les patterns extraits des corpus d'observation vont donner de précieuses indications sur les algorithmes et spécifications informatiques pertinentes à mettre en œuvre. Sans le guidage fourni par les études comportementales, les Interfaces Multimodales risquent de pâtir d'erreurs de conception.

Trois pistes complémentaires s'offrent alors aux concepteurs d'Interfaces Multimodales :

- Ils peuvent s'inspirer des données générales de la littérature en Sciences Cognitives, Psycholinguistique, Psychologie de la communication Homme-Homme. Cette solution sera discutée dans le paragraphe 3.3.1.

- Ils peuvent recourir aux résultats obtenus en Interaction Homme-Machine dans des contextes similaires à celui qu'ils étudient.
- Enfin, ils peuvent mettre en place un recueil de données centrées sur leur propre application et sur leur propre population cible. Cette dernière possibilité est évidemment la plus coûteuse en termes de ressources et de temps, mais elle peut s'avérer réellement nécessaire lorsque ni la littérature générale, ni les résultats antérieurs ne semblent transférables directement (modalités ou médias différents, contextes d'application trop éloignés, populations différentes). Ajoutons que le bénéfice d'un corpus est double : outre les indications qu'il donne sur l'utilisation des modalités, le corpus pourra être réutilisé pour entraîner le système une fois implémenté (amélioration de la grammaire vocale, de la reconnaissance de formes, des algorithmes de fusion, etc.).

Les deux paragraphes suivants présentent les principales techniques disponibles pour le recueil de données en amont de la phase de développement. Le « Magicien d'Oz » est une méthode générique de simulation de systèmes informatiques, qui peut permettre de recueillir des corpus d'interaction. Le paragraphe suivant traite de l'annotation de tels corpus. Nous évoquerons ensuite un outil complémentaire de conception, les taxonomies d'intégration des modalités.

3.2.2. Magicien d'Oz

Pour anticiper le comportement que les utilisateurs vont avoir face à un système alors que celui-ci n'est pas encore développé, il est possible d'en tester une version simulée. Si ce procédé peut se révéler utile dans toute démarche de conception, il est quasi-indispensable dans le cas des Interfaces Multimodales en entrée, comme nous l'avons exposé dans le paragraphe précédent. La simulation permet en outre une évaluation préliminaire, non pas du système, mais par exemple de l'application visée (utilité potentielle…), du modèle de dialogue, ou de divers éléments de design – la simulation est ainsi parfois utilisée dans l'évaluation des Agents Conversationnels (voir le paragraphe 2.3).

Le Magicien d'Oz est une technique ancienne de simulation des systèmes techniques avant leur mise en service effective. Elle est couramment utilisée en conception informatique (Dahlbäck *et al.*, 1993). La technique consiste à charger un compère humain de contrôler à distance l'interface de l'utilisateur de sorte que l'interaction soit fluide et que le système semble réellement fonctionnel. Les utilisateurs ne sont informés qu'*a posteriori* que le système était simulé (pour des raisons éthiques il faut toujours, en fin d'expérimentation, dévoiler le dispositif utilisé et les buts de l'expérimentation).

Figure 16 : Exemple typique de dispositif Magicien d'Oz (figure empruntée à l'article d'Hauptmann, 1989) : à gauche, le sujet réalise le scénario en étant filmé ; à droite, le magicien suit le comportement du sujet sur le téléviseur et gère l'interaction en conséquence.

Le compère humain (ou magicien) réalise diverses fonctionnalités à la place du système, mais cela ne signifie pas que l'on puisse mettre en œuvre un Magicien d'Oz en partant de rien. Cela suppose au contraire l'existence d'une véritable plateforme de simulation (voir la Figure 16 ci-dessus), comprenant au minimum :

- L'interface utilisateur,
- Le moyen, pour le magicien, d'avoir connaissance en continu du comportement de l'utilisateur (transmissions par caméras, par micros, par événements système…),
- L'interface magicien permettant de contrôler l'interface utilisateur, et placée en réseau avec celle-ci. Pour que la simulation soit plausible, les réactions du magicien doivent être aussi rapides que possible. Un soin particulier doit donc être apporté à la conception de son interface, en y incluant par exemple des réponses pré-encodées et des accès directs aux fonctionnalités (Oviatt *et al.*, 1992 ; Dahlbäck *et al.*, 1993). Ceci suppose d'avoir au préalable élaboré un modèle de réalisation de la tâche ou du scénario. Certains systèmes complexes peuvent nécessiter l'intervention de plusieurs magiciens pour gérer l'interaction (un pour simuler la reconnaissance vocale, un autre pour les actions gestuelles…), ce qui implique autant d'interfaces à développer et à coordonner (voir par exemple la plateforme développée par Coutaz *et al.*, 1996). Pour augmenter la crédibilité du dispositif, il est recommandé de simuler quelques erreurs de traitement (Oviatt *et al.*, 1992).

De telles plateformes permettent ainsi de recueillir de grandes quantités de données comportementales, sous différentes formes : corpus audio-visuel, événements système (*log files*), données subjectives fournies par les utilisateurs à l'issue du test (interviews, questionnaires).

3.2.3. Annotation de corpus multimodaux

Une fois le(s) corpus d'interaction recueillis, il faut savoir en exploiter toute la richesse, ce qui ne va pas de soi. L'état actuel des logiciels de traitement automatique de la parole et des images ne permet pas d'envisager à court ou moyen terme une reconnaissance et un codage automatique de comportements très variés, qui sont par ailleurs observés dans des situations très diverses (une ou plusieurs personnes, différents angles de vue, qualité inégale des enregistrements…). L'approche actuellement envisagée consiste donc généralement à annoter manuellement les corpus audio-visuels. Mais cela requiert une méthodologie rigoureuse. Un système d'annotation de ressources multimodales doit comporter les éléments suivants :

- Un ou plusieurs outils d'aide à l'annotation, par exemple Praat[14] et Anvil (Kipp, 2001),
- Des schémas de codage pertinents pour décrire le comportement, et la possibilité d'adapter ces schémas ou d'en créer de nouveaux,
- Des algorithmes et programmes de calcul de statistiques et de recherche de patterns de comportements,
- Des outils pour structurer les corpus à un haut niveau (métadata).

L'annotation de corpus est devenue un domaine de recherche à part entière dans la communauté étudiant l'Interaction Homme-Machine. Devant l'éparpillement des données et des outils (souvent développés en interne au sein des équipes de recherche), le besoin de capitaliser et de standardiser ces ressources s'est fait ressentir et a donné lieu récemment à plusieurs projets européens, parmi lesquels nous pouvons citer (voir aussi Dybkjaer & Bernsen, 2002) : le projet MATE[15] (*Multi-level Annotation Tools Engineering*), le projet ISLE[16] (*International Standards for Language Engineering*), qui comprenait notamment un groupe de travail consacré à la multimodalité (groupe NIMM, *Natural Interactivity and Multimodality*), et le projet NITE[17] (*Natural Interactivity Tools Engineering*). Plusieurs workshops dans des conférences internationales ont également été consacrés aux corpus multimodaux (Maybury & Martin, 2002 ; Martin *et al.*, 2004).

[14] http://www.fon.hum.uva.nl/praat/

[15] http://mate.nis.sdu.dk/

[16] http://isle.nis.sdu.dk/

[17] http://nite.nis.sdu.dk/

Les corpus multimodaux correctement annotés peuvent ensuite être utilisés pour les études comportementales et/ou pour fournir des données étiquetées à un système ayant des capacités d'apprentissage.

3.2.4. Taxonomies d'intégration des modalités

Le processus d'analyse des corpus peut également être enrichi par des annotations manuelles ou automatiques issues des taxonomies d'intégration des modalités qui existent dans la littérature. Le principe de ces taxonomies est de proposer différentes classifications, qui se veulent exhaustives, des combinaisons temporelles et/ou sémantiques entre les modalités. Elles peuvent être mises à profit dans la phase d'annotation comme dans les spécifications préalables au développement.

Dans le cadre de l'annotation de corpus, les taxonomies constituent une grille de référence permettant de construire un schéma de codage pertinent pour l'analyse du comportement. En effet, en l'absence de taxonomie, l'annotateur doit construire des catégories d'observations *ad hoc* tout au long du processus d'annotation (voir par exemple De Angeli *et al.*, 1998). A cet égard, les taxonomies permettent un gain de temps et une relative garantie d'exhaustivité ; elles sont également utiles pour représenter et synthétiser le comportement multimodal des utilisateurs.

Les taxonomies sont également souvent utilisées lors de la phase de spécification pour guider directement l'implémentation du système. Leur avantage est qu'elles permettent au concepteur de n'oublier aucun cas de figure dans l'intégration des modalités par le système. Leur inconvénient est qu'elles restent très abstraites par rapport à un contexte et un contenu d'interaction donné. Par exemple, une taxonomie peut indiquer au développeur que son système doit pouvoir gérer une utilisation **complémentaire** des modalités, mais sans lui donner un aperçu concret des cas qu'il va devoir prendre en compte, et de leur diversité. Un autre inconvénient des taxonomies est qu'elles ne permettent pas de distinguer entre les configurations très fréquentes, dont la prise en compte va être critique à l'utilisabilité du système (ex : utilisation complémentaire des modalités dans 93% des cas dans l'étude de Xiao *et al.*, 2002), et les cas marginaux, voire exceptionnels (ex : utilisation concurrente des modalités, presque jamais observée), sur lesquels le petit nombre d'utilisateurs concernés devrait aisément admettre une inexactitude de traitement et passer à une formulation alternative. Bien que les taxonomies soient souvent utilisées directement pour la spécification, nous pensons donc qu'elles ne sont qu'un complément à l'observation du comportement des utilisateurs.

Du point de vue des relations **temporelles** entre deux éléments (deux modalités par exemple), la taxonomie la plus complète semble être celle d'Allen (1983), qui prévoit treize configurations schématisées dans le Tableau 4.

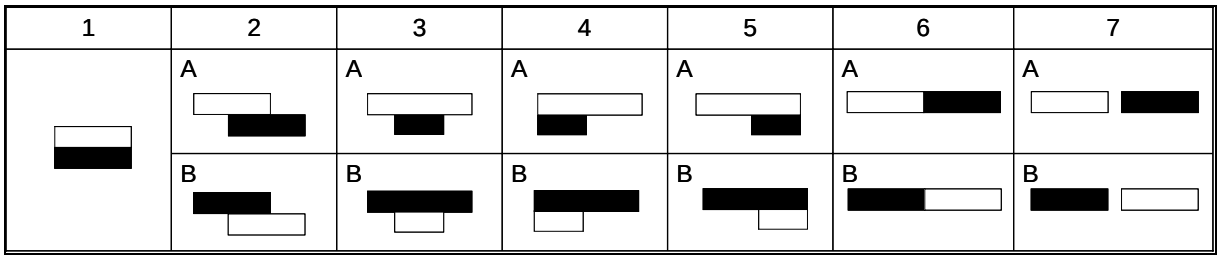

Tableau 4 : La taxonomie des relations temporelles d'Allen (1983).

Un tel niveau de précision n'est pas toujours nécessaire en analyse du comportement. Par exemple, Catinis et Caelen (1995) ont analysé uniquement l'ordre de succession des modalités, et se sont donc limités à des catégories s'apparentant aux cas 1, 2 et 7 d'Allen (soit cinq combinaisons possibles). Le cas 1 est appelé « synchronie vraie synergique », le cas 2 « asynchronie anticipative », et le cas 7 « asynchronie alternée ». Oviatt *et al.* (1997) ont utilisé les cas 1 à 5 (soit 9 combinaisons possibles) pour étudier la simultanéité des modalités, et ont étudié à part leur utilisation séquentielle (cas 7). Le cas 6 a été assimilé au cas 7, avec un délai nul entre modalités.

Certaines taxonomies spécifiques à la description des Interfaces Multimodales croisent la dimension **temporelle** avec une dimension **sémantique**. C'est le cas par exemple de la taxonomie proposée par Nigay et Coutaz (1993). La dimension temporelle comporte deux valeurs (utilisation séquentielle ou parallèle), de même que la dimension sémantique (modalités combinées ou indépendantes). Il en résulte une taxonomie à quatre cas, résumée dans le Tableau 5. Nigay et Coutaz (1993) ont aussi ajouté à cette taxonomie une dimension de degré d'abstraction, qui ne sera pas discutée ici. Le lecteur peut également se reporter aux travaux de Bellik et Teil (1992) pour une distinction entre l'utilisation des modalités et la production des énoncés.

		Utilisation des modalités (dimension temporelle)	
		Séquentielle	Parallèle
Fusion (dimension sémantique)	Combinaison	**Alterné**	**Synergique**
	Indépendance	**Exclusif**	**Concurrent**

Tableau 5 : Taxonomie d'intégration des modalités (Nigay & Coutaz, 1993).

Dans la taxonomie précédente (Nigay & Coutaz, 1993), la dimension sémantique est limitée à deux cas, combinaison/indépendance. Cette dimension peut être affinée, en particulier en explorant les différentes relations sémantiques lorsque les modalités sont combinées. C'est ce que Martin et coll. se sont attachés à faire en proposant la taxonomie TYCOON (TYpes de COOpératioN, Martin & Béroule, 1993 ; Martin *et al.*, 2001). Six types de coopération entre modalités, issus d'une étude de la multimodalité en Psychologie (Martin, 1995), sont définis dans TYCOON :

- La Complémentarité : au sein d'une même commande, différents éléments sont exprimés par différentes modalités. Les informations doivent être fusionnées pour accéder au message complet.
- La Redondance : au sein d'une même commande, le même élément est exprimé par différentes modalités. Les informations doivent être fusionnées pour obtenir un message commun et unique.
- La Concurrence : des éléments indépendants sont exprimés par différentes modalités avec un chevauchement temporel. Les modalités sont utilisées en parallèle pour réaliser différentes actions. La coopération par concurrence ne rend pas compte des cas de conflit dans lesquels, pour une même commande, les informations portées par les modalités sont incompatibles (ex : « Je prends le cube de gauche » avec sélection gestuelle du cube de droite).
- L'Équivalence : un même élément peut être exprimé par différentes modalités à des moments différents de l'interaction.
- La Spécialisation : un élément est toujours exprimé par la même modalité.
- Le Transfert : un élément exprimé par une modalité est analysé par une autre modalité.

Seuls les trois premiers types de coopération (Complémentarité, Redondance, Concurrence) peuvent caractériser la coopération des modalités dans les **constructions multimodales des utilisateurs** (lorsque les modalités d'entrée sont utilisées au sein d'une même commande, que ce soit de manière séquentielle ou simultanée). L'Équivalence et la Spécialisation concernent la coopération des modalités dans les occurrences successives des commandes. Enfin, le Transfert caractérise, du point de vue de l'utilisateur, la coopération entre modalités d'entrée et de sortie.

Signalons que Coutaz *et al.* (1995) ont fourni une définition formelle de quatre des types de coopération précédents, désignés sous le sigle CARE : Complementarity Assignment Redundancy Equivalence – l'Assignation correspondant ici à la Spécialisation dans la taxonomie de Martin *et al.* (2001). Les propriétés CARE ont été décrites d'un point de vue système pour caractériser les Interfaces Multimodales, cependant elles peuvent convenir, dans une certaine mesure, à l'annotation du comportement de l'utilisateur.

3.3. Comportement multimodal des utilisateurs

3.3.1. Interaction Homme-Homme

Nous avons été amenés, dans le chapitre précédent, à faire un rapide état de l'art sur la multimodalité entre parole et gestes des mains en communication Homme-Homme. Nous avons abordé les catégories classiques de gestes accompagnant le discours, les relations temporelles entre ces gestes et le discours verbal, ainsi que leurs relations sémantiques. Sur cette base, nous avons pu étudier le comportement multimodal des Agents Conversationnels. Dans le présent paragraphe, nous allons rediscuter ces observations pour savoir quels éléments de compréhension et de prédiction elles peuvent nous apporter sur la question de la communication Homme-Machine multimodale.

Catégories de gestes

Revenons dans un premier temps sur la classification des gestes accompagnant le discours dans la communication Homme-Homme. Cinq catégories sont identifiées dans le continuum de Kendon (1988) :

La gesticulation → les gestes langagiers → le mime → les emblèmes → les langues des signes.

La majorité des travaux cités au chapitre précédent, et plus généralement la plupart des études en Psycholinguistique, concernent la gesticulation, c'est-à-dire les mouvements spontanés des mains et des bras accompagnant la parole. Nous pouvons considérer que les gesticulations sont produites de manière automatique et involontaire (Cassell *et al.*, 1999b).

Ceci nous amène à préciser un point important de notre domaine de recherche :

- Dans le cadre de ce projet, nous restreignons notre champ d'investigation aux **Interfaces Multimodales** permettant de détecter et de traiter les **gestes volontaires**, correspondant à des **commandes explicites** de la part de l'utilisateur. Ceux-ci incluent :
    - Les entrées monomodales gestuelles (ex : sélection ou manipulation d'un objet) : dans la mesure où elles ces entrées sont indépendantes de la parole, elles peuvent tout simplement prendre le statut de **gestes manipulatoires**.
    - Les gestes volontairement intégrés par le locuteur comme constituants grammaticaux du discours (ex : « Placer le cube ici » avec un geste de localisation, indispensable à l'exécution de la commande, ou « Grand comme ça » avec une indication de taille) : ceci correspond dans le continuum de Kendon (1988) à la catégorie des **gestes langagiers**, et non des gesticulations.

- A l'inverse, la communication non verbale passive, les gestes et postures ne correspondant pas à des commandes volontaires relèvent davantage du domaine des **Interfaces Perceptives** – *Perceptual User Interfaces* (Turk & Robertson, 2000). Ces systèmes mettent en jeu des technologies de vision par ordinateur (capture et l'interprétation de la direction du regard, de la position de la tête, de la posture, des expressions faciales et des gestes des mains en trois dimensions).

Parmi les exemples cités au chapitre précédent, nous pouvons considérer que le système de Nagao & Takeuchi (1994a) est une interface perceptive (les entrées non verbales se limitent à la détection de la position des utilisateurs). Les systèmes Gandalf (Cassell & Thorisson, 1999), Rea (Cassell *et al.*, 1999b ; Cassell, 2001 ; Cassell & Bickmore, 2001) et Steve (Rickel & Johnson, 1999) sont à la fois des interfaces perceptives (détection de la position et des mouvements de la tête ; chez Rea, détection des gestes régulateurs et de battement) et des interfaces multimodales (traitement des gestes de pointage et, chez Steve, des gestes manipulatoires).

Dans ce projet, nous nous limitons à l'étude de la gestuelle active et des commandes volontaires, que nous proposons de regrouper sous le terme de **gestes opératoires** (gestes langagiers et manipulatoires). Ces gestes opératoires sont, par exemple, ceux recueillis en deux dimensions sur un écran (par l'intermédiaire d'une souris, d'un écran tactile, ou d'un stylo).

Relations temporelles entre paroles et gestes

L'étude de McNeill (1995) traitait de la synchronisation entre le discours et les gestes illustrateurs ou manipulatoires (puisque dans certains cas les sujets devaient commenter des actions qu'ils réalisaient effectivement). Etant donné qu'ils ne concernent pas uniquement la gesticulation involontaire, les résultats sont intéressants pour l'Interaction Homme-Machine multimodale. Ils ont mis en évidence l'apparition de deux types de patterns d'intégration des modalités : un pattern simultané et un pattern alterné (voir la description de cette étude dans le paragraphe 2.2.2). Mais McNeill (1995) ne nous apporte pas de statistiques sur la fréquence d'occurrence de ces deux patterns.

Les résultats de Gogate *et al.* (2000), sur la synchronisation des modalités chez les mères apprenant des nouveaux mots à leurs enfants, sont également susceptibles d'être transférés à l'Interaction Homme-Machine. Le fait que la fréquence du pattern simultané soit d'autant plus élevée que l'enfant est jeune suscite même des hypothèses intéressantes : une fréquence élevée du pattern simultané en Interaction Homme-Machine pourrait éventuellement être rapprochée d'un comportement protecteur

des utilisateurs vis-à-vis du système (simultanéité visant à souligner la relation mot/référent). Il serait alors intéressant de tester l'évolution des patterns d'intégration temporelle avec la durée d'utilisation du système et avec la perception que les utilisateurs ont de ses capacités.

Relations sémantiques entre parole et gestes

Les fonctions sémantiques des gestes en communication Homme-Homme (Scherer, 1984 ; Knapp, 2002) peuvent être confrontées aux taxonomies d'intégration des modalités développées pour l'Interaction Homme-Machine multimodale (Coutaz *et al.*, 1995 ; Martin *et al.*, 2001). Dans le Tableau 6, nous tentons d'établir une correspondance (imparfaite) entre ces différentes classifications.

<table>
<tr><th>(Scherer, 1984)</th><th>(Knapp, 2002)</th><th>(Martin et al., 2001)</th><th>(Coutaz et al., 1995)</th></tr>
<tr><td rowspan="2">Modification</td><td>Complémentarité</td><td rowspan="2">Complémentarité</td><td rowspan="2">Complémentarité</td></tr>
<tr><td>Accentuation/modération</td></tr>
<tr><td rowspan="2">Amplification,
Illustration,
Clarification</td><td>Répétition</td><td>Redondance</td><td>Redondance</td></tr>
<tr><td>Régulation</td><td></td><td></td></tr>
<tr><td rowspan="2">Signification
indépendante,
Substitution</td><td rowspan="2">Substitution</td><td>Spécialisation</td><td>Assignation</td></tr>
<tr><td>Equivalence</td><td>Equivalence</td></tr>
<tr><td>Contradiction</td><td>Conflit</td><td></td><td></td></tr>
<tr><td></td><td></td><td>Concurrence</td><td></td></tr>
<tr><td></td><td></td><td>Transfert</td><td></td></tr>
</table>

Tableau 6 : Correspondance entre les taxonomies d'intégration sémantique en communication Homme-Homme et Homme-Machine.

Les trois types de gestes testés par Cassell *et al.* (1999a) pourraient également être rapprochés de ces taxonomies : la concordance (« donner » + geste mimant l'action de donner) correspond en effet à de la redondance, et la discordance (« donner » + geste mimant l'action de prendre) correspond à un cas de conflit. Les gestes de manière (« Il le fait venir auprès de lui » + appel de la main, ou geste de saisie) sont quant à eux assimilables à une forme de complémentarité.

Par ailleurs, Cassell *et al.* (2000) ont constaté que 50% des constructions multimodales en communication Homme-Homme étaient redondantes et 50% complémentaires. Cependant, ces résultats ne peuvent être transférés à l'Interaction Homme-Machine, puisqu'ils concernent les gesticulations accompagnant le discours et non les gestes opératoires.

Enfin, dans l'expérience menée par Piwek et Beun (2002), 44% des références aux objets ont été accompagnées de gestes de désignation. Ce résultat ne renseigne pas la question des relations sémantiques entre parole et gestes, mais constitue une première prédiction sur la fréquence des constructions multimodales que nous pouvons attendre en Interaction Homme-Machine.

Conclusion

Nous avons vu dans ce paragraphe que les données concernant la communication multimodale Homme-Homme ne sont pas aisément transférables à l'Interaction Homme-Machine multimodale. Elles nous fournissent certains éléments de prédiction, mais nous incitent surtout à étudier l'Interaction Homme-Machine multimodale comme un mode de communication à part, qui demande à être étudié en situation. Dans le paragraphe suivant, nous présentons les observations spécifiques recueillies dans ce domaine.

3.3.2. Interaction Homme-Machine

Grâce aux techniques de simulation et d'annotation de corpus évoquées précédemment, des données empiriques sur l'utilisation des Interfaces Multimodales ont pu être publiées. Dans ce paragraphe, nous allons détailler plusieurs études menées par l'équipe de Sharon Oviatt[18], parce qu'elles fournissent incontestablement les analyses les plus poussées en la matière. Nous présenterons ensuite de manière synthétique des résultats obtenus par d'autres auteurs. Nous décrirons également une étude que nous avons menée sur l'utilisation d'un guide de programmes télévisés multimodal. L'ensemble de ces données sera discuté en fin de paragraphe.

Etudes menées par l'équipe d'Oviatt

D'une manière générale, l'étude de l'utilisation des Interfaces Multimodales peut répondre à deux objectifs principaux : (1) évaluer l'utilité de la multimodalité en la comparant à des situations d'interaction monomodales, et, si celle-ci est avérée, (2) analyser le comportement multimodal des utilisateurs et en extraire des patterns pour aider au développement d'Interfaces Multimodales fonctionnelles. La première expérimentation que nous rapportons (Oviatt, 1996a) vise le premier objectif. Oviatt a comparé le comportement de dix-huit utilisateurs adultes dans trois conditions : condition monomodale gestuelle (utilisation d'un stylo sur une tablette-écran portable : écriture,

[18] http://www.cse.ogi.edu/CHCC/Personnel/oviatt.html

dessins, pointages) ; condition monomodale verbale (utilisation de la parole) ; condition multimodale (utilisation libre du stylo et de la parole ; Figure 17). Dans tous les cas le scénario consistait à manipuler des cartes et des informations géographiques (tâches de sélection de biens immobiliers et de mise à jour de cartes) à l'aide d'un système simulé en Magicien d'Oz.

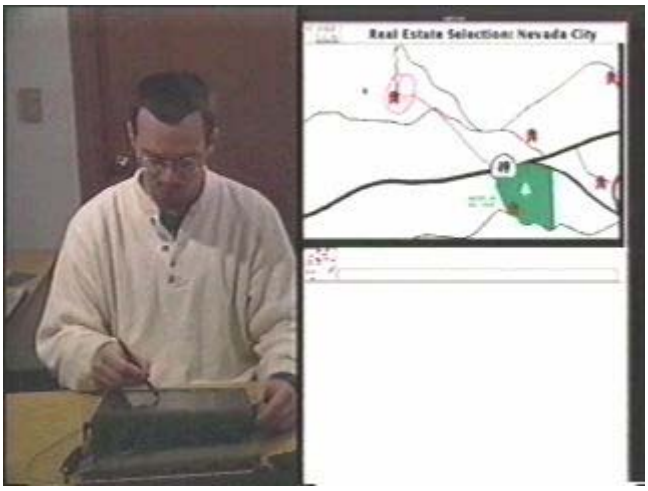

Figure 17 : Un utilisateur face au système multimodal (interaction par la parole et/ou par un stylo sur l'écran). A droite, une vue de son écran (Oviatt, 1996a).

Oviatt a obtenu les résultats suivants :

- Le nombre moyen de mots par phrase est significativement inférieur en condition multimodale (4,8 mots) par rapport à la condition monomodale verbale (6,2 mots).
- Les incorrections et auto-corrections dans l'expression (incluant les corrections du contenu, les faux départs, les répétitions, les pauses du type « euh… ») représentent globalement 2% des mots. Oviatt compare ce taux avec ceux qu'elle avait précédemment obtenus dans une autre étude (Oviatt, 1995). Cette comparaison suggère que l'interaction spatiale (manipulation de cartes) engendre plus d'incorrections que d'autres types d'interactions (domaine numérique, c'est-à-dire réalisation de calculs ; et domaine verbal, c'est-à-dire complétion de formulaires). Le taux d'incorrections est corrélé positivement à la longueur moyenne des phrases, il est donc significativement inférieur en condition multimodale par rapport à la condition monomodale verbale.
- Les phrases sont plus longues et les incorrections plus nombreuses lorsque les utilisateurs doivent faire des descriptions géographiques.
- Les scénarios sont résolus plus rapidement en condition multimodale par rapport aux deux conditions monomodales. Ce bénéfice temporel n'apparaissait pas dans des scénarios touchant aux domaines numérique et verbal.
- Le nombre d'erreurs de l'utilisateur qui aboutissent à un échec de la communication (par exemple lorsqu'il confond Est et Ouest, et provoque donc une action opposée à son intention)

est significativement inférieur en condition multimodale par rapport à la condition monomodale verbale. La condition monomodale gestuelle est intermédiaire et ne diffère pas significativement des deux autres conditions.

- Enfin, 95% des utilisateurs ont dit avoir préféré l'interaction multimodale et 5% l'interaction gestuelle. Aucun n'a déclaré avoir préféré l'interaction monomodale verbale. La Figure 18, reproduite de l'article d'Oviatt (1996a), montre que ces préférences, ainsi que l'utilisation effective des modalités, varient en fonction du contenu de l'interaction. Pour construire la partie droite du graphique, Oviatt a relevé, pour chaque utilisateur, le type d'interaction utilisé en condition multimodale (c'est-à-dire lorsque l'utilisateur avait le choix) : interaction uniquement verbale, uniquement gestuelle, ou utilisation de chaque modalité au moins une fois (interaction multimodale). Il est donc à noter que pour une tâche verbale (consistant à remplir des formulaires), environ 40% des utilisateurs avaient déclaré préférer la modalité verbale et l'avait exclusivement utilisée durant tout le scénario. Les scénarios verbaux examinés par Oviatt (1995) concernaient des tâches d'inscription à une conférence et de location de voiture ; le contenu de ces tâches consistait principalement en des noms propres et des informations temporelles.

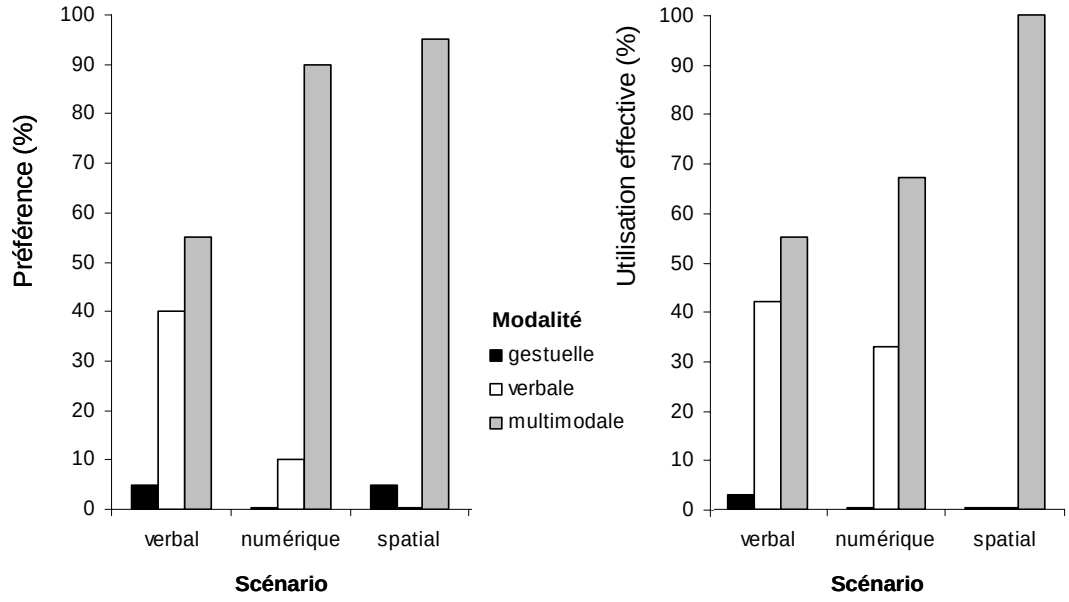

Figure 18 : Résultats obtenus par Oviatt (1996a). A gauche sont représentées les données de préférence pour les modalités en fonction du scénario d'utilisation. A droite, l'utilisation effective des modalités lors de la condition multimodale (où l'utilisateur choisit librement ses modalités d'interaction). Le scénario verbal consistait à remplir des formulaires, le scénario numérique à réaliser des calculs, le scénario spatial à manipuler des cartes.

L'étude d'Oviatt (1996a) montre donc clairement les avantages qu'il existe à utiliser la multimodalité dans un contexte spatial impliquant la manipulation de cartes et d'informations géographiques.

L'interaction est en effet plus rapide, moins complexe à traiter pour le système, engendre moins d'erreurs, et est plébiscitée par les utilisateurs. Un autre intérêt de cette étude est d'avoir mis ces résultats en relation avec des données obtenues dans d'autres types de scénarios (tâches numériques et verbales), montrant ainsi que la supériorité de la multimodalité est plus ou moins massive en fonction du domaine d'application du système. Nous retiendrons que pour une tâche verbale, la différence entre interaction verbale et multimodale est moins marquée.

L'objectif de l'étude d'Oviatt (1996a) était d'examiner l'utilité d'une Interface Multimodale pour la manipulation d'informations géographiques. Celle-ci étant établie, Oviatt *et al*. (1997) se sont penchées sur la manière dont parole et geste ont été combinés dans ce corpus. En effet, s'il existe dans le comportement des utilisateurs des patterns stables d'intégration des modalités, leur identification permettra de développer un système d'autant plus robuste. Ces analyses ont abouti aux résultats suivants :

- C'est la modalité verbale qui a globalement été la plus utilisée (63,5% des commandes), 19% des commandes ont été multimodales (combinaison de la parole et du geste au sein d'une même commande) et 17,5% gestuelles. La plupart des commandes multimodales (86%) concernent des actions nécessitant une localisation spatiale (ajouter un élément, le déplacer, etc.). Les actions de sélection représentent 11% des commandes multimodales. Dans 100% des cas, le stylo a été utilisé pour transmettre une information spatiale (localisation, taille, forme et nombre des objets).
- L'intégration des modalités a été simultanée (c'est-à-dire qu'il existait un chevauchement temporel) dans 56% des cas. Dans ces constructions simultanées, le geste précède la parole significativement plus souvent que l'inverse (57% *vs*. 14%). Le pattern le plus fréquent (35%) est : geste précède – parole clôture (cas 2 de la taxonomie d'Allen, 1983 ; voir paragraphe 3.2.4). Dans les constructions séquentielles, le geste précède la parole encore plus massivement (99%) ; le délai moyen entre la fin du geste et le début de la parole est de 1,4 secondes et n'excède jamais 4 secondes. Notons que les utilisateurs de cette expérimentation devaient signaler à l'aide du stylo chaque fois qu'ils faisaient un input verbal (système de *« click-to-speak »*). Un tel dispositif peut épargner 12,4% de mots inintelligibles. Cependant, il est incontestable que cette contrainte a dû influencer l'intégration des modalités. Oviatt *et al.* (1997) ont vérifié sur 8 utilisateurs avec un dispositif de parole sans contrainte (sur un système fonctionnel et non plus simulé) que la proportion de constructions simultanées et séquentielles était comparable, ainsi que l'ordre d'utilisation des modalités. Les auteurs ne précisent cependant pas si la proportion globale de constructions multimodales a également été maintenue. Etant donné que les cas où le *click-to-speak* était situé sur l'objet évoqué verbalement avaient été comptabilisés comme constructions multimodales, il est possible que le nombre total de ces constructions ait été légèrement biaisé.

- En ce qui concerne l'utilisation du stylo dans les constructions multimodales, les figures géométriques (rectangles, lignes) représentent 48% des cas, puis viennent les symboles (croix, flèches, 28%), les simples pointages (17%) et enfin les mots écrits (7%).
- L'utilisation de déictiques combinés aux gestes a également été étudiée : 41% des commandes multimodales contenaient un déictique, la relation temporelle le liant au geste qui le désambiguïse était séquentielle dans la majorité des cas (délai moyen de 1 seconde), le geste précédant toujours le déictique.
- Les informations apportées par chaque modalité sont majoritairement **complémentaires** les unes aux autres. Oviatt *et al*. ont relevé une certaine redondance entre parole et geste dans seulement 2% des cas de coopération.

Les principaux résultats que cette étude nous apporte concernent la fréquence des commandes multimodales (19%), et les patterns d'intégration observés : le geste précède presque toujours la parole. Il y a chevauchement entre les deux modalités dans un peu plus de la moitié des cas, mais nous supposons que cela est en partie lié au mode d'utilisation du stylo : si les pointages (utilisation très brève du stylo) avaient été plus nombreux, il y aurait sans doute eu davantage de constructions séquentielles. Ces résultats ont été utilisés par la suite dans le cadre du développement du système QuickSet (Cohen *et al.*, 1997) : par exemple, l'interpréteur multimodal de QuickSet a été programmé de manière à ce que l'intégration multimodale ne soit réalisée que lorsque qu'un geste est suivi d'une commande verbale dans un court délai. Il n'y a pas intégration entre un geste et la commande verbale qui le précède. Enfin, nous retiendrons de l'étude d'Oviatt *et al*. (1997) que les deux modalités ont été utilisées pour transmettre des informations **complémentaires**.

Les analyses sur les patterns observés ont été complétées par la suite (Oviatt, 1999b), et il s'est révélé possible de classer les utilisateurs en deux catégories bien distinctes : ceux qui, de manière stable tout au long du scénario, ont combiné leurs modalités d'interaction selon un pattern séquentiel (64% de l'échantillon testé), et ceux qui les ont combinées selon un pattern simultané. Pour chaque individu, le pattern prédominant a été utilisé dans 90% des constructions multimodales, et est presque toujours apparu dès la première d'entre elles, rendant ainsi possible de prédire très tôt quel pattern va prédominer. Un tel résultat suggère qu'il serait pertinent que les systèmes puissent détecter le pattern prédominant des utilisateurs et s'y adapter pour améliorer l'interaction (par exemple diminuer le délai d'attente pour les utilisateurs suivant un pattern simultané).

Dans des recherches ultérieures, l'équipe d'Oviatt (Xiao *et al.*, 2002 ; 2003) a étudié la stabilité des patterns multimodaux en fonction de l'âge des utilisateurs. Dans une première étude (Xiao *et al.*, 2002), 24 enfants de 6 à 10 ans ont testé un logiciel pédagogique simulé avec Interface Multimodale (parole et stylo sur tablette-écran portable), traitant de biologie marine (système *I SEE! Immersive*

Science Education for Elementary kids). L'interface comportait en sortie un petit agent conversationnel représentant un dauphin et différents animaux marins avec qui il était également possible de converser (Figure 19).

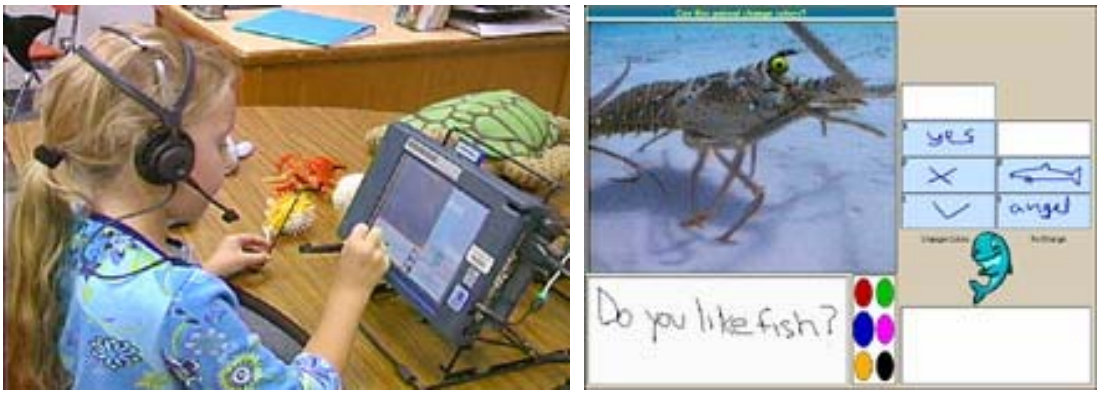

Figure 19 : Une jeune utilisatrice du système *I SEE!* et une vue de son écran.

Les analyses ont donné les résultats suivants :

- Parmi les messages en entrée, 80% étaient verbaux, 10% multimodaux et 10% gestuels. Dans 93% des constructions multimodales, le stylo a été utilisé pour « gribouiller » (pas de sémantique apparente). Parmi les constructions restantes (avec interprétation possible de l'entrée gestuelle), 93% impliquaient une utilisation complémentaire des modalités et 7% une utilisation redondante.
- Les enfants ont davantage recouru à la multimodalité lorsqu'ils devaient rectifier un message incorrect qu'ils venaient de produire.
- Au niveau de la synchronisation des modalités, les enfants ont majoritairement suivi un pattern d'intégration simultané : 70% du total des constructions sont simultanées et ce pattern est dominant chez 77% des enfants. Chaque individu a effectivement un pattern prédominant qui apparaît dès la première construction multimodale.
- Les enfants sont significativement plus nombreux que les adultes (cf. échantillon d'Oviatt, 1999b) à suivre un pattern d'intégration simultané.
- Dans 97% des constructions séquentielles, le geste précède la parole avec un délai moyen de 0,75 secondes qui n'excède pas 2 secondes. Ce délai moyen s'élève à 1,1 secondes si seules les constructions avec geste interprétable sont considérées.

Les patterns d'intégration multimodale des enfants présentent donc de nombreuses similitudes avec ceux des adultes (existence de patterns prédominants prédictibles, précédence du geste sur la parole dans les constructions séquentielles, utilisation complémentaire des modalités). Les enfants suivent cependant plus souvent un pattern simultané que les adultes (77% *vs*. 36% respectivement). Xiao *et al*.

(2002) ne précisent pas l'ordre des modalités dans les intégrations simultanées chez les enfants. Au niveau des constructions séquentielles, les enfants intègrent les modalités plus rapidement (délai maximum de 2 secondes, contre 4 chez les adultes) : le délai d'attente du système pourrait ainsi être réduit pour les applications destinées aux enfants, et l'interaction pourrait donc être plus dynamique. La différence d'utilisation du stylo et de taux de constructions multimodales entre adultes et enfants peuvent être attribuées aux contextes d'utilisation différents de ces deux études (spatial *vs*. conversationnel).

Les patterns d'intégration multimodale des seniors ont aussi été examinés (Xiao *et al.*, 2003) : quinze adultes âgés de 66 à 86 ans ont testé une Interface Multimodale simulée (utilisant parole et stylo sur tablette-écran) dans un scénario de gestion de crue sur carte géographique. Les analyses ont mis en évidence que :

- 87% des seniors observent un pattern d'intégration prédominant (ce qui représente une proportion moins élevée que chez les adultes plus jeunes et les enfants). Le pattern prédominant, lorsqu'il existe, est simultané dans 92% des cas (soit 80% de l'échantillon total), et est prédictible dès la première construction multimodale dans 85% des cas. Lorsque les modalités sont utilisées de manière séquentielle, le stylo précède la parole dans 64% des cas. Le délai moyen dans les intégrations séquentielles est de 1 seconde, et il est inférieur à 2 secondes dans 90% des cas. Ce délai augmente significativement avec la difficulté de la tâche (de 1 à 1,5 secondes).
- 80% des seniors se parlent à eux-mêmes à un moment ou à un autre de la réalisation du scénario. Le taux de paroles auto-adressées est corrélé positivement au taux d'erreurs commises dans le scénario.

Dans une autre série d'expériences (Oviatt *et al.*, 1998 ; 2003), l'équipe d'Oviatt a testé dans quelle mesure le comportement multimodal des utilisateurs est flexible aux conditions de passation. Oviatt *et al*. (1998) se sont intéressés à l'adaptation du comportement de l'utilisateur en cas d'erreur ou d'échec dans le système de reconnaissance. Le système simulé était une interface de type formulaire permettant de s'inscrire à un congrès et de réserver une voiture. Vingt utilisateurs ont eu à réaliser plusieurs tâches de ce type en utilisant librement la parole et le geste (stylo sur tablette-écran). Des erreurs répétées étaient simulées lors de la réalisation de certaines de ces tâches (deux à six erreurs de suite selon les essais). Les résultats montrent que, hors erreur, la parole a été utilisée à 82% et le stylo à 18%. En résolution d'erreur, l'utilisation de la parole est tombée à 70% et celle du stylo est montée à 30%. La probabilité de changement de modalité d'entrée est de 0,14 à la première erreur et augmente jusqu'à 0,39 lors de la quatrième à sixième erreur répétée. Les auteurs ont également étudié en détail tous les changements linguistiques et paralinguistiques apparus en résolution d'erreur (contenu lexical, longueur des phrases, pauses entre les mots, volume, hauteur, intonation de la voix). L'utilisation

combinée de la parole et du geste n'est ici que très rapidement évoquée : les auteurs signalent que moins de 1% des corrections d'erreurs ont impliqué l'utilisation simultanée et redondante des deux modalités.

Enfin, à l'aide d'une procédure expérimentale très sophistiquée (Oviatt *et al.*, 2003), la persistance des patterns prédominants a été mise à l'épreuve : dans cette expérience, le taux d'erreurs simulées en provenance du système était manipulé dans le but de renforcer le pattern opposé à celui des utilisateurs. Par exemple, si un pattern séquentiel est détecté en début de scénario, un taux de 40% d'erreurs de traitement va être simulé pour toutes les constructions séquentielles. A l'inverse, les constructions simultanées vont être gratifiées de 0% d'erreurs de traitement. Douze adultes ont été soumis à ce protocole : l'Interface Multimodale simulée (parole et stylo) permettait de réaliser des tâches de gestion de crue sur une carte géographique. Les résultats ont montré que :

- Onze utilisateurs /12 ont suivi un pattern dominant simultané. La prédominance chez les adultes du pattern séquentiel (observée par Oviatt, 1999b) n'est donc pas retrouvée. Pour les constructions séquentielles, le stylo précède la parole dans 83% des cas.
- Aucun utilisateur n'a modifié son pattern d'intégration multimodale malgré les 40% d'erreurs simulés. Bien au contraire, les utilisateurs ont **accentué** leur pattern prédominant : pour ceux ayant un pattern simultané, la durée de chevauchement a significativement augmenté au cours de la passation et le décalage entre modalités a significativement diminué. Réciproquement, pour l'utilisateur ayant un pattern séquentiel, le délai entre modalités s'est allongé. Cette accentuation apparaît également en fonction de la complexité de la tâche (le niveau de difficulté augmentait progressivement au cours de la passation). Oviatt *et al.* (2003) rapprochent ces résultats de la théorie de la Gestalt.

En résumé, nous retiendrons que l'équipe d'Oviatt a étudié de manière extensive et détaillée :

- L'utilisation de la multimodalité : 19% chez des adultes lors d'une tâche spatiale (Oviatt *et al.*, 1997), 10% chez des enfants dans un scénario conversationnel (Xiao *et al.*, 2002). La multimodalité est plus rapide et diminue le taux d'erreurs ; la syntaxe langagière est simplifiée et les incorrections sont moins fréquentes en condition multimodale (Oviatt, 1996a). Dans une étude ultérieure, Oviatt *et al.* (2004) ont montré que, dans une même application, l'augmentation de la charge mentale provoque une augmentation de l'utilisation de la multimodalité.
- L'intégration **temporelle** des modalités : il existe des patterns prédominants individuels, et prédictibles. Au total, 70% des utilisateurs (Oviatt, 1999b ; Xiao *et al.*, 2002 ; Oviatt *et al.*, 2003 ; Xiao *et al.*, 2003) suivent un pattern simultané (chevauchement entre parole et geste dans les constructions multimodales). Dans les constructions séquentielles, le geste précède la parole (64% à 99% selon les échantillons testés). Les délais mesurés entre modalités (0,75 à

1,4 secondes selon les échantillons, délai toujours inférieur à 4 secondes) constituent une base pour le réglage des seuils temporels de fusion dans les systèmes multimodaux.

- L'adaptation du comportement en cas d'erreur du système : lorsque les erreurs se répètent, la probabilité que l'utilisateur change de modalité augmente légèrement (Oviatt *et al.*, 1998). Pour les constructions multimodales, le pattern d'intégration prédominant est renforcé en cas d'erreur de la part du système (Oviatt *et al.*, 2003).

Nous remarquons que l'intégration **sémantique** des modalités a été moins approfondie : seuls deux types de coopération sont évoqués (complémentarité/redondance) sans être illustrés. La complémentarité entre modalités semble être massivement la plus fréquente, entre 93% et 99% des cas de constructions multimodales (Oviatt *et al.*, 1997 ; Oviatt *et al.*, 1998 ; Xiao *et al.*, 2002).

S'il semble évident qu'Oviatt et ses collaborateurs ont fourni les analyses les plus fines sur le comportement spontané des utilisateurs face à une Interface Multimodale, nous pouvons tout de même trouver dans la littérature d'autres données empiriques qui viennent tantôt corroborer, tantôt contredire les résultats d'Oviatt. Le paragraphe suivant présente une vue synthétique de ces études.

Etudes complémentaires

Le Tableau 7 (réparti sur les pages suivantes) présente les principales observations (classées par ordre alphabétique de leurs auteurs) relatives à l'utilisation d'Interfaces Multimodales variées. Nous avons volontairement restreint cette revue de question aux interfaces utilisant la parole **orale**. En effet, certains auteurs (Huls & Bos, 1995 ; Petrelli *et al.*, 1997 ; Trafton *et al.*, 1997 ; De Angeli *et al.*, 1998) ont étudié la multimodalité entre parole **écrite** (clavier ou stylo) et dispositif de pointage (souris ou stylo). Or, par le fait même qu'elle ne peut être simultanée au geste (puisqu'elle requiert naturellement la main dominante), l'utilisation de la parole écrite ne saurait être comparée à l'utilisation de la parole orale.

Par ailleurs, une des recherches citées dans le Tableau 7 a été menée en communication Homme-Homme, mais avec l'objectif spécifique d'obtenir des données pour l'Interaction Homme-Machine. Il s'agit de l'étude de Catinis et Caelen (1995), dans laquelle le sujet s'adresse non pas à un système informatique, mais à un assistant humain pour lui faire faire des actions sur un système. Nous pensons cependant qu'il ne s'agit pas réellement ici de communication Homme-Homme dans la mesure où plusieurs contraintes étaient imposées au sujet : le vocabulaire de commande qui lui était permis était limité, et la syntaxe restreinte afin qu'il n'y ait aucune ambiguïté dans les commandes verbales. De plus, l'objet de l'interaction se trouvait sur l'écran d'ordinateur et non dans l'espace conversationnel ;

enfin, il n'y avait pas d'interaction entre le sujet et l'assistant : l'assistant devait exécuter les actions prescrites mais ne parlait pas au sujet. Cet assistant n'était d'ailleurs pas un sujet lui-même mais au contraire un expérimentateur entraîné. Pour toutes ces raisons, nous avons classé les résultats de cette expérience parmi les données d'Interaction Homme-Machine.

Signalons également que les résultats de Zanello (1997), cités dans le Tableau 7, ont été obtenus avec des systèmes simulés, mais pour lesquels les utilisateurs étaient avertis de la simulation (contrairement aux expériences en Magicien d'Oz dans lesquelles les utilisateurs pensent interagir avec un système fonctionnel).

Référence	**Dispositifs d'entrée**	**Application**	**Principaux résultats**
(Ando *et al.*, 1994)	Micro, écran tactile	Aménagement d'intérieur, décoration	40 sujets. En Magicien d'Oz, les sujets préfèrent utiliser la modalité verbale par des phrases (syntaxe normale). Avec le système fonctionnel, ils préfèrent utiliser des noms de commandes.
(Bell *et al.*, 2000a)	Micro, souris	Carte immobilière + agent animé	16 sujets. La manière dont l'information est présentée (complémentarité verbale/graphique « cet appartement » + animation de l'icône *vs*. spécialisation verbale « l'appartement jaune » sans animation) influence le comportement des utilisateurs (adoption de la même stratégie).
(Bellik, 1995)	Micro, écran tactile, souris	Dessin	16 sujets. 91% de constructions multimodales. La parole précède le geste dans 62% des cas. Il y a simultanéité dans 30% des cas. La parole est utilisée en mots isolés (72%) plutôt qu'en flot continu (28%). Les sujets utilisent soit l'écran tactile, soit la souris, rarement les deux dans un même scénario. La parole est utilisée pour les noms de commande et de couleurs, l'écran tactile et la souris pour les arguments de forme et de position.
(Bourguet & Ando, 1998)	Micro, datagloves (pour gestes 3D)	Question/réponse, navigation dans un plan, dessin	10 sujets. Le geste de pointage est synchronisé soit avec le mot-clé de la commande (généralement un nom), soit avec le déictique.
(Caelen & Bruandet, 2001)	Micro, clavier, souris	Navigateur Internet	10 sujets. La navigation classique (réaction) est remplacée par un dialogue (coopération). L'interaction dialoguée est considérée comme bonne à 55,6% (mauvaise à 33,3%). 47% des utilisateurs trouvent ce système meilleur qu'un formulaire traditionnel du web (41% le trouvent moins bon).

Référence	**Dispositifs d'entrée**	**Application**	**Principaux résultats**
(Catinis & Caelen, 1995)	Micro, gestes des mains sur l'écran, clavier, souris	Dessin	26 sujets. 85% des sujets préfèrent l'interaction multimodale (parole et gestes des mains sur l'écran). 67% des constructions multimodales sont complémentaires. Dans 67% des cas le déictique est prononcé simultanément au geste, dans 32% le geste précède le déictique. L'usage de la redondance varie de en fonction de la nature des commandes : les commandes critiques comme « effacer » sont les plus redondantes (42%), puis viennent les commandes de déplacement (environ 20%) et enfin les commandes usuelles de dessin (moins de 10% de redondance).
(Costantini *et al.*, 2002)	Micro, stylo	Informations touristiques	35 sujets. Préférence pour la condition multimodale par rapport à verbale. Dans les constructions multimodales, les gestes suivent presque toujours la parole.
(Corradini & Cohen, 2002)	Micro, capteurs sur la tête et sur le bras dominant	Jeu en environnement virtuel	11 sujets. Utilisation du geste pour des actions de manipulation plus que de désignation. Les sujets préfèrent l'interaction multimodale plutôt que monomodale.
(Denda *et al.*, 1997)	Micro, écran tactile	Cartes touristiques	10 sujets. Pas de différence de temps de réalisation du scénario entre condition mono- et multimodale.
(Hauptmann, 1989)	Micro, gestes 3D	Manipulation d'objets graphiques	36 sujets. 71% de constructions multimodales, 13% gestuelles, 16% verbales. Structure syntaxique très simple (5 mots/commande en moyenne). 58% des sujets préfèrent l'interaction multimodale, 20% le geste seul, 22% la parole seule.
(Kehler, 2000)	Micro, stylo	Cartes touristiques	13 sujets. 65% des références aux éléments de l'interface sont faites par la parole seule. 21% sont des constructions multimodales où la parole est complémentaire au geste et 14% sont des constructions redondantes.

Référence	**Dispositifs d'entrée**	**Application**	**Principaux résultats**
(Kehler *et al.*, 1998)	Micro, stylo	Cartes touristiques	17 sujets. 51% de commandes verbales, 35% gestuelles, 14% multimodales.
(Mignot *et al.*, 1993 ; Zanello, 1997)	Micro, écran tactile	Contrôle de process, Aménagement de pièces	8 sujets. Constructions multimodales chez 6 sujets /8. La redondance est majoritaire. Analyses complétées par Zanello (1997) : 40% de constructions multimodales, 43% monomodales verbales, 17% monomodales gestuelles. 58% de constructions redondantes, 42% de complémentaires. En complémentarité, 83% des gestes portent sur un objet et/ou un lieu.
(Qvarfordt & Jönsson, 1998)	Micro, clavier, souris	Horaires de bus	12 sujets. En condition multimodale, les erreurs sont moins nombreuses et l'utilisation du système est plus optimale (meilleure anticipation des fonctionnalités).
(Robbe, 1998)	Micro, écran tactile	Aménagement d'intérieur	8 sujets. Les contraintes verbales (lexique restreint à 100 mots) et gestuelles imposées aux sujets ne suscitent aucun problème d'utilisabilité.
(Robbe-Reiter *et al.*, 2000)	Micro, écran tactile	Aménagement d'intérieur	8 sujets. Langage multimodal contraint. Lors de la première session, 58% des commandes sont verbales, 33% gestuelles, 9% multimodales (avec une grande variabilité interindividuelle). A la troisième session deux semaines plus tard, l'utilisation du geste est passée à 66%, la multimodalité n'a guère évolué. L'utilisation de contraintes semble favoriser la monomodalité.
(Rugelbak & Hamnes, 2003)	Micro, stylo sur PDA	Cartes touristiques	7 experts en IHM : 3/7 n'ont fait aucune construction multimodale. Pour les 4 autres : 87,5% des constructions sont complémentaires, et dans 81% les sujets pointent à la fin de la phrase, simultanément au déictique.

Référence	Dispositifs d'entrée	Application	Principaux résultats
(Siroux *et al.*, 1995)	Micro, écran tactile	Informations touristiques	Le geste est spontanément peu utilisé, les sujets y recourent parfois pour corriger des erreurs mais pas en premier choix.
(Sturm *et al.*, 2002)	Micro, écran tactile	Horaires de train	25 sujets. La version graphique (utilisée avec clavier et souris) est significativement plus rapide (temps de résolution du scénario) que la version multimodale, elle-même plus rapide que la version vocale (téléphone) du service. La version graphique est également préférée par les utilisateurs.
(Zanello, 1997)	Micro, souris	Dessin	7 sujets. 20% de constructions multimodales, 63% monomodales gestuelles, 17% monomodales verbales. 88% des constructions sont complémentaires, 8% redondantes et 4% concurrentes. Le geste précède la parole dans 99,5% des cas.
(Zanello, 1997)	Micro avec push-to-talk, souris	Visiophone, carnet d'adresses	14 sujets. 9% de constructions multimodales, 73% monomodales gestuelles, 18% monomodales verbales. 24% de constructions complémentaires, 56% de redondantes, 20% de concurrentes. Le geste précède toujours la parole.

Tableau 7 : Récapitulatif de résultats comportementaux sur l'utilisation d'Interfaces Multimodales (classement alphabétique par référence).

Notre analyse de l'utilisation d'un guide de programmes télévisés multimodal

Dans le cadre d'un projet de conception d'une interface de télévision multimodale (projet RNRT-TVi[19]), nous avons été amenés à recueillir un corpus d'interaction avec un guide de programmes télévisés (Buisine & Martin, 2002). Notre objectif était double : (1) tester la pertinence de la multimodalité dans cette application, et (2) recueillir un premier corpus d'interaction multimodale dans la perspective de développer un système fonctionnel. Nous avons donc choisi de comparer la réalisation de scénarios dans trois conditions d'interaction : une condition monomodale verbale, une condition monomodale gestuelle, et une condition multimodale intégrant les deux précédentes.

Utilisateurs. Six membres du département Mécanique du LIMSI, non informés du projet, ont été recrutés. Il s'agissait de doctorants, quatre hommes et deux femmes, d'âge moyen 26 ans et 10 mois.

Matériel. L'expérimentation a été réalisée avec une interface existante, celle du service de recherche multicritère de programmes télévisés disponible sur le site Internet de Canalsatellite[20] (voir la Figure 20). Le bandeau supérieur, le menu « Les Thèmes », et le menu latéral, présents sur toutes les pages du site de Canalsatellite, ne font pas partie du service de recherche de programme. Celui-ci fonctionne à l'aide d'un menu horizontal (options situées sous le titre « Tous les programmes » dans la Figure 20), d'une série de six menus déroulants pour la recherche multicritère et de deux moteurs de recherche (par mots-clés et par chaîne).

Pour les besoins du test, les utilisateurs naviguaient dans cette interface :

- A l'aide d'un stylo directement sur l'écran (conditions gestuelle et multimodale),
- A l'aide de la parole (conditions verbale et multimodale).

La reconnaissance et la compréhension de la parole ont été simulées selon la technique du Magicien d'Oz (voir le paragraphe 3.2.2 pour une introduction à cette méthode). Toutes les fonctionnalités de la souris étaient accessibles par le stylo : les menus et boutons de l'interface étaient donc activables normalement. Dans le cas où un utilisateur essayait d'écrire dans un des champs de recherche, les principaux mots-clés liés aux scénarios, préalablement encodés sous forme de *cookies*, apparaissaient dans une fenêtre d'auto-complétion. L'utilisateur n'avait plus qu'à sélectionner une de ces

[19] http://cpn.paris.ensam.fr/tvi/

[20] http://www.canalsatellite.fr/instranet/site/proghome.jsp

propositions. Ainsi, lorsque l'utilisateur interagissait avec le stylo, il provoquait lui-même la navigation, et lorsqu'il utilisait des commandes vocales, c'est le magicien qui agissait sur l'interface.

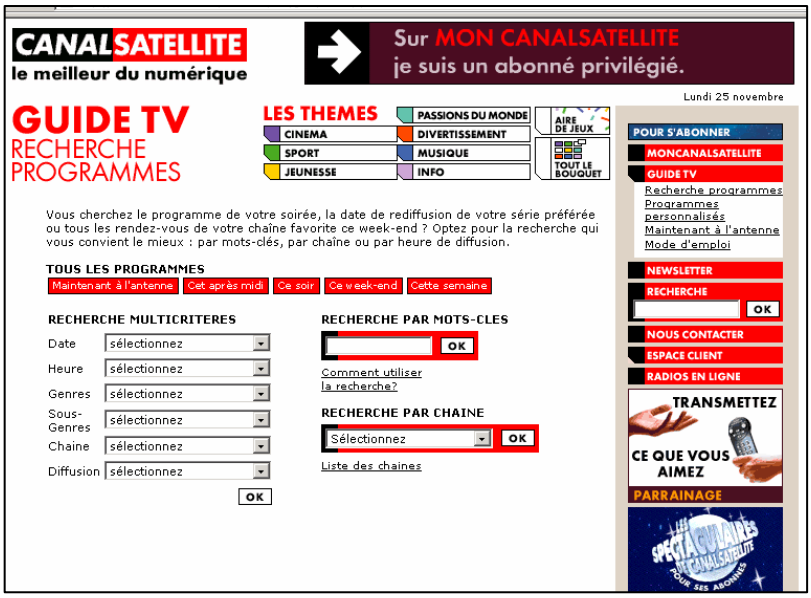

Figure 20 : Interface web de Canalsatellite utilisée pour le test.

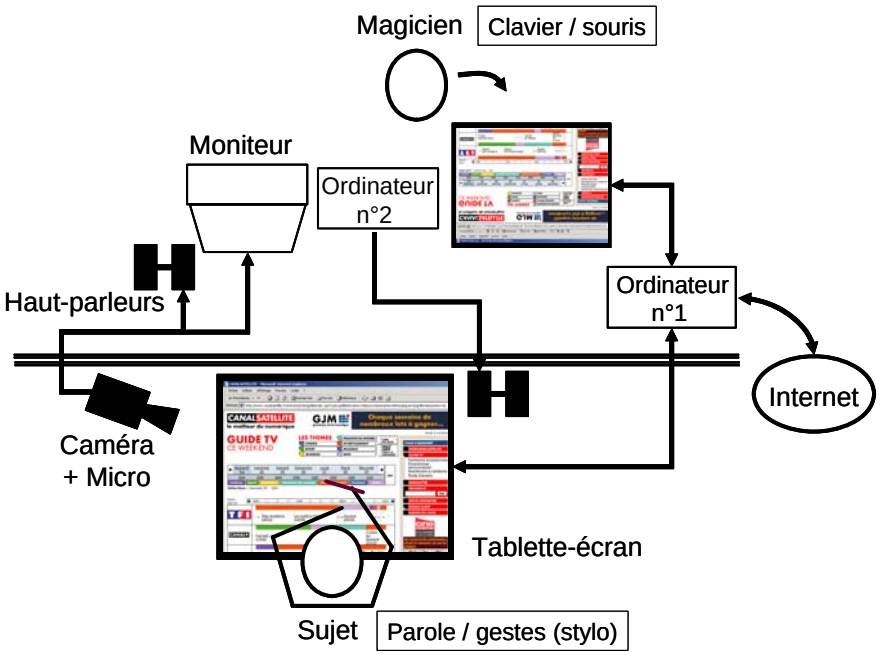

Figure 21 : Schéma du dispositif expérimental.

Le dispositif expérimental utilisé est schématisé dans la Figure 21. L'ordinateur n°1 était relié, dans une pièce, à un écran classique destiné au magicien. Il était également connecté, dans une autre pièce, à une tablette-écran Wacom Cintiq 15X permettant au sujet d'utiliser le stylo directement sur l'interface.

L'ordinateur n°1 étant relié à Internet, les deux écrans affichaient un navigateur web connecté sur le site de recherche de programmes. Les changements de l'interface étaient simultanés sur les deux écrans.

Du côté utilisateur, un micro était disposé derrière l'écran pour inciter les sujets à parler. Pour leur apporter un feedback en cas de commande impossible, deux haut-parleurs étaient disposés à côté de l'écran : le magicien pouvait déclencher, à partir de l'ordinateur n°2, un message d'erreur (« commande inexistante ») lorsque cela était nécessaire. La synthèse vocale utilisée était IBM ViaVoice[21].

Une caméra numérique, installée au-dessus de la tablette-écran, permettait un enregistrement audio-visuel du comportement de l'utilisateur en conservant son anonymat (le visage n'était pas filmé, cf. Figure 22). Un moniteur et des enceintes reliés à la caméra numérique permettaient au magicien de suivre en temps réel les paroles et les actions de l'utilisateur pour gérer l'interaction en conséquence.

Procédure. La séance débutait par l'explication des consignes et la signature du protocole d'accord pour les enregistrements vidéo. Les utilisateurs réalisaient ensuite successivement trois scénarios dans trois conditions d'interaction différentes :

- Interaction par la parole uniquement ;
- Interaction par le geste uniquement (stylo directement sur l'écran) ;
- Interaction multimodale (possibilité d'utiliser la parole et/ou le stylo).

L'ordre de passation des conditions et les combinaisons scénario/condition étaient contrebalancés. Les trois scénarios de recherches de programmes télévisés étaient les suivants :

- Scénario 1 : « Trouvez la date de dernière diffusion du magazine Forum Terre, qui passera lundi 25 novembre à 21h20 sur Encyclopedia » [Réponse : jeudi 28 novembre à 14h40]
- Scénario 2 : « Quelles chaînes diffusent un film de Science Fiction mardi 26 novembre ? » [Réponse : Ciné Cinéma Classic, Action, Ciné FX]
- Scénario 3 : « Quel est le programme diffusé ce soir à 20h sur l'Equipe TV ? » [Réponse : La grande édition (magazine)]

Ces scénarios ont été choisis en raison de leur complémentarité par rapport aux besoins pouvant être exprimés par les utilisateurs : le scénario n°1 propose la recherche d'une information très précise,

[21] http://www-3.ibm.com/software/speech/

ciblée sur un programme en particulier. Dans ce type de démarche, le choix de ce que l'utilisateur veut regarder est préalable à la démarche de recherche. Au contraire dans les scénarios n°2 et n°3, l'utilisateur n'a pas déterminé *a priori* ce qu'il va regarder (excepté sur des critères comme l'horaire ou le genre de programme). Dans les trois cas, les stratégies de recherche pouvaient être multiples (entrée de l'ensemble des critères dès la première requête, recherche large puis resserrements successifs, utilisation des moteurs de recherche…), et le scénario n°3 offrait aussi la possibilité d'utiliser le menu horizontal (option « Ce soir »).

Afin de s'assurer que le but du scénario est atteint (pour un critère d'efficacité de l'interaction), les utilisateurs devaient inscrire leurs réponses sur un formulaire.

En fin de séance, chaque utilisateur remplissait un questionnaire de satisfaction (voir l'Annexe 1). Les questions d'ordre général ont été inspirées des questionnaires habituellement utilisés pour l'évaluation d'interfaces (Nielsen, 1993 ; Perlman, 1997). Des questions plus spécifiques à l'interface testée et à la multimodalité ont été intégrées. Le résultat des évaluations sur les échelles analogiques a été relevé en centimètres (0 à 5 cm, les valeurs élevées représentant des évaluations d'autant plus positives).

L'expérimentateur expliquait ensuite à l'utilisateur que le fonctionnement du système était en partie simulé.

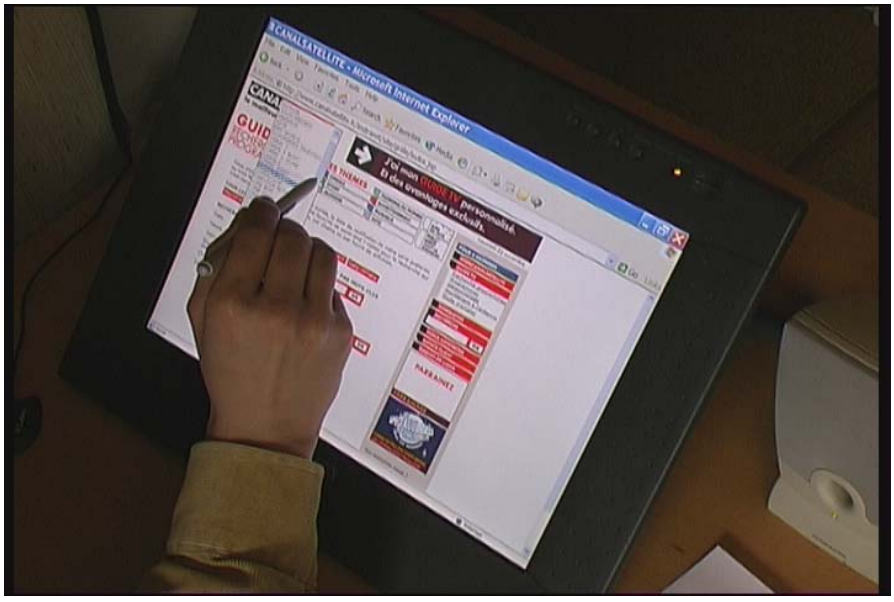

Figure 22 : Capture d'une vidéo.

Résultats. Six vidéos (une vidéo par utilisateur contenant les trois scénarios) ont été enregistrées (cf. Figure 22). La **durée moyenne** de réalisation d'un scénario a été de 3 minutes. Le test de

Friedman (analyse de variance non-paramétrique) ne met en évidence aucune différence significative dans cette durée en fonction des conditions (verbale, gestuelle, monomodale).

L'**analyse des questionnaires** montre que les évaluations sont très bonnes pour les trois conditions d'interaction. Toutes variables confondues, la satisfaction moyenne est :

- Pour la parole : 3,97/5 (écart-type = 1,06) ;
- Pour le stylo : 4,56/5 (écart-type = 0,52) ;
- Pour la condition multimodale : 4,55/5 (écart-type = 0,16).

En ce qui concerne la **préférence des utilisateurs** parmi les modes d'interaction, les résultats montrent que, en moyenne, le stylo a été placé en premier choix, la multimodalité en second et la parole en dernier. Plus précisément, trois utilisateurs ont préféré le stylo, deux utilisateurs ont préféré la parole et un dernier utilisateur a préféré la condition multimodale. Ceci tend à montrer qu'il existe des différences individuelles de préférences pour les modalités d'interaction. Les commentaires exprimés par les utilisateurs à l'issue du test révèlent que le stylo est apprécié parce qu'il constitue un mode d'interaction plus habituel.

En **condition multimodale**, la plupart des sujets n'ont utilisé qu'un seul des deux modes d'interaction proposés :

- Deux sujets ont majoritairement utilisé les commandes vocales ;
- Deux sujets ont utilisé exclusivement le stylo ;
- Un sujet a commencé le scénario par la parole puis l'a poursuivi à l'aide du stylo ;
- Un dernier sujet a utilisé principalement la multimodalité redondante (ex : « Genre : Film » et sélection simultanée de l'option Film dans le menu Genre).

La Figure 23 illustre la disparité des choix de modalités entre les utilisateurs.

Seuls deux utilisateurs ont produit des constructions multimodales : 64% d'entre elles (16/25) étaient redondantes et 36% complémentaires. Les constructions complémentaires consistaient généralement à dérouler un menu avec une modalité et à sélectionner une option de ce menu avec l'autre modalité (ex : utiliser le stylo pour dérouler le menu « chaînes » et sélectionner « Encyclopédia » par la parole).

En termes d'intégration temporelle, 68% des constructions (17/25) étaient simultanées (chevauchement temporel entre modalités).

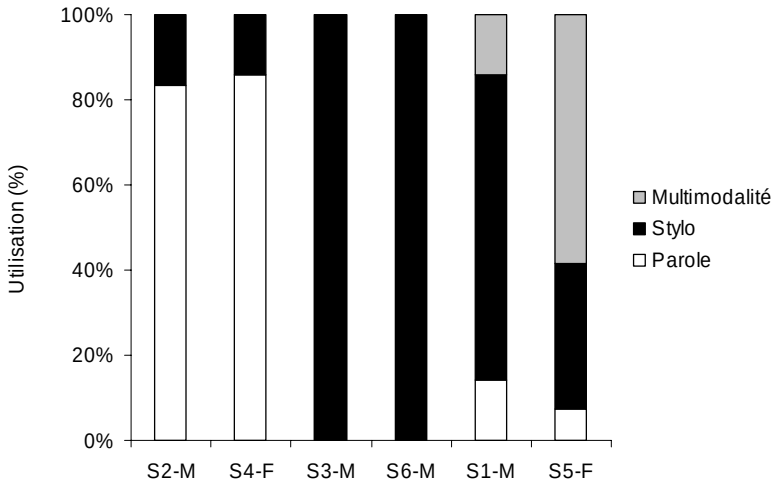

Figure 23 : Pourcentages d'utilisation des modalités pour chacun des six utilisateurs. La dernière lettre dans le nom des utilisateurs désigne leur sexe (M pour masculin, F pour féminin).

En ce qui concerne l'**utilisation des commandes vocales**, les enregistrements montrent que la syntaxe est très simplifiée : les utilisateurs n'ont pas construit de phrases pour exprimer leurs requêtes, et ont fréquemment employé des noms de commandes isolés (ex : « OK », « Sélectionner », « Annuler », « Fermer », « Retour », etc.). Par ailleurs, les expressions employées ont été fortement influencées par les labels présents dans l'interface. Par exemple, pour sélectionner la date du 25 Novembre, les utilisateurs disaient « Date » pour afficher les options du menu déroulant correspondant, puis ils lisaient « Lundi 25 Novembre ». Cette stratégie les a amenés à utiliser des termes qu'ils n'auraient peut-être pas spontanément choisis, par exemple « Genre » pour le type de programme (Film, Sport, Documentaire, etc.) et « Sous-genre » pour préciser la catégorie (Action, Rires, Suspense, etc. dans le cas des films). Chaque mot prononcé a été associé à un booléen (contenu dans l'interface oui ou non) dans l'ensemble du corpus : 68,9% (315/457 mots) des mots appartenaient à l'interface. Si à cela on ajoute les noms de commande qui ne sont pas explicitement écrits dans l'interface (« annuler », « retour », « sélection », « défiler »), ce taux atteint 71,3% (326/457). Un tel phénomène a déjà été observé par d'autres auteurs : Brennan (1996) emploie le terme de convergence lexicale pour désigner l'influence du contenu de l'interface sur le vocabulaire utilisé par les sujets. Oviatt (1996b) explique également que les interfaces structurées sous forme de formulaires incitent naturellement à faire des phrases courtes et permettent de réduire les incorrections dans le langage.

Résultats annexes : Les évaluations subjectives vis-à-vis de l'interface de Canalsatellite sont globalement bonnes. Deux utilisateurs ont tout de même signalé des problèmes de lisibilité sur certaines pages (logo des chaînes trop petits…), et un utilisateur s'est plaint d'un manque de guidage initial en raison de l'abondance d'informations sur les pages (due principalement aux menus non pertinents pour la recherche de programmes). Enfin, quelques problèmes de précision et de parallaxe

lors de l'utilisation du stylo sont parfois apparus, en raison du resserrement des options dans les menus déroulants.

En ce qui concerne les stratégies de recherches mises en œuvre par les utilisateurs, les enregistrements montrent une utilisation préférentielle des menus déroulants. Sur l'ensemble des sujets et des scénarios, la recherche par mots-clés n'a été utilisée que deux fois, la recherche par chaîne une fois, et le menu « Les Thèmes » (non spécifique au guide de programmes) une fois.

Conclusion. Eu égard au faible nombre d'utilisateurs, les analyses sur ce corpus sont limitées. Cette expérimentation apporte cependant un éclairage supplémentaire dans notre état de l'art sur l'observation du comportement multimodal des utilisateurs.

Rappelons que notre premier objectif était de tester la pertinence de la multimodalité dans cette application. Les analyses mettent en évidence qu'il n'y a pas de différence d'efficacité entre les trois conditions, et que les évaluations subjectives ont été excellentes dans tous les cas. Si l'on s'en tient à une moyenne stricte des scores de préférence, c'est le stylo qui remporte la plus grande adhésion (trois utilisateurs sur six). Mais si l'on veut satisfaire le plus grand nombre, alors c'est la multimodalité qui doit être retenue. En effet, cette solution permet à tous les utilisateurs d'interagir par leur modalité préférée, qu'elle soit gestuelle, vocale ou multimodale. Il faut cependant veiller à ce que le **recouvrement des fonctionnalités soit total** entre les modalités.

Notre second objectif était d'analyser l'usage de la multimodalité : à cet égard, nous avons vu que quatre utilisateurs sur six ont réalisé les scénarios sans produire une seule construction multimodale. Il est possible que ce soit l'interface qui ne se soit pas prêtée à la multimodalité (interface conçue pour une interaction classique par clavier et souris). Le stylo a effectivement été utilisé à la manière d'une souris (pointage, sélection d'options, activation des boutons de l'interface), et les commandes vocales se sont limitées à des noms d'action avec une syntaxe simplifiée. Ce dernier point suggère que, pour une telle application, la construction d'une grammaire vocale semble relativement simple et a toutes les chances d'être efficace pour l'utilisation que les sujets en font.

Discussion sur l'Interaction Homme-Machine multimodale

Nous pouvons à présent tenter de faire une synthèse de toutes ces études sur l'utilisation des Interfaces Multimodales :

- L'**intérêt de la multimodalité** ressort de manière stable, mais les raisons en sont multiples : meilleure efficacité (Oviatt, 1996a ; Qvarfordt & Jönsson, 1998), préférence des utilisateurs

(Hauptmann, 1989 ; Catinis & Caelen, 1995 ; Caelen & Bruandet, 2001 ; Corradini & Cohen, 2002 ; Costantini *et al.*, 2002), choix plus large dans la manière d'interagir (Buisine & Martin, 2002).

- En ce qui concerne la **fréquence d'utilisation de la multimodalité** (taux de constructions multimodales), les études recensées montrent qu'elle est plus élevée pour les applications spatiales (71% chez Hauptmann, 1989 ; 40% chez Mignot *et al.*, 1993 ; 91% chez Bellik, 1995 ; 19% chez Oviatt *et al.*, 1997 ; 14% chez Kehler *et al.*, 1998). Les applications verbales et conversationnelles ont été moins étudiées (10% chez Xiao *et al.*, 2002).
- Les résultats sur l'**intégration temporelle** des modalités semblent en faveur d'une précédence du geste sur la parole (Oviatt *et al.*, 1997 ; Zanello, 1997 ; Xiao *et al.*, 2002 ; 2003), avec chevauchement entre les modalités (Oviatt *et al.*, 2003), bien qu'un résultat opposé soit parfois rapporté (précédence de la parole sur le geste chez Bellik, 1995 et chez Costantini *et al.*, 2002).
- Enfin, les données sont assez variées sur l'**intégration sémantique** des modalités : prédominance forte de la complémentarité (Oviatt *et al.*, 1997 ; Zanello, 1997 ; Oviatt *et al.*, 1998 ; Xiao *et al.*, 2002 ; Rugelbak & Hamnes, 2003), répartition équilibrée entre complémentarité et redondance (Catinis & Caelen, 1995 ; Kehler, 2000), ou prédominance de la redondance (Mignot *et al.*, 1993 ; Zanello, 1997 ; Buisine & Martin, 2002).

L'absence de consensus général peut s'expliquer par de nombreux facteurs : la diversité des applications, des médias gestuels, la diversité des populations testées (langues et cultures différentes, catégories socioprofessionnelles…), la diversité des plateformes expérimentales (degré de simulation, efficacité des systèmes et des dispositifs Magicien d'Oz), etc. Ce constat nous inspire deux réflexions :

- Il nous semble difficile de formuler des recommandations sûres au vu de cette revue de question. Nous insisterons donc à nouveau sur l'importance d'un recueil de données spécifique à l'application et à la population visée pour chaque projet d'Interface Multimodale.
- Même dans le cas où l'implémentation de l'interface est guidée par les données comportementales, une certaine flexibilité doit être préservée : l'architecture doit permettre de traiter avec autant de précision et de succès des patterns d'intégration multimodaux variés (sur les dimensions temporelle et sémantique), ainsi que des commandes monomodales verbales ou gestuelles. Les différentes modalités doivent pouvoir être utilisées aussi bien comme des moyens d'interaction alternatifs l'un à l'autre que comme des éléments de communication qui peuvent être combinés de manière aussi naturelle qu'en interaction Homme-Homme.

Nous synthétisons dans le paragraphe suivant tous les éléments nous permettant d'anticiper le comportement des utilisateurs face à une interface multimodale intégrant un Agent Conversationnel en sortie.

3.3.3. Interaction Homme-Agent

Comme nous l'avons déjà souligné, le **comportement multimodal** des utilisateurs face aux Agents a été très peu étudié. Les deux seules études pertinentes que nous ayons trouvées sur ce sujet sont :

- L'étude de Xiao *et al*. (2002), dans laquelle des enfants interagissent avec des animaux sur le domaine de la biologie marine. Les constructions multimodales représentent 10% de ce corpus, mais une utilisation non sémantique du stylo (dans 93% de ces constructions) rend difficile l'interprétation du comportement des utilisateurs. Les constructions restantes étaient complémentaires à 93%.
- L'étude de Bell *et al*. (2000a), qui a montré que la coopération sémantique utilisée par le système (ex : complémentarité « cet appartement » + animation de l'icône *vs*. spécialisation verbale « l'appartement jaune » sans animation) influence le comportement des utilisateurs, ceux-ci adoptant la même stratégie que le système. Dans cette étude, l'Agent est une tête parlante : les informations graphiques sont donc transmises par le système (animation sur la carte) et non par l'Agent (puisqu'il ne peut faire un geste de pointage en direction de la carte). On pourrait supposer que des résultats similaires auraient été obtenus si la stratégie de présentation était attribuable à l'Agent : l'utilisation de la complémentarité entre parole et gestes de l'Agent incite l'utilisateur à utiliser la complémentarité entre parole et gestes. Cependant une telle hypothèse reste à tester.

Par ailleurs, il existe quelques résultats dans la littérature concernant l'**interaction verbale** avec les Agents. Le taux d'incorrections (*disfluencies*) – qui regroupe par exemple les faux départs, les répétitions, ou les « euh… » – est une variable importante pour les systèmes de dialogue, car difficile à gérer. Les travaux d'Oviatt (1995) ont montré que les systèmes multimodaux, parce qu'ils amènent naturellement les utilisateurs à faire des phrases plus courtes, font diminuer ce taux d'incorrections par rapport à celui de la communication Homme-Homme. Or, Oviatt et Adams (2000) comparant le taux d'incorrections dans le discours des enfants face à des Agents et face à un adulte, ont montré que les incorrections étaient beaucoup moins nombreuses face aux Agents. Oviatt et Adams (2000) en concluent que les enfants ont adopté une attitude différente face aux Agents (du type Interaction Homme-Machine) et face aux adultes (Interaction Homme-Homme).

Enfin, Bell *et al*. (2000b) ont étudié le comportement des utilisateurs face à un système dialogique monomodal (conversation téléphonique) et un système multimodal (interaction avec un Agent Conversationnel et une carte cliquable). La confrontation de ces deux corpus montre notamment que le système multimodal engendre davantage de comportements sociaux et conversationnels, ce que les

auteurs attribuent à la présence visuelle de l'Agent. Dans une précédente étude (Bell & Gustafson, 1999), les auteurs avaient relevé une proportion de 40% de comportements sociaux dans un corpus d'interaction verbale avec un Agent Conversationnel intégré à un système délivrant des informations touristiques aux passants.

3.4. Conclusion

Dans un souci de symétrie de l'interaction avec les Agents Conversationnels, nous souhaitons permettre aux utilisateurs humains d'utiliser plusieurs modalités de communication. Dans ce chapitre, nous avons donc introduit et illustré le domaine des Interfaces Multimodales en entrée. Nous avons exposé brièvement quelques méthodes de conception dont nous pourrons nous inspirer. Pour alimenter nos réflexions sur l'Interaction Homme-Machine multimodale, nous avons discuté quelques observations sur la communication Homme-Homme, puis nous avons mené une revue de question sur l'utilisation des Interfaces Multimodales. A l'issue de ce chapitre, il semble difficile d'anticiper le comportement multimodal que les utilisateurs vont adopter face aux Agents, de savoir s'il aura davantage les caractéristiques de la communication Homme-Homme ou Homme-Machine. La question du taux de constructions multimodales et de la coopération des modalités reste également ouverte et nécessite indiscutablement de nouvelles données d'observation.

4. Synthèse de l'état de l'art et problématique

4.1. Les Agents Conversationnels Multimodaux Bidirectionnels

Nous souhaitons développer des Agents Conversationnels avec lesquels les utilisateurs pourront communiquer en utilisant plusieurs modalités. Ce choix de conception repose sur la notion de **symétrie** dans les canaux de communication des deux partenaires de l'interaction. Soulignons cependant que, dans le cadre de notre projet, la symétrie restera limitée : en sortie du système, les Agents Conversationnels auront la possibilité d'utiliser de multiples modalités pour transmettre de l'information aux utilisateurs (parole, gestes opératoires, gesticulation, expressions faciales, déplacement…). En revanche, en entrée du système, les utilisateurs ne disposeront que de deux modalités : parole et gestes opératoires en deux dimensions à la surface de l'écran.

Certains systèmes existants (fonctionnels ou simulés) ont déjà réalisé cette bidirectionnalité de la communication multimodale, mais très peu de données sont disponibles sur l'utilisation et l'évaluation de ces systèmes.

4.2. Le comportement multimodal des Utilisateurs

Notre étude du comportement multimodal des utilisateurs a pour objectif de contribuer spécifiquement au projet NICE, et en particulier au développement de l'interface multimodale en entrée. Rappelons que l'application visée est un jeu éducatif mettant en scène l'auteur H.C. Andersen et des personnages de ses contes. Le scénario comprend à la fois des aspects conversationnels et des aspects orientés tâche

(ex : problème à résoudre). L'application est destinée à des enfants, adolescents et jeunes adultes de 9 à 18 ans.

La prise en compte des données comportementales le plus tôt possible dans le projet vise à optimiser le développement de l'interface multimodale en entrée : la phase de programmation devrait en effet être plus efficace si elle s'appuie sur les patterns multimodaux spontanés des utilisateurs et sur un corpus d'exemples recueillis dans une situation proche de l'interaction avec le futur système. Mais de telles données comportementales n'existent pas à ce jour dans la littérature.

En effet, les patterns comportementaux décrits dans le domaine des Interfaces Multimodales en entrée concernent l'interaction avec des interfaces sans Agent (voir le chapitre 3). Or, nous pouvons supposer que le comportement des utilisateurs sera influencé par le fait de s'adresser à des Agents Conversationnels et non à une interface graphique classique, non personnifiée. Les quelques travaux antérieurs sur l'interaction verbale avec des Agents suggèrent qu'elle constitue un intermédiaire entre la communication Homme-Homme et l'Interaction Homme-Machine : l'insertion de thèmes sociaux dans l'interaction (Bell *et al.*, 2000b) tend à montrer que les Agents sont « plus que des machines », mais le fait que les utilisateurs épurent spontanément leur langage (diminution de la fréquence des incorrections, augmentation des répétitions et sur-articulation) montre qu'ils ne font tout de même pas l'amalgame avec des humains (Oviatt & Adams, 2000). A un niveau pragmatique pour le système NICE, nous pouvons faire l'hypothèse que l'**interaction verbale** sera plus riche face à des Agents que face à des interfaces non personnifiées : alors que la syntaxe verbale semble simplifiée dans l'Interaction Homme-Machine (Oviatt, 1995, 1996a, 1996b), nous pensons retrouver avec les Agents une syntaxe de type communication Homme-Homme.

Si nous pouvons donc faire de premières hypothèses sur les comportements verbaux des utilisateurs face au système NICE, aucune donnée ne permet d'anticiper les caractéristiques de leurs **constructions multimodales**. Plusieurs facteurs fondamentaux distinguent en effet le projet NICE des systèmes habituellement étudiés, ce qui empêche de transposer les résultats de la littérature. Parmi ces facteurs, citons :

- La présence d'**Agents** tout d'abord, dont on peut supposer qu'elle incitera à l'interaction verbale.
- Le **genre** de l'application est différent de celui des systèmes classiquement étudiés dans la littérature : nous nous intéressons en effet à un scénario de jeu, incluant des aspects conversationnels et des aspects orientés tâche avec des objets graphiques. Nous supposons que ce contexte pourra avoir une influence sur l'interaction gestuelle et multimodale.
- Notre **population** cible n'est pas non plus celle qui a été la plus testée dans les études antérieures : il existe en effet très peu de données sur le comportement multimodal des

enfants. A notre connaissance, les seules observations sur ce sujet sont celles recueillies avec le système *I SEE!* (Xiao *et al.,* 2002). Or nous avons vu que ce corpus avait été parasité par l'utilisation non sémantique du stylo par les enfants (gribouillages). De nouvelles données sont donc nécessaires pour compléter celles-ci. Par ailleurs, notre système étant destiné à des joueurs d'âges variés, cela nous donnera l'occasion de comparer le comportement d'enfants et de jeunes adultes sur une même application, ce qui semble ne jamais avoir été fait auparavant.

Il apparaît donc nécessaire de recueillir et d'analyser le comportement multimodal d'utilisateurs enfants, adolescents et adultes face à des Agents Conversationnels. Plus précisément, nous nous intéressons à l'**intégration temporelle et sémantique** des modalités, ces deux dimensions pouvant apporter des bénéfices importants lors du développement du système. L'étude de l'intégration temporelle peut s'appuyer sur les méthodes et modes d'analyse employés dans la littérature sur les interfaces multimodales sans Agent. L'intégration sémantique étant en revanche beaucoup moins bien documentée dans les travaux antérieurs, notre étude sera susceptible d'apporter une originalité et de nouvelles connaissances au domaine de recherche sur les interfaces multimodales en entrée.

A titre de validation, nous souhaitons également identifier quels sont les **bénéfices de la multimodalité** en entrée, du point de vue des utilisateurs, pour interagir avec des Agents Conversationnels dans le contexte du projet NICE.

La première étape de notre travail a été réalisée grâce à un dispositif de Magicien d'Oz simulant un contexte d'interaction proche de celui du projet NICE. Puis nous avons testé le premier prototype du système fonctionnel, sur deux scénarios différents. Ces études, ainsi que l'ensemble des données recueillies, seront décrites dans le chapitre 5.

4.3. Le comportement multimodal des Agents

La problématique exposée ci-dessus (étude du comportement multimodal des utilisateurs) a impliqué une recherche bibliographique sur la multimodalité entre parole et gestes dans la communication Homme-Homme et dans la communication Homme-Machine. Nous avons notamment distingué les gestes opératoires (dont la caractéristique principale est d'être volontairement intégrés comme éléments de communication) et la gesticulation (automatique et involontaire). Seuls les gestes opératoires, lorsqu'ils sont combinés au discours verbal, forment des **constructions multimodales** telles que nous les étudions chez les utilisateurs du système NICE.

Cet état de l'art a naturellement fait émerger des parallèles entre le comportement des utilisateurs et celui des Agents. Si on regarde les travaux sur le comportement multimodal des Agents à la lumière de la distinction entre gestes opératoires et gesticulation, on s'aperçoit que la catégorie des **gestes opératoires** a été principalement développée chez les Agents pédagogiques. La production de gestes déictiques en relation avec les éléments du discours verbal a en effet été étudiée chez les Agents Cosmo (Lester *et al.*, 1997b, 1999b), PPP Persona (André *et al.*, 1999) et Steve (Johnson *et al.*, 2000). Steve est également un des seuls Agents capables de manipuler les objets de son environnement (Rickel & Johnson, 1999).

Par ailleurs, nous avons vu que les constructions multimodales pouvaient être caractérisées selon deux dimensions principales : les relations temporelles entre modalités, et leurs relations sémantiques. Il semble qu'il soit aujourd'hui possible de contrôler très précisément les relations temporelles entre le discours et les gestes opératoires des Agents. En revanche, la question des **relations sémantiques** n'a pas ou peu été traitée. Les seuls travaux sur le contrôle de ces relations (Cassell & Prevost, 1996 ; Cassell, 2001) concernent la gesticulation et non les gestes opératoires : Cassell propose d'utiliser la gesticulation redondante pour marquer l'emphase (Cassell & Prevost, 1996) et de distribuer la redondance et la complémentarité pour augmenter le réalisme (Cassell, 2001) – sans toutefois évaluer l'influence effective de ces manipulations.

Il ne semble pas exister non plus de résultats expérimentaux sur les relations sémantiques entre parole et gestes opératoires en communication Homme-Homme : à nouveau, les seules indications disponibles (Cassell *et al.*, 1999a ; Goldin-Meadow *et al.*, 1999) concernent la gesticulation et non les gestes opératoires.

A l'issue de cet état de l'art, nous constatons donc que les Agents Conversationnels sont capables de produire de la parole et des gestes opératoires, qu'ils sont également capables de combiner ces modalités selon plusieurs types de coopération sémantique (notamment la redondance et la complémentarité), mais nous ne connaissons rien de l'**influence de ces types de coopération** : sont-ils perçus par les utilisateurs ? Y a-t-il un type de coopération qui soit plus efficace que les autres, qui augmente la qualité de la communication ? Quelles dimensions sont affectées par ces types de coopération multimodale ?

Nous avons décidé de mener des expérimentations pour éclaircir ces questions. Nous nous sommes placés pour cela dans un contexte pédagogique, car il s'agit du champ d'application principal des gestes opératoires pour les Agents. Par ailleurs, si la « qualité de la communication » est toujours un facteur important, elle est sans doute plus cruciale en situation pédagogique que dans un contexte

simplement conversationnel. En situation pédagogique, cette notion de qualité de la communication peut être traduite en différents types de variables :

- Des variables subjectives : recueil des impressions et jugements des utilisateurs, par exemple comment les utilisateurs évaluent la qualité des explications, ou la sympathie des Agents.
- Des variables objectives : la situation pédagogique justifie de recueillir des mesures de performance, notamment la quantité d'informations mémorisées par les utilisateurs.

Ces expérimentations sur l'influence du comportement multimodal des Agents seront exposées au chapitre 6.

5. Le comportement multimodal des Utilisateurs face aux Agents

5.1. Magicien d'Oz avec Agents 2D, scénario orienté tâche

Conformément aux aspects méthodologiques évoqués dans le paragraphe 3.2, nous avons décidé de mettre en place un recueil de corpus en Magicien d'Oz en amont de l'implémentation du système NICE (en 2002, 1ère année du projet), avec un échantillon proche des utilisateurs finaux (enfants, jeunes adultes). Cette démarche a pour objectif d'optimiser le développement de l'interface en fondant celui-ci sur l'observation du comportement des utilisateurs.

Questions initiales

Nous souhaitions en premier lieu évaluer l'apport de la multimodalité en entrée dans l'interaction avec des Agents. Les données de la littérature (voir le chapitre 3) nous laissent supposer que la multimodalité sera préférée, mais il semble tout de même pertinent de tester à nouveau cette hypothèse puisqu'elle n'a pas été validée avec des Agents.

Par rapport aux autres Interfaces Multimodales, les Agents sont des interlocuteurs qui sont censés devoir se rapprocher d'un mode de communication Homme-Homme. Il n'est donc pas impossible que l'interaction uniquement verbale en entrée soit appréciée par les utilisateurs. En revanche, nous avons choisi de ne pas tester une condition d'interaction uniquement gestuelle. En effet, le système NICE implique des aspects conversationnels qui ne sont pas *a priori* réalisables avec le seul stylo, sauf à utiliser l'écriture. Or, d'une part ce moyen d'interaction peut être complexe pour les enfants, voire inhibiteur, et d'autre part il ne rentre pas dans le cadre du projet, où la reconnaissance d'écriture manuelle n'est pas prévue.

Nous avons donc décidé d'opérationnaliser et de tester deux conditions d'interaction : une condition monomodale verbale et une condition multimodale (parole et stylo sur tablette-écran).

L'autre objectif majeur de cette étude est d'analyser le comportement multimodal des utilisateurs pour guider le développement du système fonctionnel. Etant donné que le LIMSI est plus spécifiquement chargé de l'implémentation des modules gestuels et multimodaux en entrée dans le projet NICE, nos questions sont les suivantes : y a-t-il des patterns multimodaux qui émergent lors de l'interaction avec des Agents ? Quels sont ces patterns ? Sont-ils les mêmes pour des enfants et de jeunes adultes ? Sont-ils les mêmes que ceux observés avec les Interfaces Multimodales sans Agents ?

Nous avons enregistré puis analysé des vidéos de l'interaction entre des utilisateurs (enfants et jeunes adultes) et des Agents dans le cadre d'un jeu. A ce stade du projet NICE, les graphismes des Agents 3D n'avaient pas encore été créés, et le scénario exact n'avait pas encore été fixé par le consortium. Nous avons donc imaginé un scénario simple, impliquant principalement des aspects orientés tâche, et quelques aspects conversationnels : le jeu consiste à rendre service aux Agents en allant trouver l'objet qui manque à chacun et en le leur apportant. Nous avons utilisé pour cela des Agents en deux dimensions (2D), développés en interne au LIMSI[22]. Les graphismes[23] du jeu comportaient quatre décors, quatre Agents et dix-huit objets déplaçables (Figure 24).

Les Agents 2D développés au LIMSI

Pour animer ces Agents 2D, nous disposons d'un catalogue d'images des différentes parties du corps (Figure 25) : chaque geste, chaque expression faciale, et même chaque mouvement des lèvres requiert une image supplémentaire. Le comportement des Agents est donc limité par nature. Plus précisément, notre base de comportements pour cette première expérimentation contenait 95 images pour chacun des quatre Agents. Une des contraintes de cette version de nos Agents était qu'ils ne pouvaient se déplacer dans l'environnement du jeu.

[22] Le logiciel a été développé par Sarkis Abrilian (Abrilian *et al.*, 2002). Des démonstrations sont disponibles en ligne : http://www.limsi.fr/Individu/martin/research/projects/lea/

[23] Les graphismes ont été réalisés par Christophe Rendu.

Figure 24 : Les quatre Agents et les quatre environnements du jeux.

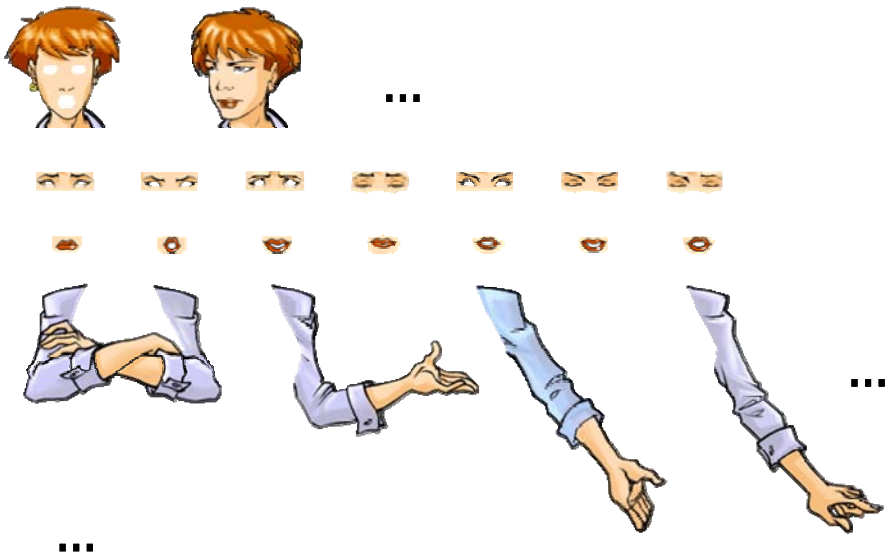

Figure 25 : Quelques images extraites du catalogue de l'Agent Léa.

Pour l'expérience, le comportement multimodal des Agents a été spécifié manuellement dans des fichiers XML contenant le texte des répliques verbales (envoyé ensuite vers la synthèse vocale IBM ViaVoice) et la description de l'animation (succession de positions pour chaque partie du corps). Chaque script peut être utilisé indifféremment pour les quatre Agents. Nous avons pris soin, en construisant ces fichiers de configuration, d'associer aux répliques verbales des gestes de chacune des grandes catégories de McNeill (1992) : gestes iconiques, métaphoriques, déictiques, battements, ainsi que des emblèmes. Le Tableau 8 présente quelques exemples de comportements multimodaux des Agents (répliques verbales associées à des gestes des bras) utilisés lors de l'expérience. Pour définir ces combinaisons multimodales, nous nous sommes aidés de données issues de la psycholinguistique en langue française (Calbris & Montredon, 1986 ; Calbris & Porcher, 1989).

Type de geste	**Exemples**	
	Réplique verbale	**Geste associé**
Emblème	Bonjour !	Signe de la main.
Iconique	C'est un grand livre rouge.	Avant-bras écartés, désignent un « grand » objet.
	Je veux une part comme ça.	Mains forment un triangle.
Métaphorique	Je vais réfléchir.	Une main sur le menton, l'autre sur la taille.
	Je ne sais pas.	Bras écartés, paumes vers le haut.
Déictique	Où veux-tu aller ?	Montre les différentes portes d'un geste large.
	C'est peut-être celle-ci.	Pointe une plante.
Battement, ou autre	Ça fait 120 ans que je suis sur terre et je suis très fatigué, je voudrais retourner dans ma lampe magique.	Avant-bras droit et gauche levés alternativement.
	Tu as besoin de quelque chose ?	Une main sur la taille, l'autre levée.

Tableau 8 : Exemples de séquences comportementales (parole accompagnée de gestes des mains).

Scénario du jeu

Au démarrage, l'action se situe dans le couloir d'une maison où six portes de différentes couleurs sont disponibles. Un génie est présent dans cet environnement et interpelle le joueur : il lui explique que pour retrouver sa lampe magique, le joueur doit l'aider en réalisant trois vœux pour les habitants de cette maison. Seules trois des six portes permettent d'accéder à une pièce, dans laquelle se trouvent à chaque fois un Agent et trois objets. Ces pièces sont : une cuisine, une bibliothèque et une serre. Le

joueur doit donc naviguer dans les pièces, rencontrer les Agents, leur demander leur vœu et le réaliser. Celui-ci consiste toujours à rapporter à l'Agent un objet manquant dans la pièce où il se trouve (apporter une lampe ou un livre à l'Agent dans la bibliothèque, apporter un gâteau ou un café à l'Agent dans la cuisine, apporter une plante ou un arrosoir à l'Agent dans la serre).

Afin de susciter les constructions multimodales de la part de l'utilisateur, plusieurs objets du même type sont disponibles dans chaque pièce et l'utilisateur doit choisir le bon en fonction de sa forme, de sa taille ou de sa couleur.

Dispositif de Magicien d'Oz

Le dispositif de Magicien d'Oz était composé de deux ordinateurs (Figure 26) : l'ordinateur n°1 assurait la présentation du jeu aux utilisateurs et l'ordinateur n°2 la gestion de l'interaction par le Magicien. L'utilisateur et le Magicien se trouvaient dans des pièces différentes.

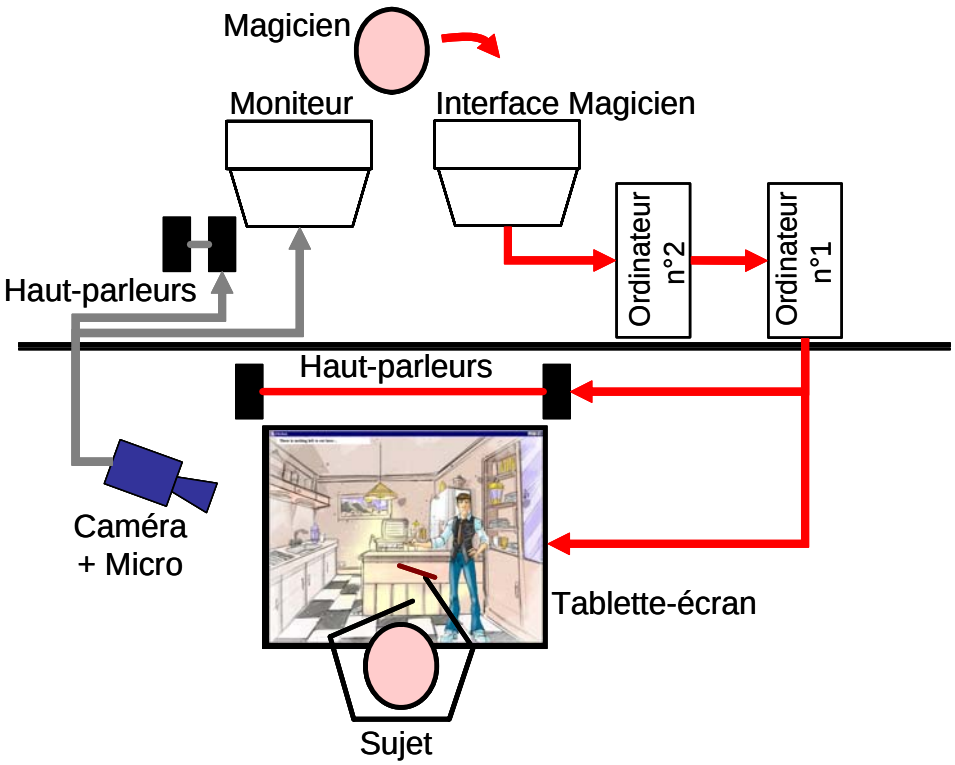

Figure 26 : Schéma du dispositif de Magicien d'Oz.

Dispositif Utilisateur. L'ordinateur n°1 était relié à une tablette-écran (Wacom Cintiq 15X) permettant une interaction directe sur l'écran par l'intermédiaire d'un stylo. Deux haut-parleurs assuraient la diffusion de la voix des Agents ; toutes les répliques verbales des Agents étaient également transcrites en haut de l'écran de l'utilisateur. Un micro était disposé derrière l'écran pour

inciter les utilisateurs à parler. La reconnaissance vocale, l'interprétation du langage, des gestes et de leurs combinaisons multimodales étaient cependant simulées par le Magicien.

Enregistrement vidéo. Une caméra numérique, installée au-dessus de la tablette-écran, permettait un enregistrement audio-visuel du comportement de l'utilisateur en conservant son anonymat (le visage n'était pas filmé).

Dispositif Magicien. Un moniteur et des enceintes reliés à la caméra numérique permettaient au Magicien de suivre en temps réel les paroles et les actions de l'utilisateur. Pour y répondre, il disposait d'une interface (Figure 27) lui permettant de modifier l'environnement du jeu (changement de lieu, déplacement d'un objet) ou le comportement des Agents (déclenchement de paroles et de comportements non-verbaux).

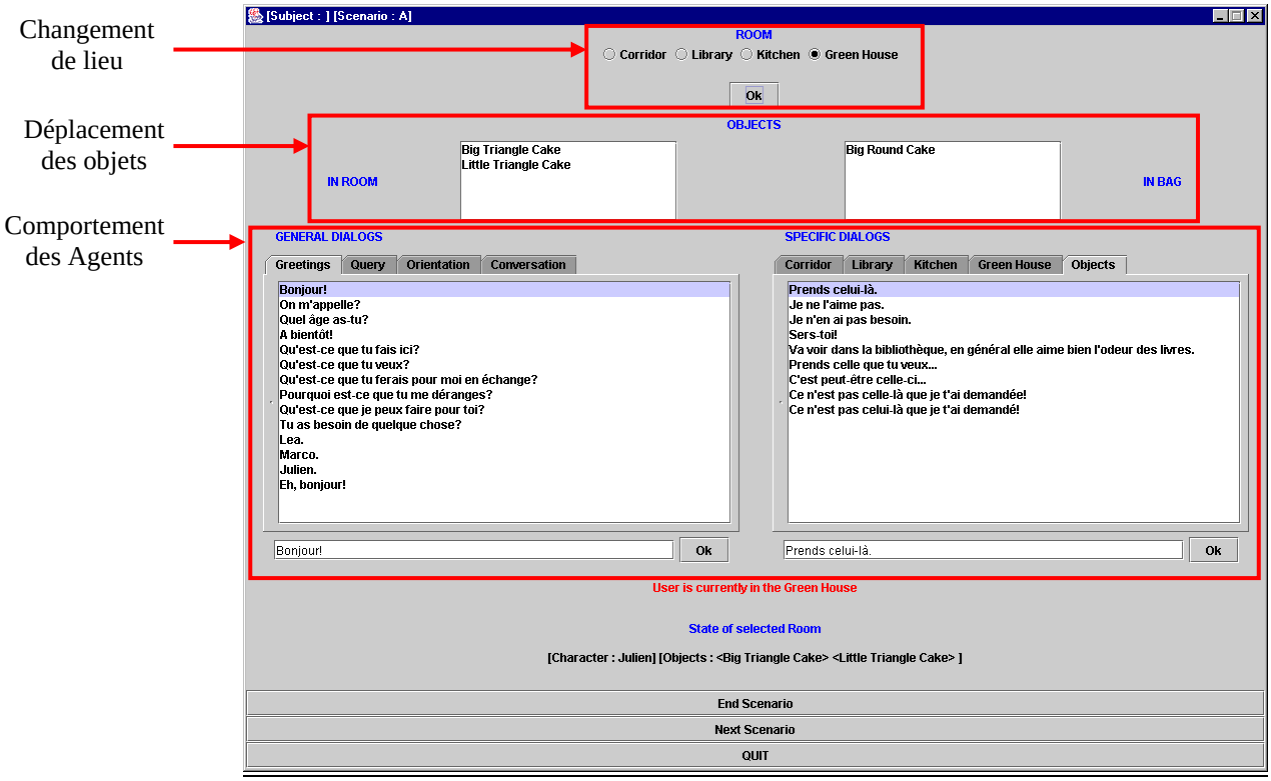

Figure 27 : Interface du Magicien.

Quatre-vingt trois séquences comportementales possibles pour les Agents avaient été anticipées et pré-encodées afin de guider l'utilisateur, de dialoguer avec lui et de répondre à ses requêtes les plus probables compte tenu du scénario de jeu. Chaque séquence comportementale était composée d'une réplique verbale et d'une animation de l'Agent en conséquence (animation de la tête de l'Agent, de ses yeux, de la direction de son regard, de sa bouche et de ses bras). Les séquences de comportements

multimodaux configurées dans des fichiers XML pouvaient être déclenchées depuis l'interface du Magicien, où elles étaient organisées en plusieurs onglets dans deux grandes catégories (dialogues généraux, dialogues spécifiques). Outre ces séquences comportementales pré-encodées, le Magicien avait également la possibilité, au cours de l'expérience, de taper une réplique non prévue et de l'associer au comportement non verbal de la réplique de son choix dans la base.

La gestion de l'interaction (action du magicien puis réaction du système, avec analyse du fichier XML et déclenchement de la synthèse vocale) impliquait un délai relativement élevé entre l'action de l'utilisateur et la réponse de l'Agent (délai de l'ordre de 2 secondes). Eu égard à cette contrainte initiale, nous avons décidé de limiter le nombre de simulations d'erreurs venant entraver encore davantage l'interaction. Nous avons seulement simulé parfois des erreurs d'interprétation, lorsque les utilisateurs étaient trop vagues dans leurs requêtes. Par exemple, un des objets que les utilisateurs devaient trouver était « le grand livre rouge ». S'ils demandaient « le livre rouge », sans plus de précision, ils obtenaient le plus souvent le mauvais, c'est-à-dire le petit livre rouge. Ils étaient alors obligés de corriger leur requête en étant plus précis.

Méthode

Utilisateurs. Dix-huit utilisateurs ont participé à cette expérience. Ils se répartissent en deux groupes :

- Huit adultes recrutés parmi les étudiants et le personnel du LIMSI. Un sujet masculin, qui avait deviné qu'il s'agissait d'une simulation, a dû être exclu de l'analyse. Celle-ci a donc porté sur 7 utilisateurs adultes : 3 hommes (âge moyen de 24 ans 6 mois, $\sigma = 2,5$ ans) et 4 femmes (âge moyen de 26 ans 9 mois, $\sigma = 7,5$ ans).
- Dix enfants recrutés parmi les enfants du personnel et au sein d'un centre de loisir : 7 garçons (âge moyen de 11 ans et 6 mois, $\sigma = 1,9$ ans) et 3 filles (âge moyen de 9 ans et 4 mois, $\sigma = 0,6$ ans).

Nous avons pu vérifier *a posteriori* que ces deux groupes (adultes, enfants) étaient équivalents dans leur fréquence d'utilisation des jeux vidéos (en moyenne une à deux fois par semaine). Ils étaient également équivalents sur l'auto-évaluation de leur expertise dans l'utilisation d'ordinateurs (niveau moyen situé entre « Intermédiaire » et « Expert »).

Procédure. L'expérience s'est déroulée au LIMSI. La séance débutait par la signature du protocole d'accord pour les enregistrements vidéo (accord parental pour les mineurs, voir l'Annexe 2). Le scénario de jeu était ensuite expliqué, puis les utilisateurs devaient réaliser successivement deux

scénarios dans deux conditions différentes :

- Une condition monomodale verbale, dans laquelle les utilisateurs devaient réaliser l'ensemble du scénario par la parole,
- Une condition multimodale, dans laquelle les utilisateurs pouvaient interagir par la parole, par le stylo directement sur la tablette-écran, ou par ces deux modes de manière combinée.

Soulignons qu'aucune indication ou suggestion n'était donnée sur la manière d'interagir avec les Agents, pas même un exemple (pas d'exemple de phrase, pas d'exemple d'utilisation du stylo). L'expérimentatrice se contentait de présenter les dispositifs physiques.

Les deux scénarios réalisés successivement étaient équivalents mais les personnages n'étaient pas affectés aux mêmes pièces, les vœux et les objets disponibles étaient différents. L'ordre de passation des conditions (monomodale / multimodale) était contrebalancé au sein de chaque groupe d'utilisateurs.

A la fin de chaque scénario, l'utilisateur donnait une évaluation subjective de la qualité de l'interaction à l'aide de quatre échelles analogiques (voir l'Annexe 3) : facilité, efficacité perçue, caractère agréable, facilité d'apprentissage. Le résultat a été relevé en centimètres (0 à 5 cm, les valeurs élevées représentant des évaluations d'autant plus positives).

Une fois les deux scénarios achevés, quelques données individuelles était relevées : niveau d'expertise dans l'utilisation d'ordinateurs (1. peu expérimenté(e), 2. intermédiaire, 3. expert(e)), préférences concernant les jeux vidéos et fréquence de jeu habituelle (1. jamais, 2. moins d'une fois par semaine, 3. une fois par semaine, 4. deux ou trois fois par semaine, 5. davantage). A la fin de la séance, l'expérimentatrice expliquait que le système avait été simulé et en donnait les raisons.

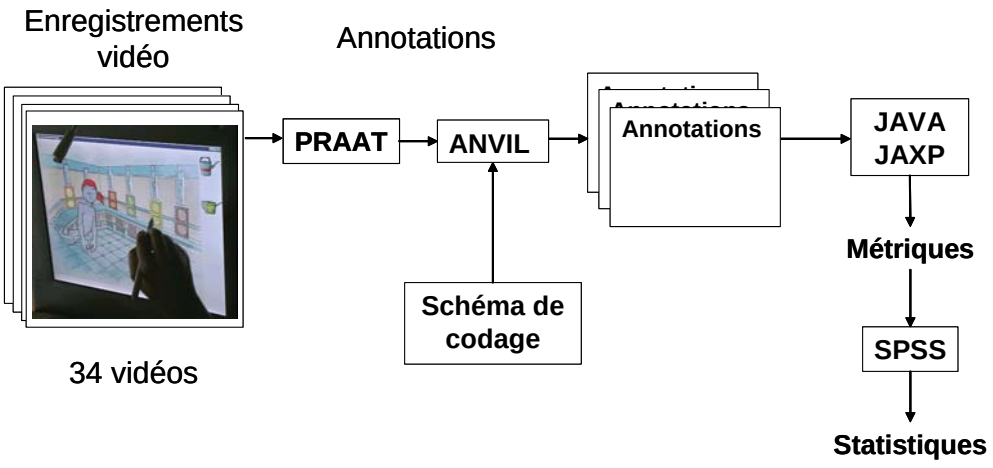

Figure 28 : Schéma d'annotation et d'analyse des enregistrements vidéo.

Annotation des vidéos. Chacun des 17 participants ayant réalisé deux scénarios, 34 vidéos (représentant un corpus total de 3h30) ont été enregistrées puis annotées. La Figure 28 résume l'ensemble du processus d'annotation et d'analyse des données. Dans la suite de ce paragraphe, nous allons développer certains points de ce processus, en expliquant en particulier en quoi consistent les deux phases d'annotation avec les outils Praat et Anvil.

Praat[24] est un logiciel d'analyse du son, que nous avons utilisé pour transcrire les verbalisations des utilisateurs[25]. Le signal sonore provenant des vidéos a été importé dans Praat, puis a été segmenté manuellement, sur la base à la fois du contenu auditif et de la forme de l'onde sonore (voir la piste supérieure de la Figure 29). Enfin, les mots associés à chaque segment ont été transcrits (piste inférieure de la Figure 29). Cette phase de transcription est indispensable lorsque l'on souhaite analyser les verbalisations (catégories morpho-syntaxiques employées, longueur et complexité des expressions), ainsi que leurs relations avec l'utilisation d'une autre modalité (relations temporelles et sémantiques).

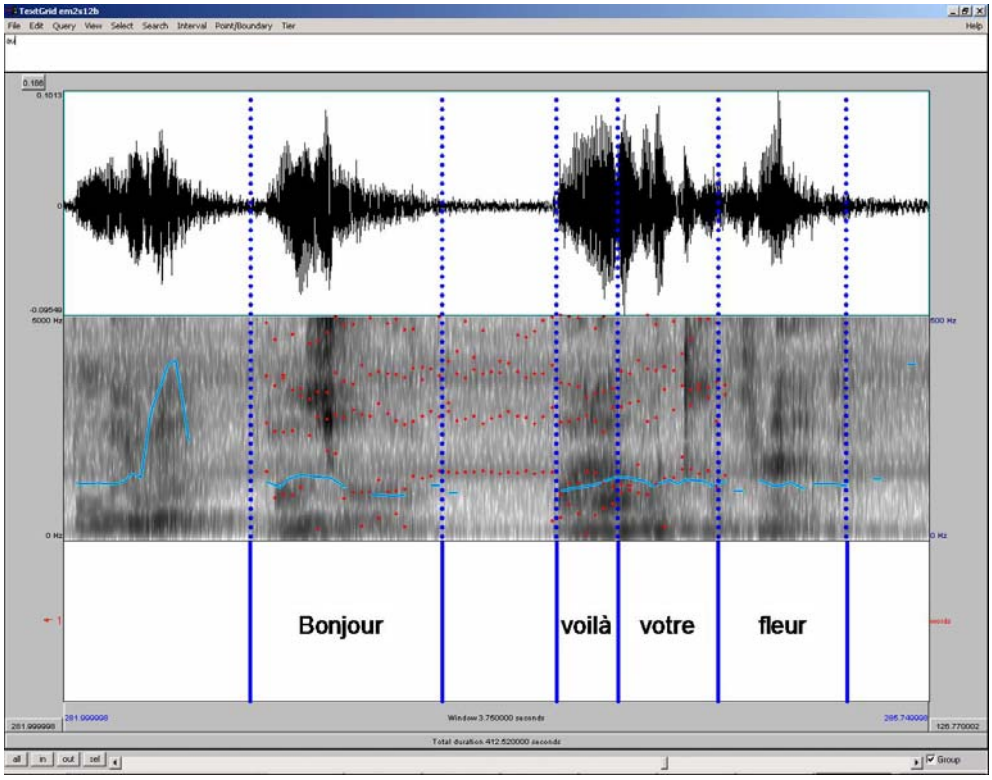

Figure 29 : Copie d'écran de Praat.

[24] http://www.fon.hum.uva.nl/praat/

[25] Une grande partie des transcriptions ont été réalisées par Frédéric Cloarec.

Le logiciel Anvil (Kipp, 2001) a ensuite été utilisé afin de compléter les annotations[26]. Un schéma de codage sous format XML permet de lister les comportements que l'on souhaite annoter et de visualiser sur différentes pistes leur décours temporel. Comme le montre la Figure 30, le schéma de codage que nous avons défini contenait sept pistes, décrites ci-dessous.

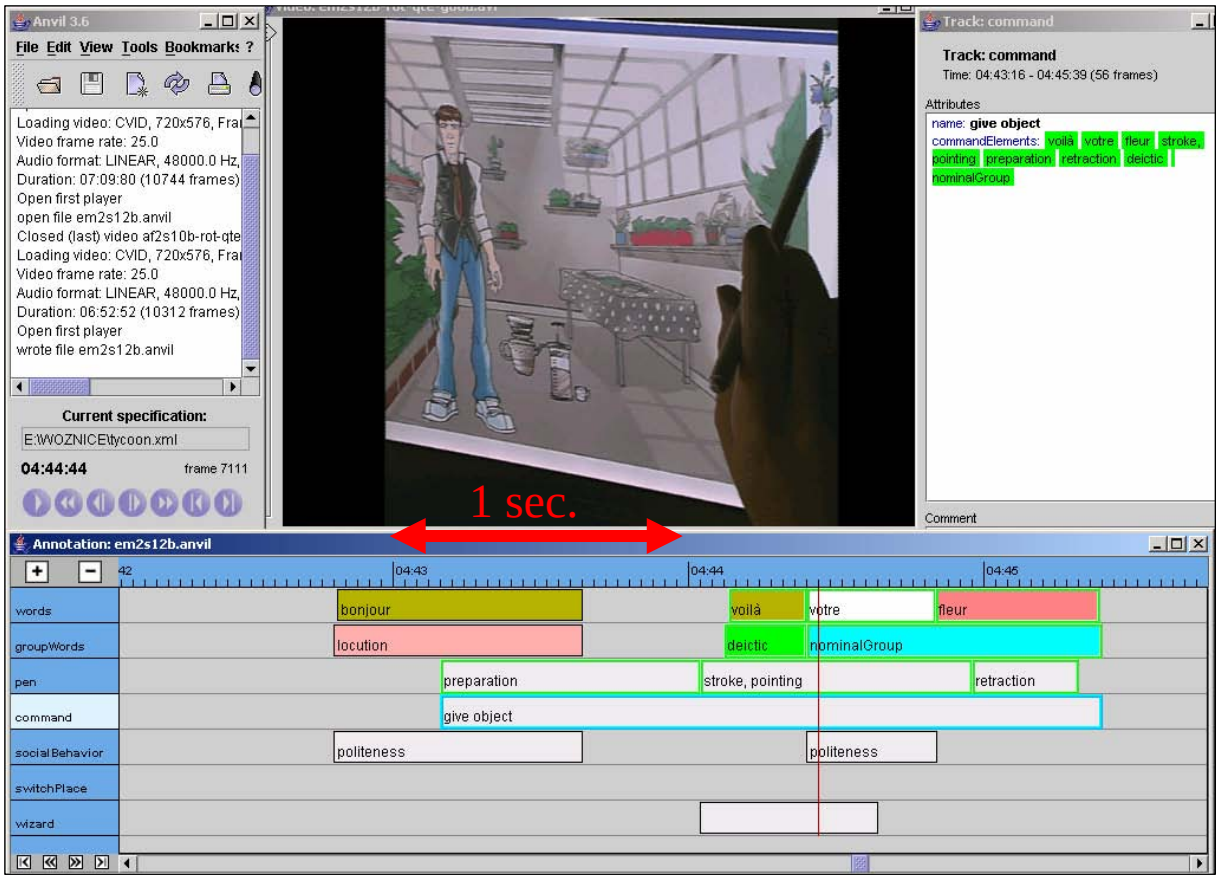

Figure 30 : Copie d'écran d'Anvil.

Piste 1 *words* : Cette piste contient les **verbalisations** des utilisateurs importées de Praat [« Bonjour, voilà votre fleur » dans la Figure 30].

Cette piste nous a permis d'étudier de manière simplifiée la **syntaxe**, en créant une variable binaire : syntaxe naturelle vrai ou faux. Toutes les formulations assimilables à de la communication Homme-Homme ont été codées « syntaxe naturelle », quel que soit leur degré d'élaboration (« tiens » ou « puis-je prendre le livre rouge s'il vous plaît »…). Les formulations elliptiques (« prendre livre », « sortir ») ont été codées « syntaxe non naturelle ».

Piste 2 *groupWords* : Dans cette piste nous avons annoté les **unités grammaticales**, c'est-à-dire les regroupements des mots de la piste précédente selon quatre catégories :

[26] Une partie des annotations ont été réalisées par Clotilde Le Quiniou (Le Quiniou, 2004).

- Locutions (expressions invariables) [« Bonjour » dans la Figure 30],
- Références déictiques : il s'agit d'unités linguistiques renvoyant à l'acte d'énonciation, ininterprétables hors du contexte spatio-temporel créé par le discours ou le dialogue (voir Lyons, 1977). Dans notre corpus, nous avons restreint l'étude des références déictiques aux éléments verbaux produits de manière combinée avec des gestes (ex : « ici » ou « celui-là » accompagnés de gestes désignant une localisation ou un objet). [dans la Figure 30, « voilà » est une référence déictique liée à l'action de donner un objet, qui est produite par le geste],
- Groupes nominaux [« votre fleur » dans la Figure 30],
- Groupes verbaux [pas d'exemple dans la Figure 30].

Piste 3 *pen* : Contient l'annotation des **gestes** effectués avec le stylo : annotation de la durée totale du geste, incluant les différentes phases de réalisation (préparation, contact ou *stroke,* rétraction). La partie *stroke* des gestes était ensuite codée selon sa forme (pointage, encerclement, ligne, flèche, exploration) [geste de pointage dans la Figure 30].

Les pistes 1 et 3 nous ont permis par la suite de relever qui de l'Agent ou de l'utilisateur prenait l'**initiative** dans le dialogue. Cette variable nous a semblé intéressante pour évaluer le scénario de jeu. Pour quantifier la prise d'initiative, nous avons dénombré toutes les fois que l'un ou l'autre des interlocuteurs engageait l'échange (parlait ou agissait en premier), ou changeait de sujet (par la parole et/ou par le geste).

Piste 4 *command* : Les **commandes** correspondent au type d'action qui résulte des comportements observés. Cinq grands types de commandes permettant de réaliser le scénario ont été identifiés dans le corpus : Entrer dans une pièce, Sortir de la pièce, Demander un vœu, Prendre un objet, Donner un objet. Dans la Figure 30, les mots « voilà votre fleur » et le geste de pointage sur la fleur constituent la commande « Donner un objet », dont l'annotation couvre la durée totale de la phrase et du geste réunis et est reliée à chacun des mots ainsi qu'au geste de pointage.

A partir de l'annotation des commandes, nous avons pu relever les constructions multimodales et étudier l'intégration des modalités :

- Intégration **temporelle** : simultanée lorsqu'il y a chevauchement des modalités, séquentielle sinon ; nous avons aussi mesuré le délai entre modalités dans les constructions séquentielles.
- Intégration **sémantique** : nous nous sommes référés aux trois types de coopération caractérisant les constructions multimodales dans la taxonomie de Martin *et al.* (2001 ; voir le paragraphe 3.2.4) : la Redondance, la Complémentarité et la Concurrence [la Figure 30 présente un exemple de complémentarité].

Piste 5 *socialBehavior* : Nous avons annoté des comportements que nous qualifierons de sociaux : les **feedbacks** (confirmations de la part de l'utilisateur sur ce que vient de dire/faire l'agent) et les éléments de **politesse** (« bonjour », « s'il te plaît », etc.). Dans les deux cas il doit s'agir d'éléments verbaux qui ne sont pas indispensables à la réalisation du scénario. Par exemple, lorsque l'expression « au revoir » sert à signifier que l'utilisateur veut quitter la pièce, elle a valeur de commande, et n'entre donc pas dans la catégorie politesse. Par ailleurs, les expressions qui étaient à la fois un feedback et un élément de politesse (typiquement « merci ») ont été systématiquement annotées comme politesse. La Figure 30 illustre deux types de politesse : la locution « bonjour » et le vouvoiement. Une seule annotation de politesse était insérée lorsqu'un utilisateur vouvoyait l'Agent pendant tout le scénario.

Piste 6 *switchPlace* : Des marqueurs ont été placés sur cette piste à chaque changement de lieu afin de diviser le scénario en séquences (une par pièce).

Piste 7 *wizard* : Enfin, la dernière piste a servi à mesurer le temps de réaction du magicien. Par exemple dans la Figure 30, nous avons mesuré le délai entre le début du pointage et le moment où la fleur est déplacée vers le personnage : ce délai est ici égal à une seconde.

En sortie d'Anvil, les fichiers d'annotation se présentent sous un format XML qui n'est pas directement exploitable : il faut utiliser un programme supplémentaire pour extraire de ces fichiers des métriques pouvant ensuite faire l'objet de calculs statistiques. Pour cela, nous avons utilisé un programme développé en interne au LIMSI[27]. Une fois extraites des fichiers d'annotations, les métriques recueillies ont été organisées en variables, qui ont enfin été soumises à des analyses statistiques à l'aide du logiciel SPSS[28].

[27] Programme développé par Jean-Claude Martin.

[28] http://www.spss.com/

Résultats

Les analyses que nous avons réalisées portent sur les variables suivantes :

Variables Indépendantes :

- L'âge des utilisateurs (variable inter-groupe),
- La condition (mono- *vs*. multimodale ; variable intra-groupe).

Etant donné le faible nombre d'utilisateurs et le déséquilibre dans la répartition hommes / femmes, nous n'avons pas inclus la variable Sexe dans nos analyses.

Variables Dépendantes Comportementales :

Générales :

- Durée de réalisation du scénario,
- Nombre de commandes,
- Utilisation de chaque modalité : parole, stylo, multimodalité (en % de commandes),
- Prise d'initiative (en % de tours de parole).

Comportement verbal :

- Nombre de mots par phrase,
- Constructions syntaxiques « naturelles » (en % de constructions contenant de la parole),
- Nombre de comportements sociaux.

Comportement gestuel :

- Forme des gestes (en % de gestes).

Comportement multimodal :

- Intégration sémantique : Redondance, Complémentarité, Concurrence (en % de constructions multimodales),
- Intégration temporelle : Simultanée, Séquentielle (en % de constructions redondantes + complémentaires),
- Délai séparant les modalités,
- Utilisation de références déictiques verbales (en % de constructions complémentaires).

Variables Dépendantes Subjectives :

- Facilité de l'interaction,
- Efficacité,
- Caractère agréable,
- Facilité d'apprentissage.

Comparaison des deux conditions d'interaction. Nous avons en premier lieu étudié les différences entre les deux conditions d'interaction (mono- *vs.* multimodale). Cette variable a influencé la **durée de réalisation du scénario** ($F(1/16) = 12{,}905$; $p = 0{,}002$) : les scénarios multimodaux ont été achevés plus rapidement ($m = 307{,}80$ sec ; $\sigma = 88{,}71$) que les scénarios monomodaux ($m = 437{,}19$ sec ; $\sigma = 129{,}42$). Le **nombre de commandes** nécessaires à la réalisation d'un scénario est, lui, resté stable quelle que soit la condition ($m = 37$ commandes ; $\sigma = 11{,}8$).

La multimodalité a par ailleurs favorisé la **prise d'initiative** de la part des utilisateurs ($F(1/16) = 8{,}551$; $p = 0{,}010$). La Figure 31 montre que cet effet a touché de la même manière les adultes et les enfants.

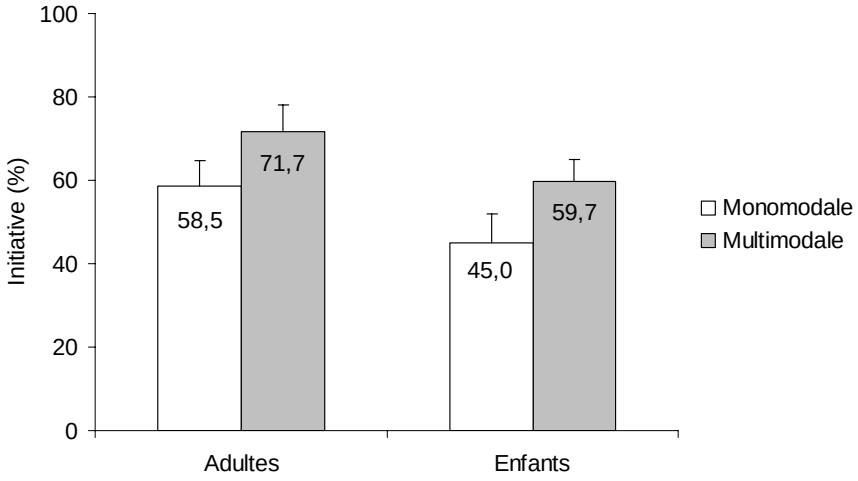

Figure 31 : Pourcentages de prise d'initiative pour les deux groupes (adultes, enfants) en fonction de la condition d'interaction (mono- *vs.* multimodale).

Les utilisateurs ont trouvé que la condition multimodale était **plus facile** ($m = 4{,}27$ /5 ; $\sigma = 0{,}81$) que la condition monomodale ($m = 3{,}67$; $\sigma = 1{,}20$). Cet effet est significatif ($F(1/16) = 6{,}304$; $p = 0{,}023$), et la Figure 32 suggère qu'il est principalement dû aux évaluations des enfants.

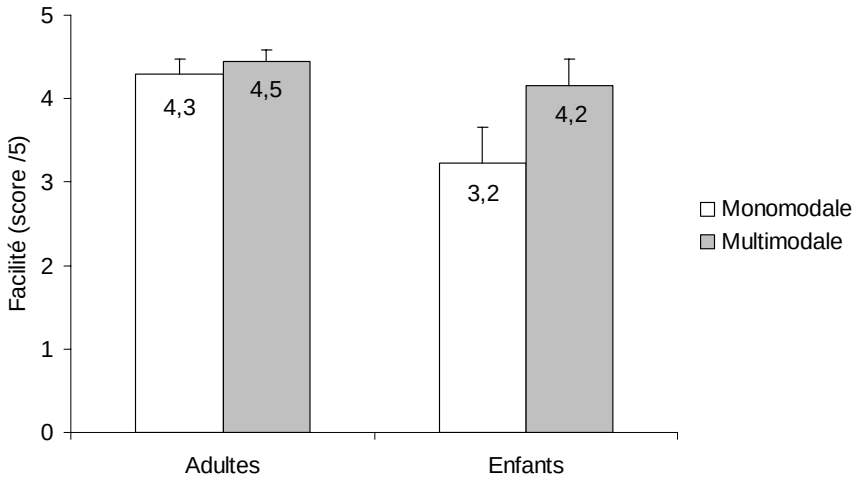

Figure 32 : Score de facilité perçue pour les deux groupes (adultes, enfants) en fonction de la condition d'interaction (mono- *vs.* multimodale).

Les autres variables subjectives que nous avons recueillies dans les questionnaires ne permettent pas de différencier les deux conditions d'interaction (mono- *vs.* multimodale). Globalement, les utilisateurs ont trouvé l'interaction **efficace** (score moyen = 4,12 /5 ; $\sigma = 0,96$) quelle que soit la condition. Ils ont également **apprécié** de la même manière les deux conditions (score moyen = 4,34 /5 ; $\sigma = 0,74$) et ont donné des scores de **facilité d'apprentissage** globalement élevés (score moyen = 4,31 /5 ; $\sigma = 0,81$).

Comportement verbal. Nos sujets ont utilisé en moyenne **6 mots par commande** ($\sigma = 4,53$). De manière surprenante, cette valeur ne diminue pas en condition multimodale. Elle ne semble pas non plus être influencée par l'âge des utilisateurs. Nous avons par ailleurs calculé un pourcentage global de **91% de syntaxe naturelle** ($\sigma = 15,6$), c'est-à-dire de formulations assimilables à de la communication Homme-Homme. Cette valeur semble elle aussi stable quelle que soit la condition et l'âge des utilisateurs.

Enfin, nous avons relevé en moyenne **12 comportements sociaux par scénario** ($\sigma = 7$), ce qui correspond à une fréquence d'environ 2 comportements sociaux par minute. Ils se répartissant en 70% de marques de politesse (ex : bonjour, au revoir, s'il te plaît, merci) et 30% de feedbacks (ex : oui, d'accord, OK, j'y vais, je sais où elle est, etc.). Ces données moyennes ne sont pas influencées par la condition d'interaction ni par l'âge des utilisateurs.

Comportement gestuel. Le Tableau 9 présente les pourcentages d'utilisation (et écarts-types) des cinq formes de gestes relevées dans le corpus. Rappelons ici que nous avions fait en sorte que les

consignes n'influencent pas les formes de gestes, puisque nous mentionnions juste aux sujets qu'ils pouvaient « utiliser le stylo directement sur l'écran » mais sans donner d'exemple ou montrer de geste.

Le pointage apparaît être le type de geste le plus massivement utilisé, en particulier par les adultes. Les enfants ont eu un comportement gestuel plus varié, incluant notamment des encerclements, des lignes (par exemple pour relier un objet avec la zone où l'utilisateur souhaite le placer), et des comportements d'exploration où le geste n'a pas de forme déterminée et parcourt l'environnement.

		Pointages	Encerclements	Lignes	Exploration	Flèches
Adultes	m	85,2%	4,4%	0,5%	0%	9,9%
	σ	*37,6*	*11,6*	*1,5*	*0*	*26,2*
Enfants	m	64,5%	16,5%	10,3%	8%	0,7%
	σ	*30,5*	*21,6*	*23,7*	*18,8*	*2,1*
Echantillon total	m	73%	11,5%	6,3%	4,7%	4,5%
	σ	*34,1*	*18,7*	*18,5*	*14,7*	*16,8*

Tableau 9 : Formes de gestes observées, pourcentages et écarts-types d'utilisation dans le corpus.

Utilisation des modalités. Globalement, la **proportion d'utilisation des modalités** se répartit comme suit : 47,53% des commandes ont été réalisées avec le stylo uniquement, 31,28% ont été réalisées avec la parole uniquement, et 21,19% ont été réalisées de façon multimodale en faisant un usage combiné de la parole et du stylo. Tous nos utilisateurs ont fait au moins une construction multimodale.

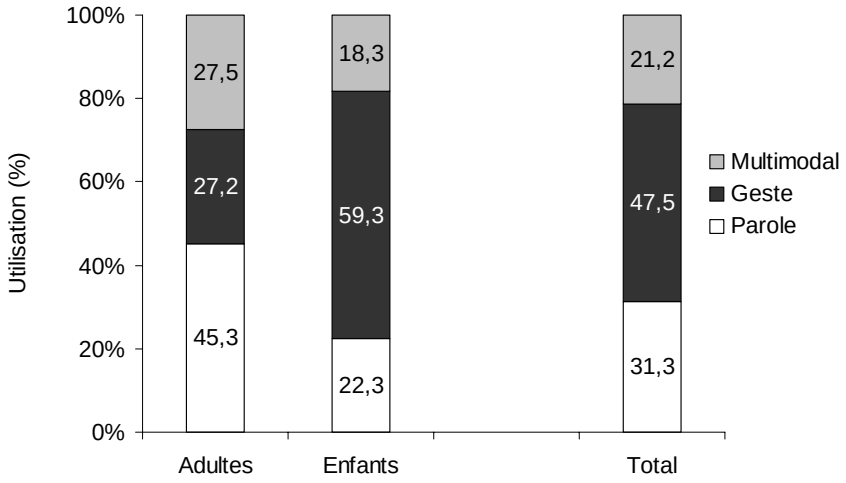

Figure 33 : Pourcentage d'utilisation des modalités par les adultes et les enfants, et pour tout l'échantillon.

Comme le montre la Figure 33, la modalité d'interaction prédominante pour les adultes a été la parole, alors que pour les enfants il s'agissait clairement du geste. Le Tableau 10 fournit les écarts-types associés aux pourcentages moyens d'utilisation des modalités.

		Parole	Geste	Multimodalité
Adultes	m	45,3%	27,2%	27,6%
	σ	*17,7*	*23,9*	*18,9*
Enfants	m	22,3%	59,4%	18,3%
	σ	*16,3*	*17,9*	*13,7*
Echantillon total	m	31,3%	47,5%	21,2%
	σ	*20,3*	*26,5*	*16,4*

Tableau 10 : Pourcentages et écarts-types d'utilisation des modalités.

La prédominance d'une modalité d'interaction est encore plus accentuée si l'on considère la prise d'initiative dans les échanges. La Figure 34 montre en effet que, lors de la prise d'initiative (engager ou réorienter l'interaction), les adultes ont utilisé la parole dans 62% des cas et les enfants ont utilisé le geste dans 67% des cas (voir le Tableau 11 pour les écarts-types).

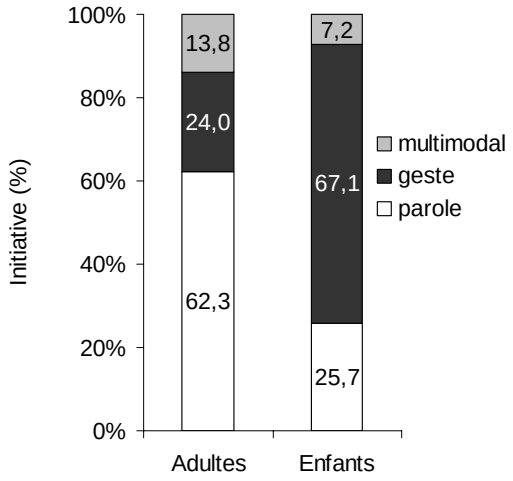

Figure 34 : Modalités par lesquelles les utilisateurs prennent l'initiative dans l'interaction.

		Parole	Geste	Multimodalité
Adultes	m	62,3%	24%	13,8%
	σ	*16,3*	*18,7*	*12,7*
Enfants	m	25,7%	67,1%	7,2%
	σ	*27,5*	*28,6*	*8,9*

Tableau 11 : Pourcentages et écarts-types d'utilisation des modalités pour la prise d'initiative.

Nous avons également étudié l'utilisation des modalités en fonction des commandes, dans le but de savoir si certaines actions suscitaient plus que d'autres l'utilisation de certaines modalités. La Figure 35 montre effectivement qu'il existe des différences importantes de modalité d'interaction en fonction des commandes. Par exemple, la commande « Demander un vœu », qui correspond à l'aspect le plus conversationnel de notre scénario, a été presque exclusivement réalisée par la parole. Dans quelques rares cas observés chez des enfants, cette commande a été réalisée de façon multimodale complémentaire (ex : [pointage sur l'Agent] puis « Que veux-tu ? ») ou concurrente (ex : « Quel est votre vœu ? » + [exploration]).

Pour la commande « Sortir », il n'existait pas d'élément graphique (porte…) incitant à utiliser le stylo pour sortir d'une pièce. La manière la plus facile de sortir était donc de le demander à l'Agent. Cependant, certains utilisateurs (à 90% des enfants) ont cherché un moyen de sortir avec le stylo, principalement en explorant le bas de l'écran au niveau du sol des pièces.

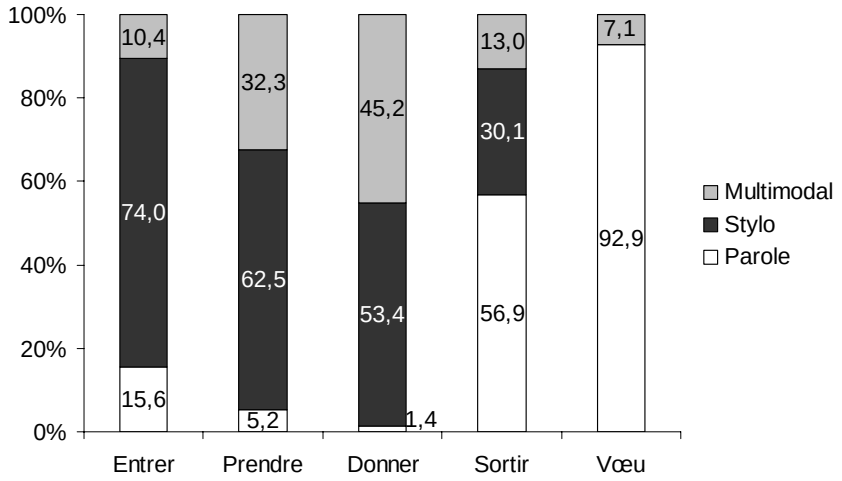

Figure 35 : Pourcentage d'utilisation des modalités en fonction de l'action réalisée.

Les commandes « Prendre » et « Donner un objet » ont été majoritairement réalisée à l'aide du stylo, mais ont également suscité un grand nombre de constructions multimodales : les utilisateurs prévenaient les Agents lorsqu'ils prenaient un objet, ou leur demandaient la permission, et leur signalaient par la parole lorsqu'ils leur donnaient un objet. Les utilisateurs ont ainsi spontanément agrémenté ces actions orientées tâche d'aspects conversationnels et sociaux. En revanche, la commande « Entrer » a été massivement réalisée par le geste seul, peut-être parce que les utilisateurs ne voyaient pas le génie comme un des habitants de la maison, à qui il fallait demander l'autorisation avant d'agir.

Intégration sémantique. Pour analyser l'intégration sémantique dans les constructions multimodales, nous avions initialement envisagé trois catégories :

- La **Redondance** (ex : « Je voudrais aller dans la pièce rose s'il te plaît. » + [pointage sur la porte rose]),
- La **Complémentarité** (ex : « Tiens prends-le. » + [pointage sur un livre]),
- La **Concurrence** (ex : « Bonjour » + [pointage sur un gâteau] ou « La cafetière électrique » + [exploration au niveau du bureau]).

Mais nous avons été amenés à construire deux catégories *ad hoc* en complément des précédentes :

- La **Complémentarité dialoguée** : il s'agit de cas où les utilisateurs annoncent ou demandent l'autorisation verbalement avant de réaliser une action par le geste (ex : « Je peux prendre du gâteau ? » Réponse de l'Agent : « Sers-toi. » [l'utilisateur pointe sur une part de gâteau]). Les modalités sont ici complémentaires mais, dans la mesure où les utilisateurs attendent l'acquiescement de l'Agent pour agir, nous qualifions la complémentarité de « dialoguée ». Ce type d'intégration représente 10% des constructions multimodales (12 constructions /117).
- La **Répétition multimodale** : face à une erreur ou une absence de réaction du système, l'utilisateur peut être amené à répéter une commande. Dans 86% des cas, la répétition a été faite dans la même modalité que la commande initiale, mais dans 14% des cas, l'utilisateur a changé de modalité lors de la répétition. Plutôt que de diviser ce type de commandes et de considérer deux entrées monomodales, nous les avons comptabilisées dans les constructions multimodales.

Au total, la répartition des 117 constructions multimodales dans les cinq catégories d'intégration sémantique est présentée dans la Figure 36.

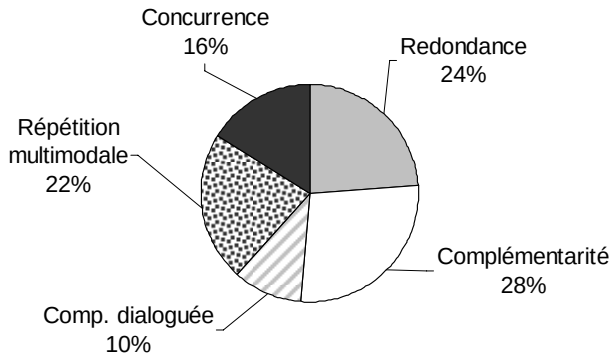

Figure 36 : Intégration sémantique des constructions multimodales.

Les **répétitions multimodales** sont intéressantes car elles nous apprennent quelle modalité est sollicitée en cas d'échec dans la communication. Nous les avons donc classifiées en fonction de la modalité initiale et de la modalité de répétition. La Figure 37 donne le pourcentage d'utilisation de chacune des 7 combinaisons possibles. De manière surprenante, c'est la parole qui est le plus souvent sollicitée comme modalité de secours, en totalisant 47,6% d'utilisation comme modalité de répétition. La multimodalité totalise quant à elle 38,1% d'utilisation en cas de répétition. Les répétitions interviennent en moyenne après un **délai de 1,9 secondes** (0,9 à 4,25 sec, délai médian de 2,1 sec, $\sigma = 0,8$).

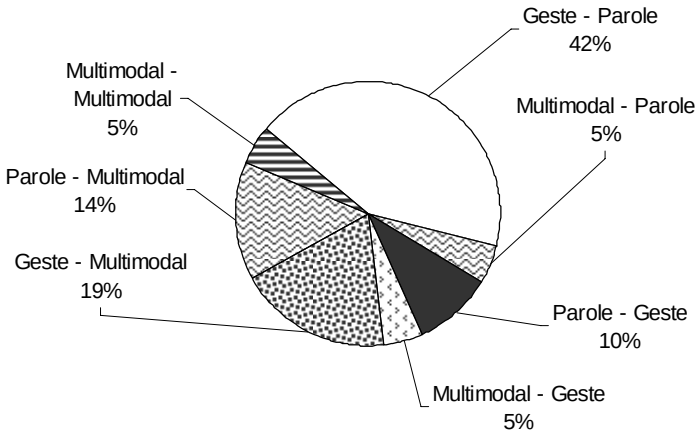

Figure 37 : Combinaisons Modalité initiale – Modalité de répétition dans les Répétitions multimodales.

Intégration temporelle. Pour certaines des catégories que nous venons de décrire, la question de l'intégration temporelle ne se pose pas : en effet, la Concurrence est par nature simultanée (c'est-à-dire qu'il y a un chevauchement temporel entre les modalités) et la Complémentarité dialoguée ainsi que la Répétition multimodale sont par nature séquentielles (sans chevauchement). L'étude de l'intégration temporelle concerne donc les deux catégories restantes : la Redondance et la

Complémentarité, celles pour lesquelles les informations portées par les deux modalités doivent être fusionnées. Il est apparu que ces constructions étaient à **73% simultanées** et à 27% séquentielles. Cette proportion est la même pour la Redondance et la Complémentarité.

62% des constructions complémentaires contiennent un **déictique verbal**. La proportion de simultanéité entre parole et geste dans ces constructions avec déictique atteint 86%. Si l'on regarde plus précisément l'intégration temporelle des références déictiques, on constate qu'il y a chevauchement entre le geste et la référence déictique dans 57% des cas. Le geste précède la référence dans 24% des cas et la référence précède le geste dans 19% des cas.

Nous avons ensuite recherché l'existence de **patterns dominants** chez les utilisateurs ayant produit 2 constructions à fusionner ou plus (soit 10 utilisateurs : 5 adultes et 5 enfants). Parmi les adultes, quatre ont suivi un pattern dominant simultané (avec un taux d'utilisation moyen de 86% de simultanéité), et le cinquième n'a suivi aucun pattern (50% de chaque type d'intégration). Parmi les enfants, deux avaient un pattern simultané (taux d'utilisation moyen de 92%) et les trois autres suivaient un pattern séquentiel (taux d'utilisation moyen de 89%). Dans 89% des cas, le pattern dominant est apparu dès la première construction multimodale.

Nous avons étudié de plus près les **constructions séquentielles** : nous avons relevé un **délai moyen de 0,6 secondes** (entre 0,1 et 1,8 sec) entre la fin de la première modalité et le début de la seconde. D'un point de vue système, un autre délai est important : celui entre la fin de la première modalité et la fin de la seconde modalité. En effet c'est le délai que le système devra attendre pour disposer de toutes les informations sur la construction multimodale : le module de fusion aura alors reçu le contenu sémantique de la parole et celui du geste. Ce délai s'élève à 1,3 secondes (valeur maximale 2,5 sec).

En ce qui concerne l'**ordre de succession des modalités**, il y a autant de constructions séquentielles dans lesquelles la parole précède le geste (50%) que l'inverse (50%).

Profil multimodal Adultes / Enfants. Dans la plupart des données précédentes, en particulier sur l'intégration sémantique, nous avons considéré le corpus global sans distinguer adultes et enfants. Dans le but d'identifier des profils de comportement multimodal pour ces deux groupes, nous avons réalisé une Analyse en Composantes Principales (voir Wolff, 2003, pour un exposé sur cette méthode d'analyse). Nous y avons inclus les variables suivantes :

- Nombre TOTAL de commandes,
- Utilisation de la PAROLE (%),
- Utilisation du GESTE (%),

- Utilisation de la MULTIMODALITE (%),
- Utilisation de la REDONDANCE (%),
- Utilisation de la COMPLEMENTARITE (%),
- Utilisation de la CONCURRENCE (%),
- Utilisation de la complémentarité dialoguée COMP-D (%),
- Utilisation de la REPETITION multimodale (%),
- Pourcentage de constructions simultanées SIM (sur le nombre total de constructions),
- Pourcentage de constructions séquentielles SEQ (sur le nombre total de constructions).

La matrice de corrélation nous indique que les corrélations positives les plus fortes sont :

MULTIMODALITE et SIM	r = 0,82
REDONDANCE et SIM	r = 0,69
MULTIMODALITE et REDONDANCE	r = 0,59

Cela signifie que les quelques sujets ayant beaucoup utilisé la multimodalité ont eu tendance à intégrer les modalités de façon simultanée et redondante. Les autres patterns temporels et sémantiques que nous avons observés sont donc sans doute attribuables à des sujets ayant moins utilisé la multimodalité.

Les corrélations négatives les plus fortes sont :

PAROLE et GESTE	r = -0,78
GESTE et SIM	r = -0,66
GESTE et MULTIMODALITE	r = -0,64

Les sujets ayant beaucoup utilisé le geste ont donc eu tendance à utiliser peu la parole, la multimodalité et la simultanéité.

A l'issue de l'Analyse en Composantes Principales, nous avons retenu 4 axes expliquant 78,3% de la variance totale du nuage de points[29]. Les variables les mieux représentées par ces 4 axes sont le GESTE (qualité de la représentation = 0,91), la REPETITION (0,84), la MULTIMODALITE (0,83), la COMPLEMENTARITE (0,83) et la COMP-D (0,83). La variable la moins bien représentée[30] est SEQ (0,55).

La Figure 38 et la Figure 39 représentent le nuage des variables dans les plans 1-2 et 3-4 respectivement. Pour l'interprétation de chaque axe, nous avons retenu les variables dont la

[29] Un total de 70% de variance expliquée est généralement considéré comme acceptable.

[30] En dessous de 0,6 la qualité de représentation est considérée comme faible. Cela signifie que l'interprétation de ces variables devra s'accompagner de précautions importantes.

contribution à la variance de l'axe était supérieure à la contribution moyenne. Ainsi, l'axe 1 oppose MULTIMODALITE – REDONDANCE – SIM d'un côté et GESTE – REPETITION de l'autre côté. Cet axe représente donc la distinction entre les individus ayant spontanément beaucoup utilisé la multimodalité et ceux qui ont plutôt utilisé le geste et n'ont recouru à la multimodalité qu'en cas d'échec de la communication. L'axe 2 oppose GESTE – CONCURRENCE avec PAROLE – COMP-D. Nous interprétons cet axe comme une opposition Action/Dialogue, l'Action étant représentée par l'utilisation fréquente du geste, parfois même en parallèle de la parole (concurrence), et le Dialogue étant associé à l'utilisation fréquente de la parole, y compris pour annoncer l'action gestuelle (complémentarité dialoguée).

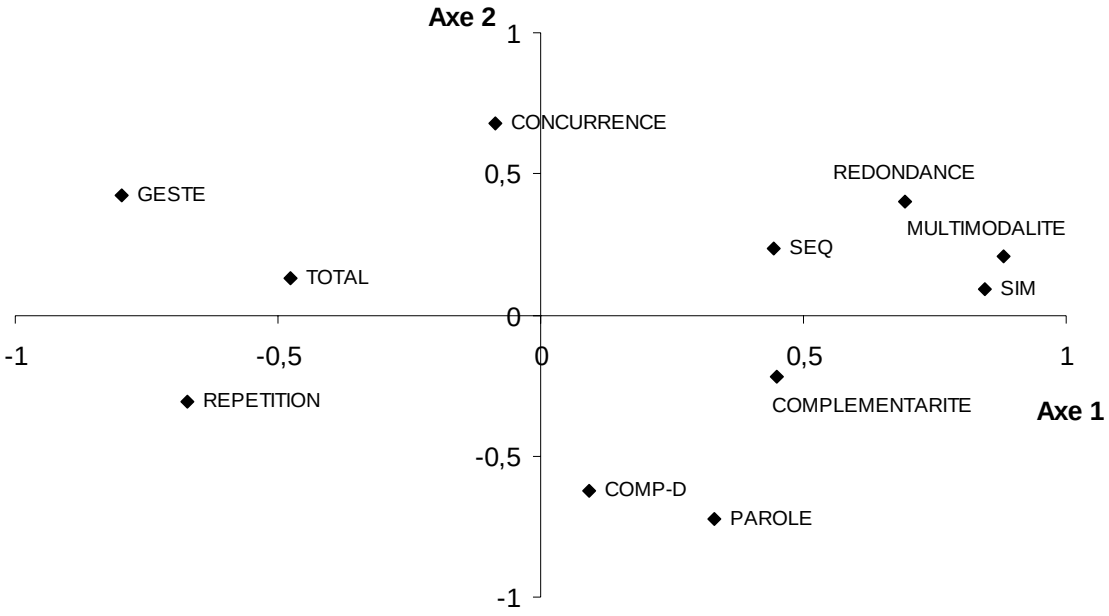

Figure 38 : Nuage des variables dans le Plan 1-2.

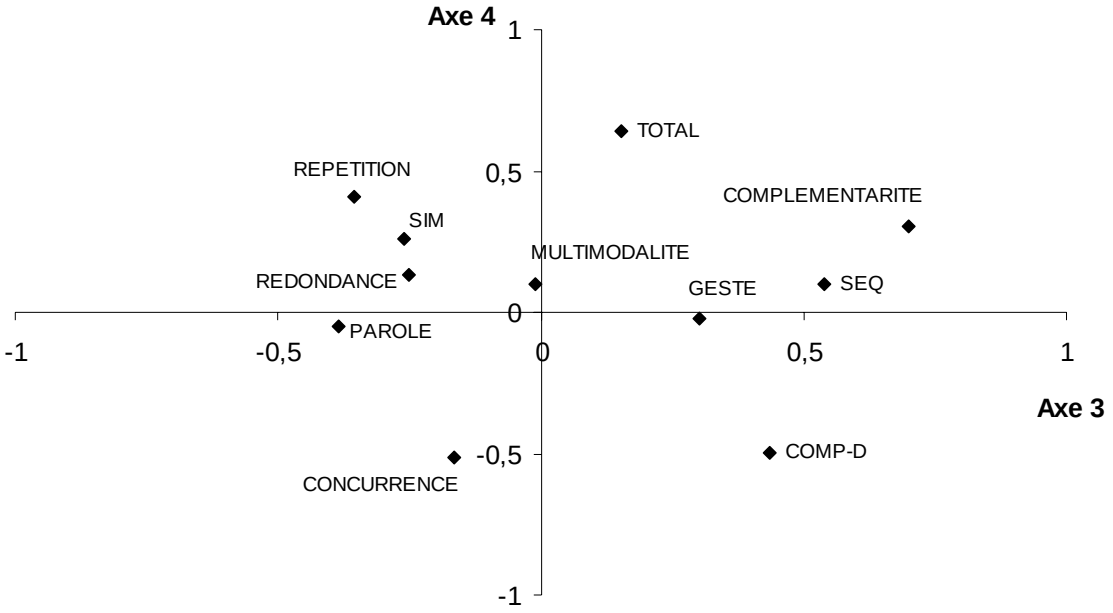

Figure 39 : Nuage des variables dans le Plan 3-4.

L'axe 3 oppose quant à lui COMPLEMENTARITE – COMP-D – SEQ avec la PAROLE. Les constructions complémentaires et/ou séquentielles peuvent donc correspondre à une économie de parole pour les individus ayant peu utilisé la modalité verbale. Enfin, l'axe 4 oppose TOTAL – REPETITION, qui traduisent une certaine inefficacité de l'interaction, avec COMP-D – CONCURRENCE, qui peuvent être vus comme des modes d'interaction optimisés, dans la mesure où les exemples de concurrence que nous avons observés concernaient une utilisation parallèle des deux modalités, c'est-à-dire une manière de faire deux choses à la fois.

Cette analyse du nuage des variables nous permet de présenter le nuage des points moyens des Adultes et des Enfants en fonction des axes tels que nous venons de les interpréter. La Figure 40 nous montre que les Adultes sont situés du côté de la multimodalité spontanée et du dialogue, alors que les enfants sont du côté du geste et de l'action. Les enfants ont tendance à recourir à la multimodalité en cas d'échec dans la communication ou bien pour effectuer simultanément deux actions concurrentes. La Figure 41, même si elle montre que Adultes et Enfants s'opposent moins sur les axes 3 et 4, semble confirmer que les enfants cherchent à économiser la parole et à optimiser l'interaction en utilisant les modalités de façon complémentaire ou concurrente. Les adultes restent du côté de la parole et d'une interaction un peu moins performante.

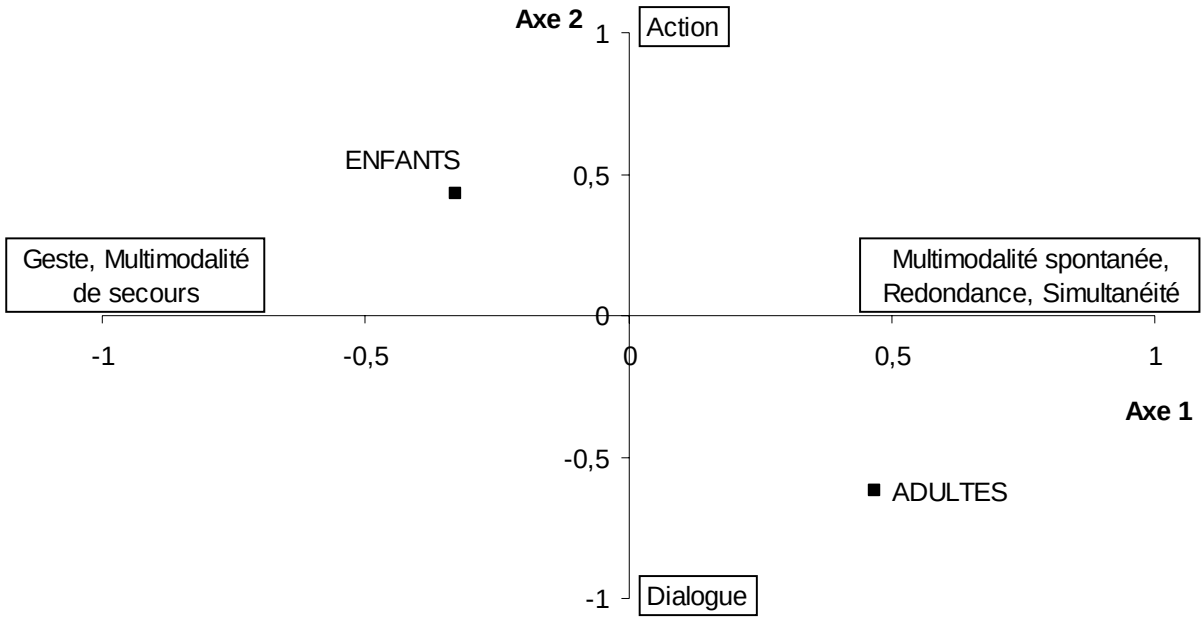

Figure 40 : Nuage des points moyens des Adultes et des Enfants dans le plan 1-2.

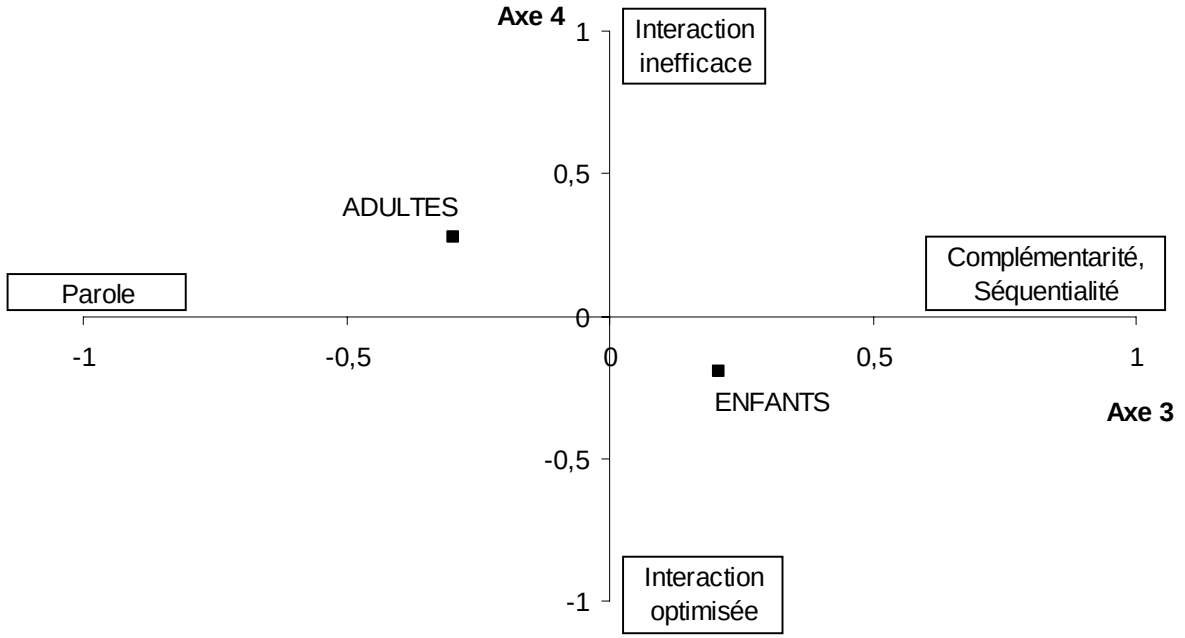

Figure 41 : Nuage des points moyens des Adultes et des Enfants dans le plan 3-4.

Discussion

Comparaison des deux conditions d'interaction. Nous avons vu que la multimodalité avait diminué la durée totale de réalisation du scénario. Cet effet témoigne sans doute d'une augmentation de l'efficacité d'interaction en condition multimodale, mais on pourrait s'interroger ici sur sa pertinence. En effet, étant donné notre objectif de concevoir une application conversationnelle, nous pourrions au contraire chercher à ce que les utilisateurs passent le plus de temps possible avec les Agents, pour les amener à dialoguer en dehors de la stricte réalisation du scénario (Ruttkay *et al.*, 2002). En cela nous pourrions, à l'extrême, choisir de limiter les entrées utilisateur à la modalité verbale. Cependant, nous avons également vu que la multimodalité augmentait la prise d'initiative de la part des utilisateurs, et augmentait également la facilité perçue, en particulier pour les enfants. Il semble donc que la condition multimodale offre le meilleur compromis : elle permet une interaction efficace et sera sans doute plus attrayante pour les enfants qu'un système uniquement verbal.

Pour illustrer la prise d'initiative par le geste, nous pouvons citer l'exemple de deux des enfants que nous avons observés qui pointaient parfois l'Agent avec le stylo lorsqu'ils arrivaient dans une nouvelle pièce : ce comportement peut être interprété comme une amorce non verbale au dialogue. A l'inverse en condition monomodale, les enfants attendaient fréquemment que ce soit l'Agent qui engage la conversation. Une autre illustration de l'intérêt de la modalité gestuelle peut être trouvée dans les comportements d'exploration qui sont apparus exclusivement chez les enfants : par exemple les enfants essayaient parfois de trouver un moyen de sortir d'une pièce à l'aide du stylo, ou bien parcouraient l'environnement graphique sans but apparent.

D'autres résultats semblent en outre indiquer que la multimodalité ne nuit en aucun cas aux aspects conversationnels : le nombre de mots par commande ne diminue pas, la syntaxe reste naturelle (proche de la communication Homme-Homme), les comportements sociaux sont aussi fréquents qu'en condition monomodale verbale. Soulignons enfin un autre effet positif de la condition multimodale : celui d'avoir homogénéisé les évaluations subjectives de facilité d'interaction. Alors que la condition monomodale faisait ressortir des différences entre adultes et enfants, la condition multimodale a gommé ces différences et remporté l'adhésion de tous.

Nous pouvons donc conclure à l'intérêt de la multimodalité, qui, au vu de nos résultats, apporte efficacité et facilité, en plus de constituer un important vecteur de convivialité pour les enfants. Nous retrouvons ainsi les résultats obtenus sur l'utilisation d'interfaces sans Agents (voir le paragraphe 3.3.1).

Comportement verbal. Nos sujets ont spontanément utilisé une **syntaxe verbale** naturelle, c'est-à-dire équivalente à un contexte de communication Homme-Homme. Cette syntaxe semble être plus élaborée que celle utilisée pour les interfaces sans Agents (voir Bellik, 1995 ; Robbe, 1998 ; Buisine & Martin, 2002). Le nombre moyen de 6 mots par commande semble légèrement supérieur à celui rapporté par d'autres auteurs avec des interfaces sans Agents (5 mots chez Hauptmann, 1989 ; 4,8 chez Oviatt, 1996a), et cette valeur n'a pas diminué lors de l'introduction de la multimodalité, contrairement à ce qui a été observé par Oviatt (1996a). Ces résultats pourraient être attribués soit au fait que nos utilisateurs interagissaient avec des Agents, soit aux aspects conversationnels de notre scénario – qui sont tout de même liés à la présence d'Agents. Bickmore et Cassell (2004) formulent une hypothèse similaire (l'interaction avec des Agents se rapproche davantage de la communication Homme-Homme que de l'interaction Homme-Machine).

Les données concernant les répétitions multimodales suggèrent par ailleurs que la parole a été perçue comme une modalité relativement fiable, puisqu'elle rend compte de la majorité des changements de modalité en cas d'échec de l'interaction. Rappelons ici que nous n'avons pas simulé d'erreurs de reconnaissance, et que ce choix a sans doute influencé la perception des utilisateurs quant à l'efficacité de l'interaction verbale.

En ce qui concerne les comportements sociaux, en particulier les marques de **politesse**, nous pensons qu'ils font partie des spécificités de l'interaction avec des Agents. Reeves et Nass (1996) ont développé la théorie *Media Equation* selon laquelle nous aurions des comportements sociaux similaires avec les systèmes informatiques (même non personnifiés) qu'avec des personnes réelles. Ils ont ainsi retrouvé dans le comportements d'utilisateurs face à des ordinateurs certaines des lois qui gouvernent nos relations sociales. Par exemple, lorsqu'un système demande l'avis des utilisateurs sur sa propre performance (par exemple « Que pensez-vous de ma prestation ? »), les utilisateurs donnent un avis plus positif et plus homogène à ce système que lorsque c'est un système tiers qui pose la question (« Que pensez-vous de la prestation de cet ordinateur ? »). Cette réaction correspond à celle que nous aurions face à des humains (par politesse, nous donnons un avis plus positif sur quelqu'un lorsque nous nous adressons à lui-même que lorsque nous nous adressons à un tiers). Les résultats de Reeves et Nass (1996) témoignent d'une manifestation implicite de politesse à l'égard des ordinateurs. Mais la différence avec les comportements que nous avons observés est que dans notre corpus la politesse est explicite, formulée par la parole. A l'inverse, quand les utilisateurs testés par Reeves et Nass (1996) ont été informés des résultats, ils se sont tous défendus d'avoir anthropomorphisé l'ordinateur et d'avoir cherché à être poli avec lui.

Comportement gestuel. Nos résultats concernant l'utilisation de la **modalité gestuelle** seule contrastent fortement avec les données de la littérature : dans notre corpus, l'utilisation du geste atteint 47,5% des entrées utilisateurs, et 59,3% si l'on considère le groupe d'enfants seul. La plupart des études sur les Interfaces Multimodales mentionnent une utilisation relativement faible du geste seul (13% chez Hauptmann, 1989 ; 17% chez Mignot *et al.*, 1993 ; 17,5% chez Oviatt *et al.*, 1997 ; 35% chez Kehler *et al.*, 1998 ; 33% chez Robbe-Reiter *et al.*, 2000). Seule Zanello (1997) a trouvé des proportions plus importantes (63% et 73%).

Ces études concernaient toutes des adultes ; la seule qui ait touché un public d'enfants (Xiao *et al.*, 2002) rapporte un taux de 10% d'utilisation du stylo seul, mais n'est pas comparable avec la situation que nous avons étudiée. En effet, l'application Xiao *et al.*, (2002) était exclusivement conversationnelle et ne comportait pas d'objets graphiques à manipuler. Nos résultats sont donc assez atypiques, et nous semblent difficiles à interpréter à la lumière de la littérature recensée. Une piste éventuelle d'explication pourrait résider dans le fait que nous avons testé un **jeu**. A notre connaissance, ce type d'application n'a pas été étudié dans le domaine des Interfaces Multimodales, et il se pourrait que les enfants aient transféré leurs habitudes de jeu (interaction principalement gestuelle dans les dispositifs actuels de jeux vidéo) dans notre test. Le comportement d'exploration relevé chez certains enfants vient renforcer cette hypothèse. Nous pensons donc que le lien fort entre interaction gestuelle et situation de jeu chez les enfants d'aujourd'hui est à prendre en compte dans la conception du système NICE.

La manière dont le stylo a été utilisé est restée relativement classique : l'utilisation principale a été la sélection d'éléments graphiques. Les adultes ont utilisé le pointage à 85%, les enfants ont eu des comportements un peu plus variés (encerclements, traçage de lignes). Deux difficultés sont tout de même à envisager pour le développement des modules gestuels : les gestes formés de plusieurs segments (*multistroke gestures*) comme par exemple les flèches (4,5% des gestes observés), sont plus difficiles à reconnaître pour un système automatique. Les comportements d'exploration (4,7% des gestes observés) sont quant à eux difficiles à interpréter du point de vue du système.

Comportement multimodal. Dans notre corpus, l'utilisation de la **multimodalité** atteint 21,2%. Comme nous l'avons souligné dans le chapitre 4, le contexte d'utilisation que nous avons étudié est nouveau par rapport aux données de la littérature : des tâches conversationnelles ont rarement été mises en place, et jamais dans le cadre de d'un jeu comportant de surcroît des aspects orientés tâche.

D'après notre analyse de l'utilisation des modalités en fonction des commandes, les actions orientées tâche avec une affordance graphique (action sur un objet visible), sont réalisées par le geste, ou par la multimodalité lorsqu'elles impliquent des aspects sociaux (prévenir, attirer l'attention, demander l'autorisation). Les actions orientées tâche sans affordance graphique (ex : sortir d'une pièce quand la porte n'est pas visible) sont traitées par la parole et parfois, chez les enfants, par le geste (exploration). Les aspects conversationnels sont réalisés par la parole (sauf les quelques cas d'activation non verbale précédemment cités).

Concernant l'**intégration sémantique**, nous avons relevé autant de constructions redondantes que complémentaires. Notre analyse des résultats antérieurs (voir le paragraphe 3.3.1) avait montré qu'il n'existait pas de consensus à ce propos. En revanche, un résultat surprenant par rapport à la littérature sur l'Interaction Homme-Machine multimodale est le pourcentage non négligeable (16%) de constructions concurrentes dans notre corpus. Celles-ci ont presque toutes été produites par des enfants (18 des 19 constructions concurrentes ont été produites par 4 enfants et la 19ème construction a été produite par un adulte). Aucune des études antérieures que nous avons recensées ne rapporte d'exemples de constructions concurrentes. Nous pouvons supposer qu'il s'agit d'une spécificité du comportement multimodal des enfants dans une telle situation (comprenant à la fois des enjeux orientés tâche et des aspects conversationnels).

Au niveau **temporel**, les modalités ont majoritairement été intégrées de façon simultanée (73%), c'est-à-dire avec un chevauchement temporel entre parole et geste. Ce résultat est conforme aux données de la littérature (Oviatt *et al.*, 2003). Notre corpus confirme également l'existence de patterns dominants, avec une fiabilité (88,3%) qui semble comparable à celle relevée par Oviatt (90% chez Oviatt, 1999b ; 93,5% chez Xiao *et al.*, 2002 ; 88,5% chez Xiao *et al.*, 2003). Le taux de prédictibilité (89%) est également comparable aux valeurs obtenues antérieurement (100% chez Oviatt, 1999b ; 92,3% chez Xiao *et al.*, 2002 ; 85% chez Xiao *et al.*, 2003). En revanche, avec autant de cas de précédence du geste sur la parole que l'inverse, nous ne vérifions pas l'ordre habituellement observé (précédence du geste chez Oviatt *et al.*, 1997 ; Zanello, 1997 ; Xiao *et al.*, 2002 ; 2003). Plusieurs facteurs peuvent être responsables de cette différence : le fait que les utilisateurs s'adressent à des Agents Conversationnels (mais le système testé par Xiao *et al.*, 2002, comportait aussi des Agents), le scénario d'interaction (à la fois conversationnel et orienté tâche), ou encore la langue d'interaction (bien que les résultats de Zanello, 1997, aient été obtenus avec des utilisateurs français).

Profil multimodal Adultes / Enfants. Enfin, notre travail permet pour la première fois de comparer le **comportement multimodal de jeunes adultes à celui d'enfants** sur la même application. Nos résultats indiquent que les adultes semblent utiliser plus de constructions multimodales (27,5%) que les enfants (18,3%). Nous pouvons également considérer que nos données

sur l'intégration des modalités chez les enfants sont inédites, dans la mesure où la seule étude antérieure portant sur un groupe d'enfants (Xiao *et al.*, 2002) faisait état de 93% de gestes inexploitables sémantiquement (gribouillages) dans les constructions multimodales.

Néanmoins, il y a eu relativement peu de spécificités du comportement multimodal des enfants dans notre corpus. Concernant l'**intégration sémantique**, nous avons relevé une tendance des enfants à produire des commandes concurrentes : les enfants dissocient parfois le canal verbal et le canal gestuel pour réaliser deux actions à la fois. Le traitement de ce type de constructions sera un défi pour le futur système NICE. Concernant l'**intégration temporelle**, notre corpus contenait une majorité de constructions simultanées, mais les constructions séquentielles provenaient principalement des enfants. Nous avons relevé autant de patterns dominants simultanés que séquentiels parmi les enfants, alors que les adultes suivaient presque tous un pattern dominant simultané.

Les différences que nous avons notées entre adultes et enfants nous semblent relever moins de la multimodalité que de leur **comportement de jeu**. En effet, l'Analyse en Composantes Principales a confirmé que les Adultes se sont davantage positionnés du côté de la conversation, alors que les enfants étaient plutôt du côté de l'action et du geste.

Plateforme expérimentale. Enfin, signalons que notre dispositif a globalement été efficace, puisqu'il a permis de mener à bien l'expérimentation et le recueil de corpus. Tous nos utilisateurs ont pu réaliser le scénario avec succès, et ils ont cru interagir avec un système fonctionnel (un seul utilisateur, exclu de l'analyse par la suite, avait deviné qu'il s'agissait d'une simulation). Le seul problème qui ait pu perturber l'interaction est que le délai de réponse du système (magicien + synthèse vocale) était trop élevé (2,2 sec en moyenne). Il est parfois arrivé que des utilisateurs fassent une action puis, en l'absence de réponse immédiate de la part de l'Agent, répètent cette même commande, et soient alors interrompus par la réponse tardive de l'Agent.

Pour de prochaines expérimentations, nous pourrons nous efforcer d'améliorer l'interface magicien, ou éventuellement de recourir à un magicien supplémentaire. Cette seconde solution pourrait rendre l'interaction plus dynamique, mais elle pose aussi le problème de la coordination des magiciens et de la cohérence de leurs réponses. Par ailleurs, il nous semble que le mode conversationnel engendre nécessairement un délai plus élevé de la part du (des) magicien(s) que la gestion d'une interface graphique classique. Une autre solution pour rendre l'interaction plus fluide (même si elle reste relativement lente) serait d'introduire des feedback et des signaux non verbaux de régulation des tours de parole (Cassell, 2001) dans le système simulé, par exemple :

- Faire en sorte que l'Agent hoche la tête occasionnellement lorsque l'utilisateur parle,

- Lui faire regarder l'objet que l'utilisateur pointe (attention conjointe),
- Lui faire faire quelques mouvements, avec le regard vers le haut pendant le délai d'attente,
- Lui faire regarder droit devant quand il commence à parler.

L'introduction de tels feedbacks serait à étudier pour les expériences de simulation, mais également pour le futur système, même si les délais de réponse sont réduits.

Conclusion

Nos recommandations pour l'implémentation du prototype I du système NICE sont synthétisées dans le Tableau 12.

Module concerné	**Recommandations**
Traitement du langage	Modèle de langage inspiré de la communication Homme-Homme.
Reconnaissance de gestes	Utilisation du corpus pour entraîner la reconnaissance de gestes.
Fusion multimodale	La multimodalité représente 21,2% des constructions.
	Nécessité d'un traitement sémantique pour distinguer les constructions complémentaires (6,8% du total des entrées) ou redondantes (5,8%) des constructions concurrentes (modalités à traiter séparément ; 3,8% du total des entrées).
	Délai pour l'intégration temporelle des constructions séquentielles : • Jusqu'à 1,8 sec entre la fin de la première modalité et le début de la seconde. A partir de 2 sec de délai, il est probable que ce soit une répétition. • Jusqu'à 2,5 sec entre la fin de la première modalité et la fin de la seconde. • Traitement des deux ordres (parole puis geste, geste puis parole).

Tableau 12 : Recommandations pour le prototype I.

5.2. Evaluation du prototype I avec Agent 3D, scénario conversationnel

Le système NICE comprend deux univers et deux scénarios différents :

- L'un des univers est le bureau de Hans Christian Andersen, dans lequel les utilisateurs le rencontrent en personne et discutent avec lui. Cette partie est essentiellement conversationnelle. Les utilisateurs peuvent interroger Andersen sur sa vie, ses contes, les objets présents dans son bureau ; ils peuvent aussi aborder quelques uns de leurs propres centres d'intérêt (jeux vidéos, films ou livres comme Harry Potter…).
- Le second univers est le monde des contes de fées, dans lequel les utilisateurs peuvent rencontrer certains personnages des contes d'Andersen et accomplir un scénario d'aventure (orienté tâche) avec eux.

Les tests utilisateurs décrits dans ce paragraphe concernent la partie conversationnelle, avec le personnage d'Andersen (Figure 42), du prototype I.

Figure 42 : H.C. Andersen dans son bureau.

Fonctionnement du Prototype I

Le prototype I a résulté de l'intégration de la première version des modules développés par les différents partenaires du projet NICE (intégration réalisée en 2003, en fin de 2ème année du projet). Le

LIMSI était chargé de l'implémentation[31] de trois modules : le module de Reconnaissance de Gestes, celui d'Interprétation du Geste, et le module de Fusion Multimodale.

Le module de **Reconnaissance de Geste** reçoit les gestes en 2D entrés par les utilisateurs à l'aide de dispositifs variés (écran tactile, souris, souris gyroscopique). Pour cette première version du système, trois formes de gestes étaient reconnues : le pointage, l'encerclement et le traçage de lignes.

La forme reconnue, et ses coordonnées, sont ensuite envoyées au module d'**Interprétation du Geste**, dont le rôle consiste à mettre en correspondance ce geste avec les informations reçues du moteur 3D sur les éléments graphiques qui étaient visibles au moment où le geste a été produit. Par exemple, si un cercle a été tracé, et qu'un objet se trouvait à ce moment dans la zone encerclée, le module d'Interprétation du Geste place cet objet en premier dans la liste des objets sélectionnés. Le moteur 3D n'envoie cependant pas les coordonnées de tous les éléments graphiques, mais seulement de certains objets définis comme sélectionnables. Pour le prototype I, le bureau d'Andersen contenait de nombreux objets, dont 21 sélectionnables. A titre d'illustration, la Figure 42 laisse entrevoir six de ces objets sélectionnables : cinq tableaux accrochés au mur et la plume posée sur le bureau d'Andersen.

L'objet retenu par le module d'Interprétation du Geste est ensuite envoyé au module de **Fusion Multimodale**, dans lequel sont intégrées les entrées verbales et gestuelles. Dans le prototype I, la fusion était réalisée sur un simple critère temporel : lorsqu'une entrée verbale et un geste étaient produits dans une même fenêtre temporelle, ils étaient fusionnés. La fenêtre temporelle était ici définie par le délai entre la fin de la première modalité et la fin de la seconde modalité : pour les tests utilisateur, ce délai a été paramétré, avec l'accord de nos partenaires, à 3 secondes.

Méthode

Les tests utilisateur pour le prototype intégré ont été réalisés en collaboration avec l'un de nos partenaires, le NIS-Lab[32] (Natural Interactive Systems Laboratory), et se sont déroulés au Danemark, sur le site du NIS-Lab.

Utilisateurs. Quatorze utilisateurs danois ont réalisé le test en Anglais :

- Neuf garçons, d'âge moyen 13 ans 9 mois (σ = 2,4 ans),
- Cinq filles, d'âge moyen 14 ans 6 mois (σ = 2,5 ans).

[31] Développement assuré par Sarkis Abrilian sous la direction de Jean-Claude Martin.

[32] http://www.nis.sdu.dk/

Matériel. Les tests ont eu lieu en parallèle dans deux salles différentes. Le dispositif était le même dans les deux salles, à l'exception du média gestuel : une des deux salles était munie d'un écran tactile, l'autre d'une souris. Dans la salle équipée de l'écran tactile, une caméra numérique filmait l'interaction (enregistrement audio du dialogue et gros plan sur l'écran tactile).

Pour le LIMSI, il s'agissait réellement de tests utilisateurs dans la mesure où les trois modules que nous souhaitions évaluer étaient fonctionnels. Cependant, sur l'ensemble du prototype, un module devait encore être simulé à ce stade du projet : celui de la reconnaissance vocale. Deux magiciens (un pour chaque salle de test) se trouvaient donc dans une autre pièce, écoutaient les paroles des utilisateurs et les retranscrivaient en temps réel. Le reste du traitement (interprétation de ces transcriptions, traitement des entrées gestuelles, gestion du dialogue, sortie verbale, animation…) était assuré par le système.

Outre les entrées verbales (par **micro casque**) et les gestes de désignation en 2D (par l'intermédiaire de l'**écran tactile** ou de la **souris**), les utilisateurs pouvaient aussi **naviguer** dans l'environnement de jeu à l'aide du clavier : la touche F2 était dédiée aux changements de caméra (vue aérienne, vue de face d'Andersen, etc.) et les flèches permettaient de faire bouger Andersen lui-même (avancer / reculer, tourner à droite / à gauche).

Procédure. Les consignes étaient données en Danois. Les utilisateurs étaient simplement informés de tous les moyens d'interaction (entrée verbale en Anglais, gestes sur l'écran, navigation) et du fait qu'Andersen les attendait dans son bureau pour discuter avec eux. Il n'y avait pas d'instructions plus précises sur le contenu de l'interaction.

Après 15 minutes d'interaction libre, l'expérimentateur donnait une liste de tâches, et les utilisateurs pouvaient en réaliser quelques unes au choix dans les 20 minutes suivantes :

- Engager la conversation sur deux contes,
- Engager la conversation sur deux tableaux et deux autres objets dans le bureau,
- Engager la conversation sur la famille d'Andersen (père, mère, grand-père),
- Trouver l'âge d'Andersen,
- Lui demander s'il a un conte préféré et lequel,
- Trouver quels jeux il aime,
- Parler à Andersen des jeux que vous aimez ou connaissez,
- Choisir un conte et demander à Andersen de donner la morale de l'histoire,
- De quels contes Andersen est-il capable de parler ?

- Parler à Andersen de vos centres d'intérêt,
- Pouvez-vous aller dans le monde des contes de fées ?

En fin de séance, les utilisateurs étaient interviewés en Danois sur les thèmes suivants :

- Quelques unes de leurs caractéristiques personnelles : préférences en termes de jeux vidéos, fréquence de jeux.
- Evaluation de l'interaction : facilité d'interaction avec Andersen, opinion sur l'interaction verbale, sur l'interaction gestuelle, niveau de compréhension de la part du système, qualité de la synthèse vocale, de l'animation, perception globale du système.
- Evaluation du contenu et amélioration du système : aspects ludiques, aspects pédagogiques, améliorations à envisager, intérêt de l'interaction verbale et gestuelle.

Les utilisateurs recevaient deux tickets de cinéma en remerciement de leur participation.

Données recueillies. A l'issue de ces tests, nous avons récupéré trois types de données :

- Les traces de l'interaction (*log files* ou fichiers de *log*) sauvegardées par les trois modules du LIMSI (Reconnaissance des Gestes, Interprétation des Gestes, Fusion Multimodale) : nous avons récupéré trois fichiers bruts pour chacun des 14 utilisateurs. Au total, ces fichiers retracent 8h30 d'interaction.
- Les enregistrements audio-visuels de l'interaction dans la salle munie de l'écran tactile (6 utilisateurs) : ce corpus vidéo couvrent environ 2h d'interaction (118,7 minutes), soit 40% des tests avec l'écran tactile (ou 25% de la totalité des tests).
- Les commentaires des utilisateurs sur l'interaction gestuelle, recueillis au cours des interviews et traduits par nos partenaires.

Les fichiers de *log* ont été dépouillés à l'aide d'un programme développé en interne au LIMSI[33] et les données extraites ont ensuite été soumises à des analyses statistiques sous SPSS.

Un échantillon de ces données provenant du système a également été confronté aux enregistrements vidéo de l'interaction, afin d'évaluer très précisément les traitements effectués par nos trois modules. Pour cela, nous avons importé dans Anvil (Kipp, 2001) à la fois les extraits vidéo et les événements systèmes afin de synchroniser et de visualiser en même temps ces deux types d'information. Les fichiers de *log* ont donc été convertis en fichiers d'annotation Anvil et insérés dans quatre pistes :

- Une piste contenant la transcription des paroles de l'utilisateur,

[33] Programme développé par Jean-Claude Martin.

- Une piste pour la Reconnaissance du Geste, signalant la survenue des gestes et, à chaque fois, la forme reconnue,
- Une piste pour l'Interprétation du Geste, qui indique quel objet a été sélectionné par le système,
- Une piste pour la Fusion Multimodale, montrant si parole et geste ont été fusionnés par le système, ou pas.

Nous avons ensuite agrémenté ce schéma de codage d'options supplémentaires afin de signaler et catégoriser les erreurs du système chaque fois que nous en observions : pour chaque élément provenant des fichiers de *log* (à l'exception des transcriptions des entrées verbales), nous avons ajouté un booléen Erreur (permettant de signaler les erreurs), une option Cause (pour expliciter la nature de l'erreur) et un champ texte pour commenter l'erreur, ou donner la forme ou l'objet correct qui aurait dû être reconnu.

Ces traitements nous ont permis d'évaluer :

- La Reconnaissance de Gestes : y a-t-il des gestes qui n'ont pas été détectés ? Les formes de gestes ont-elles été reconnues correctement ?
- L'Interprétation des Gestes : tous les gestes reconnus ont-ils été interprétés ? Les objets retenus par le système correspondent-ils à ceux sélectionnés par l'utilisateur ?
- La Fusion Multimodale : a-t-elle est réalisée de manière pertinente ?

Ce nouveau corpus nous a en outre permis d'analyser le comportement multimodal des utilisateurs de manière plus fine que ne le faisait le système (en particulier au niveau des relations sémantiques entre les modalités).

Résultats

Traitements effectués par le système. Les fichiers de *log* nous informent en premier lieu sur le nombre d'événements dans chacun de nos modules :

- 558 entrées gestuelles ont été traitées par les modules de Reconnaissance et d'Interprétation du Geste, ce qui correspond à un peu plus d'un geste par minute.
- 1049 entrées ont été traitées par le module de Fusion Multimodale, parmi lesquelles il y avait :
    - 854 (soit 81,4%) entrées monomodales verbales,
    - 188 (soit 17,9%) entrées monomodales gestuelles,
    - 7 (soit 0,7%) entrées multimodales.

Un premier constat s'impose face à ces chiffres : l'écart entre le nombre de gestes produits par les utilisateurs (558) et ceux traités au niveau du module de Fusion (195 si l'on tient compte des commandes mono- et multimodales) est flagrant. **65% des gestes n'ont pas été traités** comme éléments de communication. L'analyse de notre échantillon d'enregistrements vidéo (détaillée plus bas) montre que cela provient principalement de ce que les utilisateurs ont fait de nombreux gestes en direction d'objets non sélectionnables. Ces objets étaient invisibles pour le module d'Interprétation du Geste, qui a donc rejeté ces tentatives. De manière concordante, cinq utilisateurs (trois filles, deux garçons) ont mentionné dans les interviews à l'issue des tests qu'ils auraient souhaité que davantage d'objets soient sélectionnables.

Les pourcentages d'**utilisation des modalités** dans le module de Fusion Multimodale sont donc biaisés. Si les gestes non interprétés sont réintroduits dans ces données, les pourcentages d'utilisation de la parole et du geste se rééquilibrent sensiblement : 61% pour la parole, 39% pour le geste. L'utilisation combinée des deux modalités sera abordée plus bas.

Nous avons ensuite étudié l'influence des facteurs interindividuels sur ces données. Le **sexe des utilisateurs** ne semble avoir joué que sur le nombre d'entrées monomodales verbales : les garçons ont en effet produit plus d'entrées verbales (en moyenne 68,9 commandes par scénario, $\sigma = 17,1$) que les filles ($m = 46,8$; $\sigma = 30,4$).

Il semble par ailleurs qu'il y ait une influence du **dispositif physique** (écran tactile *vs.* souris) sur le comportement gestuel : les utilisateurs munis de la souris ont produit plus de gestes (en moyenne 52,2 par scénario, $\sigma = 19,6$) que ceux qui avaient un écran tactile ($m = 30,6$; $\sigma = 16,9$). Si l'on considère les pourcentages d'utilisation des modalités (pourcentages corrigés tenant compte des gestes vers les objets non sélectionnables), il apparaît que les utilisateurs munis de souris ont utilisé le geste et la parole à proportion équivalentes (52,2% et 47,8%, $\sigma = 22,9$) alors que les utilisateurs de l'écran tactile ont majoritairement interagi par la parole (69,7%, $\sigma = 8,6$; voir la Figure 43).

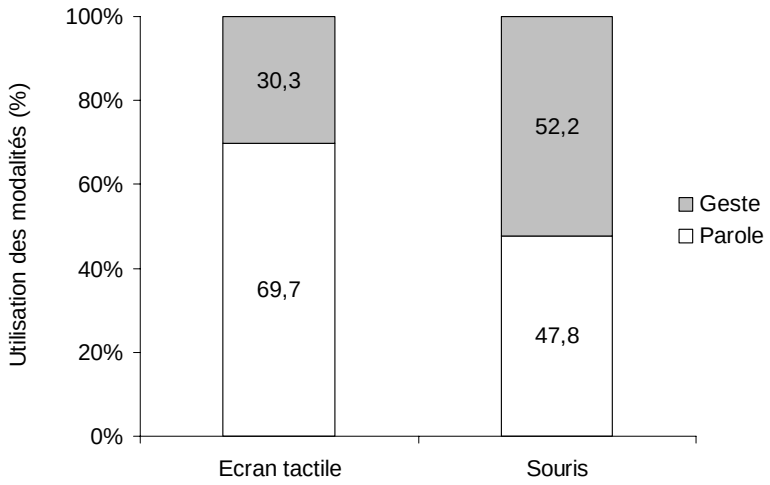

Figure 43 : Pourcentages d'utilisation de la parole et du geste en fonction du média gestuel.

En ce qui concerne les **formes de gestes**, les fichiers de *log* indiquent que 43,7% des gestes ont été reconnus comme des pointages, 32,8% comme des encerclements et 23,5% comme des lignes. Ces proportions ne sont influencées ni par le sexe des utilisateurs, ni par le média gestuel (écran tactile *vs*. souris). La Figure 44 récapitule les pourcentages d'utilisation des formes de gestes pour tout l'échantillon et pour chaque média gestuel les écarts-types sont disponibles dans le Tableau 13.

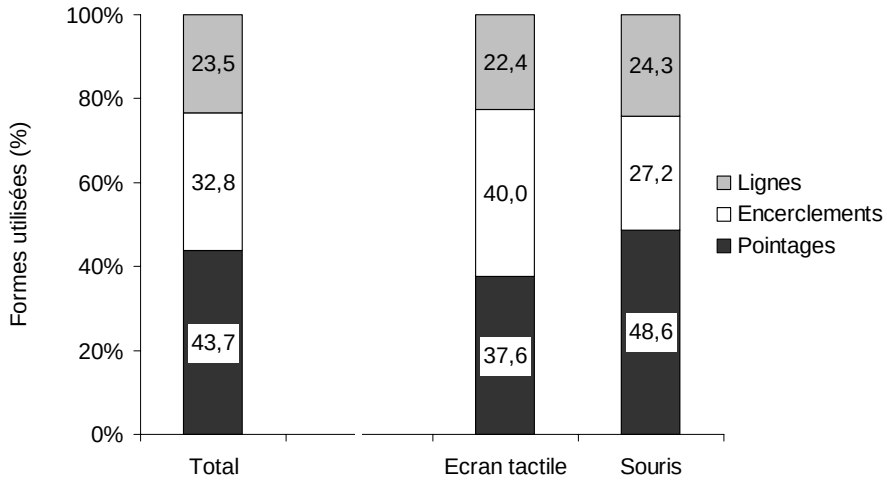

Figure 44 : Pourcentage d'utilisation des formes de gestes pour la totalité des utilisateurs (à gauche) et en fonction du média gestuel (à droite, différences non significatives).

		Pointages	Encerclements	Lignes
Total	m	43,7%	32,8%	23,5%
	σ	*3,6*	*2,7*	*1,1*
Ecran tactile	m	37,6%	40%	22,4%
	σ	*3,5*	*7,1*	*2,2*
Souris	m	48,6%	27,2%	24,3%
	σ	*9*	*4,4*	*2*

Tableau 13 : Pourcentages et écarts-types d'utilisation des formes de gestes.

Nous avons également établi, à partir des fichiers provenant du module d'Interprétation du Geste, la liste des **objets sélectionnés** dans l'ensemble du corpus. Ceci avait pour but de connaître les objets les plus demandés par les utilisateurs. Il est apparu que dans 57% des cas, l'interprétation du geste avait échoué (aucun objet sélectionné à l'issue du traitement), ce qui rejoint en partie nos précédentes observations sur le traitement des gestes. Mais les objets non sélectionnables ne sont pas seuls en cause dans l'absence de traitement des entrées gestuelles : celles traitées avec succès par le module d'Interprétation du Geste sont en effet plus nombreuses (239) que celles parvenues au module de fusion (195). Ceci témoigne d'un dysfonctionnement d'un de nos modules ou de l'application graphique. L'évaluation à proprement parler des modules sera détaillée dans le prochain paragraphe.

Parmi les objets effectivement sélectionnés, ceux qui ont eu le plus de succès sont : le portemanteau d'Andersen (sur lequel étaient accrochés un manteau et des chapeaux), qui a été sélectionné dans 14,2% des cas ayant abouti ; la chaise à bascule (13,8% des cas aboutis) ; puis les tableaux accrochés au-dessus du bureau (9,6% à 12,9% chacun).

Enfin, nous avons recherché les **constructions multimodales** détectées par notre système. Etant donné qu'il n'y en a eu que 7 sur l'ensemble des fichiers de *log*, nous les avons toutes entièrement retranscrites dans le Tableau 14. De manière surprenante, nous pouvons constater que dans toutes ces constructions, les modalités coopèrent par **concurrence** : à chaque fois, la parole et le geste sont utilisés pour aborder des sujets différents. Dans le paragraphe suivant, nous discuterons à nouveau certaines de ces constructions à la lumière des enregistrements vidéo.

Référence utilisateur	**Paroles et interprétation sémantique**	**Objet sélectionné par le geste**
21/01-3PM Boy Mouse	*hm i do not know* [user_opinion:negative] [verb:know]	Picture-of-Jenny-Lind
22/01-2PM Boy TactileScreen	*oh i remember that now* [user_opinion:general] [verb:remember] [diectic:that]	Picture-of-Little-Mermaid
22/01-2PM Boy TactileScreen	*Mm* [no_semantics]	Picture-of-Jenny-Lind
22/01-3PM Boy Mouse	*can tell me about your dad and your mom* [question:general] [user_intent:listen] [family:father] [family:mother]	Picture-of-Jonas-Collin
22/01-3PM Boy Mouse	*can you tell me about your dad and your mom and your grandpa* [question:yes/no] [hca_old] [family:father] [family:mother]	Picture-of-Jonas-Collin
22/01-3PM Boy Mouse	*okay goodbye i will go* [user_opinion:positive] [verb:visit] [greeting:ending]	Picture-of-Jenny-Lind
22/01-4PM Boy TactileScreen	*it is a bout a duck who is not as pretty as the other ones and eh but in the end it becomes much prettier in the other ones* [user_opinion:general] [greeting:ending] [fairytale:ugly_duckling] [number:1]	Picture-of-Jenny-Lind

Tableau 14 : Retranscription des constructions multimodales relevées par notre système.

Evaluation des modules. Les données exposées ci-dessus nous renseignent sur le comportement des utilisateurs, à condition que nos modules aient fonctionné correctement. Dans une seconde phase d'analyse, nous avons comparé systématiquement les données issues des fichiers de *log* avec l'échantillon d'enregistrements vidéo dont nous disposions.

Nous avons ainsi identifié, pour chacun des trois modules, les erreurs de traitement et la cause de ces erreurs. Le Tableau 15 présente une synthèse des erreurs relevées pour le module de **Reconnaissance de Gestes**.

Type d'erreur	**Nombre d'erreurs**
Non détection de geste	1
Forme erronée	7
Gestes avec segments multiples	7
Nombre total d'erreurs	15
Nombre total de gestes dans le corpus vidéo	117
Taux d'erreur	**12,8%**

Tableau 15 : Erreurs de traitement du module de Reconnaissance de Gestes.

Les erreurs sur les formes reconnues concernaient principalement des gestes de pointage reconnus comme des lignes (lorsque l'utilisateur avait légèrement bougé en pointant). Ce type d'erreur n'a entraîné aucune perturbation dans la suite du traitement par les autres modules.

Des gestes à segments multiples (*multistroke gestures*) ont été observés lorsque les utilisateurs essayaient d'encercler un objet et formaient le cercle en plusieurs fois. Chaque fragment de cercle a été traité séparément alors qu'il faisait partie, du point de vue de l'utilisateur, d'un seul et même geste. Ceci a provoqué un enchaînement d'erreurs dans les traitements consécutifs : chaque fragment de geste a été associé à l'objet visé et, dans le cas où cet objet était sélectionnable, a donné lieu à une réponse de la part du système. Il y a donc parfois eu répétition d'une même réponse suite à une seule entrée de la part de l'utilisateur.

Signalons par ailleurs que nous n'avons relevé dans ce corpus vidéo aucun geste particulièrement complexe : pas de flèches ni de croix, pas de sélection multiple d'objets (encerclement de plusieurs objets en même temps ou pointages successifs).

Le Tableau 16 récapitule les erreurs que nous avons observées dans l'**Interprétation des Gestes**. La plupart d'entre elles proviennent de l'absence de détection d'objets sélectionnables. Nous avons fourni

aux développeurs la liste des objets posant problème (principalement deux des tableaux accrochés au mur, qui étaient bien visibles aux utilisateurs mais disparaissaient de la liste reçue du moteur 3D sous certaines configurations de caméra).

Type d'erreur	**Nombre d'erreurs**
Absence de traitement	1
Non reconnaissance d'objets sélectionnables	24
Un objet pour un autre	2
Un objet à la place d'un non sélectionnable	3
Nombre total d'erreurs	30
Nombre total de gestes dans le corpus vidéo	117
Taux d'erreur	**25,6%**
Gestes vers des objets non sélectionnables	43 (38,8% des gestes dans le corpus vidéo)

Tableau 16 : Erreurs de traitement du module d'Interprétation du Geste.

Nous avons également fourni à l'ensemble des partenaires la liste des objets non sélectionnables que les utilisateurs ont cherché à sélectionner, avec le nombre de tentatives. Parmi ceux-ci, les plus demandés ont été : la feuille manuscrite posée sur le bureau, la bibliothèque et les tableaux disposés autour de la bibliothèque.

Type d'erreur		**Nombre d'erreurs**
Absence de traitement	Erreur du module d'Interprétation	24
	Autre raison	2
Fusion non pertinente (temps de transcription par le magicien)		2
Absence de fusion	Erreur du module d'Interprétation	1
	Objet non sélectionnable	5
	Temps de transcription par le magicien	2
Réponses multiples (geste avec segments multiples)		4
Nombre total d'erreurs		40
Nombre total d'événements dans le corpus vidéo		268
Taux d'erreur		**14,9%**

Tableau 17 : Erreurs de traitement du module de Fusion Multimodale.

Enfin, nous avons examiné les erreurs du module de **Fusion Multimodale** (rapportées dans le Tableau 17). La plupart sont dues à des erreurs du module d'Interprétation du Geste (non détection d'objets sélectionnables), d'autres à des erreurs du module de Reconnaissance (gestes à segments multiples),

d’autres enfin aux références à des objets non sélectionnables. Les erreurs de fusion sont quant à elles imputables au dispositif de Magicien d’Oz : le délai introduit par le temps de transcription du magicien a différé toutes les commandes verbales au-delà de la fenêtre temporelle de fusion, ou dans une fenêtre incorrecte. Le comportement multimodal des utilisateurs sera plus précisément abordé dans le paragraphe suivant.

Enfin, un autre type de problème lié aux traitements du module de Fusion concerne l’interruption du discours d’Andersen – il était prévu que cette fonctionnalité ne soit pas implémentée dans le prototype I. Les conséquences de cette impossibilité d’interrompre ont cependant été plus importantes que nous l’avions imaginé. Nous avons en effet observé en assistant aux tests (sans avoir retrouvé ce phénomène dans le corpus vidéo) que parfois des utilisateurs sélectionnaient par le geste un ou plusieurs objets pendant qu’Andersen était en train de parler. Il s’est avéré que toutes ces requêtes étaient stockées par le système, et étaient ensuite traitées au fur et à mesure : Andersen y répondait donc, mais après un délai important, ce qui perturbait beaucoup l’interaction et rendait Andersen indisponible au dialogue pendant toute cette période. L’utilisateur n’avait d’autre possibilité que d’attendre que toutes les requêtes aient été traitées pour pouvoir reprendre l’interaction.

Comportement multimodal des utilisateurs. Dans ce paragraphe, nous allons rapporter nos observations sur le comportement multimodal des utilisateurs dans le **corpus vidéo**.

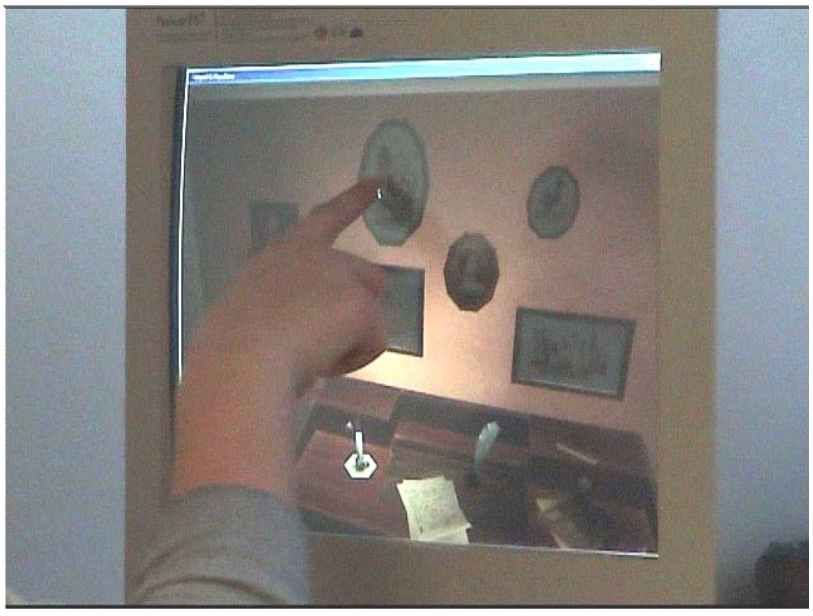

Figure 45 : Exemple de fusion non pertinente. L’utilisateur dit *« Oh I remember that now »* puis pointe le tableau représentant la petite sirène, après un délai de 6 secondes.

Revenons tout d'abord sur les constructions multimodales détectées par le système et présentées dans le paragraphe précédent (Tableau 14) : dans les vidéos, nous avons retrouvé deux d'entre elles, qui se sont avérées être toutes les deux des fusions non pertinentes. Par exemple, un des utilisateurs, après avoir écouté l'histoire de la petite sirène, a dit *« Oh I remember that now »*. Puis, après un délai de six secondes, il a pointé le tableau représentant la petite sirène (Figure 45). Dans cet exemple, le temps que l'entrée verbale soit transcrite par le magicien et envoyée au module de Fusion Multimodale, elle est arrivée dans la fenêtre temporelle suivant la survenue du geste (l'entrée verbale est arrivée une seconde après la fin du geste).

A l'inverse, la Figure 46 illustre un cas où les modalités n'ont pas été fusionnées alors qu'elles auraient dû l'être. L'utilisateur a encerclé le tableau du Colisée et a demandé *« What's this? »* deux secondes plus tard. Cependant, en raison de la transcription manuelle par le magicien, l'entrée verbale est arrivée cinq secondes après la fin de l'entrée gestuelle, soit en-dehors de la fenêtre temporelle de fusion. Signalons tout de même que, dans cet exemple, le tableau du Colisée a été correctement sélectionné lors de l'interprétation du geste et Andersen y a répondu par une anecdote sur le Colisée. L'erreur de fusion est donc probablement passée inaperçue aux yeux de l'utilisateur.

Figure 46 : Exemple de fusion manquante. L'utilisateur encercle le tableau du Colisée et demande, après un délai de 2 secondes, *« What's this? »*.

La Figure 47 présente un autre exemple de construction multimodale non détectée par le module de Fusion. Dans ce cas l'utilisateur a encerclé une statue représentant deux personnages et a demandé *« Do you have anything to tell me about these two? »*. Le traitement a échoué dès l'interprétation du geste puisque cette statue ne fait pas partie des objets sélectionnables. Le module de Fusion

Multimodale a donc traité l'entrée verbale comme monomodale, ce qui a conduit à une absence de réponse à la question de l'utilisateur.

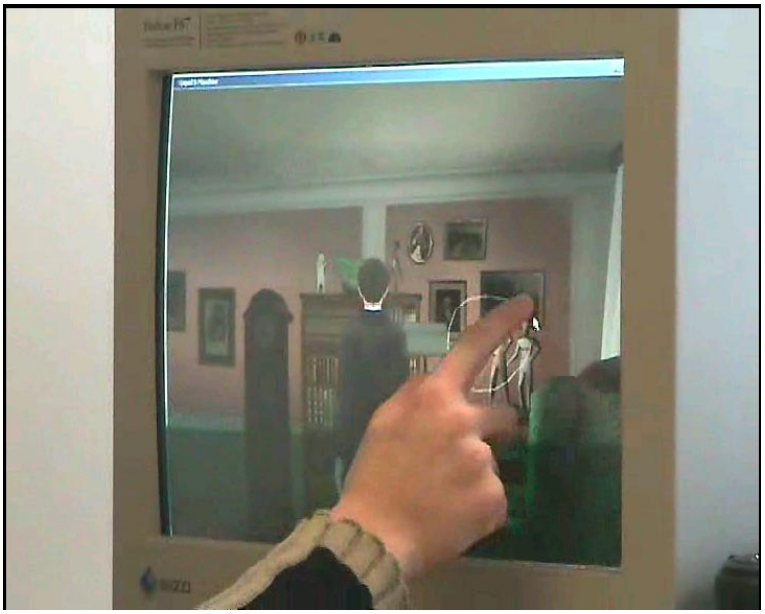

Figure 47 : Exemple de fusion manquante. L'utilisateur encercle un objet non sélectionnable et demande *« Do you have anything to tell me about these two? »*.

Au total, huit constructions multimodales (sur 306 entrées) ont été relevées dans le corpus vidéo – aucune d'entre elles n'ayant été effectivement détectée par le système. L'utilisation des modalités dans le corpus vidéo se répartit donc ainsi :

- 186 (soit 60,8%) entrées monomodales verbales,
- 112 (soit 36,6%) entrées monomodales gestuelles,
- 8 (soit 2,6%) entrées multimodales.

Les constructions multimodales sont retranscrites dans le Tableau 18 : on y voit que la moitié d'entre elles sont des constructions simultanées, l'autre moitié sont séquentielles avec précédence du geste sur la parole. En terme d'intégration sémantique, six de ces constructions (soit 75%) témoignent d'un usage **complémentaire** des modalités.

Ordre des modalités	**Délai (1) entre modalités**	**Délai (2) entre modalités**	**Objet sélectionné**	**Forme du geste**	**Entrée verbale et interprétation**	**Intégration sémantique des modalités**
Geste – parole	1 sec.	2 sec.	Tableau du Colisée (sélectionnable)	Cercle	*"What's this?"* [question:general] [deictic:this]	Complémentarité
Simultanéité	X	X	Tableau de la mère d'Andersen (sélectionnable)	Cercle	*"What's that picture?"* [question:general] [objectinstudy:picture]	Complémentarité
Simultanéité	X	X	Chapeau (sélectionnable)	Cercle	*"I want to know something about your hat."* [question:general] [objectinstudy:hat]	Redondance
Geste – parole	2 sec.	4 sec.	Statue de 2 personnages (non sélectionnable)	Cercle	*"Do you have anything to tell me about these two?"* [question:yes/no] [user_intent:listen] [number:2]	Complémentarité
Simultanéité	X	X	Statue de 2 personnages (non sélectionnable)	Point	*"What are those statues?"* [question:general]	Complémentarité
Geste – parole	1 sec.	4 sec.	Tableau au-dessus de la bibliothèque (non sélectionnable)	Cercle	*"Who is the family on the picture?"* [question:person] [family:general] [objectinstudy:picture]	Complémentarité
Geste – parole	0,1 sec.	3 sec.	Tableau au-dessus de la bibliothèque (non sélectionnable)	Cercle	*"Who is in that picture?"* [question:person] [objectinstudy:picture]	Complémentarité
Simultanéité	X	X	Vase (non sélectionnable)	Cercle	*"How old are you?"* [question:general] [hca_age]	Concurrence

Tableau 18 : Description des constructions multimodales observées dans le corpus vidéo. Le délai (1) a été mesuré entre la fin de la première modalité et le début de la seconde ; le délai (2) a été mesuré entre la fin de la première modalité et la fin de la seconde.

Comportements de navigation. Enfin, nous avons étudié dans ce corpus vidéo une troisième modalité dont nous ne disposions pas dans le corpus précédent : la navigation dans l'environnement. Cette modalité ne sert pas à communiquer, mais elle correspond tout de même à une action de l'utilisateur.

Rappelons que les utilisateurs, pour la navigation, avaient la possibilité de contrôler les déplacements d'Andersen (à l'aide des flèches du clavier) et de suivre ceux-ci (Figure 48).

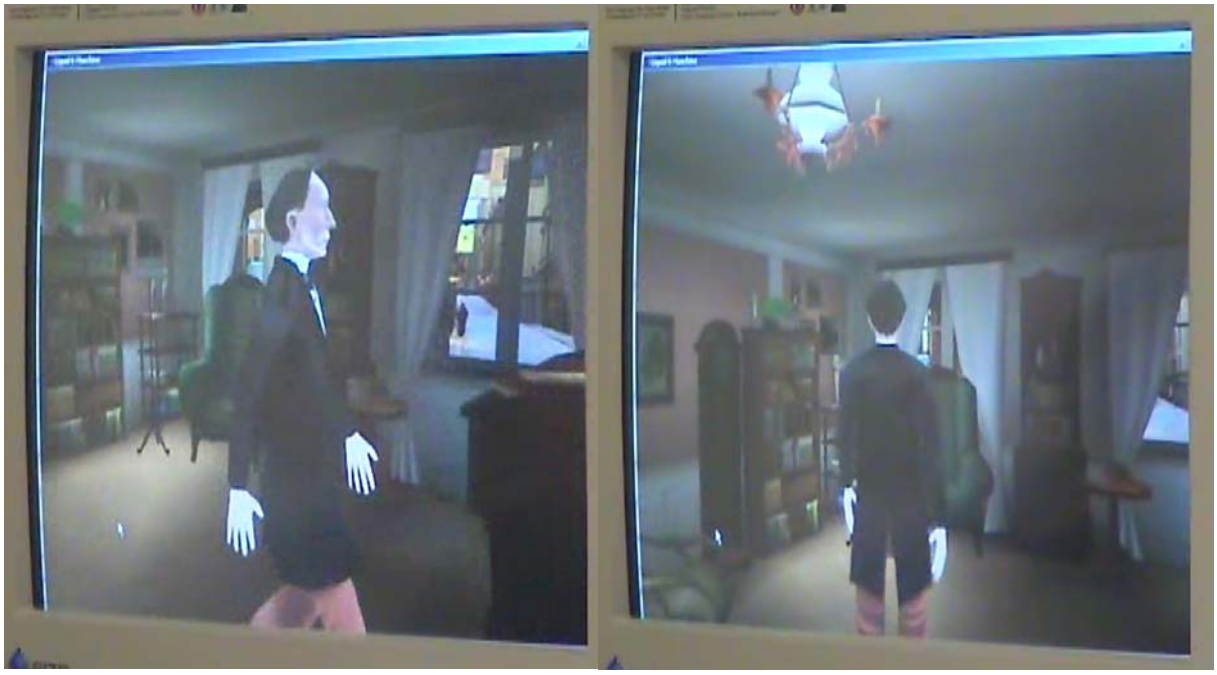

Figure 48 : Navigation dans l'environnement par contrôle et suivi des déplacements d'Andersen.

En assistant aux tests, il nous avait semblé que certains utilisateurs naviguaient beaucoup et vite, et pouvaient naviguer et converser en même temps. Il s'avère en effet que la navigation représente 32% de la durée du corpus vidéo (38,2 minutes). La plupart du temps, la navigation est réalisée pendant qu'Andersen raconte des anecdotes, mais il arrive également que les utilisateurs naviguent lorsqu'ils sont eux-mêmes en train de parler : 40,3% des entrées monomodales verbales ont été produites pendant la navigation.

Le mode de navigation par contrôle des déplacements d'Andersen a également eu la conséquence de présenter Andersen de dos la plupart du temps, remettant ainsi fortement en question la situation de conversation dite en face-à-face. Par ailleurs, nous pensons que cela nuit à la crédibilité de l'Agent, puisque cela tend à découpler son comportement verbal (converser) et son comportement non verbal (se déplacer, tourner le dos à l'utilisateur). Le fait de tourner le dos empêche aussi l'utilisateur de profiter pleinement du comportement non verbal de gesticulation d'Andersen (Figure 49).

Figure 49 : Comportement non verbal d'Andersen, de dos. L'image de gauche est un haussement d'épaules illustrant la réplique *« What do you think ? »* ; l'image de droite illustre *« My grand father was crazy alas. »*

Discussion

Comportement des utilisateurs. Par rapport à nos observations initiales dans le cadre du Magicien d'Oz, la proportion d'**utilisation des modalités** s'est inversée : dans ces tests utilisateurs, c'est la parole qui a été le moyen d'interaction privilégié. Un tel résultat pouvait être attendu, compte tenu de la nature entièrement conversationnelle du scénario d'interaction.

Nous avons observé un effet du sexe des participants sur l'utilisation de la **modalité verbale** : les garçons ont produit plus d'entrées verbales que les filles. Cet effet peut sembler surprenant : les données classiques en psychologie cognitive seraient plutôt en faveur du pattern opposé – plus grande aisance verbale chez les femmes (Kimura, 1999). Cependant cette hypothèse classique ne s'applique peut-être pas dans le contexte d'un jeu et avec des utilisateurs jeunes. Il est possible que les filles aient été plus intimidées que les garçons par cette situation, et aient donc moins interagi par la parole.

L'**interaction gestuelle**, même si elle est plus faible que dans le scénario orienté tâche, n'est cependant pas à négliger, puisqu'elle a tout de même représenté plus d'un tiers des entrées des utilisateurs (39% dans les fichiers de *log* ; 36,6% dans le corpus vidéo). Cette proportion est d'autant plus importante que l'utilisation du geste a été peu renforcée dans ces tests : les entrées gestuelles se sont soldées par un échec dans 65% des cas dans ce premier prototype (objets non sélectionnables,

erreurs d'interprétation, etc.). Compte tenu de ces obstacles et de la nature du scénario, nous considérons que l'intérêt des jeunes utilisateurs pour l'interaction gestuelle s'est confirmé.

Nous avons vu que le dispositif avait eu une influence sur l'utilisation de la modalité gestuelle : la souris semble faciliter l'interaction gestuelle par rapport à l'écran tactile. Cet effet est sans doute attribuable à l'habitude et à la grande expérience que les utilisateurs ont de la souris. L'effet de nouveauté en faveur d'un dispositif inhabituel comme l'écran tactile n'a pas dû être assez fort pour compenser cet avantage de la souris.

L'intérêt des jeunes utilisateurs pour l'**exploration** est lui aussi confirmé, étant donné la part importante du temps d'interaction consacré à la navigation (32%). Celle-ci a également engendré une sorte de concurrence dans les entrées des utilisateurs, puisque 40,3% des entrées monomodales verbales ont été produites pendant la navigation. Cette concurrence ne provoque aucune difficulté de fusion multimodale, puisque les comportements de navigation ne sont pas traités comme des éléments de communication, mais elle témoigne à nouveau de la tendance des jeunes utilisateurs à mener plusieurs actions en parallèle – tendance que nous avions déjà relevée dans la précédente étude.

L'utilisation de la **multimodalité** dans ces tests a été extrêmement faible : 2,6% au vu du corpus vidéo. A quoi pouvons-nous attribuer ce résultat ? Il est peu probable que cela soit dû à la nature conversationnelle du scénario, puisque celle-ci n'a pas empêché les utilisateurs de produire de nombreuses entrées gestuelles. On aurait même pu s'attendre à ce qu'un scénario conversationnel incite les utilisateurs à privilégier les entrées multimodales par rapport aux entrées monomodales gestuelles. En effet, les gestes ne correspondaient pas ici à des actions ou des commandes de la part de l'utilisateur, mais à l'orientation du dialogue avec Andersen, à l'introduction de nouveaux sujets de conversation : sélectionner un objet ne signifiait rien d'autre que « Parle-moi de cet objet », n'avait aucune autre conséquence concrète.

Une autre piste pour expliquer le faible taux d'utilisation de la multimodalité peut être de considérer la langue maternelle des utilisateurs (les Danois produisent-ils moins de constructions multimodales ?) ou le fait qu'ils n'interagissaient justement pas dans leur langue maternelle (puisque nos utilisateurs danois devaient parler anglais pour ce test). Ces deux hypothèses sont intéressantes, mais nous ne pouvons les étayer d'aucune donnée de la littérature : à notre connaissance, de telles problématiques n'ont jamais été étudiées.

D'autres différences entre le premier et le second corpus, telles que le rendu des graphismes (2D dans le Magicien d'Oz, 3D dans le prototype) constituent des pistes d'explication peu probables, car il est difficile d'imaginer pour quelle raison les graphismes 3D inhiberaient la multimodalité.

Enfin, une dernière hypothèse nous semble susceptible d'expliquer la quasi-absence de constructions multimodales dans ce corpus. Ce premier prototype avait le mérite d'être presque entièrement fonctionnel, mais il faut reconnaître qu'il ne fonctionnait « pas très bien ». Dans ce paragraphe, nous avons développé les problèmes relatifs aux modules gestuels et multimodal, mais nous devons signaler que les utilisateurs ont également rencontré des difficultés dans l'interaction verbale (c'était aussi l'objectif de ces tests que de mettre en évidence les *bugs* du système). Nous pensons alors que les utilisateurs, constatant ces faiblesses, ont pu adopter spontanément un comportement protecteur envers le système, en évitant les constructions multimodales. Par contraste, dans l'expérience en Magicien d'Oz, l'interaction était certes limitée, mais plus fluide, ce qui a sans doute amélioré la perception que les utilisateurs avaient des capacités du système, et les a donc moins incités à restreindre leurs comportements multimodaux.

Nous pensons que les faiblesses de ce premier prototype, dans la mesure où elles sont perçues par les utilisateurs, peuvent également expliquer l'absence de **comportements sociaux** dans ce corpus : nous n'avons pas analysé finement le comportement verbal des utilisateurs, mais nous pouvons tout de même signaler que le corpus vidéo ne contenait aucune marque de politesse, contrairement à ce que nous avions observé dans l'étude précédente. Cette hypothèse, et en particulier l'idée que les utilisateurs adaptent leur comportement multimodal aux capacités perçues du système, sera davantage développée dans le chapitre 7.

Plateforme expérimentale. Compte tenu du dispositif semi-fonctionnel et semi-simulé, notre module de fusion n'a pas pu fonctionner correctement pour ces tests. En effet, la simulation de la reconnaissance vocale a décalé toutes les entrées verbales, et, étant donné que le seul critère de fusion était temporel, il y a eu 100% d'erreurs pour la fusion. Cet effet regrettable ne pouvait être évité, faute de solution technique disponible. En effet, la fenêtre temporelle de fusion était déjà relativement importante (3 secondes). L'augmenter encore n'aurait pas écarté la possibilité de fusion non pertinente, et aurait en revanche certainement entravé l'utilisabilité du système, puisque le délai de 3 secondes entre la fin d'une entrée et le déclenchement de la réponse du système a déjà été perçu comme trop long (comme l'ont signalé cinq des utilisateurs lors des interviews).

Une des solutions pour éviter les fusions non pertinentes aurait été d'introduire un critère sémantique : le système aurait ainsi pu rejeter toutes les constructions concurrentes et ne pas fusionner les modalités. Cette fonctionnalité n'était cependant pas prévue pour le prototype I.

Conclusion

Nous avons synthétisé dans le Tableau 19 nos recommandations pour le prototype II ainsi que les améliorations qu'elles devraient apporter.

Modules gestuels et multimodal	
Recommandations pour le prototype II	**Améliorations attendues**
Résoudre les problèmes de détection des objets sélectionnables.	Disparition de 80% des erreurs du module d'Interprétation du Geste. Des modifications ont été rapidement apportées à ce module après les tests utilisateurs et des bêta-tests ont validé la disparition d'environ 75% des erreurs.
Fusionner les entrées gestuelles produites dans la même localisation spatiale et dans la même fenêtre temporelle.	Traitement correct des gestes à segments multiples : ne représentent que 10% des erreurs du module de Fusion mais perturbent fortement l'interaction (répétitions de la part d'Andersen).
Fournir un feedback à tous les gestes. Le module de Fusion devrait recevoir un signal quand un objet non sélectionnable a été sélectionné. Andersen pourrait répondre par exemple « Essaie d'autres objets ». Si l'objet fait partie d'une construction multimodale, Andersen pourrait aussi reprendre l'entrée verbale (ex : [« C'est quoi ça ? » + objet non sélectionnable] entraîne la réponse « De quoi parles-tu ? »).	Facteur fondamental d'utilisabilité.
Ajout d'un critère sémantique pour la fusion (fonctionnalité déjà prévue pour le prototype II).	Disparition de 100% des fusions non pertinentes.
Autres modules	
Recommandations pour le prototype II	**Améliorations attendues**
Permettre l'interruption d'Andersen (fonctionnalité déjà prévue pour le prototype II).	Amélioration de l'interaction.
Ajout d'objets sélectionnables (en se fondant sur la liste que nous avons établie).	Renforcement de l'interaction gestuelle, diminution de la frustration (probable) des utilisateurs.

Changement du mode de navigation.	Renforcement de la crédibilité de l'Agent (cohérence comportement verbal et non verbal) et de la notion de face-à-face.

Tableau 19 : Recommandations et améliorations attendues pour le prototype II.

5.3. Evaluation du prototype I avec Agent 3D, scénario orienté tâche

Nous allons maintenant rapporter les données issues des tests utilisateurs menés sur le prototype I du second univers de NICE : le monde des contes de fées – et son introduction qui se déroule dans le bureau d'Andersen en l'absence de celui-ci. Ce prototype comprenait deux modules du LIMSI : le module de Reconnaissance du Geste et le module d'Interprétation du Geste.

Figure 50 : Cloddy Hans dans le bureau d'Andersen devant la Machine à fabriquer des Contes.

Scénario d'interaction

Cloddy Hans (inspiré du personnage « Hans le Balourd » d'Andersen) a un excellent fond mais est un peu simplet. En l'absence d'Andersen, il doit veiller sur la Machine à fabriquer des Contes (Figure 50). Pour faire une surprise à Andersen, il veut créer un nouveau conte. Il demande à l'utilisateur de l'aider à se servir de la Machine, principalement en lui indiquant quels objets vont dans quels tuyaux.

L'utilisateur peut donner des instructions verbales et désigner des objets, mais pas manipuler directement ceux-ci.

Méthode

Les tests utilisateurs ont été réalisés par nos partenaires de Teliasonera[34] en Suède, qui nous ont ensuite transmis les fichiers de *log* de nos deux modules.

Utilisateurs. Dix utilisateurs ont réalisé le test en Suédois :

- Quatre garçons, d'âge moyen 14 ans (σ = 1,4 ans),
- Six filles, d'âge moyen 13 ans (σ = 1,6 ans).

Matériel. L'univers de Cloddy Hans était projeté sur un écran géant. Les utilisateurs étaient équipés d'un micro casque pour les entrées verbales, et le dispositif d'entrée gestuel était une souris gyroscopique (souris optique sans fil qui n'a pas besoin d'être posée sur un support, est manipulable en l'air à la manière d'un pointeur laser). Dans ce prototype, les utilisateurs n'avaient pas la possibilité de naviguer dans l'environnement : Cloddy Hans se déplaçait lui-même de manière autonome, sans que les utilisateurs aient le contrôle sur ce point.

Le système était partiellement simulé (plusieurs magiciens avaient la possibilité d'intercepter les erreurs du système et de les corriger en cours de traitement).

Données recueillies. Nos partenaires nous ont transmis 20 fichiers de *log* (2 par utilisateur).

Résultats

Comportement gestuel. Nous avons relevé dans ces fichiers 343 **entrées gestuelles**. Le nombre de gestes produits n'est pas influencé par le sexe des utilisateurs, de manière concordante avec ce que nous avons observé dans le scénario précédent.

L'analyse de la **forme des gestes** montre que la majorité d'entre eux (75%) sont reconnus comme des pointages. Les pourcentages respectifs d'encerclements et de lignes sont de 13% et 12%. La proportion de gestes de pointages est ainsi supérieure à celle observée dans le scénario conversationnel (44% ; $F(1/22) = 5,44$; $p = 0,029$). La Figure 51 permet de comparer les formes de gestes utilisées en fonction du dispositif (écran tactile, souris ou souris gyroscopique).

[34] http://www.teliasonera.com/

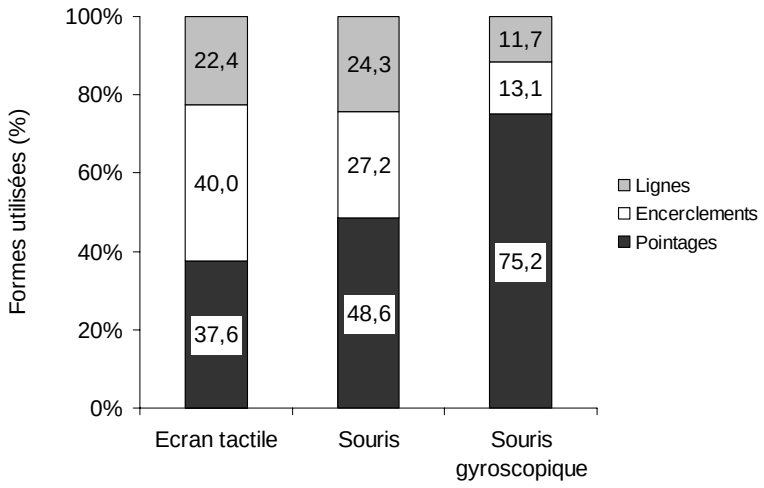

Figure 51 : Pourcentage d'utilisation des formes de gestes en fonction du média gestuel.

Comportement multimodal. Nos partenaires de Teliasonera nous ont fourni leurs résultats d'observation sur le comportement multimodal des utilisateurs dans ce scénario. L'utilisation des modalités s'est répartie ainsi :

- 85,2% d'entrées monomodales verbales,
- 2,1% d'entrées monomodales gestuelles,
- 12,6% de constructions multimodales.

Les **commandes verbales** étaient formées en moyenne de 3,6 mots. Environ 5,5% d'entre elles concernaient l'interaction sociale avec Cloddy Hans (« Comment vas-tu ? », etc.).

Parmi les **constructions multimodales**, environ 40% étaient complémentaires (dont une grande majorité contenaient un référent déictique, ex : « celui-là » + sélection d'un objet), 50% redondantes (ex : « prends le couteau » + sélection du couteau) et 10% conflictuelles (ex : « prends le couteau » + sélection de la hache). Soulignons qu'il s'agit bien de conflit et non de concurrence : les utilisateurs n'ont pas cherché ici à faire plusieurs choses en même temps, mais se sont juste trompés dans la dénomination ou dans la sélection de certains objets.

Dans les **commandes gestuelles**, nos partenaires ont relevé environ 60% de références aux objets présents dans l'environnement et 40% de gestes sans intention communicative évidente (gestes de type exploratoires).

Discussion

A première vue, l'**utilisation des modalités** dans ce corpus est très surprenante par rapport aux deux autres études présentées dans ce chapitre. En effet, le scénario étant ici orienté tâche, nous nous attendions à un taux d'utilisation du geste important. Or les commandes monomodales gestuelles n'ont représenté que 2,1% des entrées des utilisateurs. De plus, il apparaît qu'une proportion importante (40%) des commandes gestuelles a été utilisée pour explorer l'environnement et non pour réaliser le scénario.

Nous pensons que la faible utilisation du geste dans ce scénario résulte de la conjonction des facteurs suivants :

- La souris gyroscopique est un média inhabituel, peut-être difficile à utiliser.
- La tâche consistait à déplacer des objets depuis un meuble jusqu'à un des tuyaux de la machine à fabriquer des contes. La simple sélection d'un élément graphique (objet ou tuyau) n'était donc peut-être pas suffisante pour constituer une commande ou une action en soi.
- Les utilisateurs ne pouvaient pas eux-mêmes déplacer des objets, leurs actions constituaient toujours des instructions adressées à Cloddy Hans qui, seul, pouvait déplacer les objets.
- Les utilisateurs ne contrôlaient pas la navigation : ils ne pouvaient donc pas s'approcher d'un objet pour le sélectionner précisément avec la souris. Par ailleurs, lorsque Cloddy Hans était auprès de l'étagère aux objets, ils n'avaient pas non plus la possibilité de déplacer le champ pour aller pointer un tuyau. Ils devaient forcément transiter par une instruction verbale.

L'ensemble de ce scénario incitait donc les utilisateurs à interagir par la parole, ce qui explique la proportion importante de commandes monomodales verbales, et l'utilisation plus importante de la multimodalité que de la monomodalité gestuelle.

Compte tenu de ces particularités du scénario, il nous semble difficile de juger si l'utilisation de la **multimodalité** (12,6%) a été importante ou pas. En revanche, pour faire un parallèle avec les tests utilisateurs sur le système Andersen, nous pouvons souligner le fait que, même si le fonctionnement du prototype était imparfait, cela n'a pas forcément été perçu par les utilisateurs. En effet, le dispositif de test comprenait deux magiciens dont le rôle était de surveiller le fonctionnement des modules et, en cas d'erreurs, de corriger celles-ci à la volée de sorte que la réponse de l'Agent soit tout de même appropriée.

Nous pensons que les différences constatées dans les **formes de gestes** utilisées par rapport au prototype Andersen sont liées au dispositif gestuel fourni aux utilisateurs (souris gyroscopique *vs*. écran tactile et souris). Etant donné que la souris gyroscopique est un média inhabituel, et qu'elle devait ici être utilisée sur un écran géant à une certaine distance, il est probable que le pointage se soit avéré le moyen le plus facile de sélectionner avec précision des objets, en particuliers de petits objets.

Conclusion

Cette étude nous a apporté peu de données concrètes pour l'amélioration de nos modules dans l'univers de Cloddy Hans. En particulier, les seuls fichiers de *log* ne nous ont pas permis de savoir à quel point la Reconnaissance et l'Interprétation des Gestes avaient correctement fonctionné. En revanche, nous avons recueilli des indices intéressants relatifs à l'influence du scénario d'interaction et du média utilisé sur le comportement des utilisateurs.

5.4. Conclusion

Les études présentées dans ce chapitre nous ont permis d'apporter trois types de contributions à l'étude du comportement multimodal des utilisateurs face à des Agents Conversationnels : (1) une contribution méthodologique, de par la démarche et les outils que nous avons employés, (2) une contribution opérationnelle concrète, appliquée au développement du système NICE, et enfin (3) des résultats de portée plus large, susceptibles d'être exploités pour d'autres applications, se rapprochant plus ou moins du système NICE (Agents Conversationnels, interfaces multimodales et/ou applications ludiques destinées aux enfants et adolescents). Dans ce paragraphe, nous dressons un bilan de ces différentes contributions.

Apports méthodologiques

La démarche générale adoptée pour le développement informatique des modules confiés au LIMSI (reconnaissance et interprétation des gestes, fusion multimodale) a consisté à en fonder l'implémentation sur des données comportementales. L'objectif principal était d'optimiser ainsi le processus de conception en anticipant l'utilisation qui allait être faite de l'application. Cette démarche n'est pas nouvelle, mais elle mérite tout de même d'être soulignée, parce que de nombreux systèmes informatiques souffrent aujourd'hui encore d'un manque de prise en compte du point de vue, du comportement et de l'avis des utilisateurs.

Cette ligne générale de conduite du projet nous a amenés à employer des méthodes spécifiques, telles que la simulation en Magicien d'Oz et l'annotation comportementale fine (qui met elle-même en jeu des outils particuliers). Pour l'étude en Magicien d'Oz, nous avons entièrement annoté toutes les modalités à partir de rien. Pour l'analyse des premiers tests utilisateurs, nous avons pu considérablement alléger ce processus laborieux en important les fichiers de *log* du système comme des annotations afin de les confronter aux vidéos.

Dans l'étude en Magicien d'Oz, face au nombre important de variables recueillies, nous avons également eu le souci de chercher à résumer celles-ci afin d'en dégager des profils comportementaux. Ceux-ci nous ont apporté des informations nouvelles, de plus haut niveau d'abstraction que celles portées par chaque variable considérée individuellement. L'Analyse en Composantes Principales, que nous avons utilisée pour cela, n'est pas non plus une méthode nouvelle, mais c'est sa mise en œuvre dans ce champ de recherche qui est originale.

Contribution à la conception du système NICE

Dans la première phase du projet, nous avons mené un recueil de données avec un système simulé, des Agents développés en interne au LIMSI et un scénario approximant au mieux la future situation d'interaction, telle que nous la connaissions à ce moment-là. Voici ce que cette étude a concrètement apporté :

- Le corpus de gestes que nous avons recueilli (gestes produits avec un stylo) a permis d'entraîner le module de reconnaissance de formes (qui fonctionne sur la base d'un réseau de neurones),
- L'analyse des relations temporelles entre modalités dans les constructions multimodales a permis de déterminer les paramètres temporels de la fusion multimodale,
- L'analyse des relations sémantiques entre modalités a guidé l'élaboration des algorithmes de fusion multimodale.

Par la suite, les tests utilisateurs menés sur la première version du prototype ont permis :

- De fournir de nouveaux corpus de gestes, avec des médias variés (souris, écran tactile, souris gyroscopique), pour améliorer la reconnaissance de formes,
- De mettre en évidence des formes avec segments multiples (ex : cercle formé en plusieurs fois) qui nécessitent un traitement particulier,
- D'améliorer la communication des modules entre eux, dans le but d'augmenter l'utilisabilité : par exemple, le module de fusion est désormais averti de toutes les entrées gestuelles, même lorsque leur interprétation échoue, ce qui donne la possibilité de générer des feedbacks à tous

les gestes. De même, le module de fusion est averti dès qu'un geste ou une phrase débute (*startOfGesture* et *startOfSpeech* respectivement), en plus de recevoir les informations en fin de traitement. Ceci permet d'affiner le traitement de l'intégration temporelle des modalités.

Notre contribution ne s'est cependant pas limitée aux tests utilisateurs. L'implication de l'ergonomie tout au long du projet nous a en effet permis de contribuer au développement du système presque au jour le jour : nous avons pu intégrer la vision et l'intérêt de l'utilisateur en de nombreux points.

Nous pouvons par exemple citer le problème de la reconnaissance des gestes formés de segments multiples (notamment les gestes en forme de croix). Si on respecte strictement l'architecture du système, ce traitement incombe au module de reconnaissance de gestes. Cependant, le fait de mettre en relation des segments multiples implique obligatoirement un délai de traitement. Par exemple, pour pouvoir reconnaître une éventuelle croix, le module de reconnaissance de gestes aurait été obligé d'introduire un délai d'attente (d'environ 1 sec.) à chaque fois qu'une ligne était tracée – attente d'une deuxième ligne pour décider si la forme est une ligne ou une croix. Le problème est que ce délai, parce qu'il aurait été interne à la reconnaissance de gestes, aurait dû être additionné aux délais suivants (délais destinés à détecter les sélections multiples, par exemple des pointages successifs, et les constructions multimodales – ces délais combinés s'élevant à 3 sec.). Au total, la réponse du système après la sélection d'un objet par un geste linéaire n'aurait été déclenchée qu'au bout de 4 secondes (au lieu de 3 secondes pour tout autre forme de geste).

Afin d'optimiser les différents délais, nous avons proposé de transférer la reconnaissance des formes à segments multiples au module d'interprétation du geste. Ainsi, l'efficacité du point de vue de l'utilisateur s'est retrouvée privilégiée par rapport à la cohérence architecturale du système. Ce choix de conception répond également à un souci de pragmatisme : alors que la croix est une forme qui n'a jamais été observée dans les corpus (mais dont il est tout de même important d'implémenter le traitement), les lignes peuvent représenter jusqu'à 25% des corpus gestuels – il nous a donc semblé important de privilégier le traitement des lignes par rapport à celui des croix.

Ces différents résultats et recommandations ont contribué au développement du **prototype II du système NICE**, dont l'évaluation fait partie des perspectives de cette thèse.

Résultats d'observation

Les études citées nous ont également apporté de nombreuses données de plus haut niveau concernant l'interaction et la réalisation des scénarios du système NICE. Un des résultats les plus massifs de ces études a trait à l'importance de l'**interaction gestuelle** pour les enfants et les adolescents dans ce

contexte. Nous avons vu dans l'étude en Magicien d'Oz que la possibilité d'utiliser le geste a augmenté la prise d'initiative dans l'interaction et la facilité perçue d'utilisation du système. Les enfants ont beaucoup plus recouru au geste que les adultes, et l'utilisation de cette modalité a atteint des proportions surprenantes par rapport à ce que la littérature rapporte : 59,3% dans la première étude, 36,6% dans la seconde (qui utilisait pourtant un scénario conversationnel). Le troisième corpus, avec seulement 2,1% d'entrées monomodales gestuelles, constitue une exception, que nous proposons d'expliquer par les nombreuses contraintes posées sur les moyens d'interaction dans ce scénario (manipulation directe et navigation impossibles).

Nous pensons que l'importance de l'interaction gestuelle pour les enfants et adolescents est liée au domaine d'application auquel le projet NICE appartient : le jeu. L'interaction étant principalement gestuelle dans les dispositifs actuels de jeux vidéo, nous supposons que les utilisateurs du système NICE vont avoir tendance, si on leur en donne la possibilité, à transférer leurs habitudes d'interaction dans le système NICE.

Les fonctions du geste telles qu'elles sont apparues dans les corpus analysés sont sans doute également liées aux jeux vidéo courants : le geste permet d'agir directement (cf. corpus Magicien d'Oz), d'explorer l'environnement (dans les trois corpus), de mener des actions en parallèle de la conversation (Magicien d'Oz et prototype Andersen).

En ce qui concerne l'**interaction verbale** avec les Agents, la syntaxe utilisée était naturelle (c'est-à-dire proche de la communication Homme-Homme, par opposition à un langage de commande elliptique) dans les trois corpus. Les utilisateurs ont également considéré les Agents comme des partenaires sociaux (Magicien d'Oz et prototype Cloddy Hans).

Enfin, les données accumulées sur le **comportement multimodal** peuvent être synthétisées ainsi :

- Nous avons globalement observé autant de constructions simultanées que séquentielles (Magicien d'Oz et prototype Andersen),
- Un délai intermodal maximum de 2 secondes semble raisonnable pour traiter les constructions séquentielles (Magicien d'Oz et prototype Andersen),
- Les proportions de constructions redondantes et complémentaires sont approximativement équivalentes (les trois corpus),
- Une population d'utilisateurs jeunes peut avoir tendance à produire des constructions concurrentes, que le système devra pouvoir gérer (Magicien d'Oz principalement).

Le champ d'application de ces quelques données dépasse le projet NICE : nous pensons qu'elles peuvent bénéficier à la conception de systèmes plus ou moins proches de celui que nous avons étudié.

Les interfaces multimodales destinées aux enfants et adolescents, les systèmes munis d'Agents Conversationnels, ou tout simplement les applications ludo-éducatives sont des exemples de systèmes pour lesquels nos résultats constitueront peut-être des indications intéressantes.

Enfin, ces données nous ont ouvert des **perspectives** sur l'exploitation du comportement multimodal comme variable d'évaluation. Ces perspectives seront développées au chapitre 7.

6. Le comportement multimodal des Agents

Le comportement multimodal, tel que nous l'avons étudié chez nos utilisateurs dans le chapitre précédent, correspond à la combinaison d'éléments verbaux et de gestes opératoires. Rappelons que, du point de vue du locuteur, les gestes opératoires ont la caractéristique d'être volontairement intégrés comme éléments de communication (comme par exemple l'indication de taille qui accompagne la phrase « Grand comme ça »). Du point de vue de l'auditeur, les gestes opératoires peuvent donc avoir une importance cruciale pour la compréhension du message.

Certains Agents pédagogiques sont capables d'intégrer des gestes opératoires dans leur discours verbal : c'est le cas par exemple de Cosmo (Lester *et al.*, 1997b, 1999b), qui peut aller pointer de manière non ambiguë un serveur particulier dans son environnement virtuel et produire une phrase comme « Ce serveur est surchargé de connexions ». Steve (Rickel & Johnson, 1999), quant à lui, est capable de dire « Il faut ouvrir la valve n°3 » tout en pointant celle-ci, puis de réaliser l'action d'ouverture de la valve pour montrer comment faire.

Ces deux exemples semblent naturels et *a priori* efficaces dans un contexte pédagogique. Mais la relation sémantique qui lie les modalités n'est pas la même dans les deux cas. Dans la phrase de Cosmo, parole et geste coopèrent en effet par complémentarité (les deux modalités sont nécessaires pour accéder à la totalité du message), alors que dans celle de Steve, elles sont redondantes (la totalité du message est exprimée par chaque modalité). Comment tenir compte de cette dimension – la coopération sémantique – lorsqu'on spécifie le comportement multimodal d'un Agent pédagogique ?

Nous pouvons envisager deux grands types d'approche pour la spécification du comportement des Agents :

- Une approche *bottom-up*, consistant à analyser le comportement spontané d'enseignants humains dans des contextes similaires, puis à transférer les patterns observés aux Agents (démarche adoptée par Cassell *et al.*, 2000, dans le domaine de l'immobilier). Cette approche vise à assurer un comportement « naturel » aux Agents, c'est-à-dire proche du comportement humain.
- A l'inverse, une approche *top-down* pourrait chercher à améliorer au maximum la communication, en choisissant la ou les coopérations sémantiques les plus efficaces. S'il existe des raisons évidentes justifiant que les Agents imitent les humains, les Agents n'en sont pas moins des outils devant remplir leur rôle avec les meilleures performances possibles (Gratch *et al.*, 2004).

Cette seconde possibilité implique au préalable d'avoir évalué l'efficacité de chacune des coopérations sémantiques. Pour éclairer cette question, nous proposons dans le présent chapitre de contrôler le comportement multimodal des Agents afin d'observer les effets produits sur la transmission d'informations aux utilisateurs. Ceux-ci ont été placés pour cela dans une situation de type pédagogique, c'est-à-dire que leur objectif était de mémoriser les informations exposées par les Agents. L'expérimentation présentée ci-dessous a été menée en deux parties : une phase préliminaire et une phase répliquée.

6.1. Expérimentation préliminaire

L'objectif de cette première étude était d'obtenir une indication rapide sur l'existence, ou pas, d'un effet du comportement multimodal des Agents sur la qualité de la communication. Manquant de repères en la matière, nous souhaitions savoir si des séquences d'animation courtes permettaient de mettre en évidence un tel effet.

Nous avons tout d'abord déterminé les conditions de coopération sémantique à tester. Comme nous l'avons vu précédemment, les Agents pédagogiques utilisent – même si leurs auteurs n'explicitent pas ces stratégies – la redondance et la complémentarité. Nous savons par ailleurs que les modalités rentrent parfois en concurrence dans le discours des enseignants humains (gestes discordants observés par Goldin-Meadow *et al.*, 1999), ce qui perturbe l'apprentissage du côté des élèves. Notre but étant cependant d'optimiser la coopération sémantique entre modalités, nous avons préféré choisir comme condition contrôle non pas la concurrence, mais une situation où toutes les informations sont données par la parole et aucune par le geste. Ainsi les modalités ne coopèrent pas mais ne sont pas pour autant en contradiction : nous avons nommé cette condition « spécialisation par la parole » en référence à la

taxonomie de Martin *et al*. (2001). Ce type de communication peu illustrée par le geste correspond également au « style de discours élaboré » (Rimé, 1985), déjà évoqué au paragraphe 2.2.2.

Nous avons donc testé les trois conditions multimodales – Redondance entre parole et geste, Complémentarité, et Spécialisation par la parole – dans de courtes présentations techniques (explications sur le fonctionnement d'un objet) faites par des Agents Conversationnels. Par ailleurs, pour permettre un protocole intra-sujet (où chaque utilisateur voit les trois conditions), nous avons utilisé trois Agents différents : cela nous a permis en outre de tester l'impact de l'apparence des Agents. Les critères de qualité de la communication que nous avons fixés pour cette étude étaient : le rappel des informations présentées par les Agents, et les impressions subjectives des utilisateurs.

Méthode

Sujets. Dix-huit adultes recrutés au sein du LIMSI ont participé à cette expérience : 9 hommes (âge moyen 30 ans et 8 mois, σ = 8,3 ans) et 9 femmes (âge moyen 29 ans et 2 mois, σ = 8,8 ans).

Figure 52 : Léa présentant le logiciel.

Matériel. Pour cette expérience, nous avons utilisé les Agents 2D développés au LIMSI. Leur environnement était seulement composé d'un tableau blanc sur lequel était projetée l'image d'un objet (voir par exemple la Figure 52). Les Agents devaient expliquer brièvement le fonctionnement de cet

objet. Pour cela, ils pouvaient **parler**[35], faire **gestes iconiques** (mimer la taille ou la forme de certains éléments) ainsi que des **gestes déictiques** en direction de l'image (les Agents pouvaient se déplacer pour aller désigner un point précis de l'objet).

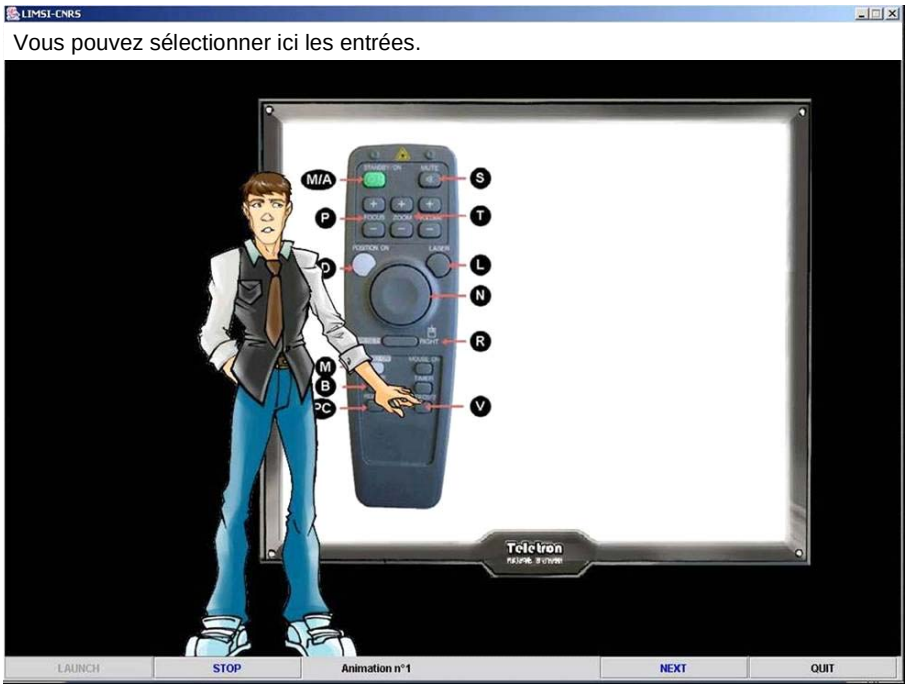

Figure 53 : Marco présentant la télécommande.

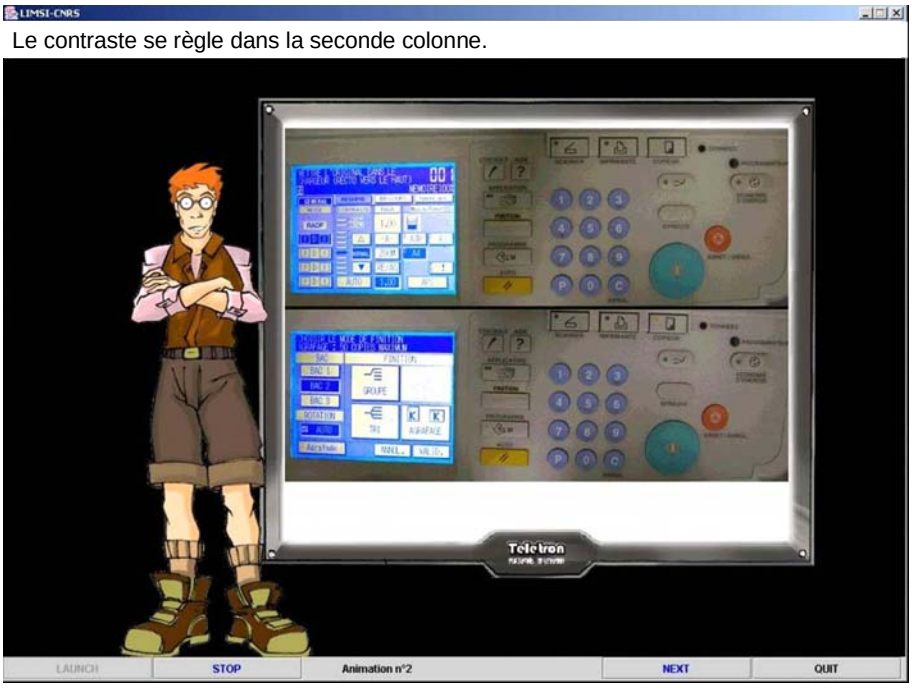

Figure 54 : Jules présentant le photocopieur.

[35] Synthèse vocale IBM ViaVoice.

Les animations étaient présentées sur un écran de 19 pouces (de résolution 1024 par 768 pixels). Deux haut-parleurs diffusaient la synthèse vocale. En plus d'être synthétisées, les explications verbales des Agents étaient retranscrites phrase par phrase sous la barre de titre de l'application. Les Agents n'étaient pas interactifs dans cette expérience, car nous devions contrôler précisément leur comportement.

Les **objets** présentés par les trois Agents étaient : un logiciel de traitement vidéo (Figure 52) une télécommande de vidéo projecteur (Figure 53), et l'interface d'un photocopieur (Figure 54). Les explications techniques consistaient en une description, en dix points, des fonctionnalités des boutons et des options des menus de chaque objet. Les présentations duraient entre 60 et 75 secondes.

Le comportement multimodal des agents a été spécifié dans un langage XML de bas niveau, et les trois stratégies de présentation ont été opérationnalisées de la manière suivante :

- **Redondance** (Figure 55) : le maximum d'informations est donné à la fois par la parole (position, forme, taille des éléments) et par des gestes des mains (gestes déictiques vers l'image ou gestes iconiques).
- **Complémentarité** (Figure 56) : par rapport aux explications redondantes, la moitié des informations sont données par la parole, et l'autre moitié par des gestes des mains (gestes déictiques vers l'image ou gestes iconiques). Par exemple, le locuteur peut désigner un objet par la parole et donner des informations (taille, forme, localisation…) par un geste des mains sans mentionner ces informations verbalement.
- **Spécialisation par la parole** (Figure 57) : toutes les informations sur l'objet sont données par la parole. Les gestes des mains se limitent à des gestes de battement.

Le nombre total de mouvements des mains a été maintenu constant dans les trois conditions : par exemple, les gestes déictiques du scénario redondant qui n'étaient pas repris dans le scénario complémentaire étaient remplacés par des gestes de battement.

Précisons aussi que les trois Agents avaient exactement les mêmes capacités verbales et non verbales (par exemple, tous trois pouvaient regarder sur leur gauche, faire des gestes de pointage avec la même posture, etc.) et que les scripts utilisés étaient les mêmes pour les trois Agents (ex : un script unique était utilisé pour la présentation redondante du photocopieur, quel que soit l'Agent).

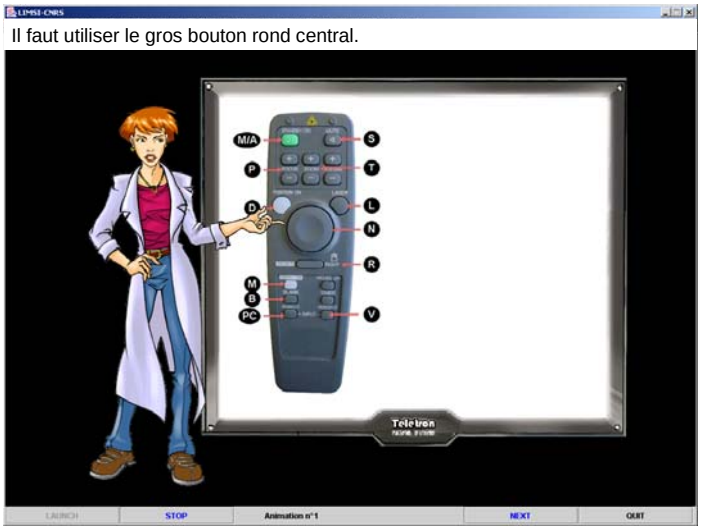

Figure 55 : Stratégie de redondance.

Figure 56 : Stratégie de complémentarité.

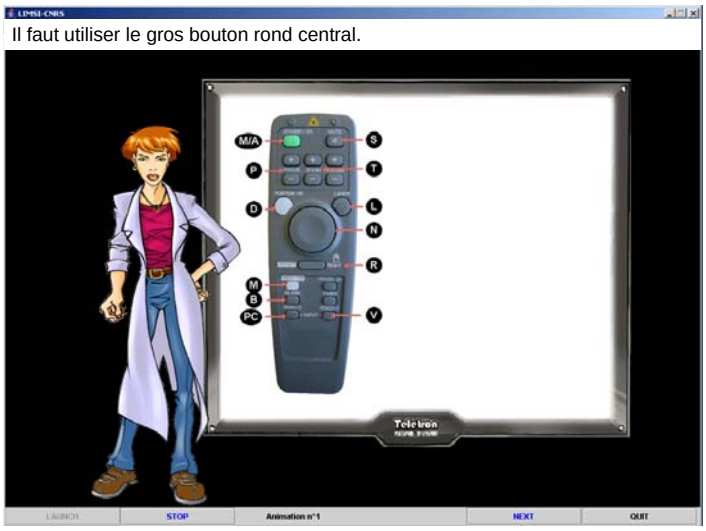

Figure 57 : Stratégie de spécialisation par la parole.

Procédure. Les combinaisons entre Stratégie multimodale, Agent et Objet ont été déterminées à l'aide d'un carré latin. Ce type de protocole permet de tester l'effet de chacune des trois variables en limitant le nombre de répétitions par rapport à un plan factoriel complet : chaque utilisateur a vu toutes les conditions en trois séquences alors qu'il en aurait fallu 27 dans un plan factoriel. Cependant, un inconvénient du carré latin est qu'il ne permet pas, avec le petit nombre de sujets de cette expérimentation préliminaire, de tester les interactions entre ces trois facteurs intra-sujets.

L'ordre de présentation des Stratégies, des Agents et des Objets a également été contrebalancé grâce au carré latin.

La passation de l'expérience débutait par l'explication des consignes : les utilisateurs étaient d'emblée prévenus qu'ils auraient à rappeler les informations présentées par les Agents. Ils visionnaient ensuite trois animations chacun. Puis l'expérimentatrice leur fournissait les trois images (télécommande, logiciel, photocopieur) utilisées dans les présentations : sur cette base, les utilisateurs effectuaient un rappel oral de ce qu'ils avaient retenu de ces trois objets. Etant donné que les explications étaient organisées en dix points, l'expérimentatrice notait la performance de rappel sur 10 à l'aide d'une grille, avec une précision n'allant pas en dessous du demi point.

Enfin, les utilisateurs remplissaient un questionnaire dans lequel ils devaient classer les trois Agents sur les critères suivants :

- Qualité de l'explication,
- Confiance inspirée par l'Agent,
- Sympathie de l'Agent,
- Personnalité de l'Agent : les trois Agents avaient-ils la même personnalité, lequel avait la personnalité la plus forte ?
- Expressivité de l'Agent.

Résultats

La performance de rappel et les évaluations subjectives ont été soumises à des analyses de variance avec le Sexe des utilisateurs en facteur inter-sujet. Pour chaque variable dépendante, l'analyse a été successivement réalisée avec la Stratégie multimodale, l'Agent, puis l'Objet en variable intra-sujet.

Toutes les analyses ont été effectuées avec SPSS.

Performance. La performance moyenne sur le rappel d'informations est de 6,45/10. Un effet principal du Sexe des utilisateurs sur la quantité d'informations rappelées apparaît en tendance ($F(1/16) = 4{,}174$; $p = 0{,}058$), montrant que les femmes ont rappelé un peu plus d'informations (7,1/10) que les hommes (5,8/10).

La stratégie multimodale des Agents n'a pas influencé le rappel, mais un effet principal de l'Agent est proche de la significativité ($F(2/32) = 3{,}215$; $p = 0{,}053$). La Figure 58 suggère que le rappel a été légèrement meilleur avec Marco, moyen avec Léa, et légèrement moins bon avec Jules.

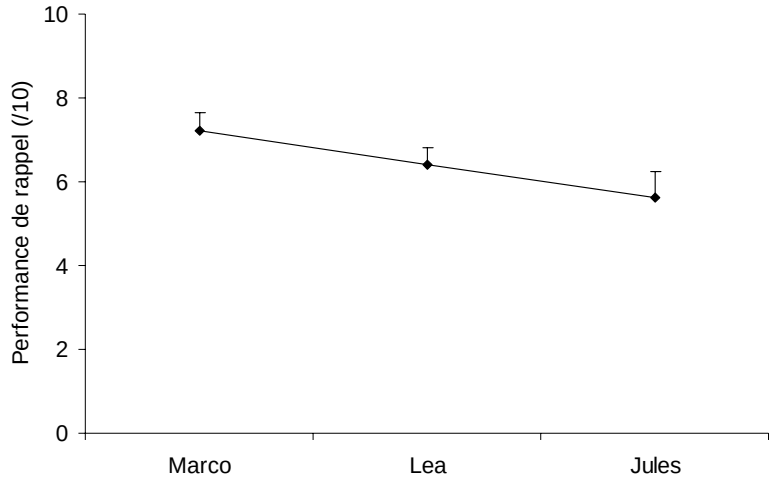

Figure 58 : Performance moyenne de rappel en fonction de l'Agent qui a présenté les informations.

Aucun effet principal de l'Objet n'apparaît, mais nous relevons une interaction entre Objet et Sexe des utilisateurs ($F(2/32) = 5{,}150$; $p = 0{,}012$). L'effet de l'Objet est significatif chez les femmes ($F(2/16) = 9{,}838$; $p = 0{,}002$) mais pas chez les hommes ($F(2/16) = 0{,}683$; N.S.). Les femmes ont produit un meilleur rappel sur le photocopieur que sur les deux autres objets. Cet effet, qui constitue un biais dans notre expérience, pourrait provenir d'une plus grande familiarité des femmes de notre échantillon vis-à-vis de cet objet, bien que nos deux groupes d'utilisateurs soient homogènes sur la catégorie socio-professionnelle.

Variables subjectives. L'analyse de la variable **Qualité de l'explication** indique un effet principal de la Stratégie multimodale ($F(2/32) = 5{,}469$; $p = 0{,}009$; Figure 59). Les Agents ayant utilisé une stratégie redondante et complémentaire ont obtenu des scores équivalents ($F(1/16) = 1{,}000$; N.S.) mais ont été jugés meilleurs que les Agents ayant une stratégie spécialisée (respectivement $F(1/16) = 13{,}474$; $p = 0{,}002$ et $F(1/16) = 4{,}102$; $p = 0{,}060$).

L'interaction entre Stratégie multimodale et Sexe des utilisateurs est également significative ($F(2/32) = 4{,}980$; $p = 0{,}013$; Figure 60). L'effet de la Stratégie est significatif chez les hommes ($F(2/16) = 19{,}000$; $p<0{,}001$) mais pas chez les femmes ($F(2/16) = 0{,}757$; N.S.). Les hommes ont évalué les Agents redondants comme meilleurs que les autres ($F(1/8) = 12{,}000$; $p = 0{,}009$ avec les Agents complémentaires et $F(1/8) = 100{,}000$; $p<0{,}001$ avec les Agents spécialisés). Les hommes semblent également avoir donné un score plus élevé aux Agents complémentaires qu'aux Agents spécialisés ($F(1/8) = 4{,}000$; $p = 0{,}081$).

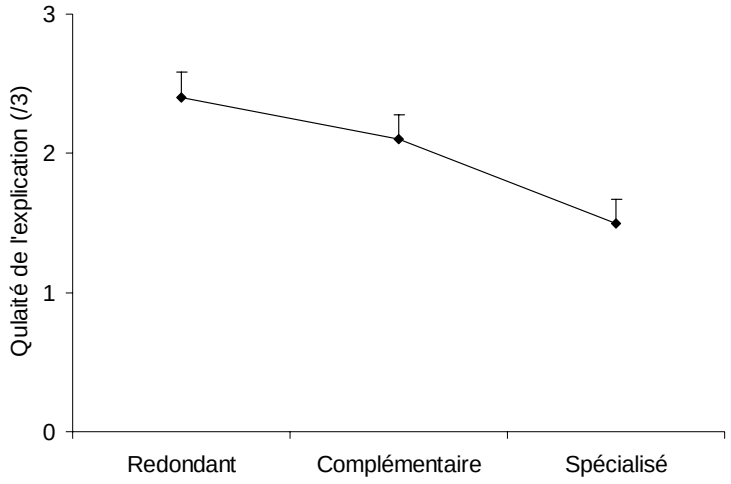

Figure 59 : Evaluation subjective de la qualité de l'explication en fonction de la stratégie multimodale des Agents.

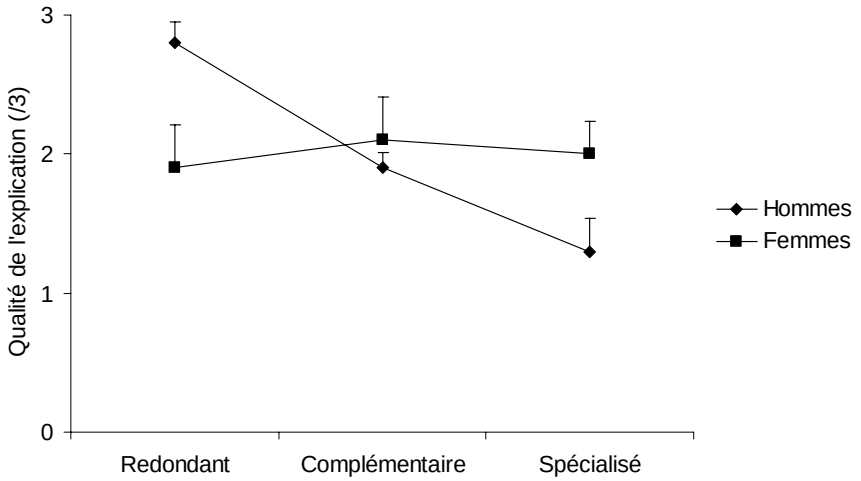

Figure 60 : Evaluation de la qualité de l'explication en fonction de la stratégie multimodale, chez les hommes et chez les femmes de notre échantillon.

L'Agent et l'Objet n'ont pas eu d'effet sur l'évaluation de la qualité de l'explication.

Les évaluations subjectives de **Confiance** accordée aux Agents ne révèlent aucun effet principal de la Stratégie multimodale. En revanche, l'interaction entre Stratégie et Sexe des utilisateurs est significative ($F(2/32) = 3,735$; $p = 0,035$). De manière similaire aux évaluations de qualité de l'explication, l'effet de la Stratégie tend à être significatif chez les hommes ($F(2/16) = 2,868$; $p = 0,086$) alors qu'il ne l'est pas chez les femmes ($F(2/16) = 2,500$; N.S.). Ces deux variables dépendantes (Confiance et Qualité de l'explication) se révèlent en effet corrélées positivement (coefficient de corrélation linéaire entre 0,630 et 0,757 ; $p<0,005$ pour les trois stratégies).

L'Agent et l'Objet n'ont pas influencé les scores de confiance.

L'analyse de la **Sympathie** ne met en évidence aucun effet de la Stratégie multimodale. En revanche, un effet principal de l'Agent ($F(2/32) = 3,328$; $p = 0,049$; Figure 61) montre que Marco et Léa ont obtenu des scores équivalents ($F(1/16) = 0,471$; N.S.), alors que Jules a été moins apprécié ($F(1/16) = 6,479$; $p = 0,022$ avec Marco et $F(1/16) = 3,390$; $p = 0,084$ avec Léa). Le sens de cet effet semble rejoindre celui de la performance de rappel, mais ces deux variables ne sont en réalité pas corrélées.

Le Sexe des utilisateurs n'a pas influencé l'effet précédent. De plus, si les scores de Marco et Jules sont moyennés, aucune interaction n'apparaît entre le sexe des utilisateurs et le sexe des Agents.

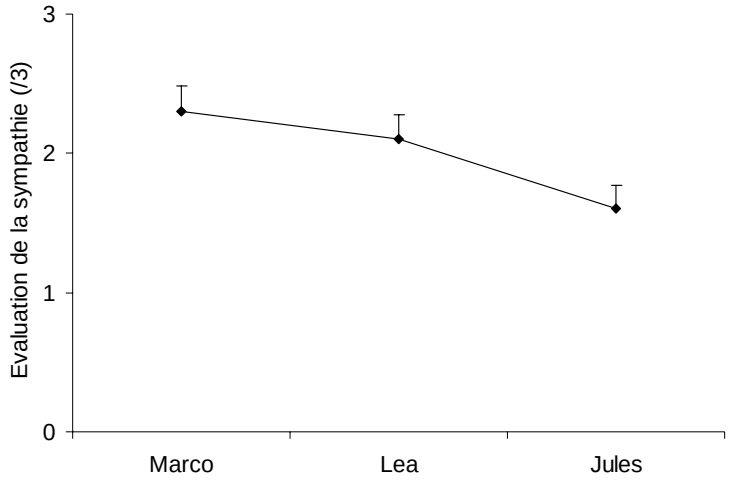

Figure 61 : Evaluation de la sympathie des trois Agents.

Enfin, les variables de **Personnalité** et d'**Expressivité** n'ont révélé aucun effet des facteurs Stratégie, Agent ou Objet.

Discussion

Effet de la Stratégie multimodale. L'objectif premier de cette expérience était de savoir s'il était possible de mettre en évidence un effet des stratégies multimodales (Redondance, Complémentarité, Spécialisation par la parole) dans de courtes présentations faites par des Agents Conversationnels. Le fait que les stratégies aient influencé l'évaluation de la qualité de l'explication et de la confiance est pour nous un résultat très encourageant.

Avant de discuter plus précisément les résultats, il nous semble important de souligner que les stratégies n'ont pas été perçues consciemment par les utilisateurs. Dix utilisateurs sur 18 (5 hommes et 5 femmes) ont remarqué des différences dans la manière dont les Agents présentaient les informations : ils ont noté que certains allaient pointer l'image. Mais aucune remarque formulée par les utilisateurs ne laisse supposer qu'ils ont perçu une différence entre Redondance et Complémentarité. Ces observations sont cohérentes avec le modèle « figure – fond » de Rimé (Rimé & Schiaratura, 1991) qui explique que le comportement non verbal du locuteur reste généralement à la périphérie du champ attentionnel de l'auditeur.

L'effet principal de la Stratégie multimodale sur la **Qualité de l'explication** est significatif, mais l'interaction avec le Sexe des utilisateurs suggère qu'il est principalement dû aux évaluations des utilisateurs masculins. Ceux-ci ont en effet montré une préférence claire – bien qu'inconsciente – pour les Agents redondants. A l'inverse, les évaluations des femmes ne font ressortir aucune préférence entre les stratégies.

Cet **effet du Sexe** des utilisateurs était inattendu et demande à être confirmé avec un échantillon plus important. Il soulève néanmoins des hypothèses intéressantes sur la manière dont hommes et femmes réagissent face aux Agents. Il est peu probable que l'absence d'effet de la stratégie chez les femmes soit due au fait qu'elles aient porté une attention moindre aux Agents : en effet, les femmes ont remarqué les différences de comportement (pointages, déplacements) dans les mêmes proportions que les hommes. Elles ont également fait tout autant de commentaires sur les Agents (sur leur apparence, sur les détails de leur tenue vestimentaire) que les hommes. La littérature sur le décodage et la reconnaissance des comportements non verbaux ne permet pas davantage d'expliquer nos résultats, puisqu'il est généralement rapporté que les femmes ont une meilleure capacité à décoder le non verbal que les hommes (Feldman *et al.*, 1991). Enfin, il semble qu'aucune différence homme / femme n'ait été décrite dans le domaine de la reconnaissance du mouvement biologique (voir Giese & Poggio, 2003).

En revanche, nous pouvons tenter de rapprocher notre effet de certains travaux classiques en Psychologie cognitive, qui font état d'une tendance chez les hommes à privilégier les traitements visuo-spatiaux, alors que les femmes utilisent davantage les traitements auditivo-verbaux (voir Kimura, 1999). Par exemple, une des différences les plus stables concerne les tests dits de rotation mentale, dans lesquels il faut imaginer ce que deviennent des figures si on en modifie l'orientation : dans ce type de tests, les hommes ont systématiquement de meilleures performances que les femmes, ce qui suggère qu'ils ont plus de facilité à former et à manipuler des représentations mentales visuelles. Réciproquement, les femmes ont de meilleures performances dans les tests dits d'aisance verbale, où il faut par exemple énumérer en un temps limité le maximum de mots commençant par une lettre donnée ; elles ont également plus de facilités à apprendre une liste de mots lus à voix haute, ce qui pourrait s'expliquer par un meilleur accès aux étiquettes verbales et à leurs représentations phonémiques[36].

Pour en revenir à notre expérience, nous pouvons supposer que, pour juger de la qualité de l'explication, les hommes ont donné de l'importance – même inconsciemment – aux informations visuelles et spatiales apportées par les gestes des mains. A l'inverse, nous pouvons supposer que les femmes se sont basées uniquement sur le contenu verbal pour juger de la qualité de l'explication, ce en quoi elles n'ont pas distingué les trois stratégies multimodales.

Cette différence entre hommes et femmes demande à être répliquée et approfondie pour pouvoir éventuellement être reliée à un modèle de fonctionnement cognitif. Pour cela, il serait intéressant de pouvoir mettre en évidence la même interaction entre Stratégie multimodale et Sexe sur la **Performance de rappel**. Dans la présente expérience, les stratégies multimodales n'ont eu aucun effet sur la quantité d'informations rappelées. Le nombre moyen d'informations rappelées (6,45/10) permet d'écarter l'hypothèse d'un effet plafond : les présentations étaient certes courtes et simples, mais les utilisateurs ont eu un niveau de performance moyen quelle que soit la stratégie. En revanche, il est possible que l'absence d'effet de la stratégie sur la performance provienne d'un manque de précision de notre mesure. Rappelons que les utilisateurs faisaient un rappel oral – qui n'était pas enregistré – et que l'expérimentatrice cochait sur une grille les informations rappelées. Cette grille contenait l'ensemble du discours des Agents, divisé en 10 informations-clés, et la précision de notation n'allait pas en dessous du demi point. Il est possible qu'avec une mesure plus fine nous aurions pu mettre en

[36] L'origine des différences cognitives entre hommes et femmes est généralement attribuée à la répartition des rôles dans nos sociétés ancestrales (Kimura, 1999) : les aptitudes spatiales étaient plus nécessaires aux hommes qui, pour aller chasser, devaient naviguer sur de longues distance et retrouver ensuite leur chemin, et qui devaient également fabriquer des outils. Les aptitudes verbales étaient plus nécessaires aux femmes, chargées de l'éducation des enfants.

évidence une variation des performances en fonction des stratégies multimodales. Nous adopterons donc par la suite une mesure de performance plus discriminante.

Enfin, les évaluations subjectives de **Confiance** ont produit le même effet d'interaction entre Stratégie multimodale et Sexe des utilisateurs que pour la variable Qualité de l'explication. Il s'est avéré que les évaluations de confiance et de qualité de l'explication étaient corrélées. Nous voyons une ambiguïté dans ce résultat. En effet, le concept de confiance recouvre plusieurs acceptions, et la question posée aux utilisateurs était très large et générale (« En quel personnage avez-vous le plus confiance ? »). Il est possible que les utilisateurs aient interprété la question dans le contexte restreint des présentations faites par les Agents, et l'aient compris comme « Quel Agent a donné les explications les plus fiables ? ». Or notre but avec cette question était de dépasser le cas particulier des présentations et d'étudier de manière plus générale la capacité des Agents à instaurer un climat de confiance, notamment exploitable dans les applications de type e-commerce. Pour mesurer cette variable de manière plus fiable, nous utiliserons donc des questions plus ciblées, plus précises, du type, par exemple, « Donneriez-vous votre numéro de carte bleue à cet Agent ? »

Effet de l'apparence des Agents. L'apparence des Agents n'a influencé ni l'évaluation de la qualité de l'explication, ni les scores de confiance. Mais elle a eu un impact sur la sympathie ressentie par les utilisateurs : Marco et Léa ont en effet été jugés plus sympathiques que Jules, les évaluations n'étant liées ni au sexe des utilisateurs, ni à celui des Agents.

Figure 62 : Marco, Léa et Jules, souriants.

Comment expliquer les préférences des utilisateurs envers Marco et Léa ? La Figure 62 montre nos trois Agents avec leur expression faciale la plus souriante. Il se trouve que le sourire de Marco a été dessiné plus large que le sourire des deux autres, et les commentaires des utilisateurs en fin d'expérience ont montré qu'ils avaient apprécié ce trait de Marco. Les commentaires sur Léa ont été plus mitigés, à cause de sa blouse blanche : pour certains utilisateurs, cela la rendait plus jolie et plus sérieuse, mais d'autres utilisateurs l'ont trouvée trop stricte. De précédentes recherches (McBreen *et al.*, 2001) ont déjà souligné l'importance des vêtements portés par les Agents sur la perception des utilisateurs. Enfin, il semble que ce soient les lunettes de Jules qui aient influencé le jugement des utilisateurs : le fait que ses yeux soient peu visibles à travers ses lunettes a été négativement perçu. De

plus, sa position de repos consistait à croiser les bras, et plusieurs utilisateurs ont trouvé cela déplaisant. Nous pourrons modifier cette dernière caractéristique pour l'expérience suivante (les autres traits des Agents tels que le style vestimentaire ou le port de lunettes ne sont malheureusement pas aisément modifiables chez nos Agents 2D).

L'apparence des Agents semble également avoir influencé la performance de rappel. Comme nous l'avons déjà signalé, nous espérons obtenir par la suite des données plus précises sur le rappel. Cependant, ce n'est pas la première fois que ce type d'effet apparaît dans une évaluation d'Agents Conversationnels. En effet, Moreno *et al.* (2002), en testant neuf Agents pédagogiques, ont trouvé que leur efficacité variait avec leur apparence, mais sans pouvoir relier cet effet avec aucune des variables subjectives relevées par ailleurs (sympathie, compréhensibilité, crédibilité, qualité de la présentation, synchronisation de l'animation). Nous espérons apporter un éclairage sur cette question lors de notre seconde étude.

Autres résultats. Nos variables de **Personnalité** et d'**Expressivité** n'ont apporté aucun effet. Cela est surprenant parce que nous attendions un effet du comportement gestuel sur la perception de la personnalité. Une fois de plus, il se peut que ce soit la formulation de la question (« Quel personnage a la personnalité la plus forte ? ») qui soit responsable de l'absence d'effet. Pour l'expérience suivante, nous laisserons une question ouverte pour que chacun définisse soi-même la personnalité des trois Agents. Il serait également intéressant de relier la personnalité des utilisateurs avec leurs jugements de la personnalité de l'Agent. Dans cette expérience préliminaire, il semblait illusoire de rechercher un tel lien, eu égard à notre faible nombre d'utilisateurs. Mais avec un échantillon plus important, nous espérons mettre en évidence des effets intéressants en relevant une ou deux dimensions de personnalité chez les utilisateurs.

Enfin, nous supposons que le biais que nous avons relevé concernant l'**Objet** de la présentation disparaîtra si nous recrutons de nouveaux utilisateurs hors du LIMSI. En effet, le biais concernait l'interface du photocopieur, or il s'agissait précisément photocopieur du LIMSI, qui appartenait donc à l'environnement quotidien des utilisateurs interrogés dans cette première expérience. Nous espérons donc que nos trois objets apparaîtront équivalents dans l'expérience suivante.

Conclusion

Notre cahier des charges pour la seconde expérience peut être résumé par les points suivants :

- Nous pouvons réutiliser notre matériel expérimental, puisqu'il nous a déjà permis de mettre en évidence de nombreux effets.

- Nous devons augmenter l’effectif de l’échantillon testé, en maintenant l’équilibre entre utilisateurs masculins et féminins pour étudier l’effet du sexe.
- Il est nécessaire de mettre au point une mesure fine pour évaluer la mémorisation des informations présentées.
- Nous allons inclure des questions plus ciblées pour mesurer la confiance inspirée par les Agents.
- Nous allons laisser une question ouverte pour que chaque utilisateur définisse la personnalité des Agents avec ses propres mots.
- Nous souhaitons relever une ou deux dimensions de personnalité chez les utilisateurs et mettre celles-ci en relation avec les résultats sur les autres variables.
- Il faut modifier la position par défaut de Jules (diminuer la fréquence de la position bras croisés).

6.2. Expérimentation répliquée

Dans cette expérience, nous avons réutilisé les présentations techniques du test précédent (avec les objets : télécommande, logiciel, photocopieur) pour opérationnaliser à nouveau les variables :

- Stratégie multimodale : Redondance, Complémentarité, Spécialisation.
- Apparence de l'Agent : Marco, Léa, Jules.

Pour évaluer la performance mnésique, il nous a semblé intéressant de distinguer les aspects visuo-spatiaux et auditivo-verbaux de la mémorisation, dans le but de pouvoir éventuellement rapprocher nos résultats des travaux antérieurs sur les différences cognitives Hommes / Femmes. Nous avons donc décidé de demander deux types de rappel aux utilisateurs : un rappel graphique (les utilisateurs doivent dessiner les objets de mémoire), et un rappel verbal (sur les fonctionnalités présentées pour chaque objet). Pour garder une trace de ces rappels et permettre ainsi une évaluation précise de la performance, nous avons demandé aux utilisateurs de les faire sur papier.

Quant à l'évaluation de la personnalité des utilisateurs, nous avons pensé que la dimension d'Introversion / Extraversion pouvait être intéressante à relever. En effet, d'après la définition d' Eysenck & Eysenck (1971) :

> *« [Les extravertis] ont tendance à être expansifs, impulsifs et non inhibés, à avoir de nombreux contacts sociaux et à prendre part souvent à des activités de groupe. (…) L'extraverti typique (…) préfère rester en mouvement et agir, a tendance à être agressif et à perdre son sang froid rapidement. (…) »*

A l'inverse,

> *« L'introverti typique est le genre d'individu tranquille, effacé, introspectif, plus amateur des livres que des gens. Il a tendance à prévoir, ne s'engage pas à la légère et se méfie des impulsions du moment. (…) Il contrôle étroitement ses sentiments, se conduit rarement d'une manière agressive et ne s'emporte pas facilement. »*

La dimension d'Introversion / Extraversion est ainsi en partie liée au comportement non verbal des individus. Nous avons donc pensé qu'elle pouvait entrer en jeu dans les évaluations subjectives demandées aux utilisateurs dans notre expérience : par exemple, nous pouvons supposer que les individus extravertis vont avoir tendance à préférer les Agents mobiles et faisant des gestes en direction de l'image (Agent redondants et complémentaires). Nous avons donc décidé de relever, dans la présente expérience, le score de chacun de nos utilisateurs sur le continuum Introversion / Extraversion.

Méthode

Sujets. Cette expérience a été réalisée avec 108 participants[37], recrutés à l'Institut de Psychologie de l'Université de Paris 5 :

- 54 hommes : âge moyen 26 ans 8 mois (σ = 9,2 ans).
- 54 femmes : âge moyen 23 ans 1 mois (σ = 5,6 ans).

Matériel. Les passations se sont déroulées à l'Institut de Psychologie. Les animations ont été présentées sur un écran d'ordinateur 17 pouces en résolution 1024*768, avec un haut-parleur pour la synthèse vocale.

Procédure. Le plan expérimental (détaillé en Annexe 4) est un carré gréco-latin à mesures répétées (voir Myers, 1979 ; Keppel, 1991). La taille de notre échantillon va nous permettre cette fois de tester les interactions entre facteurs intra-sujets, notamment l'interaction entre Stratégie et Apparence de l'Agent.

Les consignes stipulaient que trois personnages allaient successivement apparaître pour donner de courtes explications techniques sur différents objets : il était juste demandé aux utilisateurs d'écouter attentivement les présentations, étant précisé qu'à la fin, ils auraient à restituer les informations présentées.

Chaque utilisateur visionnait les trois animations. Suite à cela, le recueil des données se faisait en quatre étapes :

- C'est tout d'abord le **rappel graphique** qui était demandé : l'utilisateur devait dessiner de mémoire les trois objets présentés.
- Ensuite, l'utilisateur devait effectuer le **rappel écrit indicé** : l'expérimentateur fournissait les images utilisées dans les présentations, et l'utilisateur devait à partir de celles-ci donner par écrit le maximum d'informations dont il se souvenait.
- Après ces deux formes de rappel, l'utilisateur remplissait un questionnaire d'**évaluations subjectives** (voir l'Annexe 5), qui contenait les items suivants :
    - « Quel personnage donne les meilleures explications ? » (Classer les 3 personnages)
    - « Quel personnage est le plus sympathique ? » (Classer)
    - « Quel personnage est le plus expressif ? » (Classer)

[37] Les passations ont été assurées par Marianne Najm (Najm, 2004) et Fabien Bajeot (Bajeot, 2004).

- « Définissez en un mot la personnalité de chacun. » (Question ouverte pour chaque personnage)
- « Si votre ordinateur tombait en panne, suivriez vous les instructions de ces personnages pour le dépanner ? » (Oui ou Non pour chaque personnage)
- « Donneriez-vous votre numéro de carte bleue à ces personnages ? » (Oui ou Non pour chaque personnage)
- Les utilisateurs étaient en outre invités à expliciter leur(s) critère(s) de jugement pour chacune des évaluations subjectives, et à laisser des remarques libres. Nous leur demandions aussi s'ils avaient remarqué des différences dans la manière dont les trois Agents donnaient leurs explications.

- Les **données individuelles de personnalité** étaient relevées en toute fin de séance. Chaque utilisateur remplissait l'EPI (Eysenck Personality Inventory), qui mesure de manière standard l'Introversion / Extraversion.

Analyse des données. Dans ce paragraphe, nous décrivons la manière dont nous avons dépouillé et analysé les données de cette expérience :

- Pour quantifier le **rappel graphique**, nous avons défini une grille listant, pour chaque objet, les éléments picturaux globaux et locaux le constituant. Nous avons abouti à une grille en 13 points pour les trois objets. Nous avons ensuite noté la présence (et non le positionnement correct) de ces éléments dans les dessins des utilisateurs. La précision de notation pouvait aller jusqu'au quart de point pour chaque élément de la grille. La performance globale sur chaque objet a été enfin exprimée en pourcentage d'éléments reproduits.
- Le **rappel écrit** a été noté de la même manière : les informations données par les Agents sur les fonctionnalités des objets ont été listées au sein d'une grille de dépouillement. Nous sommes arrivés à un total de 32 informations pertinentes pour la télécommande, et 33 pour le logiciel et le photocopieur. Nous avons ensuite dénombré ces éléments dans les feuilles de rappel écrites par les utilisateurs. La précision de notation pouvait aller jusqu'au demi point. La performance finale était exprimée en pourcentage d'informations rappelées.
- Nous avons relevé dans le questionnaire les classements sur la **qualité de l'explication**, la **sympathie** et l'**expressivité**. Ces classements ont été transformés en scores (classement à la première place correspond à un score de 3, classement en seconde place à un score de 2, classement en troisième place à un score de 1).
- Les qualificatifs recueillis pour évaluer la **personnalité des Agents** ont été catégorisés comme positifs, négatifs, ou neutres, en fonction de ce qui est généralement considéré comme une qualité ou un défaut.
- Pour les deux questions de **confiance**, nous avons relevé la fréquence des réponses Oui.

- Enfin, les **données qualitatives** concernant les critères d'évaluation subjective et les remarques des utilisateurs ont été catégorisées de manière *ad hoc*.

Les résultats de l'**EPI** ont été relevés à l'aide de la grille de dépouillement standard de ce questionnaire. L'EPI mesure trois dimensions : celle d'Introversion / Extraversion, de Stabilité / Névrosisme, et une échelle de Mensonge visant à détecter les tentatives de falsification des réponses de la part du sujet. Les individus avec une note supérieure ou égale à 5 sur l'échelle de mensonge ont été éliminés de l'échantillon (pour l'étude de l'influence de la personnalité seulement) : cela représente 27 individus (12 femmes, 15 hommes). L'étude de la personnalité a donc été réalisée avec 81 de nos utilisateurs.

Le Névrosisme désigne l'hyperréaction émotionnelle et la prédisposition à la dépression : ce trait de personnalité ne nous a pas semblé pertinent pour la présente expérience, et nous ne l'avons donc pas analysé. Nous avons relevé les scores d'Introversion / Extraversion chez nos 81 utilisateurs et avons divisé cet échantillon en deux sous-groupes : la moyenne d'extraversion chez les étudiants français étant de 11,2 (Eysenck & Eysenck, 1971), les introvertis regroupaient les utilisateurs ayant une note inférieure à 11,2 et les extravertis ceux ayant une note supérieure à 11,2. Finalement, la variable Personnalité se répartit ainsi dans notre échantillon :

- Chez les femmes : 18 extraverties, 24 introverties.
- Chez les hommes : 20 extravertis, 19 introvertis.

Nous avons réalisé les calculs suivants :

- Les cinq variables numériques (rappel graphique, rappel écrit, score de qualité de l'explication, de sympathie et d'expressivité) ont été soumises à des analyses de variance (ANOVA) sous SPSS avec le Sexe des utilisateurs en variables inter-sujet et successivement la Stratégie, l'Agent et l'Objet en variable intra-sujet. Les comparaisons par paires ont été réalisées à l'aide de t de Student.
- Ces analyses ont été réitérées sur l'échantillon restreint de 81 utilisateurs en ajoutant la Personnalité (Introverti *vs*. Extraverti) en variable inter-sujet.
- En raison du plan que nous avons utilisé (le carré gréco-latin), les interactions deux à deux entre Stratégie, Agent et Objet ont été calculées en inter-sujet.
- Nous avons également examiné les corrélations entre ces cinq variables dépendantes numériques (rappel graphique, rappel écrit, score de qualité de l'explication, de sympathie et d'expressivité).

- La distribution des qualificatifs de personnalité positifs, négatifs et neutres en fonction de la Stratégie et de l'Agent a été étudiée par un calcul de Khi². L'influence de la Personnalité des utilisateurs a été testée par la même méthode.
- Pour les deux questions de confiance, le lien entre la fréquence de la réponse Oui et la Stratégie, l'Agent ou l'Objet a également été étudié à l'aide d'un Khi².
- Enfin, les catégories de critères d'évaluation ont été examinées descriptivement.

Résultats

Effets de la Stratégie multimodale sur les variables numériques. La stratégie multimodale (redondance, complémentarité, spécialisation) a influencé significativement le **rappel écrit** ($F(2/212) = 12,04$; $p<0,001$; Figure 63) : la redondance a en effet suscité un rappel meilleur que la complémentarité (48,7% contre 41,2% ; $t(107) = 4,29$; $p<0,001$) et que la spécialisation (40,7% ; $t(107) = 4,38$; $p<0,001$). La différence entre complémentarité et spécialisation n'est pas significative.

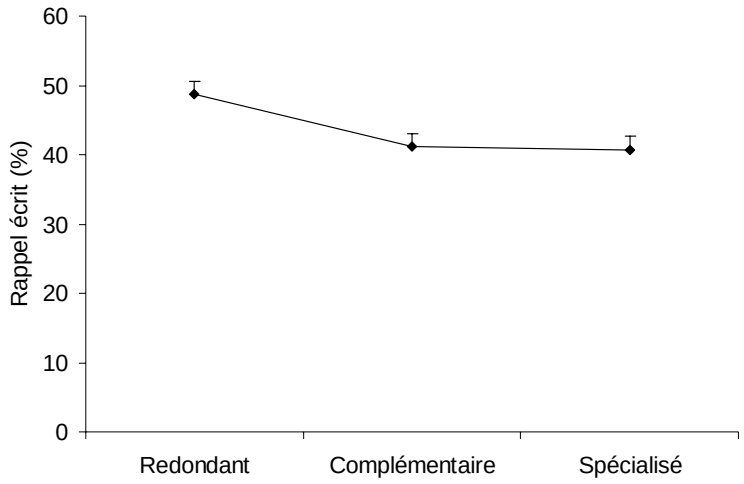

Figure 63 : Performance de rappel écrit en fonction de la stratégie multimodale des Agents (redondante, complémentaire, spécialisée).

De plus, une interaction est apparue entre le Sexe des utilisateurs, leur Personnalité et la Stratégie des Agents ($F(2/154) = 3,12$; $p = 0,047$) :

- Chez les extravertis (Figure 64, partie droite), l'interaction entre Sexe et Stratégie n'est pas significative ($F(2/72) = 0,70$; NS) : pour tout le groupe d'extravertis, la redondance est plus efficace que la complémentarité ($t(37) = 2,09$; $p = 0,044$), et la complémentarité ne se différencie pas de la spécialisation ($t(37) = 0,03$; NS).

- Chez les introvertis, l'interaction entre Sexe et Stratégie est significative ($F(2/82) = 4,16$; $p = 0,019$; Figure 64, partie gauche) :
    - L'effet de la Stratégie est en effet significatif chez les hommes introvertis ($F(2/36) = 13,46$; $p<0,001$) : la redondance est meilleure que la complémentarité ($t(18) = 2,07$; $p = 0,053$), qui est elle-même meilleure que la spécialisation ($t(18) = 4,11$; $p = 0,001$).
    - En revanche, l'effet de la Stratégie n'est pas significatif chez les femmes introverties ($F(2/46) = 2,12$; NS).

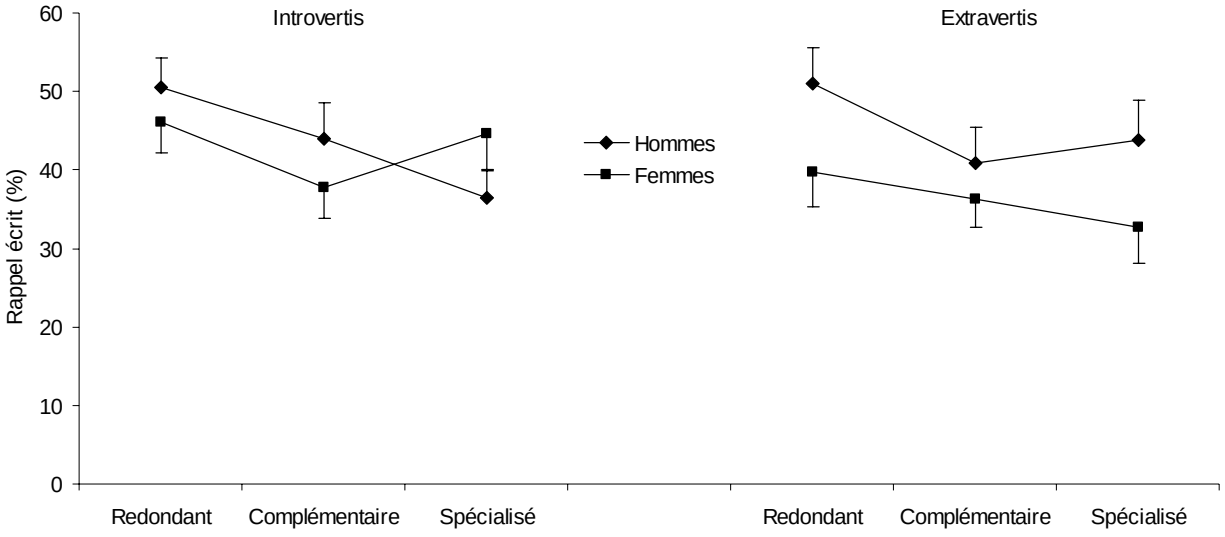

Figure 64 : Performance de rappel écrit en fonction de la stratégie des Agents (redondante, complémentaire, spécialisée), de la personnalité des utilisateurs (introvertis *vs*. extravertis) et de leur sexe (hommes *vs*. femmes).

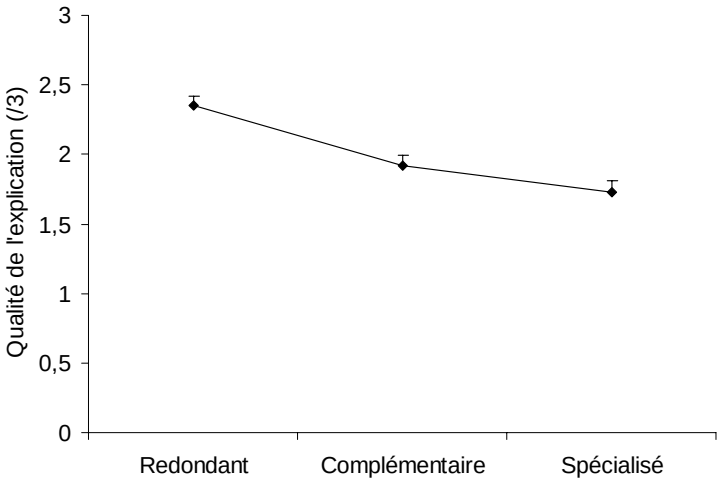

Figure 65 : Evaluation de la qualité de l'explication en fonction de la stratégie multimodale des Agents (redondante, complémentaire, spécialisée).

L'effet principal de la Stratégie est également significatif sur l'évaluation subjective de la **qualité de l'explication** ($F(2/212) = 12,01$; $p<0,001$; Figure 65). A nouveau c'est la stratégie redondante qui s'avère meilleure que les deux autres ($t(107) = 3,60$; $p<0,001$ et $t(107) = 4,85$; $p<0,001$ avec la complémentarité et la spécialisation respectivement ; pas de différence entre complémentarité et spécialisation).

Le même type d'interaction que précédemment apparaît entre la Stratégie des Agents, le Sexe des utilisateurs et leur Personnalité ($F(2/154) = 3,38$; $p = 0,037$; Figure 66) :

- L'interaction entre Sexe et Stratégie n'est pas significative chez les extravertis ($F(2/72) = 1,89$; NS) : quel que soit le sexe, la redondance est meilleure que la complémentarité ($t(37) = 2,45$; $p = 0,019$), et la complémentarité ne se différencie pas de la spécialisation ($t(37) = 0$; NS).
- Chez les introvertis, l'interaction entre Sexe et Stratégie est significative ($F(2/82) = 3,73$; $p = 0,028$). L'effet de la Stratégie est significatif chez les femmes introverties ($F(2/46) = 3,66$; $p = 0,034$) ainsi que chez les hommes introvertis ($F(2/36) = 6,07$; $p = 0,005$).
    - Chez les femmes introverties, la redondance est supérieure aux deux autres stratégies ; la complémentarité et la spécialisation ne diffèrent pas significativement.
    - Au contraire chez les hommes introvertis, la complémentarité est au même niveau que la redondance et se distingue significativement de la spécialisation.

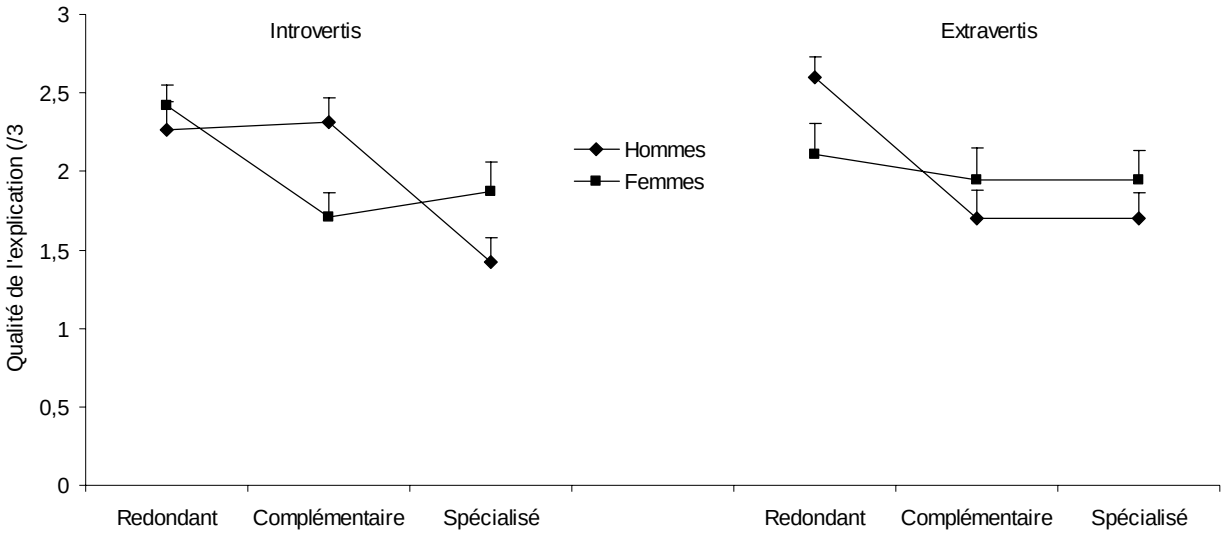

Figure 66 : Evaluation de la qualité de l'explication en fonction de la stratégie des Agents (redondante, complémentaire, spécialisée), de la personnalité des utilisateurs (introvertis, extravertis) et de leur sexe (hommes, femmes).

La Stratégie a également influencé la **sympathie** ressentie par les utilisateurs à l'égard des Agents ($F(2/212) = 6,34$; $p = 0,002$; Figure 67) : la redondance rend les Agents plus sympathiques que la complémentarité ($t(107) = 3,57$; $p = 0,001$) et que la spécialisation ($t(107) = 2,5$; $p = 0,014$). La différence entre complémentarité et spécialisation n'est pas significative.

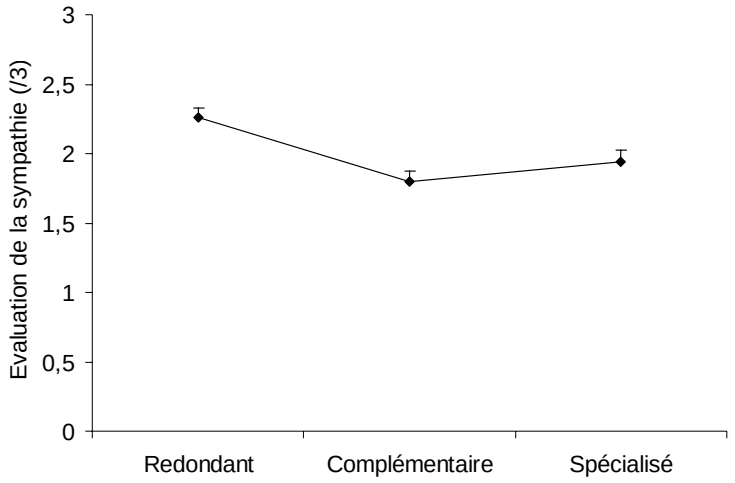

Figure 67 : Evaluation de la sympathie des Agents en fonction de leur stratégie multimodale (redondante, complémentaire, spécialisée).

Enfin, l'effet principal de la Stratégie sur l'évaluation de l'**expressivité** est significatif ($F(2/212) = 6,49$; $p = 0,002$; Figure 68) : les Agents redondants sont jugés plus expressifs que les complémentaires ($t(107) = 1,97$; $p = 0,052$) et les spécialisés ($t(107) = 3,67$; $p<0,001$). Agents complémentaires et spécialisés ne sont pas significativement différents.

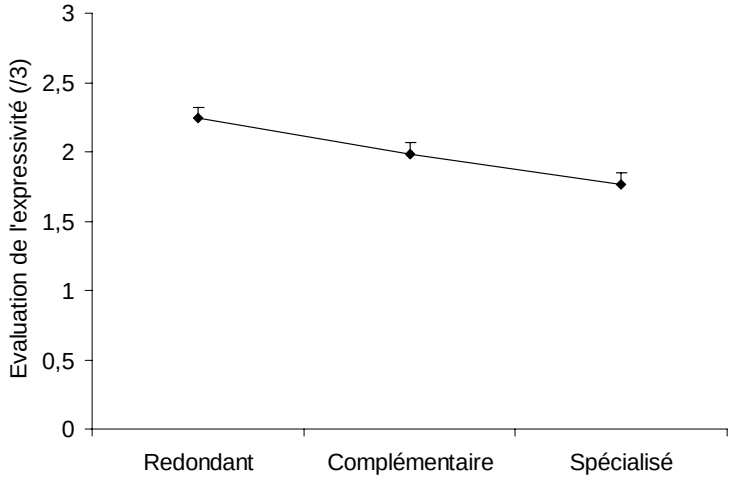

Figure 68 : Evaluation de l'expressivité des Agents en fonction de leur stratégie multimodale (redondante, complémentaire, spécialisée).

L'interaction entre Stratégie, Personnalité et Sexe apparaît une nouvelle fois ($F(2/154) = 3,19$; $p = 0,044$; Figure 69) :

- De manière similaire aux cas précédents, l'interaction entre Stratégie et Sexe n'est pas significative chez les extravertis ($F(2/72) = 0,96$; NS). Pour l'ensemble des extravertis, la redondance est meilleure que la complémentarité ($t(37) = 1,93$; $p = 0,062$), et la complémentarité ne diffère pas de la spécialisation ($t(37) = 0,12$; NS).
- L'interaction entre Stratégie et Sexe tend vers la significativité chez les introvertis ($F(2/82) = 2,48$; $p = 0,09$).
    - L'effet de la Stratégie est significatif chez les hommes introvertis ($F(2/36) = 4,89$; $p = 0,013$) : les comparaisons par paires révèlent que redondance et complémentarité ne diffèrent pas significativement mais sont toutes deux supérieures à la spécialisation ($t(18) = 1,76$; $p = 0,096$ entre redondance et spécialisation ; $t(18) = 3,39$; $p = 0,003$ entre complémentarité et spécialisation).
    - Il n'y a pas d'effet de la Stratégie chez les femmes introverties ($F(2/46) = 1,38$; NS).

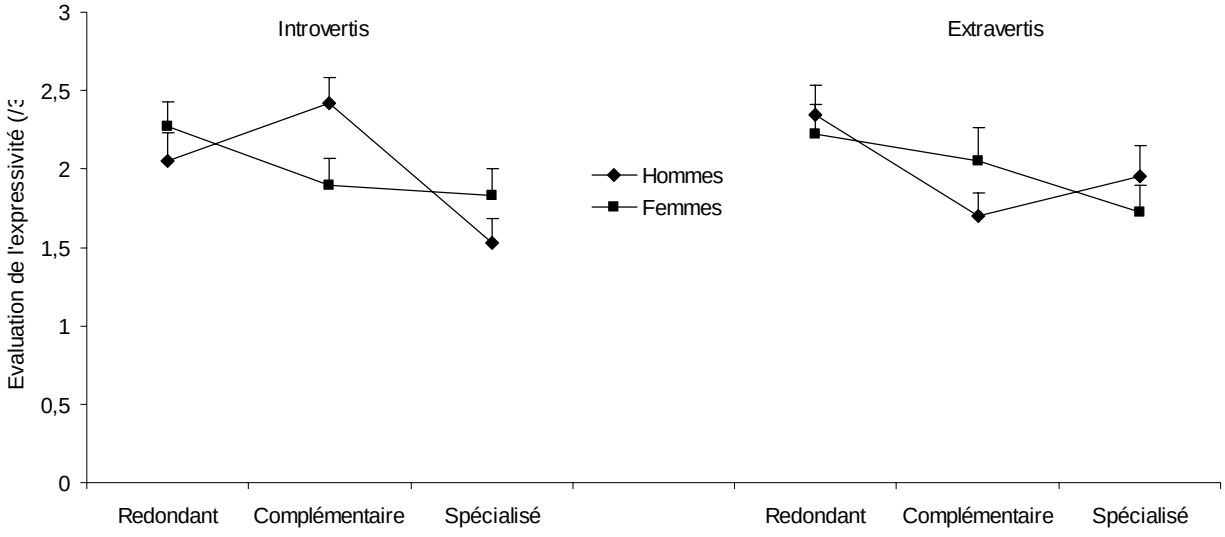

Figure 69 : Evaluation de l'expressivité des Agents en fonction de leur stratégie multimodale (redondante, complémentaire, spécialisée), de la personnalité des utilisateurs (introvertis, extravertis) et de leur sexe (hommes, femmes).

Effet de l'Agent sur les variables numériques. Le seul effet de l'apparence de l'Agent qui soit ressorti de notre analyse concerne l'évaluation subjective de la sympathie ($F(2/212) = 3,17$; $p = 0,044$; Figure 70). Les comparaisons par paires montrent que Marco a été jugé plus sympathique que Léa ($t(107) = 2,29$; $p = 0,024$) et que Jules ($t(107) = 2,15$; $p = 0,034$). Les scores de sympathie de Léa et Jules ne sont pas significativement différents.

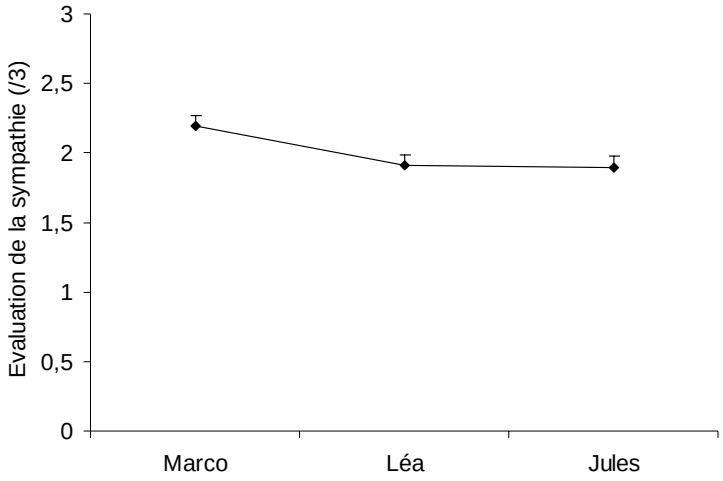

Figure 70 : Evaluation de la sympathie en fonction de l'apparence de l'Agent (Marco, Léa, Jules).

Effets de l'Objet sur les variables numériques. L'Objet est la seule variable à avoir influencé le **rappel graphique** ($F(2/212) = 42,13$; $p<0,001$). Les comparaisons par paires montrent que la télécommande a été mieux rappelée que le logiciel ($t(107) = 6,89$; $p<0,001$) et que le photocopieur ($t(107) = 9,21$; $p<0,001$). La différence entre logiciel et photocopieur n'est pas significative.

Nous avons observé le même profil de résultat sur le **rappel écrit** ($F(2/212) = 4,04$; $p = 0,019$) : la télécommande a été mieux rappelée que les deux autres objets ($t(107) = 1,90$; $p = 0,06$ et $t(107) = 3,06$; $p = 0,003$ avec le logiciel et le photocopieur respectivement).

L'effet principal de l'Objet est également significatif sur l'évaluation de la **qualité de l'explication** ($F(2/212) = 11,39$; $p<0,001$) : les explications sur la télécommande ont été jugées meilleures que celles des autres objets ($t(107) = 3,80$; $p<0,001$ et $t(107) = 4,40$; $p<0,001$ avec le logiciel et le photocopieur respectivement). De plus, nous observons une **interaction entre Objet et Sexe** ($F(2/212) = 3,45$; $p = 0,034$; Figure 71) : chez les femmes, l'explication sur la télécommande a été jugée meilleure que les deux autres ($t(53) = 4,86$; $p<0,001$ avec le logiciel ; $t(53) = 3,36$; $p = 0,001$ avec le photocopieur). Logiciel et photocopieur ne se distinguent pas. Chez les hommes, la qualité de l'explication sur le logiciel n'est pas significativement différente de celle de la télécommande, et ces deux explications sont meilleures que celle du photocopieur ($t(53) = 2,85$; $p = 0,006$ entre télécommande et photocopieur ; $t(53) = 1,96$; $p = 0,055$ entre logiciel et photocopieur).

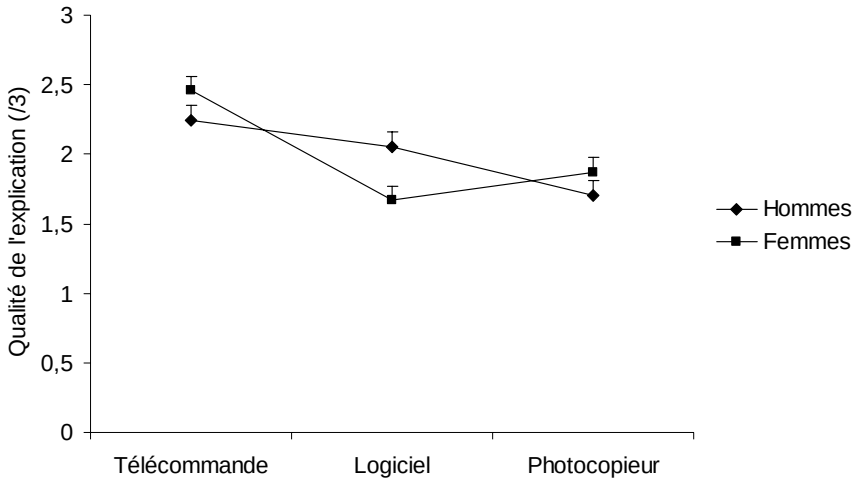

Figure 71 : Evaluation de la qualité de l'explication en fonction de l'objet (télécommande, logiciel, photocopieur) et du sexe des utilisateurs (hommes, femmes).

Enfin, il s'avère que l'objet a influencé la **sympathie** ressentie par les utilisateurs ($F(2/212) = 4,08$; $p = 0,018$) : l'Agent présentant le photocopieur a été jugé moins sympathique que ceux qui présentaient la télécommande ($t(107) = 2,84$; $p = 0,005$) ou le logiciel ($t(107) = 2,04$; $p = 0,044$). La différence n'est pas significative entre les Agents ayant présenté la télécommande et le logiciel.

Interactions entre Stratégie, Agent et Objet sur les variables numériques.

Comme annoncé précédemment, nous avons étudié les interactions deux à deux entre ces facteurs en les considérant comme des variables inter-sujet (manipulation nécessaire compte tenu de notre plan en carré gréco-latin). Nous décrivons dans ce paragraphe les résultats de cette analyse.

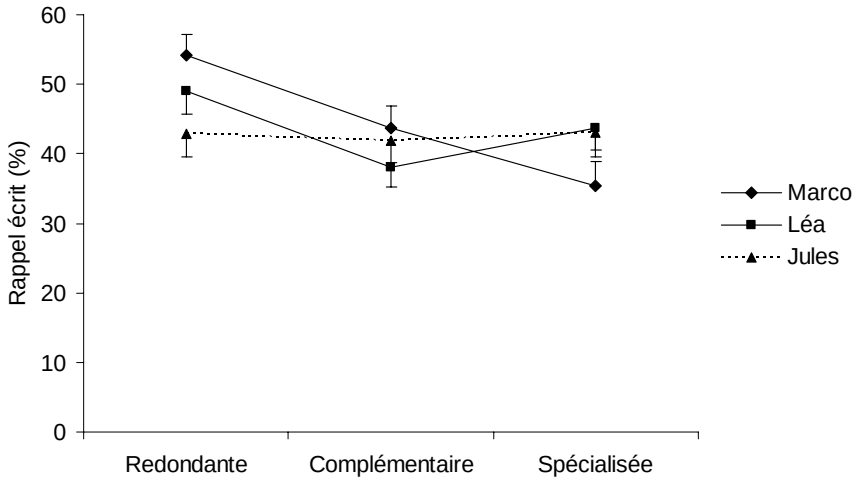

Figure 72 : Performance de rappel écrit en fonction de l'Agent ayant fait la présentation (Marco, Léa, Jules) et de sa stratégie multimodale (redondante, complémentaire, spécialisée).

Nous avons relevé une interaction entre **Agent et Stratégie** sur le **rappel écrit** (F(4/297) = 2,37 ; p = 0,053 ; Figure 72). L'effet de la Stratégie est significatif avec Marco (F(2/105) = 8,34 ; p<0,001) ainsi qu'avec Léa (F(2/105) = 2,96 ; p = 0,056), mais non significatif avec Jules (F(2/105) = 0,03 ; NS). Pour Marco, la redondance est plus efficace que la complémentarité (t(70) = 2,40 ; p = 0,019), qui tend elle-même à être plus efficace que la spécialisation (t(70) = 1,72 ; p = 0,090). Pour Léa, la redondance est plus efficace que la complémentarité (t(70) = 2,47 ; p = 0,016), mais la spécialisation se situe entre les deux (différences non significatives avec la redondance et la complémentarité).

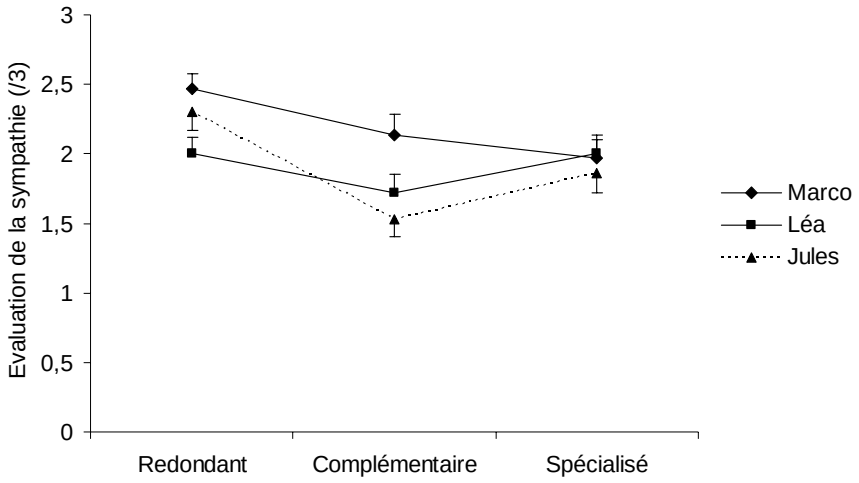

Figure 73 : Evaluation de la sympathie des Agents en fonction de leur apparence (Marco, Léa, Jules) et de leur stratégie multimodale (redondante, complémentaire, spécialisée).

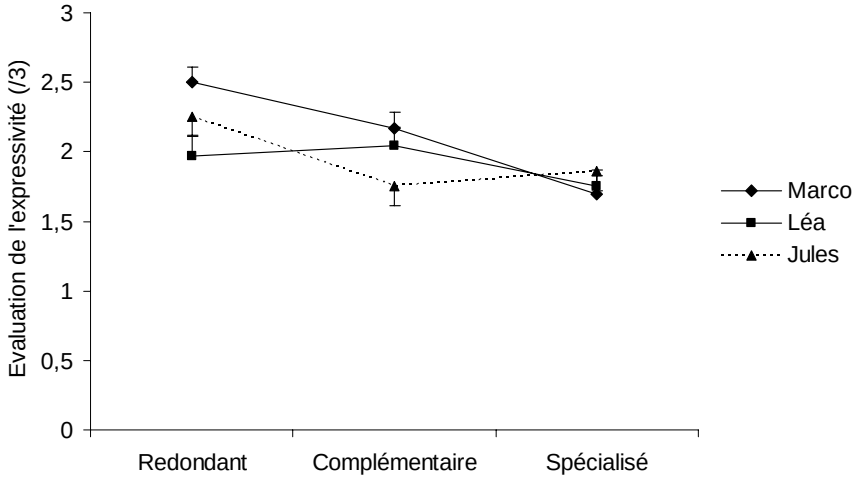

Figure 74 : Evaluation de l'expressivité des Agents en fonction de leur apparence (Marco, Léa, Jules) et de leur stratégie multimodale (redondante, complémentaire, spécialisée).

Une autre interaction entre **Agent et Stratégie** apparaît sur l'évaluation de la **sympathie** ($F(4/297) = 2,37$; $p = 0,053$; Figure 73). L'effet de la Stratégie est significatif chez Marco ($F(2/105) = 4,07$; $p = 0,020$) ainsi que chez Jules ($F(2/105) = 8,60$; $p<0,001$), mais pas chez Léa ($F(2/105) = 1,54$; NS). Pour Marco, la redondance semble être plus avantageuse que les deux autres stratégies ($t(70) = 1,89$; $p = 0,063$ avec la complémentarité). Complémentarité et spécialisation ne se différencient pas significativement. Pour Jules, la redondance est plus avantageuse que la spécialisation ($t(70) = 2,28$; $p = 0,026$), qui elle-même est meilleure que la complémentarité ($t(70) = 1,80$; $p = 0,076$).

L'interaction **Agent × Stratégie** est également significative sur l'évaluation de l'**expressivité** ($F(4/292) = 2,55$; $p = 0,039$; Figure 74). L'effet de la Stratégie est significatif pour Marco ($F(2/105) = 11,13$; $p<0,001$) ainsi que pour Jules ($F(2/105) = 3,53$; $p = 0,033$), mais pas pour Léa ($F(2/105) = 1,73$; NS). La redondance rend Marco plus expressif que la complémentarité ($t(70) = 2,03$; $p = 0,046$), qui elle-même le rend plus expressif que la spécialisation ($t(70) = 2,63$; $p = 0,011$). Quant à Jules, la redondance le rend plus expressif que les deux autres stratégies ($t(70) = 1,97$; $p = 0,053$ avec la spécialisation), sans que celles-ci se distinguent entre elles.

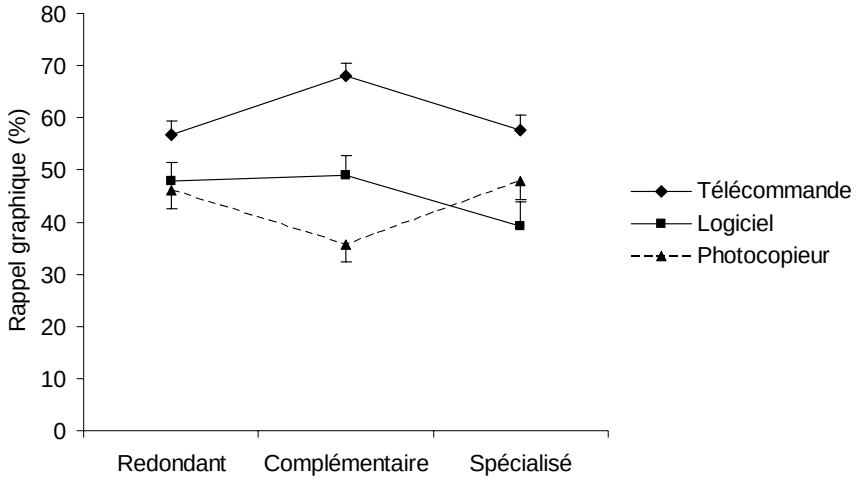

Figure 75 : Performance de rappel graphique en fonction de la stratégie des Agents (redondante, complémentaire, spécialisée) et de l'objet présenté (télécommande, logiciel, photocopieur).

Enfin, nous avons relevé une interaction entre **Stratégie et Objet** sur le **rappel graphique** ($F(4/297) = 4,68$; $p = 0,001$; Figure 75). L'effet de la Stratégie est significatif pour la télécommande ($F(2/105) = 5,74$; $p = 0,004$) ainsi que pour le photocopieur ($F(2/105) = 3,75$; $p = 0,027$), mais pas pour le logiciel ($F(2/105) = 1,84$; NS). Pour la télécommande, la meilleure stratégie est la complémentarité ($t(70) = 3,11$; $p = 0,004$ avec la redondance ; $t(70) = 2,86$; $p = 0,006$ avec la spécialisation). Pour le photocopieur, la complémentarité est au contraire la moins bonne stratégie

($t(70) = 2,21$; $p = 0,030$ entre redondance et complémentarité). Dans les deux cas, les stratégies redondante et spécialisée sont équivalentes.

Corrélations entre variables numériques. Le Tableau 20 présente les coefficients de corrélation linéaire entre nos cinq variables dépendantes numériques. Le coefficient de corrélation le plus fort lie le rappel graphique et le rappel écrit. Les autres coefficients sont relativement faibles, mais suggèrent tout de même qu'il existe un lien entre les trois variables subjectives (qualité de l'explication, sympathie, expressivité).

	Rappel graphique	Rappel écrit	Qualité de l'explication	Sympathie
Rappel graphique				
Rappel écrit	**0,581****			
Qualité de l'explication	0,197**	0,205**		
Sympathie	0,051	0,118*	**0,405****	
Expressivité	0,021	0,100	**0,361****	**0,395****

Tableau 20 : Coefficients de corrélation linéaire entre les cinq variables dépendantes numériques deux à deux. Les caractères gras indiquent les corrélations les plus fortes. Le signe ** correspond à celles qui sont significatives au seuil 0,01. Le signe * indique celles qui sont significatives au seuil 0,05.

Perception de la personnalité des Agents. Pour étudier cette variable qualitative, nous avons regroupé en des catégories séparées tous les qualificatifs positifs (ex : sympa, compétent, sérieux, ouvert, enthousiaste, intelligent, décontracté, amusant…) et négatifs (ex : froid, inexpressif, sévère, détaché...). Les qualificatifs restants ont été classifiés comme neutres, soit parce qu'ils étaient explicitement neutres, soit parce qu'ils désignaient des traits qui ne sont pas aisément identifiables comme des qualités ou des défauts (ex : classique, technique, discret, etc.). Globalement, la catégorie positive représente 56,8% des qualificatifs, la catégorie négative 29,3% et la catégorie neutre 8,3% (5,6% des cases du questionnaire n'ont pas été renseignées).

Les pourcentages de ces catégories en fonction de l'Agent (Marco, Léa, ou Jules) sont représentés dans la Figure 76. Marco semble bénéficier d'évaluations plus positives, mais le calcul du Khi² montre que la distribution des catégories de qualificatifs n'est pas significativement différente pour les trois Agents ($\chi^2(6) = 5,52$; NS).

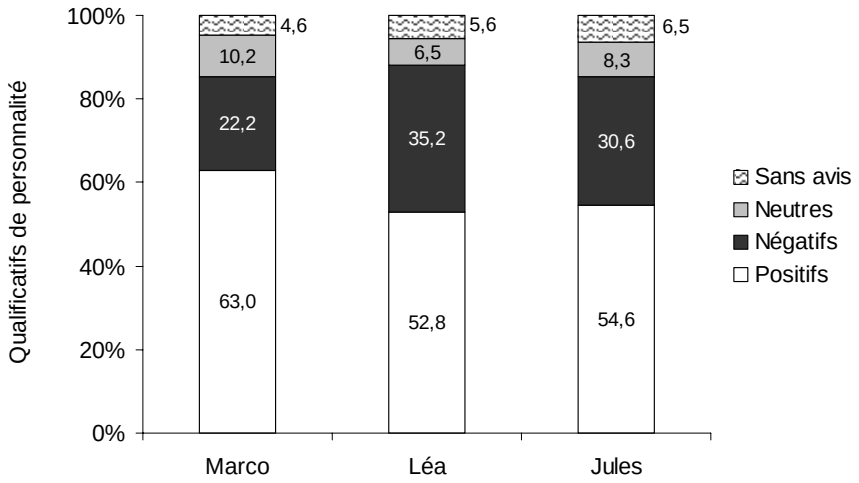

Figure 76 : Pourcentages de qualificatifs de personnalité (positifs, négatifs, neutres, sans avis) pour les trois Agents (Marco, Léa, Jules).

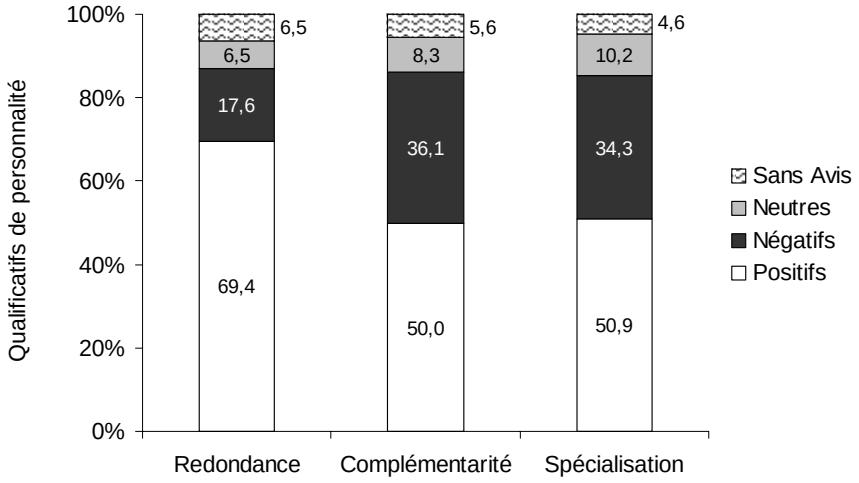

Figure 77 : Pourcentages de qualificatifs de personnalité (positifs, négatifs, neutres, sans avis) en fonction de la Stratégie adoptée par les Agents (redondante, complémentarité, spécialisation).

En revanche, la distribution des qualificatifs est significativement influencée par la Stratégie multimodale des Agents ($\chi^2(6) = 13,46$; $p = 0,036$). La Figure 77 montre que c'est à nouveau la redondance qui a été favorisée.

Pour chacune des Stratégies, nous avons comparé la distribution des qualificatifs faite par les utilisateurs introvertis et extravertis : nous n'avons pas trouvé de différence significative dans l'évaluation de la personnalité des Agents en fonction de la personnalité des utilisateurs.

Confiance vis-à-vis des Agents. Deux questions nous ont permis d'évaluer la confiance vis-à-vis des Agents. Dans la première, c'est la confiance dans une situation d'assistance qui était testée. Nous demandions aux utilisateurs s'ils suivraient les conseils des Agents s'ils avaient besoin de réparer leur ordinateur : globalement, les utilisateurs ont répondu Oui à 63,9% à cette question. Dans la seconde question, nous visions plutôt la confiance dans un contexte d'e-commerce. Nous demandions aux utilisateurs s'ils donneraient aux Agents leur numéro de carte bleue : cette fois le pourcentage de Oui a été seulement de 23,8%.

Il est apparu que ni la Stratégie multimodale, ni l'Agent, ni l'Objet n'avaient une influence significative sur la fréquence de réponses Oui. Nous verrons dans le paragraphe suivant quels sont les facteurs qui ont influencé le choix des utilisateurs.

Critères de jugement. Dans ce paragraphe, nous rapportons l'analyse des commentaires subjectifs des utilisateurs lorsqu'ils remplissaient le questionnaire : pour chacune des questions, nous leur avons demandé de préciser sur quel critère il fondaient leur réponse. Par exemple, lorsque les utilisateurs ordonnaient les trois Agents sur la qualité de leurs explications, nous leur avons demandé d'expliquer comment ils évaluaient les explications des Agents. Les critères ont été analysés globalement pour chaque question, c'est-à-dire que nous n'avons pas séparé les critères selon la réponse qui leur était associée. L'objectif de ces analyses est double :

- Nous souhaitons comparer les résultats statistiques des paragraphes précédents avec la perception subjective des utilisateurs : par exemple, nous avons mis en évidence l'influence prépondérante de la stratégie multimodale, mais nous ne savons pas si ces effets sont consciemment perçus par les utilisateurs. Réciproquement, il se peut que l'apparence des Agents soit importante aux yeux des utilisateurs même si celle-ci n'a quasiment aucun effet statistiquement significatif.
- Un autre objectif est de découvrir si les utilisateurs accordent du poids à des facteurs autres que ceux que nous avons testés. Nous espérons également que ces questions nous apporterons des précisions sur les détails de l'apparence des Agents qui sont déterminants pour les utilisateurs.

En moyenne, 81% des utilisateurs ont réussi à expliciter un ou plusieurs critères de jugement pour chacune des questions du questionnaire. La Figure 78 (Gauche) montre les catégories que nous avons formées avec les critères d'évaluation de la **qualité de l'explication**. La plupart des utilisateurs (63%) disent s'être fondés sur l'objet ou le contenu (organisation des informations verbales, vocabulaire utilisé…) pour juger de la qualité de l'explication. 11% des critères invoqués concernent l'Agent (son apparence ou sa voix). Les autres critères cités (26%) sont liés au comportement multimodal des

Agents : 16% des utilisateurs mentionnent le pointage comme critère de jugement de la qualité de l'explication. Un de nos utilisateurs (1% des critères) a exprimé l'idée de stratégie multimodale, en expliquant qu'un des Agents pointait l'image et donnait l'information verbale en même temps, un autre pointait seulement, et le dernier ne pointait jamais.

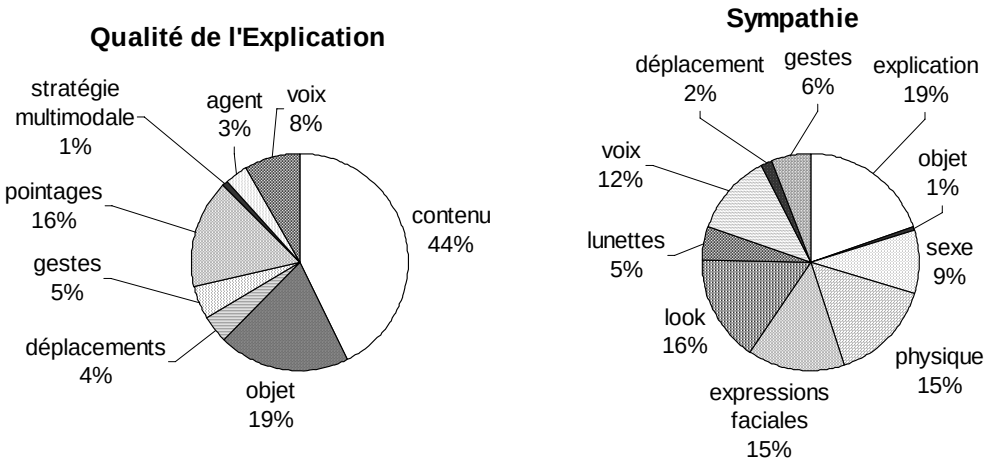

Figure 78 : Critères pour juger de la qualité de l'explication (Gauche) et de la sympathie des Agents (Droite).

Pour l'évaluation de la **sympathie** des Agents, les critères les plus fréquemment rapportés sont (Figure 78 Droite) : la qualité de l'explication, le physique de l'Agent (visage, coiffure…), les expressions faciales (en particulier les sourires) et le look (vêtements, chaussures…).

Pour l'évaluation de l'**expressivité**, le geste et les expressions faciales sont les critères qui ont été les plus fréquemment cités (Figure 79 Gauche).

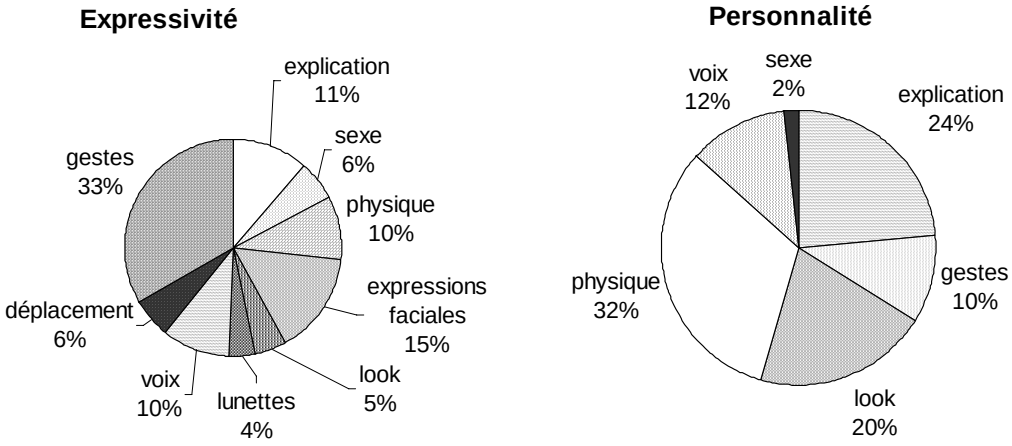

Figure 79 : Critères pour évaluer l'expressivité (Gauche) et la personnalité (Droite) des Agents.

En ce qui concerne l'évaluation de la **personnalité** des Agents (Figure 79 Droite), le critère le plus utilisé semble être le physique, qui regroupe ici le visage, les expressions faciales (en particulier le sourire), le port de lunettes, la coiffure. Les critères de qualité de l'explication et des gestes représentent tout de même à eux deux 34% des critères cités.

Enfin, la Figure 80 présente les facteurs ayant influencé les réponses des utilisateurs à propos de la **confiance** accordée aux Agents. La partie Gauche de la figure donne les critères pour la question sur l'assistance au dépannage (« Si votre ordinateur tombait en panne, suivriez vous les instructions de ces personnages pour le dépanner ? »). La plupart des utilisateurs (72%) se sont fondés sur la qualité de l'explication et sur leur perception de la compétence des Agents. Certains utilisateurs ont également expliqué que dans un tel cas ils étaient prêts à accepter de l'aide quelle qu'en soit la provenance (réponse « général », 11%), c'est-à-dire que leur réponse n'était en aucun cas liée aux Agents ou à leur comportement.

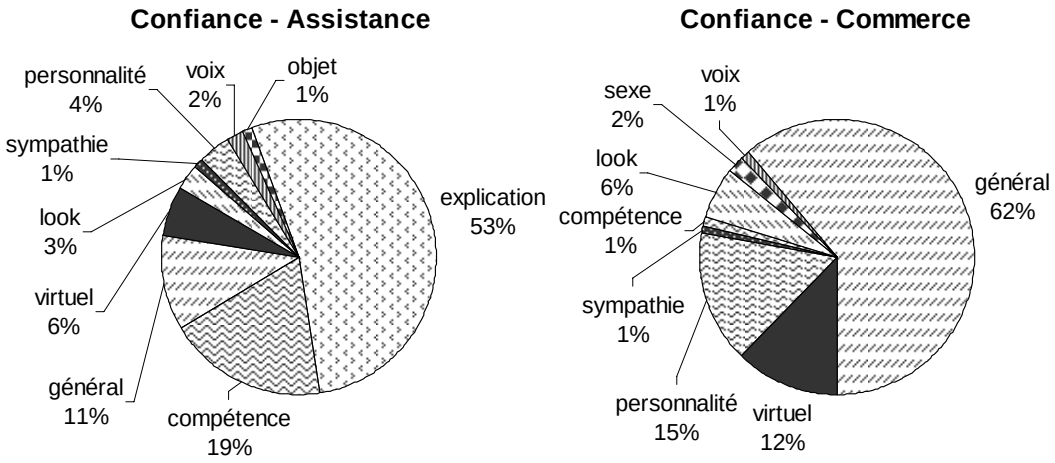

Figure 80 : Critères de décision pour la confiance assistance (Gauche) et la confiance commerciale (Droite).

Ce type de réponse, fondé sur un principe général et non influencé par la présence ou les caractéristiques des Agents, a été encore plus fréquent pour la question de confiance liée à l'e-commerce (« Donneriez-vous votre numéro de carte bleue à ces personnages ? »). Mais ces critères de décision généraux ont cette fois été en faveur de la réponse Non, c'est-à-dire du refus de confiance. Par ailleurs, 12% des utilisateurs ont justifié leur refus non par un principe contre les Agents (critère nommé « virtuel » dans la Figure 80) : ces utilisateurs ont déclaré ne pas faire confiance à des personnages virtuels.

Caractéristiques des explications. Nous présentons dans ce paragraphe les résultats de la question ouverte « Avez-vous remarqué des différences dans la manière dont les personnages donnent

les explications ? Lesquelles ? ». La Figure 81 montre que de nombreux utilisateurs ont vu des différences dans le contenu verbal de l'explication (organisation des informations, vocabulaire employé…) et dans la parole (voix, débit, élocution…). Les aspects non verbaux ont également été largement remarqués : déplacements des Agents dans l'écran, gestes. 23% des utilisateurs ont précisé que certains Agents allaient pointer l'image sur l'écran. Enfin, deux utilisateurs ont correctement formulé la notion de coopération multimodale entre le contenu verbal et les gestes de pointage, sans toutefois employer les termes de redondance ou complémentarité.

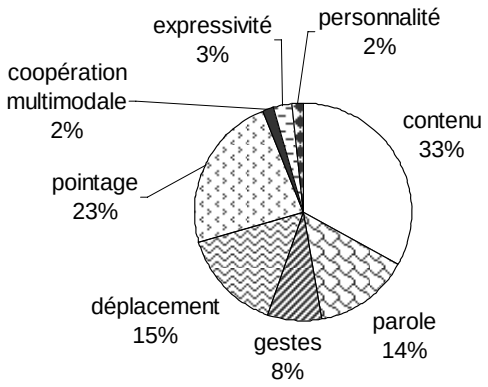

Figure 81 : Facteurs caractérisant les explications données par les Agents.

Discussion

Le Tableau 21 présente une synthèse des résultats obtenus.

Effet de la Stratégie multimodale. Les résultats de cette seconde expérience mettent clairement en évidence les bénéfices que la **redondance entre parole et geste** a apportés dans le contexte de nos présentations pédagogiques. Ces bénéfices sont apparus de manière concordante sur cinq de nos variables : la redondance a permis d'augmenter le rappel écrit des informations et la qualité perçue de l'explication, elle a rendu les Agents plus sympathiques aux yeux des utilisateurs, plus expressifs, et leur a conféré une personnalité plus positive.

En termes de performance de mémorisation, nous avons testé deux types de rappel : un rappel graphique et un rappel verbal. La Stratégie multimodale a principalement influencé le rappel verbal ; au niveau du rappel graphique, le seul effet de la Stratégie est une interaction avec l'Objet. Il semble donc que la Stratégie multimodale des Agents ait eu peu d'influence sur l'attention que les utilisateurs ont portée au support pictural de l'exposé.

	Rappel graphique	Rappel écrit	Qualité de l'explication	Sympathie	Expressivité	Personnalité	Confiance
Effet de la Stratégie multimodale		Redondance meilleure	Redondance meilleure	Redondance meilleure	Redondance meilleure	Redondance : personnalité plus positive	
Interaction Stratégie, Sexe, Personnalité		Effet du sexe chez les introvertis : seuls les hommes sont influencés par la Stratégie.	Effet du sexe chez les introvertis : chez les femmes, R meilleure ; chez les hommes, R = C.		Effet du sexe chez les introvertis : seuls les hommes sont influencés par la Stratégie (R = C).		
Effet de l'Agent				Marco plus sympathique.			
Interaction Stratégie, Agent		Pour Marco : R > C > S. Pour Léa : R > S > C. Pour Jules : pas de différence.		Pour Marco : R meilleure. Pour Jules : R > S > C. Pour Léa : Pas de différence.	Pour Marco : R > C > S. Pour Jules : R meilleure. Pour Léa : Pas de différence.		
Effet de l'Objet	Télécommande meilleure	Télécommande meilleure	Télécommande meilleure	Moins sympathique avec le photocopieur.			
Interaction Objet, Sexe			Chez les femmes, T meilleure, chez les hommes, T = L.				
Interaction Stratégie, Objet	Pour télécommande : C meilleure. Pour photocopieur : C moins bon. Pour logiciel : Pas de différence.						
Corrélation	Corrélation positive		Corrélations positives (moyennes)				
Critères d'évaluation			Contenu verbal, objet.	Explication, look, physique de l'Agent, expressions faciales.	Gestes, expressions faciales.	Physique, look, explication.	Explication, principes généraux.

Légende : R : Redondance ; C : Complémentarité ; S : Spécialisation ; T : Télécommande ; L : Logiciel ; P : Photocopieur.

Tableau 21 : Synthèse des résultats de cette expérience.

En revanche, la Stratégie multimodale a modifié la mémorisation du contenu verbal (présentation des fonctionnalités des objets) et/ou de la correspondance entre le contenu verbal et le support pictural (attribution de la bonne fonctionnalité au bon élément). La redondance a en effet permis une augmentation relative de 19% du rappel verbal (49% d'informations rappelées correctement, contre 41% en moyenne pour la complémentarité et la spécialisation). Par rapport à l'expérimentation préliminaire, ce résultat donne une valeur beaucoup plus importante à la Stratégie multimodale : il ne s'agit plus cette fois uniquement de satisfaire des évaluations subjectives, mais d'augmenter l'efficacité même d'une application pédagogique.

A noter que nous n'avons pas relevé de différence dans les performances des hommes et des femmes sur les deux types de rappel : les hommes n'ont pas eu de meilleures performances au rappel graphique ni les femmes au rappel verbal. Nous pouvons en conclure que ces deux tâches ne sont pas directement assimilables aux protocoles classiques de test des capacités visuo-spatiales et auditivo-verbales. Par exemple, il est possible que la mémorisation graphique des objets puisse s'appuyer sur un étiquetage verbal ; de plus, il est clair que la mémorisation du contenu verbal de la présentation n'est pas une tâche purement auditivo-verbale, car elle met en jeu une représentation visuelle de l'objet pour pouvoir associer les fonctionnalités présentées aux bons éléments de cet objet.

Dans l'expérimentation préliminaire, nous avions relevé un effet d'interaction entre le Sexe des utilisateurs et la Stratégie multimodale. Cet effet a été retrouvé à un moindre niveau dans la présente expérience. En effet, le Sexe est intervenu dans des interactions de second ordre mettant en jeu la Stratégie multimodale et la Personnalité des utilisateurs (introvertis / extravertis). Ces effets, qui sont apparus sur le rappel écrit, l'évaluation de la qualité de l'explication et l'évaluation de l'expressivité, peuvent être simplifiés ainsi :

- L'effet principal en faveur de la redondance se vérifie chez les sujets extravertis (hommes et femmes),
- Les hommes introvertis sont également influencés par la Stratégie, mais semblent moins différencier la redondance et la complémentarité (cf. effets sur l'évaluation de la qualité de l'explication et de l'expressivité),
- Enfin, il semble que la Stratégie multimodale ait peu d'effet sur les femmes introverties (cf. effets sur le rappel écrit et sur l'évaluation de l'expressivité).

Il semble donc que la Stratégie multimodale des Agents ait globalement eu un impact moins fort sur les individus introvertis. Des études antérieures sur ce trait de personnalité ont montré que les sujets introvertis ont de meilleures performances dans les tests de vigilance, et sont plus persévérants dans la réalisation de tâches mentales par rapport aux sujets extravertis (pour une revue de la littérature sur

l'introversion, voir Eysenck & Eysenck, 1971). Nous pouvons ainsi supposer que nos utilisateurs introvertis étaient plus concentrés lors des présentations des Agents, et ainsi n'ont pas eu besoin de l'aide apportée par la redondance multimodale. Rappelons que le contenu sémantique de la présentation était le même pour les trois Stratégies, et que celles-ci n'apportaient éventuellement qu'un soutien à l'attention des utilisateurs et une facilitation à l'encodage de l'information. Il était donc peut-être possible de compenser les différences entre les stratégies par une concentration accrue.

Pour expliquer l'effet du sexe au sein du groupe des introvertis, nous pouvons faire référence, comme pour l'expérimentation précédente, aux différence cognitives entre hommes et femmes : une préférence pour les traitements visuo-spatiaux pourrait expliquer que les hommes introvertis aient tout de même accordé de l'importance à la présence de gestes opératoires (stratégies redondante et complémentaire), alors que les femmes introverties se seraient davantage focalisées sur le contenu auditivo-verbal.

En résumé, pour ces interactions complexes, nous proposons l'interprétation suivante : les individus extravertis ont prioritairement été influencés par les conditions externes de la tâche, c'est-à-dire ici les Stratégies multimodales des Agents. Au contraire, les individus introvertis ont plutôt déployé des stratégies internes de traitement : une concentration plus forte a diminué l'impact des Stratégies multimodales et a ainsi laissé apparaître les différences cognitives entre hommes et femmes.

Contrairement à ce que nous avions prévu, nous n'avons pas observé d'effet de la Personnalité des utilisateurs sur l'évaluation de la personnalité des Agents : la plupart des utilisateurs se sont accordés à évaluer plus positivement les Agents redondants. Il est possible que cela soit dû à un manque de précision de notre mesure (codage des qualificatifs de personnalité en trois grandes catégories : positif, négatif, neutre) et qu'avec une méthode plus fine des différences seraient apparues dans les évaluations des individus introvertis et extravertis. Mais il est également possible que la redondance soit effectivement la stratégie la plus désirable socialement pour le rôle pédagogique qu'endossaient nos Agents. Nous interprétons la redondance du tuteur comme un comportement plutôt extraverti, en raison des gestes plus larges et des déplacements qu'elle occasionnait. Cette stratégie pourrait donc être la plus appréciée par tous, y compris par les apprenants introvertis.

Comme dans l'expérience préliminaire, il semble que l'effet des stratégies soit en grande partie inconscient. En effet, seuls 2 utilisateurs sur 106 ont explicité dans leurs remarques la notion de coopération entre modalités. S'il est indéniable que de nombreux utilisateurs (environ la moitié de notre échantillon, voir la Figure 81) ont remarqué que certains Agents allaient pointer l'image, ils n'ont généralement pas remarqué la différence entre redondance et complémentarité – ou ne l'ont pas explicitement formulée. D'autre part, même pour ceux qui ont noté une différence de comportement

gestuel, voire de stratégie multimodale, les utilisateurs ne se sont pas rendus compte à quel point ce facteur influençait leurs évaluations : ceux qui ont mentionné les pointages, les gestes en général ou les déplacements comme critères d'évaluation restent relativement peu nombreux (26% pour l'évaluation de l'explication, 25% pour la sympathie, 50% pour l'expressivité, 34% pour la personnalité) eu égard à la significativité des effets.

Effets de l'apparence des Agents. Il semble que la taille de notre échantillon d'utilisateurs ait globalement favorisé un nivellement des évaluations subjectives sur les trois Agents. En effet, un seul effet principal de l'apparence a subsisté : Marco a été jugé globalement le plus sympathique des trois Agents.

Les commentaires des utilisateurs sont assez concordants avec ceux que nous avions déjà recueillis lors de l'expérimentation préliminaire : les caractéristiques les plus marquantes de Marco sont son sourire et, dans une moindre mesure, son look (coiffure, vêtements). Cela lui a généralement apporté des évaluations positives. Nous retiendrons qu'un large sourire de type cartoon, bien qu'il soit peu réaliste, est un facteur important de sympathie. Ce résultat peut être rapproché des recommandations de Kohar et Ginn (1997), selon lesquelles les personnages très expressifs, parce qu'ils sont susceptibles de véhiculer une composante émotionnelle au cours de l'interaction, font de meilleurs Agents que les personnages plus réalistes et humanoïdes – cette recommandation est aussi valable pour les Agents pédagogiques, car l'engagement et l'amusement sont des facilitateurs importants de la situation d'apprentissage (Lester *et al.*, 1999a). Les éléments marquants de l'apparence de Léa sont sa voix et sa blouse blanche ; ceux de Jules sont ses lunettes, sa coiffure et son short. Ces caractéristiques de Léa et de Jules ont produit des jugements contrastés de la part des utilisateurs ; pour obtenir un personnage qui remporte l'assentiment général, il semble préférable d'adopter un Agent comme Marco, souriant et avec un look relativement neutre.

La voix des Agents est un paramètre important pour les perceptions subjectives des utilisateurs, comme l'ont souligné plusieurs recherches (voir par exemple Darves & Oviatt, 2004). Dans la présente étude, la voix a été fréquemment citée dans les critères d'évaluation (environ 10% des facteurs de décision pour chacune des questions subjectives). Or le facteur voix n'était pas contrôlé dans notre expérience : nous avons utilisé deux voix de base d'IBM ViaVoice (une voix masculine, une vois féminine), qui ont été jugées assez négativement (37% des utilisateurs se sont plaints de la qualité de la synthèse vocale). Il est donc possible qu'avec des voix de meilleure qualité les évaluations aient été légèrement différentes : nous pensons notamment qu'il y aurait eu moins de qualificatifs de personnalité négatifs. Il est également possible que Léa ait été désavantagée par ce facteur, car il y a davantage d'utilisateurs qui se sont plaints de sa voix que de celle de Marco et Jules.

Contrairement à ce que nous avions observé précédemment, l'apparence des Agents n'a pas influencé la mémorisation des informations. Ce résultat est rassurant en un sens car il permet d'espérer que les détails physiques des Agents importent peu, et qu'ils peuvent tous être des tuteurs efficaces pourvu qu'ils aient un comportement adéquat. Pour confirmer cela, il faudrait cependant tester des Agents d'apparence franchement différente (ex : Agents non humanoïdes).

En revanche, nous avons observé plusieurs effets d'interaction entre Stratégie et Agents. Ces effets, qui sont apparus sur le rappel écrit, sur l'évaluation de la sympathie et de l'expressivité, suggèrent que certaines stratégies conviennent plus que d'autres à certains Agents. Ces trois effets d'interaction sont cependant troublants car leur sens est à chaque fois différent (voir le Tableau 21) : par exemple pour Jules, sa stratégie n'a pas eu d'effet sur le rappel écrit, mais il a été jugé plus sympathique et plus expressif avec la redondance. Pour Léa c'est l'inverse, sa stratégie a influencé le rappel écrit mais pas les évaluations subjectives. Au total, ces effets d'interactions nous semblent difficiles à interpréter de manière univoque.

Effets de l'Objet. Contrairement à la précédente expérience, nous avons relevé ici d'importants effets de l'Objet de l'explication. La télécommande s'est révélée être un objet plus facile à mémoriser que les deux autres : cela a influencé le rappel graphique, le rappel écrit et l'évaluation de la qualité de l'explication. L'Objet a également influencé l'évaluation de la sympathie des Agents, puisque celui (celle) qui présentait le photocopieur a globalement été jugé comme moins sympathique que les deux autres.

Il n'est pas anormal en soi que l'Objet de l'explication joue sur la performance de mémorisation et sur les évaluations subjectives des apprenants, mais dans cette expérience notre but était de neutraliser ce facteur pour étudier de manière plus épurée l'influence des Stratégies multimodales. Nous considérons donc que ces effets biaisent nos données. Cependant, il faut également souligner que l'effet massif de la Stratégie multimodale n'a pas été entravé par ce biais : l'absence d'interaction entre Objet et Stratégie (sur toutes les variables sauf le rappel graphique, voir le Tableau 21) montre que l'effet de la Stratégie reste valable bien que le niveau de difficulté des objets n'ait pas été homogène.

La confiance accordée aux Agents. Les résultats que nous avons obtenus sur la confiance vis-à-vis des Agents montrent qu'il s'agit d'une notion faiblement influencée par les détails du comportement ou de l'apparence des Agents : les utilisateurs semblent fonder leur confiance sur une impression plus générale. Cela les a amenés à affirmer en majorité qu'ils feraient confiance aux conseils des Agents dans une situation d'assistance / dépannage, et ceci pour les trois Agents quelles qu'aient été leur Stratégie multimodale et leur apparence.

Les utilisateurs se sont également fondés sur un principe général (« je ne donne jamais mon numéro de carte bleue » ou « je n'ai pas confiance dans les ordinateurs ») pour refuser, en majorité, la confiance aux Agents dans la question e-commerce. Non seulement la Stratégie ou l'apparence des Agents n'ont pas influencé leurs réponses, mais il semble même que la présence d'Agents n'ait pas été prise en compte dans la plupart de leurs réponses.

6.3. Conclusion

Le principal résultat apporté par les deux expérimentations décrites dans ce chapitre concerne le bénéfice de la redondance entre parole et gestes opératoires chez les Agents pédagogiques. Ce bénéfice se manifeste au niveau de la performance mnésique des utilisateurs, ainsi que sur des variables d'évaluation subjective des Agents : la perception de la sympathie, de la personnalité des Agents se sont révélées être influencées par leur type de coopération multimodale.

Les effets de la redondance dans les logiciels pédagogiques ont déjà été discutés, mais ils concernaient la redondance multimédia (par exemple la présentation simultanée de la leçon en version texte et en version orale) et non le comportement multimodal des Agents. Au vu de cette littérature sur la redondance multimédia, il semblait d'ailleurs difficile de prédire les résultats de notre expérience : en effet la redondance entre texte et synthèse vocale s'est parfois avérée bénéfique (Moreno & Mayer, 2002), et parfois au contraire néfaste par rapport à la synthèse vocale seule (Kalyuga *et al.*, 1999 ; Craig *et al.*, 2002). Il semble donc que le fait de fournir des informations redondantes qui ne sont pas nécessaires à la compréhension soit parfois susceptible de saturer la mémoire de travail et donc de perturber la performance.

A notre connaissance, la redondance multimodale n'avait jamais été testée auparavant avec des Agents, pas plus qu'avec des humains. L'étude de Goldin-Meadow *et al*. (1999), présentée au chapitre 2.2.2, avait comparé l'effet de gestes concordants et discordants avec le discours des enseignants de mathématiques, mais sans traiter de la nature redondante ou complémentaire des gestes concordants. Il serait d'ailleurs assez complexe de répliquer notre expérience avec des tuteurs humains à la place des Agents, car la manipulation des stratégies de coopération multimodale nécessiterait un long travail préliminaire d'entraînement des tuteurs. Cassell *et al.* (1999a) avaient mis en place une telle procédure pour tester l'influence de la gesticulation sur la mémorisation des auditeurs. Il est d'ailleurs possible de considérer que ces auteurs ont utilisé des gestes redondants, complémentaires et concurrents, mais, dans leur expérience, les conditions redondante et complémentaire n'étaient pas comparables puisque

les informations transmises aux auditeurs n'étaient pas les mêmes dans ces deux cas (voir le chapitre 2.2.2). Bien entendu, nous ne pouvons affirmer que nos résultats seraient répliqués avec des humains, mais notre expérience avec les Agents fournit tout de même une indication forte sur l'intérêt de la redondance.

Les participants de cette expérience étaient des étudiants, c'est-à-dire de jeunes adultes de niveau d'éducation post-baccalauréat. Nos résultats pourraient donc être appliqués dans le cadre de la conception d'Agents pédagogiques destinés aux étudiants, par exemple pour l'enseignement à distance (comme le fait l'Agent Adele de Johnson *et al.*, 2003). Cependant nos résultats ne peuvent être généralisés à d'autres populations, en particulier à la population cible majeure des logiciels pédagogiques, c'est-à-dire aux enfants. Pour ces utilisateurs, nous ne pouvons que formuler l'hypothèse que les bénéfices de la redondance multimodale, s'ils sont apparus avec des étudiants, devraient apparaître également avec des enfants. Cette hypothèse mériterait bien entendu d'être testée en bonne et due forme.

Perspectives

Nous avons fait en sorte que notre protocole s'apparente à une situation pédagogique : les Agents faisaient de courts exposés magistraux et le but des utilisateurs était de retenir le maximum d'informations (rappelons qu'ils étaient d'emblée prévenus qu'ils devraient, après avoir visionné les présentations, en restituer le contenu). Pour cette situation précise, c'est la redondance qui s'est avérée la mieux adaptée et a permis d'améliorer le rappel écrit. Nous pensons que les évaluations subjectives (en particulier sur la qualité de l'explication, la sympathie et la personnalité) sont liées à ce même phénomène : nous avons vu qu'elles ne sont pas corrélées à la performance de rappel, mais nous pensons que les évaluations plus positives accordées à la redondance proviennent du fait qu'il s'agissait de la stratégie la plus appropriée à la tâche des utilisateurs. Avec une tâche différente, ce ne sont peut-être pas les Agents redondants qui auraient été les mieux jugés.

La complémentarité présente d'autres avantages : elle permet notamment d'alléger la quantité d'informations transmises par chaque modalité, et d'éviter ainsi à la fois une surcharge d'informations verbales et une quantité de gestes exagérée, qui pourrait être jugée non naturelle (Cassell & Stone, 1999). Les présentations complémentaires sont également plus courtes en temps : dans notre expérience, pour communiquer le même contenu, les séquences complémentaires étaient 20% plus courtes que les séquences redondantes et spécialisées. Si la redondance est la meilleure stratégie pour une situation de mémorisation et d'apprentissage, il doit y avoir des situations dans lesquelles la complémentarité serait la stratégie la plus appropriée. Nous pensons par exemple qu'un contexte d'assistance technique, qui est un autre rôle courant des Agents Conversationnels, pourrait être

intéressant à étudier : dans ce cas, l'objectif des utilisateurs est d'obtenir le plus rapidement possible un renseignement précis (par exemple, quel est le bouton qui permet d'accéder à telle fonctionnalité). Il est possible que ce type de situation soit plus favorable à une stratégie complémentaire et que les évaluations subjectives des utilisateurs soient modifiées en conséquence. Nous envisageons donc de compléter les données exposées dans ce chapitre sur l'effet du comportement multimodal des Agents dans une situation pédagogique par l'étude d'une situation d'assistance technique.

7. Discussion sur l'évaluation du système développé

Les travaux réalisés au cours de cette thèse, que ce soient les recherches bibliographiques ou les études expérimentales, nous ont amenés à considérer plusieurs pistes pour la mise au point de nouvelles méthodes d'évaluation pour des systèmes tels que celui du projet NICE. Ce sont ces pistes que nous souhaitons expliciter dans le présent chapitre.

Mais avant de discuter la possibilité de nouvelles méthodes d'évaluation, un rappel sur les méthodes existantes s'impose.

7.1. Méthodes d'évaluation existantes

Dans un premier temps, nous présentons les méthodes les plus classiques d'évaluation des systèmes d'information, puis nous exposons des méthodes moins courantes mais davantage ciblées pour un système tel que NICE (application destinée à un public jeune, avec une composante ludique).

Les méthodes d'évaluation de l'**Interaction Homme-Machine** peuvent être catégorisées selon différents critères : elles peuvent par exemple être classées selon les objectifs de l'évaluation (ex : diagnostic, examen de la conformité à des normes), ou selon le moment de l'évaluation (ex : maquette, version bêta, système fonctionnel). Certains auteurs distinguent approche empirique et approche analytique de l'évaluation (Kolski, 1993 ; Senach, 1993), ce qui rejoint les catégories développées par Bastien et Scapin (2001) sous les termes suivants :

- Les méthodes requérant la participation directe des utilisateurs (approche empirique) : cette catégorie comprend les tests utilisateurs, les outils logiciels, les questionnaires et entretiens.

- Les méthodes s'appliquant aux caractéristiques de l'interface (approche analytique) : cette catégorie regroupe les méthodes à base de modèles formels, le recours à l'expert, les méthodes d'inspection et les outils d'évaluation automatique.

La seconde catégorie de méthodes ne peut exister que pour les domaines disposant d'une longue expérience d'évaluations ergonomiques, de repères solides et fiables (ex : grilles de recommandations, normes), ce qui n'est pas encore le cas pour les Agents Conversationnels Animés. Les méthodes qui nous intéressent ici sont donc celles qui requièrent la participation directe des utilisateurs. Elles reposent sur deux types de variables. Des variables **objectives** :

- Temps de réalisation d'une tâche,
- Exactitude du résultat,
- Nombre et type d'erreurs commises,
- Mouvements oculaires (permettent d'estimer les stratégies de traitement et les processus attentionnels) et fréquence de clignement des yeux (indice de charge mentale),
- Indicateurs psychophysiologiques tels que les potentiels évoqués cérébraux, l'électroencéphalogramme, le rythme cardiaque ou l'activité électrodermale (donnent des indications sur le stress et la charge cognitive, sur les réactions émotionnelles),
- Evénements système sauvegardés dans les fichiers de *log* ou enregistrés sur vidéo (ex : requêtes des utilisateurs, procédures lancées, nombre de clic de souris).

Et des variables **subjectives** :

- Questionnaires (voir par exemple Chin *et al.*, 1988 ; Davis, 1989 ; Nielsen, 1993 ; Lewis, 1995 ; Lin *et al.*, 1997 ; Perlman, 1997),
- Entretiens destinés à recueillir, tout comme les questionnaires, la satisfaction, les attitudes et les opinions des utilisateurs.

Les variables précédentes, objectives et subjectives, peuvent être ou ont déjà été appliquées à l'évaluation des **Agents Conversationnels Animés.** Comme nous l'avons vu dans le paragraphe 2.3, où nous avons présenté un tableau synthétique des résultats expérimentaux concernant les Agents, les problématiques de recherche et les aspects à évaluer sur les Agents sont nombreux et hétérogènes. Dans leur ouvrage consacré à l'évaluation des Agents, Ruttkay et Pelachaud (2004) distinguent deux types d'études : les micro-évaluations sur les Agents, et macro-évaluations. Les micro-évaluations ont pour objectif d'isoler une seule caractéristique de l'Agent pour l'étudier très précisément : c'est par exemple cette démarche que nous avons adoptée au chapitre 6 pour tester l'effet des coopérations sémantiques entre modalités. A l'inverse, les macro-évaluations visent à étudier un Agent dans son ensemble (on pourrait presque dire dans son individualité), à l'échelle d'une application.

Les macro-évaluations impliquent parfois l'étude de notions abstraites (ex : évaluation de la crédibilité de l'Agent) : une des difficultés majeures est alors de trouver un ou plusieurs indicateurs mesurables ou observables permettant d'estimer de manière valide ces notions abstraites. Isbister et Doyle (2004) proposent une taxonomie associant quelques unes de ces notions abstraites avec des variables mesurables (Tableau 22).

Objectifs	**Variables dépendantes associées**
Crédibilité	Subjectives : Questionnaire auprès des utilisateurs, et auprès des experts. Interroger sur la crédibilité de l'apparence, de la voix, du langage et des réactions de l'Agent.
	Objectives : Réactions physiologiques et comportementales.
Sociabilité	Subjectives : Questionnaire sur les qualités sociales de l'Agent (amicalité, capacités communicatives…) et sur la satisfaction de l'utilisateur.
	Objectives : Réponses sociales face à l'Agent (ex : influence exercée par l'Agent).
Compétences dans un domaine défini	Subjectives : Questionnaire sur la satisfaction de l'utilisateur par rapport à l'interaction et à son résultat.
	Objectives : Résultats (performance, mémorisation…).
Utilité	Subjectives : Questionnaire (ex : le système est-il optimisé ?).
	Objectives : Efficacité de l'Agent dans un contexte réel.
	Crédibilité et sociabilité telles que définies plus haut.

Tableau 22 : Variables d'évaluation des Agents répertoriées par Isbister et Doyle (2004).

Pour des utilisateurs particuliers comme le sont les **enfants**, certaines méthodes sont plus appropriées que d'autres pour évaluer des application multimédia interactives (MacFarlane *et al.*, 2003). Parmi celles qui sont adaptées aux enfants, on trouve :

- La méthode de la pensée à voix haute (*thinking aloud*) : cela consiste à demander à l'utilisateur de commenter ses actions au cours de la réalisation d'une tâche, afin d'accéder à son mode de raisonnement. Ce commentaire peut être fait rétrospectivement, lorsque l'utilisateur est confronté à l'enregistrement vidéo de la session d'interaction (ex : « Là j'étais en train de chercher le bouton pour envoyer un *e-mail*. »).
- L'enseignement à un partenaire (*peer tutoring*) : dans ce cas, l'utilisateur doit apprendre à une autre personne (ex : un camarade dans le cas d'une évaluation avec des enfants) comment se servir d'un produit. Cette méthode fournit un contexte plus naturel que la pensée à voix haute

pour obtenir des verbalisations sur la manière dont l'utilisateur s'est servi d'une application, et sur les difficultés qu'il a éventuellement rencontrées. Une autre méthode incitant naturellement les utilisateurs à verbaliser leurs actions consiste à leur faire réaliser un scénario en binôme. Les communications que les deux personnes vont échanger durant ce travail coopératif sont susceptibles de renseigner l'expérimentateur sur les stratégies employées, les difficultés, etc.

- La méthode du journal de bord (*user diaries*) : elle peut s'appliquer lorsque le produit est fourni aux utilisateurs pour une période prolongée, pour qu'ils puissent s'en servir dans des contextes réalistes. Il leur est alors demandé de consigner leurs remarques sur l'utilisation du produit (résultats positifs, incidents, etc.) dans un journal de bord. L'attractivité du journal lui-même peut être augmentée en fournissant des outils multimédia (ex : ordinateur de poche, dictaphone, appareil photo), ce qui incite particulièrement les enfants à le remplir régulièrement.

A l'inverse, il semble que d'autres méthodes comme le *focus group* (un groupe d'utilisateurs potentiel est réuni autour d'un modérateur pour discuter de la conception d'un produit) soient moins adaptées aux enfants (MacFarlane *et al.*, 2003).

Le contexte des **jeux vidéo** apporte également des particularités supplémentaires à la situation d'évaluation. En effet, l'objectif d'un jeu est différent de celui de la plupart des applications logicielles : la clé d'une expérience divertissante ne consiste pas à utiliser l'outil le plus efficace et le plus facile. C'est même parfois l'opposé, puisque les joueurs peuvent apprécier le challenge et les difficultés, afin de les surmonter et d'obtenir *in fine* une expérience de succès. Ainsi, la stratégie de conception d'une application de jeu peut parfois impliquer la recherche d'une frustration temporaire de l'utilisateur (Keeker *et al.*, 2004), ce qui peut sembler opposé aux notions fondamentales d'utilisabilité (Johnson & Wiles, 2003). En réalité, un jeu doit à la fois être utilisable (l'utilisateur doit avoir la maîtrise de toutes les fonctionnalités) et poser des défis (davantage dans le scénario, ou *gameplay*, que dans la gestion de l'interface). Mais ce défi doit être savamment dosé, puisque l'amusement, la satisfaction finale, résultent du succès que l'utilisateur doit réussir à obtenir, malgré les difficultés. Un jeu trop difficile, un scénario impossible à achever risquerait de ne pas plaire aux joueurs. Un exemple de solution à ce problème du dosage des difficultés peut être trouvé dans les jeux de rôle en ligne (*Massive Multiplayer Online Role Playing Games*), qui offrent une multitude de niveaux de jeu, de difficulté croissante. Le sentiment de progresser incite les joueurs à continuer de jouer, et le nombre quasi-illimité de niveaux les empêche d'atteindre jamais la fin, ce qui les fidélise au plus haut point.

Les jeux vidéo peuvent bénéficier de façon complémentaire de trois types d'évaluation (Lazzaro & Keeker, 2004) :

- Les méthodes traditionnelles basées sur des grilles de recommandations, les mesures de performances et d'erreurs, les questionnaires et entretiens,
- Le recueil de mesures physiologiques pour estimer les réactions émotionnelles (pression artérielle, activité électrodermale),
- Les mesures qualitatives des réactions cognitives et émotionnelles, parmi lesquelles peuvent être trouvées la pensée à voix haute et le journal de bord. L'observation des expressions faciales et posturales est également recommandée : par exemple, face à un jeu vidéo, Kaiser *et al.* (1998) ont relevé des expressions faciales assez intenses (ex : lorsqu'un ennemi effrayant apparaît brusquement).

Ces variables permettent d'évaluer les quatre dimensions essentielles du jeu (Pagulayan *et al.*, 2002) : la qualité globale (variable de haut niveau correspondant au plaisir de jeu), la facilité d'utilisation (des média, de l'interface, des instructions), le challenge (difficultés et atteinte d'objectifs) et le rythme du jeu (niveau de vitesse, de tension et de concentration requis en fonction du jeu).

Dans la suite de ce chapitre, nous allons discuter de la possibilité de mettre en œuvre de nouvelles méthodes d'évaluation : la première que nous allons considérer utilise la multimodalité en entrée (telle que nous l'avons étudiée au chapitre 5) comme variable d'évaluation ; la seconde méthode concerne l'étude des réactions émotionnelles faciales et verbales de l'utilisateur.

7.2. Vers de nouvelles méthodes

7.2.1. La multimodalité en entrée comme indice d'évaluation

Nous avons vu au chapitre 5 que l'utilisation de constructions multimodales dans le comportement des utilisateurs avait beaucoup varié entre les trois études rapportées (de 2,6 à 21,2% des entrées des utilisateurs). Pour expliquer cette différence, nous avons écarté le facteur « nature du scénario » puisque, même dans le scénario entièrement conversationnel, la modalité gestuelle avait représenté une part importante de l'interaction (36,6% des entrées). Nous pensons que les deux hypothèses plausibles qui subsistent pour expliquer l'écart dans les taux d'utilisation de la multimodalité sont les suivantes :

- Dans l'une des études, les utilisateurs n'interagissaient pas dans leur langue maternelle (jeunes utilisateurs danois devant parler anglais), ce qui a diminué le naturel de la communication et a ainsi pu limiter la production de constructions multimodales.

- La perception que les utilisateurs ont eue des capacités du système et de son efficacité n'était pas équivalente dans les trois études : dans deux d'entre elles, l'interaction était gérée par des magiciens ; dans la troisième, elle était gérée par le prototype, et a été perturbée par des *bugs* du système. Constatant les difficultés d'interaction dans ce dernier cas, les utilisateurs ont peut-être spontanément limité les constructions multimodales afin de faciliter le traitement du système.

Les trois études diffèrent également sur l'introduction de comportements sociaux face aux Agents. Dans l'étude initiale en Magicien d'Oz, nous avons relevé de nombreuses marques de politesse, ainsi qu'une sorte de timidité de la part de certains enfants (vouvoiement, prise d'initiative par le geste plutôt que par la parole…). De même, l'interaction verbale avec Cloddy Hans contenait 5,5% de thèmes sociaux. Mais de tels comportements ne sont pas apparus dans le test du prototype I avec Andersen. Les deux mêmes hypothèses que précédemment sont tenables pour expliquer cette différence de réaction sociale entre les trois corpus :

- Il est possible que l'interaction sociale soit réfrénée par le fait d'interagir dans une langue étrangère ;
- Il est également possible que les faiblesses du système aient décrédibilisé l'Agent, et fait que celui-ci n'a pas été perçu comme un partenaire social.

Hypothèse. Dans les deux cas (utilisation de la multimodalité et interaction sociale), nous penchons davantage pour la seconde hypothèse, selon laquelle **les utilisateurs adapteraient leur comportement verbal et multimodal aux capacités du système, ou à ce qu'ils en perçoivent**. Il a été montré que le comportement multimodal varie avec la tâche (Oviatt, 1996a) et la charge mentale (Oviatt *et al.*, 2004), mais à notre connaissance l'influence des capacités perçues du système n'a pas été étudiée. Si notre hypothèse s'avère juste, il serait donc possible de renverser le statut du comportement multimodal en entrée, en le considérant non plus seulement comme un moyen de communication, mais également comme un indicateur de la perception que les utilisateurs ont du système. La multimodalité en entrée pourrait alors devenir une variable d'évaluation.

Validation envisagée. Pour tester cette hypothèse, il faudrait par exemple organiser une étude en Magicien d'Oz, avec multimodalité en entrée et Agents Conversationnels en sortie, et faire passer deux groupes d'utilisateurs :

- Un groupe dans une condition d'interaction de qualité, c'est-à-dire où l'Agent comprendra et répondra de manière appropriée aux entrées utilisateurs,
- Un second groupe en condition d'interaction dégradée, c'est-à-dire en introduisant volontairement des erreurs de traitement, des *bugs* simulés.

Les passations seront filmées afin de pouvoir, par la suite, annoter le comportement de l'utilisateur en se focalisant : sur les aspects sociaux de l'interaction verbale, et sur la proportion d'utilisation des constructions multimodales.

Sans avoir besoin de recourir à une étude en Magicien d'Oz, le test du prototype II avec Andersen devrait nous apporter un éclairage certain sur la validité de notre hypothèse. En effet, les conditions de test seront comparables à celles du prototype I (utilisateurs danois interagissant en anglais, même scénario, même environnement de passation), à ceci près que la qualité d'interaction devrait avoir augmenté substantiellement. Nous considèrerons notre hypothèse validée si le taux d'utilisation de la multimodalité apparaît largement supérieur au corpus précédent, et si des comportements sociaux émergent chez les utilisateurs.

Dans ce cas, nous pourrons conclure que ces nouveaux utilisateurs ont attribué au système une qualité supérieure à celle que les utilisateurs précédents avaient perçue. Cependant, cette évaluation ne nous renseignera pas sur les aspects ludiques de l'interaction avec Andersen. Nous proposons d'aborder ce point ci-dessous.

7.2.2. L'expression multimodale des émotions

Le système NICE étant développé avant tout pour un contexte de loisir, une partie de son succès dépendra de sa capacité à divertir, voire à amuser les utilisateurs (Johnson & Wiles, 2003). Cet aspect est l'un des plus difficiles à évaluer dans les applications ludiques. L'amusement peut être estimé par l'intermédiaire d'un questionnaire, en demandant explicitement à l'utilisateur d'auto-évaluer le niveau d'amusement que le système lui a procuré. Mais ce recueil s'expose à un biais important de désirabilité sociale (tendance pour les sujets à satisfaire l'expérimentateur), en particulier lorsque les utilisateurs sont des enfants (Hanna *et al.*, 1997 ; MacFarlane *et al.*, 2003). Un indice comportemental serait donc utile pour évaluer de manière plus fine l'amusement des utilisateurs.

Les expressions faciales sont l'une des sources d'information les plus importantes sur l'état émotionnel d'une personne (Collier, 1985). Il est alors possible qu'elles puissent traduire, au cours de l'interaction avec un système, les réactions affectives d'un utilisateur (amusement, satisfaction, énervement, ennui…). Dans l'évaluation de l'Interaction Homme-Machine, nous avons vu que ce type d'indice était parfois utilisé de manière qualitative (Lazzaro & Keeker, 2004), c'est-à-dire de manière globale et non systématique : cette forme d'utilisation pose la question de la reproductibilité du recueil et de l'analyse des expressions faciales.

Les expérimentations de psychologie et d'éthologie sur le décodage des expressions faciales (voir Manstead, 1991) sont classiquement réalisées avec un matériel de jeu d'acteurs (les personnes sont filmées en train de mimer l'expression d'une émotion) et non un corpus d'interaction spontané. Or dans une situation naturelle, nos expressions faciales sont moins marquées et sont même parfois indétectables, sauf à recueillir l'activité musculaire du visage (Collier, 1985). Les résultats[38] obtenus dans les expériences de décodage d'émotions mimées ne seraient donc pas forcément répliqués avec un corpus d'interaction naturelle.

Par ailleurs, il existe de nombreux travaux sur l'analyse automatique des expressions faciales, utilisant la vision par ordinateur et l'analyse morphologique du visage (Essa & Pentland, 1997 ; Kaiser *et al.*, 1998), mais la détection des émotions (à partir de quand considère-t-on un mouvement comme une réaction émotionnelle ?) et leur interprétation (quelle est l'émotion sous-jacente ?) constituent toujours un défi pour ces systèmes automatiques. Il existe tout de même des systèmes dans lesquels les expressions faciales sont une modalité d'entrée de l'interaction (Bianchi-Berthouze & Lisetti, 2002), et à ce titre sont détectées et reconnues automatiquement. Ces systèmes fonctionnent grâce à des réseaux de neurones, et requièrent donc une phase d'apprentissage au cours de laquelle l'utilisateur doit étiqueter lui-même ses expressions faciales. L'objectif, une fois cette phase accomplie, est de prendre en compte les expressions faciales des utilisateurs pour adapter les réactions du système et/ou le comportement d'un Agent Conversationnel (ex : répéter l'explication d'un concept si l'utilisateur montre de la perplexité). Il semble cependant que lors de l'interaction avec de tels systèmes les utilisateurs exagèrent leurs expressions faciales, et tout simplement les contrôlent volontairement. Or ce ne sont ni ces expressions contrôlées, ni celles exagérées, auxquelles nous nous intéressons ici, mais au contraire nous souhaitons analyser les réactions spontanées, d'intensité modérée.

L'utilisation d'indices comportementaux spontanés a déjà été envisagée pour l'évaluation des Agents Conversationnels : une première étude (Cerrato & Ekeklint, 2004) suggère qu'il serait possible de mettre en relation le niveau de satisfaction utilisateurs (tels qu'ils l'auto-évaluent) avec certains indices prosodiques et avec la quantité de gestes de régulation des tours de parole face à l'Agent. Mais les expressions faciales autres que celles liées à un feedback ou à la régulation du dialogue ne sont pas considérées dans cette étude.

Morkes *et al.* (1998) rapportent un type d'expérience proche de celui qui nous intéresse : ces auteurs ont filmé des utilisateurs interagissent par messagerie instantanée avec un partenaire – ce partenaire

[38] Ces résultats montrent que les six émotions primaires (joie, tristesse, peur, surprise, dégoût, colère) sont reconnues de manière fiable, même lorsque acteurs et décodeurs sont de cultures différentes.

étant simulé de sorte que ses messages soient toujours les mêmes. Dans une condition, les utilisateurs pensent interagir avec un partenaire humain (communication Homme-Homme médiatisée) ; dans une seconde condition, ils pensent interagir avec un ordinateur (interaction Homme-Machine). L'expérience porte sur les réactions des utilisateurs à l'humour de leur partenaire : les résultats montrent que les utilisateurs réagissent moins (sourient et rient moins) à une même réplique lorsque le partenaire est un ordinateur que lorsque c'est un humain. Les commentaires sociaux, c'est-à-dire les marques de politesse et d'amicalité, sont également moins nombreux dans la condition du partenaire ordinateur que dans celle du partenaire humain. Les expressions faciales recueillies dans cette expérience sont simples (elles se limitent aux sourires), mais elles sont analysées de manière quantitative, ce qui permet de comparer différentes conditions et d'obtenir ce type de résultat extrêmement intéressant. Dans le cadre de l'évaluation d'Agents, Höök (2004) a également utilisé le comptage des rires et des sourires de l'utilisateur. Nous visons nous aussi une analyse quantitative des expressions faciales, mais ne se limitant pas aux sourires.

King (1997) a réalisé une telle étude : il a filmé le visage des utilisateurs à leur insu pendant la réalisation de tâches informatiques. Ses résultats montrent que les utilisateurs produisent significativement plus d'expressions faciales lorsqu'ils réalisent des tâches (création de graphiques, édition de textes, jeux…) que lors d'une condition contrôle de manipulation d'interface. La fréquence des expressions faciales est cependant toujours inférieure à une situation de communication Homme-Homme. De plus, les analyses montrent que face à l'ordinateur, les utilisateurs manifestent les six émotions primaires (joie, tristesse, peur, dégoût, surprise, colère). Par ailleurs, une expression faciale « cognitive » apparaît lorsque la réalisation de la tâche implique un traitement complexe de la part de l'utilisateur. Ces résultats tendent à montrer qu'il est possible d'utiliser les expressions faciales pour mener des analyses quantitatives en Interaction Homme-Machine. Cependant, King (1997) ne rapporte de données concernant la fiabilité des annotations et des interprétations sur son corpus.

La problématique principale ayant motivé l'expérience ci-dessous est donc la suivante : nous aimerions savoir à quel point la détection et l'interprétation des expressions faciales dans un contexte d'interaction naturel sont reproductibles et sont indépendantes de la subjectivité du codeur.

Pour mener cette expérience, nous envisagions initialement de constituer un corpus d'expressions faciales seules, c'est-à-dire en supprimant les indices verbaux et paralinguistiques qui les accompagnent généralement dans un contexte d'interaction naturel. Mais cela a tout de suite fait émerger une question : quelle est la contribution des indices auditifs dans le décodage des émotions ? Il apparaît que la perception des émotions dans le discours verbal a, elle aussi, été peu étudiée dans un contexte d'interaction naturelle, par opposition à un jeu d'acteurs (Devillers *et al.*, 2003).

Nous avons donc finalement conçu l'expérience de manière à pouvoir tester la fiabilité inter codeurs dans l'annotation des indices faciaux et verbaux séparément, puis la fiabilité intra codeurs entre les deux modalités d'expression des émotions (faciale et verbale). Dans un premier temps, c'est un corpus d'interaction Homme-Homme que nous avons étudié, avec l'objectif de répliquer ensuite l'expérience sur un corpus d'interaction Homme-Machine. La présente étude constitue donc une étape préliminaire à l'élaboration d'une méthode d'évaluation de systèmes tels que NICE, c'est-à-dire de systèmes dont le succès repose en partie sur les émotions (par exemple l'amusement) suscitées auprès des utilisateurs.

Méthode

Figure 82 : Exemples d'extraits du corpus.

Corpus. Le corpus a été constitué de 55 extraits vidéo provenant d'émissions télévisées (Figure 82), d'une durée de quelques secondes à une minute chacun. Ces extraits ont été sélectionnés selon les critères suivants :

- La présence du visage du locuteur dans le champ (suffisamment proche de la caméra pour que les codeurs puissent analyser les expressions faciales) ;
- Le focus sur une seule personne ;
- La présence de paroles ;
- Le fait que le locuteur parle français ;
- La bonne qualité du signal sonore (pas de recouvrements entre les paroles de plusieurs personnes, etc.) ;
- Le fait que les personnes soient inconnues (pas de célébrités dans les extraits) ;
- Le réalisme de la situation (ex : passants interrogés dans la rue ; pas de jeu d'acteurs) ;
- La présence d'un ou plusieurs événements émotionnels, même subtils.

Participants. Nous avons recruté deux annotateurs – naïfs quant à l'objectif du projet – pour réaliser l'analyse des événements émotionnels dans ce corpus. Un des annotateurs (**annotateur A**) était un homme de 25 ans, de personnalité intermédiaire sur l'échelle d'introversion/extraversion – score de 11 au test d'Eysenck (Eysenck & Eysenck, 1971). L'**annotateur B** était une annotatrice, âgée de 23 ans, de personnalité extravertie (score de 18 sur l'échelle d'Eysenck).

Segmentation modalité par modalité. Nous nous intéressons aux « événements émotionnels fins », c'est-à-dire aux émotions de faible intensité, apparaissant dans un contexte réaliste, et ne se limitant pas aux grandes catégories émotionnelles franches et exclusives (joie, peur, dégoût…). Dans un premier temps, l'occurrence des émotions a été étudiée séparément dans les expressions faciales et dans le comportement verbal des individus filmés. Nous avons donc généré des fichiers vidéo avec image noire (pour l'annotation de la bande son seule) et des fichiers vidéo sans son (pour l'annotation du corpus visuel) à partir du corpus initial.

Une des méthodes d'analyse des expressions faciales les plus connues est le système FACS – *Facial Action Coding System* (Ekman & Friesen, 1978) –, qui requiert un codage précis de la morphologie des muscles faciaux (ex : position des coins de la bouche, forme des sourcils). A l'inverse, nous avons choisi ici d'annoter les expressions faciales au niveau fonctionnel et non morphologique (Manstead, 1991 ; Steininger *et al.*, 2002) : dans ce cas, une expression faciale est définie par l'impression subjective qu'elle provoque chez le partenaire dans la communication (ici l'annotateur). La tâche des annotateurs consistait donc à détecter les événements émotionnels (quels que soient leur signification

et leur intensité) et à les délimiter dans le temps. Aucun critère autre que le jugement personnel des annotateurs n'était fourni pour décider de ce qu'était un événement émotionnel.

Pour le corpus audio, les annotateurs pouvaient se fonder sur l'ensemble des indices auditifs (lexicaux, sémantiques, syntaxiques et paralinguistiques tels que les hésitations ou l'intonation) pour décider si un événement émotionnel se produit, et quels en sont les bornes temporelles.

L'annotation des deux corpus (visuel et auditif) a été réalisée sous Anvil (Kipp, 2001). Les deux annotateurs ont travaillé indépendamment l'un de l'autre. L'un des deux a commencé par la segmentation des images ; l'autre par la segmentation du corpus audio. Puis les rôles ont été inversés. A l'intérieur des corpus audio et vidéo, les extraits se succédaient dans un ordre aléatoire. Chacune des deux sous-phases (annotation audio et vidéo) débutait par un entraînement de l'annotateur sur 5 extraits du corpus qui n'ont pas été analysés.

Mise en commun de la segmentation. Une fois le corpus audio et le corpus vidéo segmentés en totalité par les deux codeurs, leurs fichiers d'annotations ont été fusionnés. Les segments communs utilisés pour la suite du protocole ont été retenus en fonction des critères suivants :

- Pour le corpus vidéo, c'est l'intersection des segments qui a été retenue. Ce choix se justifie par le fait que les expressions faciales sont composées d'indices fins et de courte durée. Les intersections d'une durée inférieure à 100 ms n'ont cependant pas été retenues en raison de leur difficulté d'interprétation.
- Pour le corpus audio, c'est au contraire l'union des éléments annotés qui a été retenue pour la suite. La détection des émotions dans un discours verbal nécessite en effet un signal plus long et plus complet que pour les expressions faciales.

Labellisation des segments, modalité par modalité. Les deux annotateurs ont ensuite dû passer en revue sous Anvil, indépendamment l'un de l'autre, les segments communs pour chacun des deux corpus et réaliser les opérations suivantes :

- Attribuer un label à l'émotion : le choix des labels était laissé entièrement libre à l'annotateur.
- Attribuer un score de valence (sur une échelle en 7 points variant de « émotion très négative » à « émotion très positive ») ;
- Attribuer un score d'intensité (sur une échelle en 7 points variant de « très faible » à « très forte »).

Analyses

Les calculs suivants ont été réalisés (sous SPSS).

Accord inter juges (alpha de Cronbach) sur :

- Le nombre d'événements émotionnels dans le corpus vidéo,
- Le nombre d'événements émotionnels dans le corpus audio,
- La durée des événements émotionnels dans le corpus vidéo,
- La durée des événements émotionnels dans le corpus audio,
- L'échelle de valence dans le corpus vidéo,
- L'échelle d'intensité dans le corpus vidéo,
- L'échelle de valence dans le corpus audio,
- L'échelle d'intensité dans le corpus audio.

Calcul du pourcentage de recouvrement temporel (intersection / union) entre :

- Les annotations du corpus vidéo par les deux annotateurs,
- Les annotations du corpus audio par les deux annotateurs.

Fiabilité intra juges (alpha de Cronbach) sur :

- Le nombre d'événements émotionnels dans les corpus audio et vidéo,
- La durée des événements émotionnels dans les corpus audio et vidéo.

Résultats

Corpus vidéo.

Détection. L'annotateur A a détecté 267 événements émotionnels dans le corpus vidéo, pour une durée totale de 219 secondes ; l'annotateur B en a détecté 362, pour une durée de 366 secondes. Pour le nombre d'événements comme pour leur durée, l'accord inter juges est très élevé[39] : $\alpha = 0,879$ et $\alpha = 0,809$ respectivement. La durée d'intersection des annotations entre juges représente 44,8% de leur union ($\sigma = 20,8$). Cette valeur pourrait sembler faible à première vue, mais il faut l'interpréter en tenant compte de la différence initiale entre les deux annotateurs : la valeur de recouvrement temporel ne pouvait en effet excéder 59,8% puisque les segments de l'annotateur A représentaient 59,8% de la

[39] Une valeur d'alpha de Cronbach supérieure ou égale à 0,80 est considérée comme élevée.

durée totale des segments de l'annotateur B. Si le pourcentage d'intersection est calculé par rapport aux segments de l'annotateur A, il atteint 84,4% ($\sigma = 20,5$).

Interprétation. La mise en commun des annotations a abouti à un total de 260 événements émotionnels, couvrant une durée de 178 secondes. Sur ces annotations, l'accord inter juges sur la valence des émotions exprimées est $\alpha = 0,709$. L'accord inter juges sur l'intensité est $\alpha = 0,510$.

Corpus audio.

Détection. L'annotateur A a détecté 108 événements (277 secondes) dans la version audio du corpus, alors que l'annotateur B en a détecté 267 (312 secondes). L'accord inter juges sur le nombre d'événements est $\alpha = 0,400$. Sur la durée des événements, $\alpha = 0,839$. Le pourcentage d'intersection vaut 35,1% ($\sigma = 20,4$) par rapport à l'union des annotations et 54,6% ($\sigma = 30,4$) par rapport aux segments de l'annotateur A.

Interprétation. 164 événements communs ont été retenus, pour une durée totale de 427 secondes. Sur ceux-ci, l'accord inter juges sur la valence des émotions est $\alpha = 0,915$ et $\alpha = 0,865$ sur leur intensité.

Comparaison audio/vidéo. La fiabilité intra juge pour l'annotateur A est $\alpha = 0,171$ pour le nombre d'événements émotionnels dans les versions audio et vidéo et $\alpha = 0,834$ pour la durée de ces événements. Pour l'annotateur B, la fiabilité vaut $\alpha = 0,765$ pour le nombre et $\alpha = 0,898$ pour la durée des événements émotionnels.

Discussion

Le Tableau 23 récapitule l'ensemble des résultats obtenus à l'issue de cette expérience.

Corpus vidéo. Les mesures de fiabilité sur la **détection** des expressions faciales sont globalement bonnes, puisque l'accord inter juges sur leur nombre et sur leur durée est supérieur à 0,8. Nous observons cependant une différence entre les deux annotateurs sur le nombre absolu d'expressions faciales détectées : l'annotateur B a en effet détecté 35% plus d'expressions faciales que l'annotateur A dans le même corpus. L'annotateur B étant une femme, cette différence peut éventuellement être rapprochée d'observations antérieures selon lesquelles les femmes auraient de meilleures capacités à décoder les expressions faciales que les hommes (Feldman *et al.*, 1991). Compte tenu de cette différence initiale entre annotateurs, la mesure d'intersection entre les annotations est assez faible.

Mais si on la rapporte aux annotations du juge A, l'intersection représente plus de 80% de leur durée totale (voir l'explication du calcul dans le paragraphe précédent).

		Annotateur A	Annotateur B
Corpus Vidéo	Nombre	267	362
	α - Nombre	**0,879**	
	Durée	219 sec.	366 sec.
	α - Durée	**0,809**	
	Intersection	**84,4%**	
	α - Valence	0,709	
	α - Intensité	0,510	
Corpus Audio	Nombre	108	267
	α - Nombre	0,400	
	Durée	277 sec.	312 sec.
	α - Durée	**0,839**	
	Intersection	54,6%	
	α - Valence	**0,915**	
	α - Intensité	**0,865**	
Comparaison Audio/Vidéo	α - Nombre	0,171	0,765
	α - Durée	**0,834**	**0,898**

Tableau 23 : Récapitulatif des résultats. Les éléments en gras indiquent une fiabilité satisfaisante.

La fiabilité sur l'**interprétation** des expressions faciales est moins bonne : 0,7 pour l'accord inter juges sur la valence des émotions et seulement 0,5 pour l'accord inter juges sur leur intensité. Il est possible que le manque de fiabilité sur l'intensité provienne de ce que, dans ce corpus, il y avait peu d'expressions faciales de forte intensité : par conséquent l'un des annotateurs a pu juger les expressions de manière absolue (c'est-à-dire en mettant uniquement des notes d'intensité faible à modérée) et l'autre de manière relative (en utilisant toute l'amplitude des notes). Mais il est également possible que l'intensité soit une dimension difficile à estimer en soi. Un examen plus approfondi des données de cette expérience ou un recueil de données supplémentaire pourraient éclairer cette question.

Pour l'objectif qui est le nôtre, nous considérons toutefois que la dimension la plus importante est la valence : si l'utilisateur manifeste une réaction visible lors de l'interaction avec un système, le point crucial en termes d'évaluation sera de déterminer si c'est une réaction positive (amusement, intérêt…) ou négative (ironie, énervement…). Une difficulté supplémentaire dans le cas de l'évaluation d'un jeu

est que parfois une émotion négative (ex : la peur) peut être considérée comme un facteur positif pour l'application, un vecteur d'attractivité. Mais ce genre d'ambiguïté ne devrait pas se produire avec le système NICE, étant donnée la nature du scénario (conversationnel et orienté tâche, n'exploitant pas d'émotions comme la peur).

Corpus audio. Le corpus audio donne un profil de résultats opposé à celui du corpus vidéo. En effet, la fiabilité sur la **détection** est faible : 0,4 seulement sur le nombre d'événements détectés. L'annotateur B a cette fois détecté presque 2,5 fois plus d'événements que l'annotateur A. L'accord inter juges sur la durée des événements est supérieur à 0,8 mais l'intersection se représente que 55% de la durée des segments de l'annotateur A.

En revanche, la fiabilité sur l'**interprétation** des émotions verbales est très forte : supérieure à 0,9 pour la valence des émotions, et supérieure à 0,8 pour leur intensité.

Comparaison audio/vidéo. Globalement, les annotateurs ont détecté moins d'événements émotionnels dans la version audio du corpus que dans la version vidéo, mais ces événements ont une durée supérieure. Les analyses intra juges montrent que la valeur de fiabilité dépend des individus : avec l'annotateur B, la fiabilité de la détection sur les deux modalités est correcte, alors qu'elle est extrêmement faible avec l'annotateur A.

Conclusion

Cette expérience nous a fourni des résultats encourageants quant à la possibilité de détecter et d'interpréter les réactions émotionnelles naturelles (c'est-à-dire non jouées) de manière reproductible. Les résultats ont montré que la détection des émotions est plus fiable sur des indices visuels, et que leur interprétation est meilleure sur les indices verbaux. Une bonne stratégie pour l'annotation d'un corpus multimodal pourrait donc consister à se fonder principalement sur les expressions faciales (indices visuels) pour la phase de détection, et de s'aider des indices verbaux (lexicaux, sémantiques, paralinguistiques) pour interpréter les segments sélectionnés. Il est possible que l'efficacité des indices verbaux soit liée au fait que les annotations étaient plus longues dans ce cas : il faudra donc laisser la possibilité à l'annotateur de prendre en compte les indices verbaux légèrement avant et après l'expression faciale pour interpréter celle-ci.

Comme nous l'avons déjà évoqué plus haut, la dimension d'interprétation la plus importante pour notre objectif est la valence des émotions. Mais nous ne pouvons pas nous en contenter dans le cadre de l'évaluation de l'Interaction Homme-Machine : il est en effet important de distinguer les émotions

positives entre elles (ex : joie, attendrissement, intérêt…) et les émotions négatives entre elles (ex : ennui, colère, moquerie…), car les recommandations ou les solutions proposées ne seront pas les mêmes dans ces différents cas. Dans le contexte d'une évaluation, il sera donc important que l'attribution de labels aux événements annotés soit elle aussi fiable. Nous envisageons de tester cette fiabilité en analysant les labels insérés par les annotateurs dans la présente expérience. La première étape de ce travail consistera à catégoriser la multitude des labels recueillis (848 en tout). Une fois que des catégories auront été déterminées, nous pourrons calculer un accord inter juges (kappa de Cohen) sur leur utilisation.

Figure 83 : Une utilisatrice à deux moments de son interaction avec Cloddy Hans (système NICE).

Nous envisageons par la suite de renouveler cette expérience avec un corpus d'interaction Homme-Machine, et plus particulièrement un corpus recueilli avec des utilisateurs du système NICE (enregistrement vidéo lors des tests du prototype II, voir la Figure 83). Ce type de corpus présente plusieurs différences avec le précédent (extraits d'émissions télévisées) :

- Dans un corpus d'interaction Homme-Machine, on peut s'attendre à ce que les événements émotionnels soient plus ponctuels et donc plus facilement détectables : dans ce cas, les émotions sont en effet générées en temps réel, en réaction immédiate à un élément de l'interaction (une réplique ou une action de l'Agent). A l'inverse dans le corpus précédent, les émotions avaient une origine antérieure à l'interaction, ou une origine endogène. Elles étaient davantage prises dans un flux, ce qui a dû rendre leur détection plus délicate.
- Une autre différence avec le corpus précédent est que dans le système NICE, la parole de l'utilisateur est une modalité d'entrée pour le système. Or, il est possible que dans ce cas

l'utilisateur modifie, volontairement ou pas, les propriétés lexicales, syntaxiques et paralinguistiques de ses paroles (voir par exemple Oviatt & Adams, 2000, pour le cas des incorrections langagières). On peut donc craindre que les indices verbaux ne soient pas si riches en interaction Homme-Machine qu'en interaction Homme-Homme et que leur contribution à l'interprétation des événements émotionnels soit minorée.

A de multiples égards, il sera donc intéressant de comparer les résultats de l'expérience précédente avec l'analyse du corpus de test du prototype II de NICE : nous faisons, entre autres, l'hypothèse que la fiabilité sur la détection sera meilleure, mais que l'interprétation des événements émotionnels sera plus complexe. Nous disposerons cependant d'une information supplémentaire, d'importance pour l'interprétation : nous aurons en effet la possibilité, à tout moment, de relier une réaction émotionnelle avec l'événement qui l'a provoquée (enregistrement vidéo de l'écran synchronisé à l'enregistrement vidéo des expressions faciales). Cela nous donnera par exemple la possibilité, face à un sourire de l'utilisateur, de décider si celui-ci traduit de l'amusement (en réponse, par exemple, à un trait d'humour de l'Agent) ou s'il traduit au contraire une réaction ironique (suite, par exemple, à une réponse inappropriée).

Au final, les réactions émotionnelles identifiées de manière univoque dans le corpus NICE (qu'elles soient positives ou négatives) pourront donner lieu à des recommandations exploitables directement dans le projet, et sans doute également intéressantes dans le contexte plus large de la conception de systèmes interactifs ludiques et d'Agents Conversationnels Animés.

7.3. Conclusion

Les études réalisées dans le cadre du projet NICE nous ont amenés à entamer une réflexion sur les méthodes d'évaluation des Interactions Homme-Machine, et à proposer dans le présent chapitre deux axes de recherche vers la mise au point de nouvelles méthodes adaptées à un système tel que celui du projet NICE :

- Un de ces axes de recherche nous a été suggéré par les données recueillies : c'est en effet l'analyse de corpus multimodaux qui a fait émerger l'hypothèse que le comportement verbal et multimodal des utilisateurs en entrée était susceptible de refléter leur perception du système (et/ou de l'Agent).
- Un second axe de recherche a au contraire été imaginé pour répondre à un manque dans le projet NICE : nos partenaires et nous-mêmes disposons en effet de nombreuses mesures permettant d'évaluer chaque composant de l'architecture, ainsi que leur intégration au sein

d’un système unifié, et les capacités d’interaction qui en résultent. Par conséquent, le défi majeur pour notre consortium – presque entièrement issu du domaine de la recherche – était peut-être la conception et l’évaluation des qualités ludiques du système. Dans la seconde partie de ce chapitre, nous avons jeté les bases d’une méthode permettant d’évaluer ces aspects.

Les réflexions et propositions présentées dans ce chapitre sont préliminaires à différents travaux d’application et de validation, qui font partie des perspectives de cette thèse.

8. Conclusion générale

Cette thèse a été réalisée dans le cadre d'un projet de recherche européen – projet NICE – portant sur la conception d'un système interactif multimodal bidirectionnel mettant en scène des Agents Conversationnels Animés. Ce chapitre de conclusion a pour objectif de récapituler de manière synthétique les contributions apportées par ce travail de thèse, sachant que les publications qui en sont issues sont listées en Annexe 6.

Du fait de la bidirectionnalité de la communication multimodale, la thèse s'est organisée autour de deux pôles symétriques : la multimodalité en entrée avec les Agents et la multimodalité en sortie de la part des Agents. Du point de vue de la multimodalité en entrée, les travaux antérieurs relataient essentiellement le comportement des utilisateurs face à des interfaces non personnifiées, c'est-à-dire sans Agents. De plus, c'est principalement l'intégration temporelle des modalités qui était étudiée. Notre contribution à ce domaine de recherche a été d'étudier le **comportement multimodal des utilisateurs face à des Agents**, et dans un contexte de jeu, ce qui constitue un type d'application jamais étudié auparavant. Nous avons également comparé un groupe d'adultes avec un groupe d'enfants sur un même système multimodal, ce qui n'avait jamais été fait à notre connaissance. Enfin, nous avons analysé finement l'intégration sémantique des modalités, ce qui est également original par rapport à la littérature sur l'interaction Homme-Machine multimodale.

En outre, ces données comportementales nous ont permis de contribuer au développement d'un système qui est aujourd'hui fonctionnel. Certaines parties du système NICE, par exemple l'algorithme de fusion des entrées multimodales et son paramétrage temporel, reposent en effet sur nos observations du comportement des utilisateurs. En cela, ces travaux constituent également une étude de cas intéressante du point de vue du processus de conception employé, fondé en partie sur le recueil de données comportementales. Enfin, les analyses réalisées ont également fait apparaître des résultats

de portée plus générale, sur l'attitude de jeu et le comportement des utilisateurs face à un système comme celui du projet NICE. Pour mémoire, la conclusion du chapitre 5 dresse un bilan complet des contributions que nous avons apportées à la l'étude du comportement multimodal des utilisateurs face aux Agents.

La multimodalité bidirectionnelle reposant sur la notion de symétrie d'interaction, nous avons nous-mêmes cherché à respecter cette symétrie dans les travaux que nous avons réalisés. En effet, les connaissances acquises sur le comportement multimodal des utilisateurs, et plus généralement le comportement multimodal humain, ont fait émerger la problématique de la spécification du **comportement multimodal des Agents**. De ce point de vue, nous avons constaté que l'état de l'art était relativement bien fourni en ce qui concerne la génération et la combinaison de la parole et des gestes de la part des Agents ; en revanche, les données d'évaluation de ces comportements multimodaux étaient très lacunaires.

Nous nous sommes ainsi intéressés à une variable apparemment jamais expérimentée auparavant : la coopération sémantique entre paroles et gestes des Agents, et son influence dans un contexte pédagogique. Les résultats ont montré que la stratégie de redondance est manifestement la plus appropriée à cette situation. Elle permet en effet d'améliorer les performances de rappel de la part des utilisateurs et, résultat inattendu, elle améliore également leurs évaluations subjectives : les Agents redondants sont par exemple jugés plus sympathiques et de personnalité plus positive que les autres. Cette étude a donc abouti à des résultats tangibles sur l'effet des coopérations multimodales dans une situation proche d'un contexte pédagogique, résultats qui sont nouveaux dans le domaine des Agents, voire même dans le domaine de la pédagogie en général. Cette expérience a par ailleurs ouvert des perspectives pour compléter ces résultats par l'étude d'autres contextes d'interaction (voir le chapitre 6).

Une autre contribution de la thèse est d'ordre **méthodologique**, que ce soit au niveau de la conception des protocoles expérimentaux (ex : protocole en carré gréco-latin à mesures répétées du chapitre 6) ou de l'analyse des données (ex : annotation du comportement, statistiques uni- et multidimensionnelles). Rappelons que les méthodes et techniques employées ne sont pas nouvelles en soi, mais que c'est leur application à ce domaine de recherche qui est originale, et qui est susceptible de faire évoluer ledit domaine. Les résultats obtenus ont également ouvert des perspectives de méthodes d'évaluation innovantes de l'Interaction Homme-Machine, qui ont été présentées au chapitre 7. En effet, l'analyse des corpus comportementaux a permis de formuler une hypothèse sur l'exploitation des constructions multimodales des utilisateurs comme variable d'évaluation – hypothèse originale au vu de l'état de l'art que nous avons mené. De manière transversale dans tout le projet NICE, nous nous sommes

également interrogés sur l'évaluation des aspects ludiques de ce système. Ceci nous a conduit à étudier la possibilité de collecter et d'interpréter les expressions faciales des utilisateurs pendant l'interaction.

L'évaluation finale du système NICE nous permettra d'apporter des éléments supplémentaires de discussion sur ces deux propositions de méthodes d'évaluation. En retour, et de manière toujours cohérente avec la notion de symétrie d'interaction, les analyses des expressions faciales des utilisateurs seront peut-être utilisables dans le champ de la génération et de la perception des expressions faciales des Agents – domaine déjà largement documenté, mais qui offre tout de même de nombreuses perspectives de recherche. Ces perspectives constituent une manière de réintégrer la multimodalité en entrée et en sortie, issue naturelle à cette thèse où ces deux aspects ont été étudiés séparément afin de poser de premières bases.

Les Agents Conversationnels Multimodaux Bidirectionnels, par leurs aspects intégratifs et ambitieux, constituent un projet de recherche à long terme, dans lequel l'évaluation et l'expérimentation occupent une place décisive. D'une part, les enjeux (notamment industriels) qu'ils représentent à moyen terme font qu'ils nécessitent de valider scientifiquement les étapes de leur progression technologique, et d'enrichir ainsi leur conception. D'autre part, les Agents Conversationnels peuvent servir d'outils d'étude de la communication humaine, mais à condition souvent de mettre en œuvre une démarche expérimentale rigoureuse. Bien que le champ d'application de nos travaux soit restreint du fait des technologies que nous avons utilisées (ex : geste 2D en entrée, Agents 2D en sortie), nous espérons avoir contribué à ces deux aspects de la recherche sur les Agents Conversationnels.

Références bibliographiques

Abrilian, S., Buisine, S., Rendu, C., & Martin, J.C. (2002). Specifying cooperation between modalities in lifelike animated agents. Proceedings of PRICAI'2002 Workshop on Lifelike Animated Agents Tools, Affective Functions, and Applications, pp. 3-8.

Allbeck, J., & Badler, N. (2002). Toward representing agent behaviors modified by personality and emotion. Proceedings of AAMAS'2002 Workshop on Embodied Conversational Agents.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communication of the ACM*, 26, pp. 832-843.

Ando, H., Kitahara, Y., & Hataoka, N. (1994). Evaluation of multimodal interface using spoken language and pointing gesture on interior design system. Proceedings of ICSLP'94, pp. 567-570.

André, E., & Rist, T. (2000). Presenting through performing: On the use of multiple lifelike characters in knowledge-based presentation systems. Proceedings of IUI'2000, pp. 1-8.

André, E., Rist, T., & Müller, J. (1999). Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*, 13, pp. 415-448.

André, E., Rist, T., Van Mulken, S., Klesen, M., & Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied conversational agents*, pp. 220-255: MIT Press.

Bajeot, F. (2004). *Evaluation du comportement multimodal d'agents conversationnels.* Mémoire de Maîtrise de Psychologie, Université de Paris V.

Ball, G., & Breese, J. (2000). Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied conversational agents*, pp. 189-219: MIT Press.

Bastien, J.M.C., & Scapin, D.L. (2001). Evaluation des systèmes d'information et critères ergonomiques. In C. Kolski (Ed.), *Environnements évolués et évaluation de l'IHM*, pp. 53-79. Paris: Hermès.

Baylor, A.L. (2003). The split-persona effect with pedagogical agents. Proceedings of AAMAS'2003 workshop on Embodied Conversational Characters as Individuals.

Bell, L., Boye, J., Gustafson, J., & Wirén, M. (2000a). Modality convergence in a multimodal dialogue system. Proceedings of Götalog 2000, Workshop on the Semantics and Pragmatics of Dialogue, pp. 29-34.

Bell, L., Eklund, R., & Gustafson, J. (2000b). A comparison of disfluency distribution in a unimodal and a multimodal speech interface. Proceedings of ICSLP'2000, pp. 626-629.

Bell, L., & Gustafson, J. (1999). Interaction with an animated agent in a spoken dialogue system. Proceedings of Eurospeech'99, pp. 1143-1146.

Bellik, Y. (1995). *Interfaces multimodales : Concepts, modèles et architectures*. Thèse de Doctorat en Informatique, Université Paris XI.

Bellik, Y., & Teil, D. (1992). Les types de multimodalités. Proceedings of IHM'92.

Benoit, C., Martin, J.C., Pelachaud, C., Schomaker, L., & Suhm, B. (2000). Audio-visual and multimodal speech-based systems. In D. Gibbon, I. Mertins & R. Moore (Eds.), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*, pp. 102-203. Boston: Kluwer Academic Publishers.

Beun, R.J., de Vos, E., & Witteman, C. (2003). Embodied conversational agents: Effects on memory performance and anthropomorphisation. Proceedings of IVA'2003.

Bianchi-Berthouze, N., & Lisetti, C.L. (2002). Modeling multimodal expression of user's affective subjective experience. *User Modeling and User-Adapted Interaction*, 12, pp. 49-84.

Bickmore, T., & Cassell, J. (1999). Small talk and conversational storytelling in embodied interface agents. Proceedings of AAAI Symposium on Narrative Intelligence, pp. 87-92.

Bickmore, T., & Cassell, J. (2004). Social dialogue with embodied conversational agents. In J.v. Kuppevelt, L. Dybkjaer & N.O. Bernsen (Eds.), *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*. New York: Kluwer Academic Publishers.

Bolt, R.A. (1980). "Put-That-There": Voice and gesture at the graphics interface. *Computer Graphics*, 14, pp. 262-270.

Bourguet, M.L., & Ando, A. (1998). Synchronization of speech and hand gestures during multimodal human-computer interaction. Proceedings of CHI'98, pp. 241-242.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. Proceedings of ISSD'96, pp. 41-44.

Buisine, S., & Martin, J.C. (2002). *Projet RNRT - Télévision interactive. Phase 2 : Définition du cahier des charges fonctionnel*: Raport Intermédiaire du LIMSI-CNRS.

Caelen, J., & Bruandet, M.F. (2001). Interaction multimodale pour la recherche d'information. In C. Kolski (Ed.), *Environnements évolués et évaluation de l'IHM*, pp. 175-205. Paris: Hermès Science Publications.

Calbris, G., & Montredon, J. (1986). *Des gestes et des mots pour le dire*. Paris: Clé International.

Calbris, G., & Porcher, L. (1989). *Geste et communication*. Paris: Hatier.

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents*, pp. 1-27. Cambridge: MIT Press.

Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interface. *AI Magazine*, 22, pp. 67-83.

Cassell, J., & Bickmore, T. (2001). A relational agent: A model and implementation of building user trust. Proceedings of CHI'2001, pp. 396-403.

Cassell, J., Bickmore, T., Vilhjalmsson, H., & Yan, H. (2001a). More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14, pp. 55-64.

Cassell, J., McNeill, D., & McCullough, K.E. (1999a). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognition*, 7, pp. 1-33.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. Proceedings of SIGGRAPH'94, pp. 413-420.

Cassell, J., & Prevost, S. (1996). Distribution of semantic features across speech and gesture by humans and computers. Proceedings of Workshop on the Integration of Gesture in Language and Speech.

Cassell, J., & Stone, M. (1999). Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems. Proceedings of AAAI'99 Fall Symposium on Psychological Models of Communication in Collaborative Systems, pp. 34-42.

Cassell, J., Stone, M., & Yan, H. (2000). Coordination and context-dependence in the generation of embodied conversation. Proceedings of International Natural Language Generation Conference, pp. 171-178.

Cassell, J., & Thorisson, K. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13, pp. 519-538.

Cassell, J., Vilhjálmsson, H., & Bickmore, T. (2001b). BEAT: the Behavior Expression Animation Toolkit. Proceedings of SIGGRAPH '01, pp. 477-486.

Cassell, J., Vilhjalmsson, H., Chang, K., Bickmore, T., Campbell, L., & Yan, H. (1999b). Requirements for an architecture for embodied conversational characters. In D. Thalmann & N. Thalmann (Eds.), *Computer Animation and Simulation '99 (Eurographics series)*, pp. 109-120. Vienna: Springer Verlag.

Catinis, L., & Caelen, J. (1995). Analyse du comportement multimodal de l'usager humain dans une tâche de dessin. Proceedings of IHM'95, pp. 123-130.

Cerrato, L., & Ekeklint, S. (2004). Evaluating users' reactions to human-like interfaces. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents*, pp. 101-124: Kluwer Academic Publishers.

Cheyer, A., & Julia, L. (1995). Multimodal maps: An agent-based approach. Proceedings of CMC'95, pp. 103-113.

Chi, D., Costa, M., Zhao, L., & Badler, N. (2000). The EMOTE model for effort and shape. Proceedings of SIGGRAPH'2000, pp. 173-182.

Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. Proceedings of CHI'88, pp. 213-218.

Cohen, P.R., Dalrymple, M., Moran, D.B., Pereira, F.C.N., Sullivan, J.W., Gargan Jr, R.A., Schlossberg, J.L., & Tyler, S.W. (1989). Synergistic use of direct manipulation and natural language. Proceedings of CHI'89, pp. 227-234.

Cohen, P.R., Johnston, M., McGee, D.R., Oviatt, S.L., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). QuickSet: Multimodal interaction for distributed applications. Proceedings of Multimedia'97, pp. 31-40.

Collier, G. (1985). *Emotional Expression*. Hillsdale: Lawrence Erlbaum Associates.

Corradini, A., & Cohen, P.R. (2002). On the relationships among speech, gestures, and object manipulation in virtual environments: Initial evidence. Proceedings of CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems.

Cosnier, J., Coulon, J., Berrendonner, A., & Orecchioni, C. (1982). *Les voies du langage. Communications verbales gestuelles et animales*. Paris: Dunod.

Costantini, E., Burger, S., & Pianesi, F. (2002). NESPOLE!'s multilingual and multimodal corpus. Proceedings of LREC'2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation, pp. 165-170.

Coutaz, J., Nigay, L., Salber, D., Blandford, A.E., May, J., & Young, R.M.Y. (1995). Four easy pieces for assessing the usability of multimodal interaction. Proceedings of Interact'95, pp. 115-120.

Coutaz, J., Salber, D., Carraux, E., & Portolan, N. (1996). NEIMO, a multi-workstation usability lab for observing and analyzing multimodal interaction. Proceedings of CHI'96, pp. 402-403.

Cowell, A.J., & Stanney, K.M. (2003a). Embodiment and interaction guidelines for designing credible, trustworthy ECAs. Proceedings of IVA'2003.

Cowell, A.J., & Stanney, K.M. (2003b). On manipulating nonverbal interaction style to increase anthropomorphic computer character credibility. Proceedings of AAMAS'2003 Workshop on Embodied Conversational Characters as Individuals.

Craig, S.D., Gholson, B., & Driscoll, D. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*, 94, pp. 428-434.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies - Why and how. Proceedings of IUI'93, pp. 193-200.

Darves, C., & Oviatt, S.L. (2004). Talking to digital fish. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents*, pp. 271-292: Kluwer Academic Publishers.

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13, pp. 319-340.

De Angeli, A., Gerbino, W., Cassano, G., & Petrelli, D. (1998). Visual display, pointing, and natural language: The power of multimodal interaction. Proceedings of AVI'98, pp. 164-173.

Dehn, D.M., & Van Mulken, S. (2000). The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies*, 52, pp. 1-22.

Denda, A., Itoh, T., & Nakagawa, S. (1997). Evaluation of spoken dialogue system for a sightseeing guidance with multi-modal interface. Proceedings of IJCAI'97 Workshop on Intelligent Multimodal Systems.

Devillers, L., Vasilescu, I., & Mathon, C. (2003). Prosodic cues for perceptual emotion detection in task-oriented human-human corpus. Proceedings of ICPhS'2003.

Duncan, L., Brown, W., Esposito, C., Holmback, H., & Xue, P. (1999). Enhancing virtual maintenance environments with speech understanding. *Boeing M&CT TechNet*, pp.

Dybkjaer, L., & Bernsen, N.O. (2002). Natural interactivity resources - Data, annotation schemes and tools. Proceedings of LREC'02.

Ekman, P., & Friesen, W.V. (1978). *Facial Action Coding System (FACS). A technique for the measurement of facial action.* Palo Alto: Consulting Psychologists Press.

Elliott, C., Rickel, J., & Lester, J. (1999). Lifelike pedagogical agents and affective computing: An exploratory synthesis. In M. Wooldridge & M. Veloso (Eds.), *Artificial Intelligence Today*, pp. 195-212. Berlin: Springer Verlag.

Essa, I.A., & Pentland, A.P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, pp. 757-763.

Eysenck, H.J., & Eysenck, S.B.G. (1971). *Manuel de l'Inventaire de Personnalité d'Eysenck*. Paris: Les Editions du Centre de Psychologie Appliquée.

Feldman, R.S., Philippot, P., & Custrini, R.J. (1991). Social competence and nonverbal behaviour. In R.S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior*, pp. 329-350: Cambridge University Press.

Fitts, P.M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47, pp. 381-391.

Giese, M.A., & Poggio, T. (2003). Neural mechanisms for recognition of biological movements. *Nature Reviews*, 4, pp. 179-192.

Gillies, M., & Ballin, D. (2003). A model of interpersonal attitude and posture generation. Proceedings of IVA'2003.

Gogate, L.J., Bahrick, L.E., & Watson, J.D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, 71, pp. 878-894.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3, pp. 419-429.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology*, 91, pp. 720-730.

Graesser, A.C., Moreno, K., Marineau, J., Adcock, A., Olney, A., & Person, N. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? Proceedings of AIED'2003, pp. 47-54.

Granström, B., House, D., & Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. Proceedings of Speech Prosody'2002.

Gratch, J., Marsella, S., Egges, A., Eliëns, A., Isbister, K., Paiva, A., Rist, T., & ten Hagen, P. (2004). *Design criteria, techniques and case studies for creating and evaluating interactive experiences for virtual humans.* Working group on ECA's design parameters and aspects, Dagstuhl seminar on Evaluating Embodied Conversational Agents (organized by Z. Ruttkay, E. André, K. Höök, W. Lewis Johnson, C. Pelachaud).

Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., & Petajan, E. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17, pp. 54-63.

Hanna, L., Risden, K., & Alexander, K.J. (1997). Guidelines for usability testing with children. *Interactions*, 4, pp. 9-14.

Hartmann, B. (2002). *The EMOTE system as a plug-in for Maya.* Applied Senior Design Project - Final Report.

Hartmann, B., Mancini, M., Buisine, S., & Pelachaud, C. (soumis). Design and evaluation of expressive gesture synthesis for Embodied Conversational Agents.

Hartmann, B., Mancini, M., & Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. Proceedings of Computer Animation'2002.

Hauptmann, A.G. (1989). Speech and gestures for graphic image manipulation. Proceedings of CHI'89, pp. 241-245.

Heinlein, R.A. (1966). *The Moon is a harsh mistress* (Titre français: Révolte sur la Lune. Edition Opta): Orb Books Publisher.

Holzman, T.G. (1999). Computer-human interface solutions for emergency medical care. *Interactions*, 6, pp. 13-24.

Höök, K. (2004). User-centered design and evaluation of affective interfaces. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents*, pp. 127-160: Kluwer Academic Publishers.

Höysniemi, J., Hämäläinen, P., Turkki, L., & Rouvi, T. (2005). Children's intuitive gestures in vision-based action games. *Communications of the ACM*, 48, pp. 44-50.

Huls, C., & Bos, E. (1995). Studies into full integration of language and action. Proceedings of CMC'95, pp. 161-174.

Isbister, K., & Doyle, P. (2004). The blind men and the elephant revisited. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents.*, pp. 3-26: Kluwer Academic Publishers.

Johnson, D., & Wiles, J. (2003). Effective affective user interface design in games. *Ergonomics*, 46, pp. 1332-1345.

Johnson, W.L., Rickel, J., & Lester, J. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, pp. 47-78.

Johnson, W.L., Shaw, E., Marshall, A., & LaBore, C. (2003). Evolution of user interaction: The case of agent Adele. Proceedings of IUI'2003, pp. 93-100.

Kaiser, S., Wehrle, T., & Schmidt, S. (1998). Emotional episodes, facial expressions, and reported feelings in human-computer interactions. Proceedings of Conference of the International Society for Research on Emotions, pp. 82-86.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13, pp. 351–371.

Keeker, K., Pagulayan, R., Sykes, J., & Lazzaro, N. (2004). The untapped world of video games. Proceedings of CHI'2004, pp. 1610-1611.

Kehler, A. (2000). Cognitive status and form of reference in multimodal Human-Computer Interaction. Proceedings of AAAI'2000.

Kehler, A., Martin, J.C., Cheyer, A., Julia, L., Hobbs, J., & Bear, J. (1998). On representing salience and reference in multimodal human-computer interaction. Proceedings of AAAI'98 Workshop on Representations for Multi-modal Human-Computer Interaction.

Kendon, A. (1988). How gestures can become like words. In F. Payatos (Ed.), *Cross-cultural perspectives in nonverbal communication*, pp. 131-141. Toronto: Hogrefe.

Keppel, G. (1991). *Design and Analysis: A researcher's handbook -- 3rd Edition*: Prentice-Hall, Inc.

Kimura, D. (1999). *Sex and Cognition*. Cambridge: MIT Press.

King, W.J. (1997). Human-computer dyads? A survey of nonverbal behavior in human-computer systems. Proceedings of PUI'97 Workshop.

Kipp, M. (2001). Anvil - A generic annotation tool for multimodal dialogue. Proceedings of Eurospeech'2001, pp. 1367-1370.

Kipp, M. (2004). *Gesture generation by imitation. From human behavior to computer character animation.* Doctoral dissertation, Saarbrücken University.

Knapp, M.L. (2002). *Nonverbal communication in human interaction*: Wadsworth - Thomson Learning.

Koda, T., & Maes, P. (1996). Agents with faces: The effects of personification of agents. Proceedings of HCI'96, pp. 98-103.

Kohar, H., & Ginn, I. (1997). Mediators: Guides through online TV services. Proceedings of CHI'97 Electronic Publications.

Kolski, C. (1993). *Ingénierie des interfaces homme-machine. Conception et évaluation*. Paris: Hermès.

Koons, D.B., & Sparrell, C.J. (1994). ICONIC: Speech and depictive gestures at the Human-Machine Interface. Proceedings of CHI'94, pp. 453-454.

Kopp, S., Sowa, T., & Wachsmuth, I. (2004a). Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In A. Camurri & G. Volpe (Eds.), *Gesture-Based Communication in Human-Computer Interaction*, pp. 436-447. Berlin: Springer-Verlag.

Kopp, S., Tepper, P., & Cassell, J. (2004b). Towards integrated microplanning of language and iconic gesture for multimodal output. Proceedings of ICMI'2004, pp. 97-104.

Krahmer, E., & Swerts, M. (2004). More about brows. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents*, pp. 191-216: Kluwer Academic Publishers.

Krämer, N.C., Tietz, B., & Bente, G. (2003). Effects of embodied interface agents and their gestural activity. Proceedings of IVA'2003.

Lazzaro, N., & Keeker, K. (2004). What's my method? A game show on games. Proceedings of CHI'2004, pp. 1093-1094.

Le Quiniou, C. (2004). *Étude du comportement multimodal d'humains en interaction avec des Agents Conversationnels.* Mémoire de Maîtrise de Psychologie, Université de Paris V.

Lee, E.J., & Nass, C. (1999). Effects of the form of representation and number of computer agents on conformity. Proceedings of CHI'99, pp. 238-239.

Lester, J., Converse, S., Kahler, S., Barlow, T., Stone, B., & Bhogal, R. (1997a). The Persona effect: Affective impact of animated pedagogical agents. Proceedings of CHI'97, pp. 359-366.

Lester, J., & Stone, B. (1997). Increasing believability in animated pedagogical agents. Proceedings of International Conference on Autonomous Agents, pp. 16-21.

Lester, J., Towns, S., Callaway, C., Voerman, J., & FitzGerald, P. (2000). Deictic and emotive communication in animated pedagogical agents. In J. Cassell, S. Prevost, J. Sullivan & E. Churchill (Eds.), *Embodied Conversational Agents*, pp. 123-154. Cambridge: MIT Press.

Lester, J., Towns, S., & FitzGerald, P. (1999a). Achieving affective impact: Visual emotive communication in lifelike pedagogical agents. *International Journal of Artificial Intelligence in Education*, 10, pp. 278-291.

Lester, J., Voerman, J., Towns, S., & Callaway, C. (1997b). Cosmo: A life-like Animated Pedagogical Agent with deictic believability. Proceedings of IJCAI '97 Workshop on Animated Interface Agents: Making Them Intelligent, pp. 61-69.

Lester, J., Voerman, J., Towns, S., & Callaway, C. (1999b). Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13, pp. 383-414.

Lester, J., Zettlemoyer, L., Grégoire, J., & Bares, W. (1999c). Explanatory lifelike avatars: Performing user-centered tasks in 3D learning environments. Proceedings of International Conference on Autonomous Agents, pp. 24-31.

Lewis, J.R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, pp. 57-78.

Lin, H.X., Choong, Y.Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, 16, pp. 267-278.

Lyons, J. (1977). *Semantics* (Vol. 2). London: Cambridge University Press.

MacFarlane, S., Read, J., Höysniemi, J., & Markopoulos, P. (2003). *Evaluating interactive products for and with children.* Tutorial Notes, Interact'2003 Conference.

Manstead, A.S.R. (1991). Expressiveness as an individual difference. In R.S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior*, pp. 385-328: Cambridge University Press.

Martin, J.C. (1995). *Coopération entre modalités et liage par synchronie dans les interfaces multimodales.* Thèse de Doctorat en Informatique et Réseaux, Ecole Nationale Supérieure des Télécommunications.

Martin, J.C., & Béroule, D. (1993). Types et buts de coopération entre modalités. Proceedings of IHM'93.

Martin, J.C., Den Os, E., Kuhnlein, P., Boves, L., Paggio, P., & Catizone, R. (2004). *Proceedings of the Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*: In conjunction with LREC'2004.

Martin, J.C., Grimard, S., & Alexandri, K. (2001). On the annotation of the multimodal behavior and computation of cooperation between modalities. Proceedings of International Conference on Autonomous Agents Workshop on Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents, pp. 1-7.

Massaro, D.W. (2003). A computer-animated tutor for spoken and written language learning. Proceedings of ICMI'2003, pp. 172-175.

Maybury, M., & Martin, J.C. (2002). *Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*: In conjunction with LREC'2002.

McBreen, H., Anderson, J., & Jack, M. (2001). Evaluating 3D embodied conversational agents in contrasting VRML retail applications. Proceedings of International Conference on Autonomous Agents Workshop on Multimodal Communication and Context in Embodied Agents, pp. 83-87.

McBreen, H., & Jack, M. (2000). Empirical evaluation of animated agents in a multi-modal e-retail application. Proceedings of AAAI Symposium on Socially Intelligent Agents.

McBreen, H., & Jack, M. (2001). Evaluating humanoid synthetic agents in e-retail applications. *IEEE Transactions on Systems, Man and Cybernetics*, 31, pp. 394-405.

McNeill, D. (1992). *Hand and Mind*: University of Chigago Press.

McNeill, D. (1995). Relationships between language and motor action revisited. In M.E. Landsberg (Ed.), *Syntactic iconicity and linguistic freezes: The human dimension*: Mouton Publisher.

Mignot, C., Valot, C., & Carbonell, N. (1993). An experimental study of future 'natural' multimodal Human-Computer Interaction. Proceedings of InterCHI'93, pp. 67-68.

Moreno, K.N., Klettke, B., Nibbaragandla, K., & Graesser, A.C. (2002). Perceived characteristics and pedagogical efficacy of animated conversational agents. Proceedings of ITS'2002, pp. 963-971.

Moreno, R., & Mayer, R.E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, 94, pp. 156-163.

Moreno, R., Mayer, R.E., & Lester, J. (2000). Life-like pedagogical agents in constructivist multimedia environments: Cognitive consequences of their interaction. Proceedings of ED-MEDIA'2000, pp. 741-746.

Moreno, R., Mayer, R.E., Spires, H., & Lester, J. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, 19, pp. 177-213.

Mori, J., Prendinger, H., & Ishzuka, M. (2003). Evaluation of an embodied conversational agent with affective behavior. Proceedings of AAMAS'2003 workshop on Embodied Conversational Characters as Individuals.

Morkes, J., Kernal, H., & Nass, C. (1998). Humor in task-oriented computer-mediated communication and human-computer interaction. Proceedings of CHI'98, pp. 215-216.

Moundridou, M., & Virvou, M. (2001). Evaluating the impact of interface agents in an intelligent tutoring systems authoring tool. Proceedings of PC-HCI'2001 Advances in Human-Computer Interaction, pp. 371-376.

Myers, J.L. (1979). *Fundamentals of Experimental Design -- 3rd Edition*. Boston: Allyn & Bacon, Inc.

Nagao, K., & Takeuchi, A. (1994a). Social interaction: Multimodal conversation with social Agents. Proceedings of AAAI'94, pp. 22-28.

Nagao, K., & Takeuchi, A. (1994b). Speech dialogue with facial displays: Multimodal human-computer conversation. Proceedings of ACL'94, pp. 102-109.

Najm, M. (2004). *Evaluation du comportement multimodal d'Agents Conversationnels.* Mémoire de Maîtrise de Psychologie, Université de Paris V.

Nielsen, J. (1993). *Usability Engineering*: Academic Press.

Nigay, L., & Coutaz, J. (1993). Design space for multimodal systems: Concurrent processing and data fusion. Proceedings of InterCHI'93, pp. 172-178.

Oviatt, S.L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9, pp. 19-35.

Oviatt, S.L. (1996a). Multimodal interfaces for dynamic interactive maps. Proceedings of CHI '96, pp. 95-102.

Oviatt, S.L. (1996b). User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimedia*, 3, pp. 26-35.

Oviatt, S.L. (1999a). Mutual disambiguisation of recognition errors in a multimodal architecture. Proceedings of CHI'99, pp. 576-583.

Oviatt, S.L. (1999b). Ten myths of multimodal interaction. *Communications of the ACM*, 42, pp. 74-81.

Oviatt, S.L., & Adams, B. (2000). Designing and Evaluating Conversational Interfaces with Animated Characters. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents*, pp. 319-343. Cambridge: MIT Press.

Oviatt, S.L., Bernard, J., & Levow, G.A. (1998). Linguistic adaptation during error resolution with spoken and multimodal systems. *Language and Speech*, 41, pp. 419-442.

Oviatt, S.L., Cohen, P.R., Fong, M.W., & Frank, M.P. (1992). A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. Proceedings of ICSLP'92, pp. 1351-1354.

Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., & Ferro, D. (2000). Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions. *Human Computer Interaction*, 15, pp. 263-322.

Oviatt, S.L., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally? Cognitive load and multimodal communication patterns. Proceedings of ICMI'2004, pp. 129-136.

Oviatt, S.L., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., & Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during Human-Computer Interaction. Proceedings of ICMI'03, pp. 44-51.

Oviatt, S.L., De Angeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. Proceedings of CHI '97, pp. 415-422.

Pagulayan, R., Keeker, K., Wixon, D., Romero, R.L., & Fuller, T. (2002). User-centered design in games. In J.A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications archive*, pp. 883 - 906. Mahwah: Lawrence Erlbaum Associates, Inc.

Pelachaud, C., & Bilvi, M. (2003). Modelling gaze behaviour for conversational agents. Proceedings of IVA'2003.

Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., & Poggi, I. (2002). Embodied contextual agent in information delivering application. Proceedings of AAMAS'2002, pp. 758-765.

Perlman, G. (1997). Practical Usability Evaluation. Based in part on Nielsen's 1993 Heuristics and Norman's 1990 Principles. http://www.acm.org/~perlman/question.cgi?form=PHUE

Petrelli, D., De Angeli, A., Gerbino, W., & Cassano, G. (1997). Referring in multimodal systems: The importance of user expertise and system features. Proceedings of ACL/EACL'97 Workshop on Referring phenomena in a multimedia context and their computational treatment.

Piwek, P., & Beun, R.J. (2002). Multimodal referential acts in a dialogue game: From empirical investigations to algorithms. Proceedings of International Workshop on Information Presentation and Natural Multimodal Dialogue.

Poggi, I. (2002). From a typology of gestures to a procedure for gesture production. In I. Wachsmuth & T. Sowa (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, pp. 158-168. Berlin: Springer.

Poggi, I., Pelachaud, C., & de Rosis, F. (2000). Eye Communication in a Conversational 3D Synthetic Agent. *The European Journal on Artificial Intelligence*, 13, pp. 169-181.

Prada, R., Vala, M., Paiva, A., Höök, K., & Bullock, A. (2003). FantasyA - The duel of emotions. Proceedings of IVA'2003, pp. 62-66.

Prendinger, H., Mayer, S., Mori, J., & Ishizuka, M. (2003). Persona effect revisited: Using bio-signals to measure and reflect the impact of character-based interfaces. Proceedings of IVA'2003.

Qvarfordt, P., & Jönsson, A. (1998). The efficiency of a multimodal timetable information system for users with different domain knowledge. Proceedings of Second Symposium on Multimodal Communication.

Reeves, B., & Nass, C. (1996). *The Media Equation. How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.

Rickel, J., & Johnson, W.L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13, pp. 343-382.

Rickenberg, R., & Reeves, B. (2000). The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. Proceedings of CHI'2000, pp. 49-56.

Rimé, B. (1985). The growing field of nonverbal behavior: A review of twelve books on nonverbal behavior and nonverbal communication. *European Journal of Social Psychology*, 15, pp. 231-247.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In R.S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior*, pp. 239-284: Cambridge University Press.

Robbe, S. (1998). An empirical study of speech and gesture interaction: Toward the definition of ergonomic design guidelines. Proceedings of CHI'98, pp. 349-350.

Robbe-Reiter, S., Carbonell, N., & Dauchy, P. (2000). Expression constraints in multimodal human-computer interaction. Proceedings of IUI'2000, pp. 225-228.

Rugelbak, J., & Hamnes, K. (2003). Multimodal interaction - Will users tap and speak simultaneously? *Telektronikk*, 99, pp. 118-124.

Ruttkay, Z., Dormann, C., & Noot, H. (2002). Evaluating ECAs - What and how? Proceedings of AAMAS'2002 Workshop on Embodied conversational agents - let's specify and evaluate them!, pp. 168-175.

Ruttkay, Z., & Pelachaud, C. (2002). Exercises of style for virtual humans. Proceedings of AISB'02 Convention Symposium on Animating Expressive Characters for Social Interactions.

Ruttkay, Z., & Pelachaud, C. (2004). *From brows to trust: Evaluating Embodied Conversational Agents*: Kluwer Academic Publishers.

Ryokai, K., Vaucelle, C., & Cassell, J. (2003). Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning*, 19, pp. 195-208.

Scherer, K.R. (1984). Les fonctions des signes non verbaux dans la communication. In J. Cosnier & A. Brossard (Eds.), *La communication non verbale*, pp. 71-100: Delachaux & Niestlé.

Senach, B. (1993). L'évaluation ergonomique des interfaces homme-machine. In J.C. Sperandio (Ed.), *L'ergonomie dans la conception des projets informatiques*, pp. 69-122. Toulouse: Octares.

Seto, S., Kanazawa, H., Shinshi, H., & Takebayashi, Y. (1994). Spontaneous speech dialogue system TOSBURG II and its evaluation. *Speech Communication*, 15, pp. 341-353.

Shaw, E., Ganeshan, R., Johnson, W.L., & Millar, D. (1999). Building a case for agent-assisted learning as a catalyst for curriculum reform in medical education. Proceedings of AIED'99, pp. 509-516.

Siroux, J., Guyomard, M., Multon, F., & Remondeau, C. (1995). Oral and gestural activities of the users in the GEORAL system. Proceedings of Workshop on Intelligence and Multimodality in Multimedia Interfaces: Research and Applications.

Steininger, S., Rabold, S., Dioubina, O., & Schiel, F. (2002). Development of the user-state conventions for the multimodal corpus in Smartkom. Proceedings of LREC'2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation.

Stone, B., & Lester, J. (1996). Dynamically sequencing an animated pedagogical agent. Proceedings of National Conference on Artificial Intelligence, pp. 424-431.

Sturm, J., Bakx, I., Cranen, B., Terken, J., & Wang, F. (2002). Usability evaluation of a Dutch multimodal system for train timetable information. Proceedings of LREC'2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation, pp. 255-261.

Thorisson, K., Koons, D.B., & Bolt, R.A. (1992). Multi-modal natural dialogue. Proceedings of CHI'92, pp. 653-654.

Towns, S., FitzGerald, P., & Lester, J. (1998). Visual emotive communication in lifelike pedagogical agents. Proceedings of ITS'98, pp. 474-483.

Trafton, J.G., Wauchope, K., & Stroup, J. (1997). Errors and usability of natural language in a multimodal system. Proceedings of IJCAI'97 Workshop on Intelligent multimodal systems.

Turk, M., & Robertson, M. (2000). Perceptual user interfaces (introduction). *Communications of the ACM - Special issue on Perceptual User Interfaces*, 43, pp. 32-34.

Van Mulken, S., André, E., & Müller, J. (1998). The persona effect: How substantial is it? Proceedings of HCI'98, pp. 53-66.

Vo, M.T., & Waibel, A. (1993). A multi-modal human-computer interface: Combination of gesture and speech recognition. Proceedings of InterCHI'93, pp. 69-70.

Wahlster, W., Reithinger, N., & Blocher, A. (2001). SmartKom: multimodal communication with a life-like character. Proceedings of Eurospeech'2001.

Walker, J.H., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. Proceedings of CHI'94, pp. 85-91.

Wolff, M. (2003). Apports de l'analyse géométrique des données pour la modélisation de l'activité. In J.C. Sperandio & M. Wolff (Eds.), *Formalismes de modélisation pour l'analyse du travail et l'ergonomie*, pp. 195-227. Paris: Presses Universitaires de France.

Wonish, D., & Cooper, G. (2002). Interface Agents: preferred appearance characteristics based upon context. Proceedings of HF'2002 Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges.

Woods, S., Hall, L., Sobral, D., Dautenhahn, K., & Wolke, D. (2003). Animated characters in bullying intervention. Proceedings of IVA'2003.

Xiao, B., Girand, C., & Oviatt, S.L. (2002). Multimodal integration patterns in children. Proceedings of ICSLP'2002, pp. 629-632.

Xiao, B., Lunsford, R., Coulston, R., Wesson, M., & Oviatt, S.L. (2003). Modelling multimodal integration patterns and performance in seniors: Toward adaptative processing of individual differences. Proceedings of ICMI'2003, pp. 265-272.

Zanello, M.L. (1997). *Contribution à la connaissance sur l'utilisation et la gestion de l'interface multimodale*. Thèse de Doctorat en Sciences Cognitives, Université de Grenoble I.

Annexes

Annexe 1 : Questionnaire pour l'évaluation d'un guide de programmes télévisés multimodal.....251

Annexe 2 : Protocole d'accord parental pour l'enregistrement vidéo des expérimentations 253

Annexe 3 : Questionnaire pour l'évaluation du Magicien d'Oz avec Agents 2D 255

Annexe 4 : Plan expérimental utilisé pour l'évaluation répliquée des stratégies multimodales des Agents : Carré gréco-latin à mesures répétées.....257

Annexe 5 : Questionnaire pour l'évaluation répliquée des stratégies multimodales des Agents 259

Annexe 6 : Publications issues de la thèse 261

Annexe 1 :
Questionnaire pour l'évaluation d'un guide de programmes télévisés multimodal

Questionnaire

Date : .....
Sujet : .....

Pour chaque question, vous devez indiquer, en plaçant une croix sur la ligne, où vous vous situez entre les deux extrêmes proposés.

Interaction par la parole uniquement

1. Impression générale :

difficile	———	facile
inefficace	———	efficace
frustrant	———	satisfaisant
désagréable	———	agréable
non intuitif	———	intuitif

2. Apprentissage du fonctionnement :

difficile	———	facile

3. Le dialogue est simple et naturel :

pas d'accord	———	d'accord

Remarques, commentaires :

.....

.....

Interaction par le stylo uniquement

4. Impression générale :

difficile	———	facile
inefficace	———	efficace
frustrant	———	satisfaisant
désagréable	———	agréable
non intuitif	———	intuitif

5. Apprentissage du fonctionnement :

difficile	———	facile

6. L'utilisation du système demande beaucoup de précision :

pas d'accord	———	d'accord

Remarques, commentaires :

.....

.....

Interaction à la fois par la parole et le stylo

7. Impression générale :

difficile	⊢——————⊣	facile
inefficace	⊢——————⊣	efficace
frustrant	⊢——————⊣	satisfaisant
désagréable	⊢——————⊣	agréable
non intuitif	⊢——————⊣	intuitif

8. Apprentissage du fonctionnement :

difficile	⊢——————⊣	facile

Remarques, commentaires :

.....

.....

9. Type d'interaction préférée *(numéroter par ordre de préférence)* :

..... interaction par la parole
..... interaction par le stylo
..... interaction à la fois par la parole et le stylo

Fonctionnement de la recherche de programme

10. J'ai facilement trouvé l'information que je cherchais :

pas d'accord	⊢——————⊣	d'accord

11. La correction des erreurs est :

difficile	⊢——————⊣	facile

12. Ce système est accessible à tous, quel que soit le niveau de compétences :

pas d'accord	⊢——————⊣	d'accord

13. Ce système est utile :

pas d'accord	⊢——————⊣	d'accord

14. Ce système permet de gagner du temps :

pas d'accord	⊢——————⊣	d'accord

15. Ce système possède toutes les fonctions correspondant à mes attentes :

pas d'accord	⊢——————⊣	d'accord

Remarques, commentaires :

.....

.....

Profil de l'utilisateur :

Age : .....

Sexe : .....

Profession : .....

Annexe 2 :
Protocole d'accord parental pour l'enregistrement vidéo des expérimentations

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS)
BP 133, 91403 Orsay Cedex, FRANCE http://www.limsi.fr

PROTOCOLE D'ACCORD

Dans le cadre du programme de recherche IST (**I**nformation **S**ociety **T**echnologies, pour Société des Technologies de l'Information) de la Commission européenne, un contrat de recherche intitulé « IST-NICE » (NICE signifiant : **N**atural **I**nteractive **C**ommunication for **E**dutainment, en français Communication Interactive Naturelle ludo-Educative) a été signé entre le CNRS et l'Union européenne.

M. Jean-Claude Martin, Maître de conférence à l'Université de Paris VIII, est responsable scientifique de ce contrat de recherche pour les travaux qui sont menés au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), laboratoire du CNRS.

Plus précisément, l'objet de cette étude consiste à observer le comportement d'enfants, d'adolescents et d'adultes face à un jeu vidéo dans lequel ils peuvent communiquer avec des personnages par l'intermédiaire de la parole et du geste. Mlle Stéphanie Buisine, recrutée au CNRS dans le cadre de ce contrat de recherche « IST- NICE », est chargée de recueillir ces données sur un support audio-visuel. Le cadrage qui a été adopté ne permet pas d'identifier les personnes.

T.S.V.P.

AUTORISATION PARENTALE

Je, soussigné(e)

Nom, Prénom : .....

Adresse *(facultatif)* : .....

.....

N° de téléphone *(facultatif)* : .....

père, mère, tuteur légal[1], de l'enfant :

Nom, Prénom : .....

(Il ne sera effectué aucun traitement de ces données nominatives.)

- autorise mon fils / ma fille[1] à participer à l'étude organisée par le LIMSI-CNRS et coordonnée par M. Jean-Claude Martin dans le cadre du projet européen IST-NICE.
- accepte que mon fils / ma fille[1] soit filmé(e).
- déclare céder tous mes droits sur l'enregistrement vidéo effectué.

J'ai la possibilité, préalablement au passage de mon fils / ma fille, de participer moi-même à l'étude :

- ☐ Je souhaite participer à cette étude. J'accepte d'être filmé(e) et de céder tous mes droits sur l'enregistrement vidéo.
- ☐ Je ne souhaite pas participer à cette étude.

Les données qui concernent mon fils / ma fille ou moi-même ne seront pas nominatives et resteront strictement confidentielles. La publication des résultats de la recherche ne comportera aucun résultat individuel. Ces données ne donneront pas lieu à une exploitation commerciale ; elles seront utilisées à des fins de recherche dans le cadre du projet IST-NICE et pourront également être utilisées à des fins d'enseignement.

Fait à ……, le ……,

Signature (précédée de la mention « lu et approuvé »).

Pour tout renseignement, contacter
Jean-Claude Martin, martin@limsi.fr, Tel. : 01.69.85.81.04.

[1] Rayer les mentions inutiles.

Annexe 3 :
Questionnaire pour l'évaluation du Magicien d'Oz avec Agents 2D

Questionnaire

Date : .....
Sujet : .....
Expérimentateur : .....

Pour chaque question vous devez indiquer, en plaçant une croix sur la ligne, où vous vous situez entre les deux extrêmes proposés.

Scénario 1

L'interaction avec le système est :

difficile ☹	⊢——————⊣	☺ facile
inefficace ☹	⊢——————⊣	☺ efficace
désagréable ☹	⊢——————⊣	☺ agréable
difficile à apprendre ☹	⊢——————⊣	☺ facile à apprendre

Scénario 2

L'interaction avec le système est :

difficile ☹	⊢——————⊣	☺ facile
inefficace ☹	⊢——————⊣	☺ efficace
désagréable ☹	⊢——————⊣	☺ agréable
difficile à apprendre ☹	⊢——————⊣	☺ facile à apprendre

Type d'interaction préférée :

☐ interaction par la parole uniquement
☐ interaction à la fois par le geste et la parole

Remarques, commentaires :

.....

.....

.....

.....

Profil de l'utilisateur :

Age : .....
Sexe : .....
Profession : .....

Niveau en informatique : peu expérimenté(e) ☐
intermédiaire ☐
expert(e) ☐

Annexe 4 :
Plan expérimental utilisé pour l'évaluation répliquée des stratégies multimodales des Agents : Carré gréco-latin à mesures répétées

Variables Indépendantes :

- Stratégie multimodale : redondante, complémentaire, spécialisée par la parole {Red ; Comp ; Spé}.
- Apparence de l'agent : {Léa ; Marco ; Jules}.
- Objet d'explication : télécommande, logiciel, photocopieur {TVc ; Log ; Pht}.
- Sexe des sujets : {Masculin ; Féminin}.

	Léa	Marco	Jules
A	Red – TVc	Comp – Log	Spé – Pht
B	Comp – Pht	Spé – TVc	Red – Log
C	Spé – Log	Red – Pht	Comp – TVc

	Marco	Jules	Léa
D	Red – TVc	Comp – Log	Spé – Pht
E	Comp – Pht	Spé – TVc	Red – Log
F	Spé – Log	Red – Pht	Comp – TVc

	Jules	Léa	Marco
G	Red – TVc	Comp – Log	Spé – Pht
H	Comp – Pht	Spé – TVc	Red – Log
I	Spé – Log	Red – Pht	Comp – TVc

Chaque utilisateur a été affecté à l'un des 9 « groupes-passation » (A à I, variable technique) et a été soumis aux trois conditions expérimentales de la ligne correspondante (en respectant leur ordre d'apparition). Les groupes-passation ont été constitués de 12 utilisateurs parmi lesquels la variable Sexe a été équilibrée.

Annexe 5 :
Questionnaire pour l'évaluation répliquée des stratégies multimodales des Agents

Questionnaire

Sujet : …… Date : ……

Pour les questions 1 à 3, vous devez classer les trois personnages (n°1 – n°2 – n°3).
Pour toutes les questions, merci de commenter au maximum vos réponses.

❶ Quel personnage donne le mieux les explications ? (n°1 : celui qui explique le mieux n°2 : intermédiaire n°3 : celui qui explique le moins bien)			
Expliquez :			
❷ Quel est le personnage le plus sympathique ?			
Expliquez :			
❸ Quel est le personnage le plus expressif ?			
Expliquez :			
❹ Les trois personnages avaient-ils la même personnalité ?	◻ Oui ◻ Non		
❺ Définissez en un mot la personnalité de chacun.			
Expliquez :			
❻ Si votre ordinateur tombait en panne, suivriez-vous les instructions de ces personnages pour le dépanner ?	◻ Oui ◻ Non	◻ Oui ◻ Non	◻ Oui ◻ Non
Expliquez :			
❼ Donneriez-vous votre numéro de carte bleue à ces personnages ?	◻ Oui ◻ Non	◻ Oui ◻ Non	◻ Oui ◻ Non
Expliquez :			

Avez-vous remarqué des différences dans la façon dont les trois personnages donnent les explications ?

.....

.....

.....

.....

.....

.....

.....

Remarques libres :

.....

.....

.....

.....

Age : ..... Sexe : .....

Connaissiez-vous déjà les fonctions :

- Du photocopieur : ❒ oui ❒ non
  Commentaires :

- De la télécommande : ❒ oui ❒ non
  Commentaires :

- Du logiciel Végas : ❒ oui ❒ non
  Commentaires :

Annexe 6 :
Publications issues de la thèse
(au 8 avril 2005)

2005

Buisine, S., Martin, J.C. (2005). Children's and adults' multimodal interaction with 2D conversational agents. Proceedings of CHI'2005, pp. 1240-1243.

Buisine, S., Martin, J.C., Bernsen, N.O. (sous presse). Children's gesture and speech in conversation with 3D characters. Proceedings of HCII'2005 (10p).

Abrilian, S., Devillers, L., **Buisine, S.**, Martin, J.C. (sous presse). EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. Proceedings of HCII'2005 (10p).

Hartmann, B., Mancini, M., **Buisine, S.**, Pelachaud, C. (sous presse). Design and evaluation of expressive gesture synthesis for embodied conversational agents. Proceedings of AAMAS'2005 (2p).

2004

Buisine, S., Abrilian, S., Martin, J.C. (2004). Evaluation of Multimodal Behaviour of Embodied Agents. In: Zs. Ruttkay, C. Pelachaud (Eds.), *From Brows to Trust: Evaluating Embodied Conversational Agents*, Chapter 8, pp. 217-238, Kluwer Academic Publishers.

Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., **Buisine, S.**, Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T.V., Stanney, K.M., Su, H., Wang, Q.Y. (2004). Guidelines for multimodal user interface design. *Communications of the ACM*, 47, pp. 57-59.

Martin, J.C., **Buisine, S.**, Abrilian, S. (2004). 2D gestural and multimodal behavior of users interacting with embodied agents. Proceedings of AAMAS'2004 Workshop on Balanced Perception and Action in ECAs (8p).

Martin, J.C., **Buisine, S.** (2004). Evaluation of cooperation between modalities in ECAs. Invited talk for the Dagstuhl Seminar on Evaluating Embodied Conversational Agents, Wadern-Dagstuhl, Germany, March 14-19.

2003

Buisine, S., Martin, J.C. (2003). Experimental evaluation of bi-directional multimodal interaction with conversational agents. Proceedings of INTERACT'2003, pp. 168-175.

Abrilian, S., Martin, J.C., **Buisine, S.** (2003). Algorithms for controlling cooperation between output modalities in 2D Embodied Conversational Agents. Proceedings of ICMI'2003, pp. 293-296.

Buisine, S., Abrilian, S., Martin, J.C. (2003). Evaluation of individual multimodal behavior of 2D Embodied Agents in presentation tasks. Proceedings of AAMAS'2003 Workshop on Embodied Conversational Characters as Individuals, pp. 22-28.

Buisine, S., Martin, J.C. (2003). Design principles for cooperation between modalities in bi-directional multimodal interfaces. CHI'2003 Workshop on Principles for Multimodal User Interface Design (position paper, 3p).

Buisine, S. (2003). Experimental evaluation of multimodal interfaces based on multimodal corpora analysis. Invited talk for the Nordic Symposium on Multimodal Communication, Copenhagen, Denmark, September 25-26.

2002

Buisine, S., Abrilian, S., Rendu, C., Martin, J.C. (2002). Towards experimental specification and evaluation of lifelike multimodal behavior. Proceedings of AAMAS'2002 Workshop “Embodied conversational agents - let's specify and compare them!”, pp. 42-48.

Abrilian, S., **Buisine, S.**, Rendu, C., Martin, J.C. (2002). Specifying cooperation between modalities in lifelike animated agents. Proceedings of PRICAI'2002 Workshop on Lifelike Animated Agents Tools, Affective Functions, and Applications, pp. 3-8.

Conception et Evaluation d'Agents Conversationnels Multimodaux Bidirectionnels

Résumé

L'objectif de la thèse est d'étudier la communication multimodale entre un agent conversationnel (personnage virtuel) et un utilisateur humain. Les deux modalités considérées sont la parole et les gestes des mains. Nous étudions, dans des contextes conversationnels et ludiques, le comportement multimodal spontané des utilisateurs face aux agents. Nous analysons l'utilisation de chacune des modalités, leurs coopérations et les différences interindividuelles liées à l'âge des utilisateurs. Symétriquement, nous évaluons différentes stratégies de coopération multimodale pour des agents dans un contexte pédagogique : nous montrons qu'elles peuvent influencer la mémorisation des utilisateurs et leurs impressions subjectives. Enfin, nous proposons d'exploiter le comportement multimodal des utilisateurs pour évaluer le comportement des agents conversationnels. Ces hypothèses ouvrent des perspectives de méthodes d'évaluation originales.

Conception and Evaluation of Bidirectional Multimodal Conversational Agents

Abstract

The goal of this PhD work is to investigate multimodal communication between conversational agents (virtual characters) and human users, with a focus on two modalities: speech and hand gesture. We examine the spontaneous multimodal behavior that users display with agents in conversational and entertaining situations. We analyze the use of each modality, cooperation between them and the inter-individual differences related to users' age. Symmetrically, we evaluate several multimodal strategies of agents in a pedagogical context: we show that these strategies are likely to influence users' memorization as well as their subjective impressions. We finally hypothesize that users' multimodal behavior can be used to evaluate agents' behavior. We thus draw some perspectives of new evaluation methods.

Discipline : Psychologie Cognitive – Ergonomie

Mots-clés :

Interaction Homme-Machine, Agents Conversationnels Animés, Interfaces Multimodales.

Laboratoires :

- *Laboratoire d'Ergonomie Informatique*
  45 rue des Saints-Pères, F-75270 Paris Cedex 06
- *LIMSI-CNRS, BP 133, F-91403 Orsay Cedex*