Dialogue homme-machine
Conception et enjeux

Frédéric LANDRAGIN

2013

Table des matières

Avant-propos

La rédaction de ce livre s'est faite en lien avec la préparation d'une habilitation à diriger des recherches. Il s'agit d'une synthèse portant sur dix années de recherche, c'est-à-dire depuis ma thèse de doctorat (Landragin, 2004), dans le domaine du dialogue homme-machine. Le but est de faire un point sur les théories, les méthodes, les techniques, les enjeux impliqués dans la conception de programmes informatiques capables de comprendre et de produire de la parole. Cette synthèse regroupe ainsi la présentation de travaux importants du domaine et une approche plus personnelle, ne serait-ce que par les choix des thématiques explorées. Comment une machine peut-elle parler, comprendre ce qu'on lui dit, et entretenir un dialogue proche du dialogue naturel entre deux humains ? Quelles sont les étapes de conception d'un système de dialogue homme-machine ? Quelles sont les capacités de compréhension, de raisonnement et d'interaction attendues pour de tels systèmes ? Comment les implémenter ? Comment s'approcher du réalisme et de la fluidité du dialogue humain ? Un système de dialogue peut-il mentir ?

Ces questions sont à l'origine de mon parcours, qui a oscillé entre linguistique et informatique, entre recherche fondamentale et développement, entre laboratoires de recherche publics et privés : INRIA, puis Thales, et actuellement CNRS. Ce sont aussi des questions qui m'ont été posées lors du cours de dialogue homme-machine que j'ai donné pendant quelques années à des étudiants de Master 2 à l'Université Paris Diderot. De fait, ce livre s'inspire en partie de mes supports de cours, et a pour vocation d'être accessible à un public ayant des notions de linguistique et de traitement automatique des langues, mais pas forcément de dialogue homme-machine.

Le but est donc d'expliciter les principaux problèmes posés par chaque étape de conception d'un système de dialogue homme-machine, et de montrer quelques pistes théoriques et techniques pour traiter ces problèmes. La présentation sera forcément réductrice par rapport à la richesse des travaux existants, mais au-delà de ce but, se trouve aussi la volonté de donner un aperçu du domaine qui puisse donner envie au lecteur d'en découvrir plus.

Le but est aussi de montrer qu'il existe toujours une école française du dialogue homme-machine, particulièrement active ces dernières années, même si elle a parfois été en perte de vitesse et si l'on a pu croire à certains moments que le dialogue homme-machine était une impasse. Cette école française se caractérise par sa pluridisciplinarité, par son implication dans différents secteurs, qu'il s'agisse de développement de systèmes (prototypes universitaires, systèmes grand public, mais aussi – on a tendance à les oublier car ils restent confidentiels – systèmes militaires), de mise en œuvre de méthodes et de campagnes d'évaluation, de conception d'architectures logicielles. Il y a une école française pour le dialogue multimodal, pour l'ergonomie, pour les agents conversationnels animés, ou encore pour l'application de techniques d'apprentissage automatique au dialogue homme-machine. Tous les liens entre ces différentes spécialités ne sont pas encore complètement réalisés, mais la dynamique générale est indéniable et encourageante.

Comme tout travail, celui présenté dans ce livre doit beaucoup aux encouragements, aux conseils, et plus généralement au partage d'un environnement de travail efficace et agréable. Pour leurs encouragements, institutionnels aussi bien que scientifiques et humains, je remercie Francis Corblin, Catherine Fuchs, Valérie Issarny, Jean-Marie Pierrel, Laurent Romary, Jean-Paul Sansonnet, Catherine Schnedecker, Jacques Siroux, Mariët Theune, Bernard Victorri, Anne Vilnat. Pour l'expérience Ozone très enrichissante lors de mon post-doctorat à l'INRIA, je remercie en particulier Christophe Cérisara, Yves Laprie, et surtout Alexandre Denis sur qui j'ai pu me reposer pour l'implémentation d'un démonstrateur mémorable. Pour l'expérience également mémorable de Thales R & T, je remercie notamment Claire Fraboulet-Laudy, Bénédicte Goujon, Olivier Grisvard, Jérôme Lard, Célestin Sedogbo. Pour le cadre exceptionnel qu'est le laboratoire Lattice, unité mixte de recherche du CNRS, je remercie, entre autres et sans répéter de noms déjà cités, Michel Charolles pour nos échanges très riches sur la référence, Shirley Carter-Thomas et Sophie Prévost pour ce qui concerne la structure informationnelle, Thierry Poibeau et Isabelle Tellier pour le traitement automatique des langues, mes collègues de bureau successifs Sylvain, Laure, Frédérique, et puis Benjamin, Denis, Fabien, Jeanne, Julie, Marie-Josèphe, Noalig, Paola, Paul, Pierre, Sylvie. Merci aussi à ceux avec qui j'ai interagi dans le cadre de l'Atala (je pense notamment à Frédérique, Jean-Luc, Patrick) et dans le cadre de mes cours de dialogue homme-machine, ainsi qu'à ceux avec qui j'ai pu entamer des collaborations, même si parfois elles n'ont pas abouti. Merci donc à Ali, Anne, Gaëlle, Jean-Marie, Joëlle, Meriam, Nathalie, Tien. Enfin, merci à Céline pour ses encouragements constants et son soutien sans faille.

Frédéric Landragin

Introduction

Le système Ozone (Issarny *et al.*, 2005) évoqué dans l'avant-propos était un démonstrateur pour un service de réservation de billets de train, dans le cadre du projet européen Ozone. Il s'agit d'une « application » (ou « tâche ») récurrente dans le domaine du dialogue homme-machine, et c'est dans ce cadre que nous allons choisir nos exemples tout au long de ce livre. Le programme informatique constituant notre démonstrateur était ainsi capable de traiter une entrée audio, de transcrire la parole capturée en texte, et de comprendre ce texte de manière à trouver une réponse adéquate. La tâche nécessitant que le système connaisse les horaires d'un ensemble de trains d'une région donnée, une base de données avait été mise en œuvre : elle permettait au système de dialogue de trouver les informations indispensables pour ses réponses, réponses qui, comme en dialogue humain, étaient émises oralement. Jusqu'ici, nous restons dans le cadre du « dialogue homme-machine oral », c'est-à-dire avec entrées et sorties vocales. Ce type de système peut être utilisable par téléphone, sans canal visuel. Dans l'idéal, le système est rapide, compréhensif et produit des réponses pertinentes, au point que l'utilisateur a l'impression de dialoguer spontanément, comme avec un interlocuteur humain.

Cependant, nous nous étions donnés comme spécification supplémentaire de faire un système « multimodal », c'est-à-dire capable de gérer à la fois la parole et des gestes de désignation effectués sur écran tactile. Le système était ainsi capable de reconnaître des gestes de pointage et de faire des liens entre ces gestes et les mots prononcés simultanément. Ce qui est valable pour les entrées du système devrait l'être aussi pour ses sorties, et c'est ainsi que nous avons conçu un système capable de gérer la multimodalité en sortie, c'est-à-dire capable de produire simultanément un énoncé vocal et un affichage sur écran. Autrement dit, une fois que le système décidait d'une réponse à donner à l'utilisateur, il pouvait choisir entre verbaliser cette réponse, l'afficher à l'écran, ou, mieux, en verbaliser une partie et en afficher une autre. C'est ce que l'on appelle une « présentation d'information multimédia ». Au-delà des problématiques du dialogue oral, nous sommes à présent dans celles du dialogue multimodal. Les systèmes concernés impliquent une « situation de communication » partagée entre

l'utilisateur humain et la machine. Cette situation partagée fait intervenir un contexte visuel (ce qui apparaît sur l'écran de l'ordinateur) et des gestes (pour l'instant très simples car ils se limitent à des contacts sur l'écran). Avec cette situation de communication, on se rapproche du dialogue humain en face à face : l'utilisateur parle face à la machine et voit une scène visuelle que la machine « voit » également.

Pour fonctionner, le programme devait donc être exécuté sur un ordinateur doté au minimum d'un microphone, d'un haut-parleur et d'un écran tactile, ce qui était moins courant en 2004 que ça ne l'est maintenant. La figure 1 montre un exemple de dialogue que ce système était capable de tenir avec un utilisateur, exemple adapté dans la mesure où le système réel était conçu pour la langue anglaise. Les tours de parole successifs sont repérés par une lettre (U pour utilisateur, S pour système) et un chiffre, de manière à les identifier facilement dans les analyses et les discussions.

	Enoncé	Action sur l'écran
S1 :	« Bonjour, je suis le système de réservation de billet de train. »	Affichage sur l'écran d'une carte géographique
U1 :	« Bonjour, je voudrais aller à Paris. »	–
S2 :	« Voici les trajets possibles. »	Apparition de deux chemins
U2 :	« Combien de temps avec ce chemin qui semble être le plus court ? »	Geste de désignation sur l'un des deux chemins
S3 :	« Vingt minutes. »	Mise en valeur du chemin désigné
U4 :	« D'accord, je réserve un aller. »	–
S4 :	…	…

Figure 1. *Exemple de dialogue homme-machine*

Un dialogue comme celui-ci est un type de « discours » – en tant que suite de plusieurs phrases liées les unes aux autres – avec la spécificité de faire intervenir deux locuteurs et non un seul. Dans le cas d'un dialogue faisant intervenir plus de deux locuteurs, on peut parler de « multilogue » ou « polylogue ». Si l'on considère la suite de mots « voici les trajets possibles », on parle de phrase tant que l'on considère les mots, leur organisation et leur sens hors contexte, c'est-à-dire en ignorant la situation dans laquelle ces mots ont été prononcés, et on parle d'énoncé justement quand on tient compte du contexte, c'est-à-dire du fait que cette phrase a été prononcée par le système S à un instant précis du dialogue, et, dans le cas présent, simultanément à une action d'affichage (ce qui permet de donner un sens précis à « voici », mot en quelque sorte dédié à une présentation d'information multimédia). En fonction du contexte, une même phrase peut ainsi être à l'origine de plusieurs énoncés.

L'exemple de la figure 1 constitue une interaction, ou incursion selon la terminologie adoptée. En S1, le système se présente, puis, de U1 à U4, le dialogue porte sur le choix d'un billet de train. L'extrait qui va de U1 à U4 constitue un « échange » : le

but défini en U1 est atteint en U4, ce qui clôt l'échange, sans pour autant clore l'interaction. Un échange fait nécessairement intervenir les deux locuteurs, et comporte plusieurs tours de parole, au minimum deux. S1, U1… U4 sont des « interventions », qui correspondent aux tours de parole. Une intervention n'implique qu'un seul locuteur, et se définit ainsi comme la plus grande unité monologale dans un échange. Une intervention peut comprendre un seul « acte de langage » (action réalisée par la parole, comme celle de donner un ordre ou celle de répondre à une question), comme en S2 et S3, ou plusieurs actes de langage, comme en S1 et U1 où le premier acte est une salutation et le second la transmission d'une information.

Fondé sur une utilisation de la langue (ou langage naturel par opposition aux langages artificiels de l'informatique), le dialogue s'étudie à l'aide des notions des sciences du langage. L'analyse des énoncés relève ainsi de la pragmatique, étude de la langue en usage. L'analyse des phrases elles-mêmes relève de la linguistique. Plus précisément, l'analyse du sens des phrases et des concepts impliqués relève de la sémantique. Au niveau de la construction de la phrase, on s'intéresse aux mots, aux unités qui constituent le lexique, aux groupes de mots, à l'ordre dans lequel ils apparaissent et aux relations qui existent entre eux, ce qui relève de la syntaxe. En dialogue oral, on s'intéresse également à la matérialisation phonique des phrases, aux proéminences, au rythme et à la mélodie, ce qui relève de la prosodie. A ces plans d'analyse s'ajoutent tous les phénomènes caractéristiques du langage naturel, notamment le fait qu'il existe une multitude de façons d'exprimer un même sens, ou encore que la langue est par essence vague, peu précise, ce qui entraîne des « ambiguïtés » (plusieurs interprétations d'un même énoncé sont possibles) et des phénomènes de « sous-spécification » (l'interprétation d'un énoncé peut rester incomplète). C'est là toute la richesse et la diversité de la langue, dont un système de dialogue en langage naturel qui se veut compréhensif doit tenir compte. La langue en situation de dialogue se caractérise aussi par une richesse et une diversité qui s'expriment notamment dans les combinaisons d'énoncés, c'est-à-dire dans la façon dont un énoncé est relié au précédent, dans la façon dont plusieurs énoncés successifs constituent un échange, et d'une manière générale dans la structure de dialogue qui se construit au fur et à mesure de l'interaction et qui constitue elle aussi un objet d'analyse. Lorsque cette structure reflète non pas un protocole rigide ou codifié mais un usage naturel de la langue, nous arrivons à une dernière définition, celle de « dialogue naturel en langage naturel ».

C'est le domaine de recherche et de développement dont il est question dans ce livre, et qui a déjà fait l'objet de nombreux ouvrages, qu'il s'agisse de présentations de systèmes, ou de théories suffisamment formelles pour en autoriser à terme des implémentations informatiques. A titre d'exemples, ici dans l'ordre chronologique, nous citerons un ensemble de livres dont la lecture s'avère utile pour ne pas dire indispensable à tout spécialiste du domaine du dialogue homme-machine : Reichman (1985), Pierrel (1987), Sabah (1989), Carberry (1990), Bilange (1992), Kolski (1993), Luzzati (1995), Bernsen *et al*. (1998), Reiter et Dale (2000), Asher et Lascarides (2003), Cohen *et al*. (2004), Harris (2004), McTear (2004), López-Cózar Delgado et Araki

(2005), Caelen et Xuereb (2007), Jurafsky et Martin (2009), Jokinen et McTear (2010), Rieser et Lemon (2011), Ginzburg (2012), Kühnel (2012). Afin de donner quelques repères et d'en appréhender les principales facettes, nous ferons un historique du domaine dans le chapitre 1.

Le domaine du dialogue homme-machine couvre plusieurs disciplines scientifiques. Nous avons mentionné l'informatique et les sciences du langage, mais nous verrons dans le chapitre 2 que d'autres disciplines peuvent apporter des théories et des points de vue complémentaires. Avec le but de concevoir une machine ayant des capacités proches de celles d'un être humain (il s'agit de se rapprocher des capacités humaines, pas d'en proposer une simulation), on peut s'inspirer de toutes les études portant de près ou de loin sur la langue et le dialogue, afin de les modéliser dans un cadre computationnel qui permette leur exploitation en dialogue homme-machine.

Ce domaine du dialogue homme-machine (désormais DHM) entretient des liens privilégiés avec d'autres domaines, notamment celui du traitement automatique des langues (TAL), dont il constitue une application essentielle, celui de l'intelligence artificielle (IA), dont il est issu et qui complète les aspects linguistiques avec les aspects liés au raisonnement et à la prise de décision, celui des interfaces homme-machine (IHM), qu'il contribue à enrichir en offrant des possibilités d'interaction vocale en complément des interactions graphique et tactile, et ceux plus récents des systèmes de questions-réponses (SQR) et des agents conversationnels animés (ACA), qui en sont des facettes – la première portant sur l'interrogation en langage naturel de grandes bases de données, la seconde sur le rendu visuel et vocal d'un avatar représentant l'interlocuteur-machine – devenues des domaines de recherche à part entière. Le domaine du DHM regroupe ainsi plusieurs problématiques, qui peuvent se répartir en trois grandes catégories :

– le traitement des signaux en entrée du système, avec la reconnaissance et l'interprétation automatique ;

– les traitements et raisonnements internes au système ;

– la gestion des messages produits par le système, donc en sortie de celui-ci, avec la génération automatique et la présentation d'information multimédia.

Selon le type de système envisagé (outil *versus* partenaire, c'est-à-dire en offrant à l'utilisateur une logique de « faire » ou de « faire-faire »), selon les modalités de communication entre l'utilisateur et le système (dialogue écrit *versus* oral), selon la part accordée à une tâche qui sous-tend le dialogue (dialogue en domaine ouvert *versus* en domaine fermé), selon l'importance donnée à la langue (dialogue privilégiant la finalité *versus* dialogue privilégiant la fluidité et le réalisme linguistiques), ces problématiques donnent lieu à de nombreuses approches et à de nombreuses façons d'implémenter. Les approches peuvent être plutôt théoriques – par exemple étendant et testant une théorie syntaxique ou pragmatique particulière – ou plutôt pratiques (privilégiant

la robustesse). Les implémentations peuvent être plutôt symboliques ou plutôt statistiques, etc. Le chapitre 3 fera le point sur ces aspects en décrivant les étapes de réalisation d'un système de DHM. Pour ce qui concerne la question des architectures logicielles, le chapitre 4 complètera et terminera la première partie du livre avec des enjeux cruciaux tels que la réutilisabilité et la conception de modèles génériques, à l'image de ce qui se fait dans le domaine des IHM.

Le traitement des énoncés en entrée du système fera l'objet de la deuxième partie, avec le chapitre 5 pour les aspects lexicaux, syntaxiques, prosodiques et sémantiques fondamentaux, le chapitre 6 pour la question de la résolution des références en contexte, et le chapitre 7 pour la reconnaissance et l'interprétation des actes de langage dans le contexte d'un dialogue. Nous passerons rapidement sur les questions de la reconnaissance automatique de la parole et des processus dits de « bas niveau », pour nous concentrer sur les processus de plus « haut niveau » qui concernent le sens des énoncés : sémantique, référence, actes de langage. Avec l'exemple U2 de la figure 1, le chapitre focalisé sur l'analyse sémantique montrera comment représenter la signification de la phrase « Combien de temps avec ce chemin qui semble être le plus court ? », phrase complexe dans la mesure où elle comprend une principale et une subordonnée, la principale étant qui plus est dépourvue de verbe. Sans analyse linguistique de ce type, un système de DHM peut difficilement être qualifié de compréhensif. Le chapitre focalisé sur la référence montrera comment l'énoncé et le geste de désignation de U2 permettent d'attribuer un « référent », en l'occurrence un trajet de train bien particulier, à l'expression référentielle démonstrative « ce chemin ». Sans cette capacité à résoudre les références, un système de DHM peut difficilement savoir de quoi il est question dans le dialogue. Le chapitre focalisé sur les actes de langage montrera comment cette intervention U2 peut s'interpréter comme un ensemble de deux actes de langage, avec un premier acte qui consiste en une question, et un second acte qui fait un commentaire à propos du trajet de train référé, commentaire qui pourra ensuite être traité de différentes manières par le système, par exemple selon qu'il s'agit effectivement ou non du chemin le plus court. Là encore, le chapitre met en avant une facette essentielle d'un système de DHM : sans cette capacité à identifier les actes de langage, un système peut difficilement trouver comment réagir et répondre à l'utilisateur.

Les traitements en interne et en sortie du système déterminent le comportement de celui-ci et font l'objet de la troisième partie du livre. Dans le chapitre 8, nous verrons comment l'identification des actes de langage permet au système de raisonner en fonction des actes identifiés, de la tâche et du dialogue déjà effectué. C'est toute la question de la mise en perspective de l'énoncé de l'utilisateur et de la détermination de la réaction à produire en retour. Plus que tous les processus étudiés dans la deuxième partie, il s'agit ici de raisonner non pas au niveau d'un seul énoncé, mais au niveau de tout le dialogue. On parle ainsi de « gestion du dialogue ». Dans le chapitre 9, nous verrons comment un système peut matérialiser la réaction qu'il a décidé de produire. C'est la question de la génération automatique de messages, question qui

prend une orientation particulière dès que l'on prend en compte un avatar (on rejoint ici le domaine des ACA), ou même tout simplement, comme nous en avons déjà parlé, la possibilité de présenter des informations à l'écran en même temps qu'un message est verbalisé.

Enfin, le chapitre 10 traitera d'un aspect qui concerne aussi bien les étapes de conception que le système final, une fois l'implémentation informatique terminée. Il s'agit de l'évaluation, question délicate dans la mesure où un système de DHM intègre des composants ayant des fonctionnalités variées, et où, comme nous l'avons vu, les types de systèmes que l'on peut envisager présentent eux-mêmes des priorités et des caractéristiques très variées. Cette question nous amènera à conclure sur le domaine du DHM, sur l'état actuel des réalisations, et sur les enjeux pour les années à venir.

PREMIÈRE PARTIE

Repères historiques et méthodologiques

Chapitre 1

Un bilan de l'évolution des recherches et des systèmes

Les systèmes de DHM semblent plus présents dans les œuvres de science-fiction que dans la réalité. Dans combien de films peut-on voir des ordinateurs, des robots, voire des réfrigérateurs ou des jouets pour enfants, parler et comprendre tout ce qu'on leur dit ? La réalité semble plus compliquée : quelques produits issus des nouvelles technologies, par exemple des téléphones portables ou des robots de compagnie, parlent et comprennent quelques mots, mais on est bien loin d'un dialogue naturel, tel que nous le promet la science-fiction depuis si longtemps.

Ce ne sont pourtant pas les idées d'application qui manquent. Instaurer un dialogue avec une machine pourrait être efficace pour obtenir des renseignements ciblés, et ce sur tout type d'informations : transports (Lamel *et al.*, 2000), commerces divers, activités touristiques ou ludiques (Singh *et al.*, 2002), fonds bibliothécaires, démarches administratives, financières (Cohen *et al.*, 2004), etc., voir (Gardent et Pierrel, 2002) et (Grau et Magnini, 2005). Le dialogue est en effet adapté à l'élaboration pas à pas d'une requête, requête qui tiendrait difficilement en un seul énoncé, ou en une commande exprimée dans un langage informatique. Ce premier domaine d'application du DHM, qui inclut les SQR, est parfois défini sous le nom de « dialogue de renseignement ». Quand le dialogue ne concerne qu'une seule thématique, par exemple le renseignement ferroviaire, on parle de dialogue en domaine fermé. Quand le dialogue peut concerner tout et n'importe quoi, par exemple l'interrogation de bases de données encyclopédiques comme l'a fait récemment IBM Watson avec une tâche de jeu télévisé, on parle de dialogue en domaine ouvert (Rosset, 2008). En reprenant l'exemple de l'introduction, un énoncé unique, sans dialogue, pourrait être le suivant : « je voudrais réserver un aller à Paris en prenant le chemin le plus court, du moins s'il prend

moins d'une demi-heure (sinon je ne réserve rien) ». L'élaboration d'un dialogue naturel est beaucoup plus souple : elle permet à l'utilisateur d'exprimer une première requête, simple, pour l'affiner ensuite en fonction de ce que la machine répond ; elle lui permet de transmettre une information en vue d'une action future, de confirmer et d'infirmer au fur et à mesure (Pierrel, 1987). Le nombre total de mots pour arriver au même résultat sera peut-être plus grand, mais la spontanéité des énoncés, leur rapidité, leur facilité de production, compenseront largement. Avec l'exemple de l'interrogation d'un annuaire de type « pages jaunes », (Luzzati, 1995) démontre un autre avantage du dialogue : l'utilisateur peut obtenir l'adresse d'un taxidermiste alors même qu'il ne connaît pas le nom de cette profession. C'est en discutant, c'est par le dialogue que l'utilisateur fait comprendre à la machine ce qu'il recherche exactement. Il y a coconstruction d'un concept commun aux deux interlocuteurs, cette coconstruction faisant l'intérêt du dialogue par rapport à l'énoncé unique ou à la requête en langage informatique.

Au-delà de la demande de renseignement ou de la consultation d'une base de données, instaurer un dialogue avec une machine peut aussi être efficace pour manier un système informatique, comme par exemple un logiciel de création numérique (dessin, traitement d'image, 3D) ou tout simplement le système d'exploitation d'un ordinateur. On peut alors imaginer qu'au lieu d'aller chercher la bonne fonction dans les nombreux menus et sous-menus du logiciel en question, l'utilisateur procède à des commandes vocales, beaucoup plus rapides et directes, du moins s'il ne s'est pas encore familiarisé avec le logiciel. Ce deuxième domaine d'application du DHM, proche de celui des IHM, est parfois défini sous le nom de « dialogue de commande ». En incluant les logiciels de développement informatique, on peut quasiment imaginer un utilisateur qui se servirait de la langue pour « programmer en langage naturel » (Luzzati, 1995). En incluant la robotique, c'est le domaine de la direction d'un robot, application phare de l'IA moderne (Gorostiza et Salichs, 2011). Par ailleurs, c'est aussi le domaine des systèmes professionnels, civils et militaires, à la conception desquels s'intéresse Thales : gestion et contrôle du trafic aérien, surveillance maritime, supervision de situation sur un terrain d'opérations, commande de systèmes dans des milieux dangereux. Ces systèmes sont à l'heure actuelle des IHM complexes, et le travail de l'équipe de recherche dans laquelle nous avons travaillé était de tester la faisabilité de leur « donner la parole ». On restait là dans du dialogue de commande à domaine fermé, mais avec de fortes contraintes de robustesse.

Dialogue de renseignement, dialogue de commande : tous les systèmes existants ne rentrent pas dans l'une ou l'autre de ces deux catégories un peu strictes. Certains systèmes permettent les deux types d'interaction, par exemple certains robots de compagnie, capables à la fois de renseigner son utilisateur et d'effectuer sur demande quelques tâches simples, comme marcher ou danser. D'autres systèmes n'ont pas vocation à renseigner ni à effectuer des tâches spécifiques. Ce sont par exemple les systèmes purement ludiques, avec l'exemple des robots de discussion sur Internet. Les exemples donnés ici sont des exemples de systèmes de DHM grand public, ou en tout

cas voués à l'être. Mais fonctionnent-ils vraiment ? En fait, ce que l'on constate en utilisant ce type de système, c'est qu'il est bien difficile d'établir un dialogue digne de ce nom. Quand un mot est reconnu et compris, ce qui n'est pas systématique, la machine tente une réponse basée sur ce mot, ou essaie de relancer le dialogue à sa façon, mais souvent sans pertinence. Comme l'écrit (Vilnat, 2005, p. 5), les systèmes de DHM ne fonctionnent que très imparfaitement, et de ce fait subissent des flots de critiques, jusqu'au « ça ne marchera jamais ! » maintes fois entendu. Les critiques proviennent avant tout des utilisateurs, qui constatent bien qu'il existe un fossé entre ce qu'ils testent et ce qu'ils espéraient, et qui considèrent parfois qu'une IHM classique est plus rapide, plus efficace, voire plus facile à utiliser, en tout cas moins déroutante. Elles proviennent également des chercheurs et développeurs issus du domaine du DHM. En effet, la quantité de travail nécessaire à la réalisation d'un système est telle que les découragements sont nombreux. La quantité de travail correspondant à une thèse de doctorat ne suffit plus, en tout cas quand le but est de réaliser un système innovant. A titre d'exemple, le dérouragement de (Guibert, 2010, p. 60) au cours de la conception d'un système nommé « A », est évocateur : « suite à l'arrêt du développement de ce système A, pris en exemple parmi d'autres, ce travail constitue en fait la chronique d'un échec annoncé de systèmes de dialogue actuels ». Nous allons voir que, quand le dialogue est dirigé par une tâche clairement délimitée, la conception d'un système de DHM performant reste possible, et a même fait de grands progrès ces dernières années. Après une section visant à rappeler quelques repères historiques (section 1.1), nous allons faire un tour rapide des fonctionnalités de plus en plus présentes sur les systèmes actuels (section 1.2), et nous en déduirons une première liste d'enjeux pour les années à venir (section 1.3).

1.1. Quelques repères historiques essentiels

Le dialogue entre l'homme et la machine est un domaine phare de l'informatique : c'est une sorte de quête du Graal, qui a été la source de développements des sciences informatiques et de vocations de chercheurs. Il se trouve que le premier système qui a fait date, Eliza (Weizenbaum, 1966), est aussi une énorme tromperie (assumée, nous le verrons ci-après). Plusieurs voies ont ensuite été prises dans la conception de systèmes de DHM sérieux : une voie proche de l'IA, avec une focalisation sur les problèmes d'interprétation et de raisonnement, et une voie qui a consisté à enrichir les systèmes de reconnaissance automatique de la parole. Les deux voies, avec leurs deux communautés séparées (Vilnat, 2005, p. 47), se sont rejointes tardivement et ont permis à nombre de systèmes de DHM cohérents de voir le jour. Ce sont ces systèmes que nous allons présenter dans cette section.

1.1.1. *Premières motivations, premiers systèmes écrits*

Une machine peut-elle penser ? En 1950, Alan Turing donne un nouvel essor à cette question récurrente dans l'histoire des technologies : il lui substitue la question

« une machine peut-elle imiter l'homme ? » et propose un jeu, ou test, basé sur l'imitation, et qui restera célèbre sous le nom de test de Turing. Dans un premier temps, l'imitation concerne un homme et une femme : le sujet du test dialogue à tour de rôle avec un homme et une femme, par le biais de feuilles tapées à la machine, et sans rien voir ni rien savoir de ses interlocuteurs successifs. L'homme a pour consigne de se faire passer pour une femme, et le sujet doit ainsi deviner qui est l'homme et qui est la femme. Dans un second temps, sans que le sujet n'en sache rien, l'homme est remplacé par une machine. Si le sujet est incapable d'identifier l'un et l'autre, c'est que la machine a réussi le test de Turing. Ce petit jeu, spécifié à une époque où il était impossible de programmer un système de DHM, a été la source de discussions innombrables, d'affirmations diverses et variées sur la nature de la machine, voire de l'humain. Ce qui nous intéresse ici, c'est le défi posé à l'informatique : arriver à programmer un système de DHM qui puisse se faire passer pour un humain. A. Turing ne donne pas beaucoup de pistes pour arriver à un tel résultat. La description de son test reste ciblée sur les conditions de l'expérimentation, et ne souligne même pas l'importance du langage et du dialogue dans cette appréhension de la pensée (Tellier et Steedman, 2009). Il n'en reste pas moins qu'aujourd'hui encore, des compétitions sont organisées (on pensera au prix Loebner) inspirées par ce test de Turing. Ce sont les années 1950 qui sont à l'origine des premières motivations de la recherche sur le DHM, mais aussi sur la recherche d'informations et sur le TAL. Il est à noter que l'association Atala, initialement « Association pour l'étude et le développement de la traduction automatique et de la linguistique appliquée », puis « Association pour le traitement automatique des langues », voit le jour en 1959.

Les années 1960 sont celles des premiers systèmes de DHM. Eliza[1] (Weizenbaum, 1966), que nous évoquions plus haut, est fascinant à plus d'un titre. Tout d'abord, c'est un système de dialogue écrit, qui fonctionne réellement, sans boucler ni s'arrêter de manière intempestive. Il est ainsi possible d'entretenir un dialogue sur des centaines de tours de parole. Ensuite, la tâche choisie est elle-même fascinante : le système est censé jouer le rôle d'un psychothérapeute non directif, c'est-à-dire qui se contente d'écouter le locuteur raconter ses problèmes (« j'ai un problème avec mes parents »), et de réagir parfois à quelques phrases (« parlez-moi de votre famille »). Le réalisme est tel que certains utilisateurs ont passé des heures à dialoguer avec Eliza, et que J. Weizenbaum a dû renoncer à ajouter ouvertement un module de sauvegarde des dialogues, face aux accusations d'espionnage et d'atteinte à la vie privée. Cette tâche présente deux avantages : d'une part celui de ne pas avoir à entretenir un dialogue

1. Le nom est issu du personnage d'Eliza Doolittle dans le film *My Fair Lady* (1964, G. Cukor), lui-même adapté de la pièce de théâtre *Pygmalion* (1914, G.B. Shaw) qui avait déjà fait l'objet d'une adaptation au cinéma. Eliza Doolittle est une fleuriste issue d'un milieu très pauvre, et est l'objet d'un pari d'un aristocrate qui pense pouvoir la faire passer pour une aristocrate rien qu'en changeant sa manière de parler. L'idée de la tromperie par la langue et le dialogue est ainsi à l'origine du nom du système.

complexe, avec par exemple de la négociation ou de l'argumentation, tout en gardant un caractère spontané et naturel, l'utilisateur pouvant dire ce qu'il veut quand il veut, et d'autre part celui d'être facile à programmer, le système n'ayant pas besoin de tout comprendre dans la mesure où des énoncés tels que « qu'est-ce qui vous fait dire cela ? » ou « je vois, continuez » suffisent amplement. De fait, et c'est aussi une cause de fascination pour les chercheurs en TAL, IA ou DHM, J. Weizenbaum a réussi à développer un système qui semble maîtriser la langue et réussir le test de Turing, alors qu'il n'aborde même pas les problématiques les plus basiques de la compréhension automatique.

En effet, tout le fonctionnement Eliza repose sur quelques règles ou heuristiques bien choisies. Le système connaît quelques mots, notamment ceux liés la famille : « parents », « mère », « père » (en anglais dans la version originale du système). C'est ainsi qu'il est capable de rebondir suite à l'énoncé « j'ai un problème avec mes parents », sauf qu'aucune compréhension n'intervient dans le processus : le système a juste détecté « parents » et répondu avec une nouvelle question portant sur la « famille », nouvelle question pour justement ne pas avoir à tenir compte du sens de l'énoncé de l'utilisateur. Le système connaît aussi les pronoms personnels désignant les deux interlocuteurs, « je », « me », « mon », « tu », « vous », « ton », ce qui lui permet de réaliser des remplacements et de construire un énoncé reprenant celui de l'utilisateur, comme « qu'est-ce qui vous fait croire que *vous* écoutez *mes* conseils », généré après « *j*'écoute *vos* conseils ». Sur cet exemple, on remarque que le système ne comprend pas grand-chose, mais qu'il est capable d'intervertir les personnes et d'encapsuler l'énoncé de l'utilisateur dans une question « qu'est-ce qui vous fait croire que », question volontairement ouverte. Les techniques mises en œuvre pour les énoncés en entrée sont la recherche de séquences de mots et la recherche de mots-clés. Celles mise en œuvre pour la génération d'énoncés en sortie du système sont la production directe de phrases types, la concaténation de séquences de mots, qu'il s'agisse de séquences types ou de séquences obtenues à partir d'un énoncé de l'utilisateur. Le système a également une ébauche de mémoire, dans la mesure où il est capable de revenir à un terme familial utilisé plusieurs tours de parole auparavant.

Quelques années après Eliza, c'est le système Parry (Colby *et al.*, 1971) qui marque les esprits avec des techniques complémentaires. La machine simule cette fois un sujet paranoïaque, lors de son premier entretien (par écrit) avec l'utilisateur qui est censé jouer le rôle d'un médecin psychiatre, comme l'est d'ailleurs l'auteur principal. La démarche scientifique revendiquée est celle de l'étude et de la modélisation de la paranoïa, jusqu'au financement qui provient en partie du *National Institute of Mental Health*, et à la méthodologie qui inclut non seulement la modélisation et le développement informatique du modèle, mais aussi son évaluation par des professionnels de la santé mentale : 25 psychiatres ont été impliqués, et la grande majorité d'entre eux (23) ont diagnostiqué le système comme paranoïaque, lui faisant ainsi passer avec succès le test de Turing. Les dialogues prennent la forme d'interviews et débutent avec des questions factuelles posées par l'utilisateur au système : nom, âge, situation. Parry

a ainsi en stock un ensemble de réponses à ces questions types : il s'appelle Frank Smith, il a 28 ans, et il est interné à l'hôpital. Sont en stock également plusieurs questions que le système est capable de poser, inversant ainsi l'orientation du dialogue : « qui êtes-vous ? », « que me voulez-vous ? », de même que des anecdotes, et notamment des propos autour d'un concept relativement bien élaboré, celui de « mafia ». Les techniques mises en œuvres sont là aussi des techniques de recherche de séquences de mots, de détection de mots-clés, de gestion des pronoms de première et de deuxième personne, mais avec plus de finesse qu'Eliza. Par exemple, le mot « peur » fait l'objet d'un ensemble de séquences prédéfinies, et des verbes comme « croire » font l'objet de traitements spécifiques. De plus, le système se caractérise par une ébauche de personnalité, ou d'« états mentaux », *via* des variables : peur, colère, méfiance. Les valeurs de ces variables augmentent ou baissent tout au long du dialogue, en fonction de ce que dit l'utilisateur. Le comportement du système évolue en conséquence : il devient par exemple agressif si la valeur de colère dépasse un certain seuil. Les règles ou heuristiques, contrairement à celles d'Eliza, se basent donc à la fois sur les énoncés de l'utilisateur et sur ces variables d'état. Parry marque une évolution des systèmes de DHM, avec les moyens techniques de l'époque : le programme, écrit dans une variante du langage Lisp, prend 35 Ko dont 14 Ko de base de données.

Les années 1970 voient venir les premiers systèmes de compréhension (à l'écrit), avec des progrès significatifs en TAL, notamment en analyse syntaxique et sémantique, et, par conséquent, les premiers véritables systèmes de DHM écrit, qui modélisent un domaine de connaissance, savent interpréter un énoncé dans ce domaine, et commencent à gérer un dialogue structuré. Ces avancées font suite à quelques travaux marquants en linguistique et en linguistique computationnelle, notamment, comme le souligne (Jurafsky et Martin, 2009, p. 892), ceux de B.J. Grosz puis de C.L. Sidner. C'est la première voie évoquée précédemment, avec deux systèmes emblématiques, Shrdlu et Gus. En parallèle, la voie des systèmes de reconnaissance de la parole fait des progrès importants, avec notamment les systèmes développés dans le cadre des projets américains ARPA (*Advanced Research Projects Agency*) : Harpy, Hearsay, Hwim. On passe ainsi de la reconnaissance de mots isolés, totalement inadaptée au DHM, à la reconnaissance de mots continus, éventuellement multilocuteurs, et avec des préoccupations qui commencent à rejoindre celles du DHM, par exemple la question des architectures logicielles pour faire communiquer plusieurs sources de connaissances à l'intérieur des systèmes, voir l'historique de (Pierrel, 1987). Nous y reviendrons au paragraphe 1.1.2 avec les premiers systèmes de DHM oraux.

Le système Shrdlu[2] (Winograd, 1972) donne un nouvel essor au DHM écrit en montrant les possibilités approfondies de compréhension et de dialogue dès lors que

2. Le nom vient de la suite de lettres E T A O I N S H R D L U, c'est-à-dire, par ordre décroissant, la suite des lettres les plus fréquentes en anglais, telles qu'elles ont été disposées verticalement au milieu du clavier de certaines machines d'imprimerie.

l'on se restreint à une tâche clairement délimitée et modélisée. Cette fois, on oublie le test de Turing et on se tourne vers des applications ciblées : la tâche consiste à faire déplacer par la machine des objets géométriques (cubes, cônes, pyramides). Elle implique l'affichage d'une scène sur écran, avec une représentation du système lui-même par une sorte de bras robot qui manipule les objets. L'utilisateur produit des énoncés tels que « prends un cube vert » ou « cherche un cube plus grand que celui que tu tiens et mets-le dans la boîte », et le système effectue les actions, dont le résultat devient visible à l'écran. Cette tâche met l'accent sur les phénomènes de « référence aux objets » : quel objet est-il désigné par « la pyramide » ? Pour interpréter correctement une telle référence, le système doit trouver parmi les objets de la scène lequel est le bon, c'est-à-dire lequel correspond à l'intention de l'utilisateur. Si deux ou trois pyramides sont visibles, le système peut ainsi répondre « je ne comprends pas de quelle pyramide il s'agit ». Après éclaircissement, il fait ce qu'il doit faire, c'est-à-dire exécuter les actions et répondre aux questions. Parmi les questions possibles, beaucoup tournent autour du monde physique des objets : « qu'est-ce que la boîte contient ? », « qu'est-ce qui supporte la pyramide ? ». A chaque fois, Shrdlu est capable d'analyser la scène, de repérer les relations spatiales entre objets, de compter et de répondre. Certes, un monde composé d'objets géométriques reste simple. Mais ce sont tous les processus de compréhension automatique mis en œuvre qui impressionnent, ainsi que la modélisation des connaissances associées : le système est capable de résoudre des références complexes, comme « un cube plus grand que celui que tu tiens », de résoudre des anaphores, comme « mets-*le* dans la boîte », d'identifier les actes de langage. Le dialogue qui en résulte va à l'essentiel. Il manque peut-être de fluidité, mais l'objectif est de satisfaire la tâche et effectivement, tout est fait dans ce sens.

Quant au système Gus (Bobrow *et al.*, 1977), dont le nom signifie *Genial Understanding System*, il fait un pas de plus dans la direction du DHM utilitaire, avec une tâche de réservation de vols. Pour la démonstration de ce prototype de recherche, la base de données ne comprend en fait qu'un seul vol en Californie. Au-delà de cette limitation, la modélisation linguistique, la modélisation informatique et les aspects méthodologiques donnent une idée de ce à quoi va ressembler le domaine du DHM pendant plusieurs dizaines d'années. Comme Shrdlu, le système est capable de résoudre les références aux objets et les anaphores, du moins quand elles concernent directement les objets de la tâche, donc les vols, les jours, les horaires. Il arrive par exemple à attribuer un référent de type date à l'expression référentielle « le vendredi » utilisée comme date de retour après la spécification de « le 28 mai » comme date d'aller. L'interprétation des énoncés de l'utilisateur déclenche une analyse syntaxique et sémantique qui peut être partielle, c'est-à-dire fonctionner sur d'autres matériaux linguistiques que des phrases complètes. Elle déclenche également une reconnaissance des actes de langage, avec notamment la compréhension des réponses indirectes à des questions. Les grands résultats des travaux linguistiques sur la structure du dialogue et sur la structure informationnelle sont exploités, ce qui conduit à la gestion par le système de nombreuses connaissances sur la langue : lexique (3 000 racines enregistrées,

ce qui dépasse les précédents systèmes), règles morphologiques, constructions syntaxiques, principes simplifiés de la structure informationnelle, patrons pour la structure du dialogue, modèle conceptuel pour les plans de voyage et les dates, et enfin concept d'agenda : structure centrale qui va permettre au système de gérer les événements et de savoir à tout moment quelle tâche effectuer. L'implémentation informatique se rationalise : les différents analyseurs linguistiques sont implémentés comme des modules indépendants, et un langage de communication entre modules est spécifié. Le fait que les modules soient indépendants permet de les tester, les corriger, les améliorer séparément. En fait, toute la conception suit une méthodologie exemplaire : les auteurs ont commencé par collecter et étudier des dialogues humains portant sur la même tâche, c'est-à-dire qu'ils ont procédé à une étude de corpus, le mot « corpus » désignant une collection de matériaux linguistiques attestés, et ils ont même mis en œuvre une expérimentation de simulation du système (ce qui s'appellera plus tard un « Magicien d'Oz »), afin de recueillir des données sur le comportement d'un utilisateur face au système qu'ils imaginaient. Les méthodes fondamentales du DHM sont posées. Bien entendu, elles sont appliquées avec les moyens de l'époque, et la lenteur des ordinateurs conduit par exemple à une attente qui va de dix à soixante secondes pour chaque énoncé, attente qui est prise en compte dans l'expérimentation de simulation, et qui s'avère bien éloignée de la rapidité et du naturel d'un dialogue humain.

1.1.2. *Premiers systèmes oraux et multimodaux*

S'il était possible jusqu'à maintenant de présenter les grandes avancées du DHM par le biais de quelques systèmes emblématiques, cela n'est plus le cas avec les années 1980. Celles-ci voient en effet venir une multitude de travaux théoriques qui font découvrir le dialogue et ses caractéristiques à de nombreux chercheurs, une multitude de prototypes et de systèmes de DHM, et notamment les premiers systèmes oraux et les premiers systèmes multimodaux. Par ailleurs, c'est aussi l'âge d'or des jeux vidéo, et le grand public découvre les jeux d'aventure avec interaction textuelle[3], autrement dit les premiers systèmes de DHM ludiques.

3. A titre d'exemple, le jeu Sram (« Mars » à l'envers), édité en 1986 par Ere Informatique, a beaucoup marqué les premiers mordus de jeux vidéo : toute l'interaction du jeu passait par des commandes écrites, ce qui amenait par exemple à taper « je veux aller vers l'ouest », pour voir alors (visualisation des étapes d'analyse réalisées par le logiciel) cet énoncé apparaître avec une mise en couleur des mots « aller » et « ouest », puis découvrir la réponse du logiciel : « vous arrivez près d'une chute d'eau », avec l'affichage d'une scène visuelle dans laquelle le joueur doit chercher des indices pour continuer sa quête. Les techniques mises en œuvre sont beaucoup plus simples que celles de Shrdlu ou de Gus, avec la détection de mots-clés et non de séquences de mots, quasiment toujours selon le schéma verbe-complément, mais les contraintes ne sont pas non plus les mêmes : le vocabulaire et les possibilités sont larges, adaptés à un usage intensif, et surtout le jeu se doit d'être robuste, fiable et intéressant.

Parmi les travaux théoriques qui ont marqué les années 1980, se trouvent les recherches menées en analyse conversationnelle (Sacks *et al.*, 1974) et en analyse du discours, ou pragmatique du discours (Roulet *et al.*, 1985 ; Moeschler, 1985). Si les objectifs de l'une et l'autre divergent, l'objet d'étude est le même, à savoir des enregistrements et des transcriptions de dialogues humains, et les observations vont amener à mieux définir les unes par rapport aux autres les notions définies dans l'introduction (tour de parole, intervention, énoncé, acte de langage) et les notions de coopération, planification, organisation conversationnelle, structure du dialogue, terrain commun, ancrage ou pertinence que nous verrons dans le chapitre 8. Ces travaux vont contribuer à la parution de nombreux articles (Allen et Perrault, 1980 ; Clark et Wilkes-Gibbs, 1986 ; Grosz et Sidner, 1986 ; Clark et Schaefer, 1989 ; Cohen et Levesque, 1990) qui vont être des sources d'inspiration pour l'ensemble de la communauté du DHM.

Les premiers systèmes oraux sont issus des progrès de la reconnaissance automatique de la parole. Pour fonctionner correctement, ils se focalisent sur des tâches bien délimitées, à l'image de Shrdlu et Gus. La société Nuance développe par exemple divers systèmes spécialisés, souvent pour du dialogue par téléphone, pour des clients tels que des banques (Cohen *et al.*, 2004). Quant aux premiers systèmes multimodaux, c'est-à-dire qui associent la reconnaissance de la parole avec la capture de gestes correspondant au départ à de simples clics souris, ils apparaissent avec un article célèbre (Bolt, 1980), qui montre que la multimodalité est bien plus efficace que la parole seule pour la référence aux objets, à partir du moment où le système de DHM met en jeu une scène visuelle. Un nouveau pan du domaine du DHM s'ouvre alors, de nouvelles questions se posent sur l'ergonomie, la spontanéité du dialogue multimodal, les interactions entre IHM et DHM, et d'une manière générale tous les apports et les contraintes de la multimodalité, voir une synthèse dans (Oviatt, 1999). Parmi ces questions, la suivante va ouvrir de nouvelles perspectives : si un système de DHM est capable de procéder à une interprétation automatique tenant compte de la multimodalité, ne doit-il pas procéder à une génération automatique tenant compte également de la multimodalité ? Avec des démonstrateurs, par exemple dans le domaine du contrôle aérien, c'est la problématique de la multimodalité en sortie qui commence à être explorée et à délimiter son propre domaine de recherche, celui des systèmes de présentation d'information multimodale (Immps, *Intelligent MultiMedia Presentation Systems*, voir chapitre 9).

Les années 1980 sont ainsi riches en remises en question. Après les premiers systèmes qui fascinent et aident à clarifier la méthodologie et les limites du DHM, place au dialogue naturel en langage naturel, avec de nouveaux objectifs tels que le traitement de la parole spontanée, la capture de gestes et ainsi l'exploitation des dispositifs d'interaction, avec tout ce que cela entraîne : gestion du contexte, adaptation au dispositif d'affichage, adaptation à l'utilisateur.

Les années 1990 continuent dans cette voie en élargissant l'éventail des fonctionnalités attendues pour un système de DHM. Ces années correspondent tout d'abord à

l'entrée dans l'ère numérique, ce qui a comme conséquences d'une part un renouveau des recherches théoriques et expérimentales sur la langue orale spontanée, et d'autre part l'introduction de techniques de programmation fondées sur d'importants calculs, notamment probabilistes, coûteux en moyens informatiques. Les recherches sur l'oral étaient jusque-là freinées par les contraintes techniques, mais le numérique va grandement faciliter l'essor des logiciels d'analyse de l'oral, la multiplication des études, et finalement un changement de point de vue sur la langue orale, celle-ci acquérant un statut d'objet d'étude à part entière, et pas seulement celui d'une facette appauvrie, voire erronée, de la langue écrite (Blanche-Benveniste, 2010). Les conséquences pour le DHM sont que les travaux ne se fondent plus seulement sur les grammaires et les règles issues de l'écrit, et intègrent petit à petit les spécificités de l'oral : corrections, reprises, incises, comme nous le verrons en 5.1.2. En revanche, l'exploitation d'une entrée vocale apporte de nouveaux problèmes aux systèmes de DHM, avec par exemple la nécessité pour l'utilisateur d'utiliser un bouton ou une pédale en même temps qu'il parle (*push-to-talk*), afin de signaler au système le début et la fin de son énoncé. Quant aux techniques de programmation, elles s'enrichissent des avancées de l'approche statistique, intègrent des probabilités calculées à partir de corpus, ce qui, comme le souligne (Jurafsky et Martin, 2009, p. 892), amorce le traitement probabiliste des actes de langage, et apporte un complément aux réalisations de systèmes de DHM, qui dépassent en qualité les précédents prototypes de recherche. Des efforts sont faits également sur l'enrichissement des méthodes de compréhension automatique, avec des approches mixtes qui combinent à la fois les techniques « ascendantes » (partant de l'énoncé, le système procède à diverses analyses pour identifier le sens et l'intention sous-jacente) et les techniques « descendantes » (partant de plans et d'intentions possibles, le système procède à diverses analyses pour déterminer quelle intention satisfait l'énoncé). Ces efforts impliquent des recherches sur la représentation des plans et des raisonnements, ce qui suppose que le système arrive à raisonner sur les croyances de l'utilisateur (Vilnat, 2005, p. 6). On voit ainsi apparaître les modèles de type BDI (*Belief, Desire, Intention*).

Parmi les systèmes des années 1990, le système Trains (Allen *et al.*, 1995) est exemplaire parce qu'il cherche à trouver des solutions à un éventail très large d'enjeux autour de la compréhension automatique, du dialogue à initiative mixte (dirigé indifféremment par le système ou par l'utilisateur), de la représentation et du raisonnement sur le temps, les actions et les événements. La tâche relève du transport, mais, contrairement à Gus ou à notre exemple présenté en introduction, elle implique plusieurs moyens de transport et gère donc les connexions entre ces moyens, les problèmes de planification, les optimisations (calculs de temps de parcours), les éventuels conflits, etc.

En France, les systèmes et publications se multiplient (Bilange, 1992 ; Guyomard *et al.*, 1993 ; Duermael, 1994 ; Luzzati, 1995 ; Sabah *et al.*, 1997 ; Grisvard, 2000) et nous retiendrons comme exemple le système Dialors de D. Luzzati, qui concerne encore une fois la réservation de billets de train. La méthodologie commence ici aussi

par une étude de corpus approfondie, en l'occurrence un corpus issu d'enregistrements de la SNCF, corpus qui a fait par ailleurs l'objet de nombreuses publications. Le système Dialors comporte un analyseur, nommé Alors, qui a pour fonction de transformer les énoncés en une représentation interne au système, et un dialogueur, Dialog, qui décide, en fonction de cette représentation, de l'action à mener : demande de clarification, réponse à la requête de l'utilisateur après consultation de la base de données des horaires des trains. Ce deuxième composant a comme spécificité d'implémenter le modèle de dialogue proposé par l'auteur, modèle qui distingue le dialogue « régissant », autrement dit le dialogue principal qui reflète la progression de la tâche, des éventuels dialogues « incidents », autrement dit les demandes de clarification et autres sous-dialogues éphémères, qui n'influent pas sur la progression de la tâche mais permettent aux interlocuteurs de bien se comprendre. Cette structuration du dialogue permet au système de procéder à des analyses fines et surtout d'évaluer en temps réel la progression de la tâche, sans nécessiter pour autant l'implémentation d'un modèle plus complexe comme le modèle hiérarchique de l'école de Genève (Roulet *et al.*, 1985), évoqué plus haut en tant qu'approche de l'analyse du discours.

1.1.3. *Systèmes actuels : multiplicité des domaines et des techniques*

Notre historique a commencé dans les années 1950 et en arrive maintenant aux années 2000. Il est plus difficile d'avoir du recul sur cette période qui inclut les systèmes actuels, d'autant plus que les travaux se sont multipliés et que les techniques ont explosé. D'une manière générale, outre l'amélioration de l'ensemble des modèles des années 1990 (Jurafsky et Martin, 2009, p. 892), les années 2000 voient venir :

– l'application au DHM des techniques informatiques d'apprentissage automatique, afin de reporter une partie des différents réglages sur des traitements de gros corpus ou sur une amélioration des performances au fur et à mesure de l'utilisation du système (Rieser et Lemon, 2011) ;

– l'essor des efforts de standardisation : W3C, ISO, TEI, DAMSL, etc. ;

– l'essor des méthodologies d'évaluation de systèmes (voir chapitre 10) ;

– la multiplication des modalités des communications avec la machine, et donc des modèles et techniques de dialogue multimodal : geste à retour d'effort ou geste haptique, gestes et postures captés par des caméras, prise en compte de la direction du regard, lecture sur les lèvres, etc. (López-Cózar Delgado et Araki, 2005) ;

– la mise en place de liens avec d'autres domaines scientifiques, par exemple la robotique (voir chapitre 10 de Garbay et Kayser, 2011), et d'autres domaines du TAL, par exemple la traduction automatique dans le cadre de systèmes de DHM capables de passer d'une langue à une autre (López-Cózar Delgado et Araki, 2005) ;

– l'essor des « boîtes à outils » et des « ateliers de génie logiciel » pour le prototypage rapide de systèmes de DHM, avec l'exemple répandu de VoiceXML, langage

normalisé pour des applications vocales relativement simples d'un point de vue linguistique ;

– l'intégration de DHM dans des plates-formes d'intercommunication plus larges, qu'il s'agisse de ce que l'on appelle l'intelligence ambiante ou d'autres aspects, par exemple liés à des architectures logicielles (Issarny *et al.*, 2005) ;

– l'essor du domaine des ACA, avec la prise en compte des émotions ;

– l'essor du domaine des SQR.

Sur ce dernier point par exemple, l'intérêt d'intégrer des capacités de dialogue à un SQR est de permettre des échanges pour préciser pas à pas la requête (Vilnat, 2005, p. 48). D'un système de question-réponse (au singulier) qui se contente de trouver un résultat à une requête, à l'image de ce que font les gestionnaires de bases de données ou à l'image d'IBM Watson qui reste contraint par les règles d'un jeu télévisé, on passe ainsi à un système de questions-réponses (au pluriel), où le dialogue vient autoriser les éclaircissements, les précisions, et surtout les questions successives portant sur un même sujet : « est-ce que ce trajet passe par Meudon ? », « est-ce le trajet le plus court pour arriver à Paris ? », « à quelle heure part-il ? ».

Un exemple d'un tel système regroupant DHM et SQR est le projet Ritel (Van Schooten *al.*, 2007 ; Rosset, 2008). L'architecture du système met en avant la gestion des questions, avec des modules dédiés à la détection de thème, la gestion du retour de l'utilisateur, la gestion des historiques, le routage des questions, la gestion des confirmations implicites et celle des demandes complémentaires. Les objectifs du projet mettent clairement en avant les performances pour le SQR autant que pour le DHM, et le projet est ainsi un pas significatif pour les systèmes de DHM en domaine ouvert, qui commencent à émerger dans les années 2000. A titre indicatif et de manière à comparer avec les chiffres déjà mentionnés dans ce chapitre, le vocabulaire de Ritel est de 65 000 mots, ce qui correspond à peu près au nombre d'entrées dans un dictionnaire de langue. Autre exemple des années 2000, le système Amitiés (Hardy *et al.*, 2006), système de DHM en domaine fermé, donne un point de comparaison par rapport aux précédents systèmes du même type et par rapport à un système en domaine ouvert comme Ritel. Amitiés a été conçu suite à une étude de corpus approfondie, sur environ 1 000 dialogues appartenant tous au domaine financier sur lequel porte l'une des tâches. Les chiffres correspondant à ce matériau sont les suivants : 30 000 phrases, pour un vocabulaire de 8 000 mots. C'est beaucoup plus que ce que Gus pouvait faire, mais ça reste bien loin des 65 000 mots de la langue.

Enfin, les années 2000 (et 2010) voient apparaître, comme nous l'avons vu au tout début de ce chapitre, les premiers systèmes de DHM grand public, incorporés à divers sites Internet, robots de compagnie, ou encore téléphones portables, agendas électroniques, systèmes de géolocalisation et autres assistants personnels. Même si la qualité n'est pas au rendez-vous, on peut imaginer que cela contribue à encourager les efforts de la communauté scientifique.

1.2. Une liste des capacités possibles d'un système actuel

Au niveau des systèmes grand public, nous l'avons dit, on est encore loin d'un dialogue naturel en langage naturel. Quelques tests des systèmes dits « à commande vocale » ou « à reconnaissance vocale » permettent de le vérifier rapidement. Par exemple, les systèmes de géolocalisation et les téléphones portables en sont encore à la détection de mots-clés : noms de ville pour les premiers, noms de destinataires pour les seconds. On est très loin de la compréhension automatique d'énoncés tels que « je veux aller à Grenoble en contournant Lyon et en évitant l'autoroute entre Saint-Etienne et Lyon », où l'utilisateur indique à la fois un point de passage et des préférences différentes pour deux parties du trajet (requête bien plus rapide à énoncer qu'à programmer directement sur le système, quand cela est possible). Il faut reconnaître cependant qu'à part des exemples tels que celui-ci, la commande vocale n'est pas souvent adaptée à l'utilisation de systèmes informatiques : c'est bruyant, on n'est jamais sûr d'être bien compris, et on pense toujours faire plus efficacement par manipulation directe du système avec une IHM classique. Contrairement à ce qu'ont affirmé certains chercheurs dans les années 1980, ce n'est pas parce qu'il y a de plus en plus d'ordinateurs et de plus en plus de données accessibles que le DHM va s'imposer en tant que nouvelle forme de communication. Comme l'écrit (Vilnat, 2005, p. 5), il faut plutôt se poser la question de savoir quelles sont les tâches pour lesquelles il est utile de mettre en œuvre un dialogue plutôt que toute autre technique d'interaction : le principal frein, c'est le faible intérêt que l'utilisateur peut retirer de l'usage d'un système de DHM.

Au niveau des prototypes de recherche, le dialogue naturel en langage naturel devient réalisable, en tout cas dans le cadre d'une tâche ciblée. C'est le cas pour le dialogue en domaine fermé, et également pour certains démonstrateurs en domaine ouvert comme IBM Watson. Il est à noter cependant que les efforts récents ont plus porté sur un élargissement des capacités des systèmes plutôt que sur un approfondissement des aspects de TAL. C'est ce que nous allons voir avec trois volets correspondant aux trois caractéristiques d'un système cognitif : le traitement des entrées (paragraphe 1.2.1), les analyses internes au système (paragraphe 1.2.2) et la gestion des sorties (paragraphe 1.2.3).

1.2.1. *Les dispositifs de capture et leur exploitation*

Le chapitre 2 de (López-Cózar Delgado et Araki, 2005) fait une liste très complète des systèmes de DHM multimodaux avec les traitements réalisés sur les entrées. Sans refaire une telle liste, citons rapidement les captures suivantes : la capture de la parole ; la lecture sur les lèvres de l'utilisateur pour aider, voire remplacer celle-ci (environnement bruité, utilisateur handicapé, chuchotement) ; la reconnaissance de l'utilisateur ; la localisation et le suivi de son visage, de sa bouche, de ses yeux et par conséquent de la direction de son regard (à la fois pour surveiller son attention par rapport au dialogue en cours et pour aider à la résolution d'une référence à un objet de la scène) ; la

capture des émotions faciales ; la capture des gestes de désignation, notamment ceux de la main, et plus généralement de tout type de geste effectué avec les mains ou le corps. Par ailleurs, nous avons déjà évoqué le geste à retour d'effort, dans le cadre d'une interaction haptique : il s'agit d'un dispositif qui gère à la fois la capture de position de la main et l'émission à destination de l'utilisateur d'une éventuelle résistance. L'intérêt est de coupler ce dispositif à une immersion dans un environnement virtuel, l'utilisateur voyant par exemple une représentation graphique de sa main en train de manipuler les objets de la scène virtuelle. Dans ce contexte, le retour d'effort prend tout son sens : il simule une perception tactile qui complète la perception visuelle.

Aucun système ne réalise tout cela simultanément et en temps réel, mais c'est un enjeu intéressant pour les chercheurs les plus technophiles de la communauté du DHM. On le voit, les possibilités sont nombreuses, et les enjeux informatiques très larges : les traitements associés à ces types de capture incluent de nombreuses problématiques relevant de la vision artificielle, du traitement du signal, des modélisations mathématiques adaptées à la représentation de configurations et de trajectoires, tout cela avec des contraintes de rapidité d'exécution, de précision, et d'abstraction en représentations manipulables efficacement par le système, afin que celui-ci puisse confronter ces représentations avec celles issues de la compréhension automatique des énoncés. Comme le montre par exemple (Bellalem et Romary, 1996) pour des trajectoires gestuelles effectuées sur écran tactile, une représentation d'un geste sous la forme d'une suite de plusieurs centaines de positions n'est tout simplement pas gérable. Il est nécessaire d'en abstraire des régularités et des instants signifiants, afin d'aboutir par exemple à une courbe qui puisse se décrire en quatre ou cinq paramètres. Si cette courbe sert ensuite à aider à la résolution d'une référence à un objet, il sera possible de la confronter avec une représentation, simplifiée elle aussi, de la scène visuelle et des objets qui y apparaissent.

Certains traitements nécessitent des dispositifs de capture spécifiques, avec les exemples immédiats du microphone pour le traitement de la parole et du clavier pour le traitement de l'écrit. Les autres traitements peuvent se faire de plusieurs façons, que l'on peut classer de la plus gênante pour l'utilisateur à la plus transparente. Un exemple de « capture gênante » est le gant de désignation que l'utilisateur doit revêtir afin que le système puisse capter la position et la configuration de sa main, ou encore le gant avec exosquelette nécessaire au retour d'effort. L'exemple de plus en plus répandu de « capture transparente » est la caméra ou le système de caméras couplées qui permet, en toute liberté pour l'utilisateur, de procéder à plusieurs traitements simultanés, par exemple le suivi du visage et la détection de la configuration de la main.

La reconnaissance automatique de la parole est un domaine de recherche à part entière, et son utilisation dans le cadre d'un système de DHM pose des problèmes supplémentaires (Jurafsky et Martin, 2009). Le principe consistant à passer d'un signal audio à une transcription suivant un code plus ou moins proche du langage écrit, nécessite plusieurs sources de données, dont les suivantes : un modèle acoustique, une

liste des mots de la langue, un dictionnaire des prononciations, et, source de données quasiment indispensable pour augmenter les performances, un « modèle de langage ». Ce modèle est construit à partir d'analyses statistiques de corpus. En apportant la notion de contexte (un, deux ou trois mots précédents), il permet au système de calculer des probabilités et de retenir les hypothèses les plus probables pour le mot (ou autre unité) en cours de reconnaissance. Dans le cadre d'un système de dictée vocale, le modèle de langage est construit à partir de calculs réalisés sur des textes issus de la littérature ou de la presse écrite. On maximise la taille de ces textes, de façon à affiner la modélisation de la langue en termes de successions possibles de mots. Dans le cadre d'un système de DHM, la construction de ce modèle nécessite de bien choisir les corpus exploités pour les calculs statistiques : ce n'est pas forcément pertinent de ne retenir que des transcriptions de dialogues oraux, mais disposer de dialogues proches de ceux attendus pour le système est un avantage certain, quitte à devoir gérer plusieurs modèles de langage. Par ailleurs, l'alternance des interventions de l'utilisateur et du système apporte une contrainte supplémentaire : les probabilités pour un énoncé de l'utilisateur dépendent de ce que le système vient de dire. Les modèles de langage doivent ainsi tenir compte de l'état du dialogue, et deviennent de ce fait plus difficiles à gérer.

Une autre difficulté s'ajoute par rapport à la dictée vocale : si le résultat de celle-ci consiste en un texte écrit correspondant à ce qui est prononcé, le résultat du module de reconnaissance de la parole dans un système de DHM peut être beaucoup plus détaillé. Il peut tout d'abord comporter plusieurs hypothèses de reconnaissance, de manière à ce que les modules suivants fassent un choix en fonction de leurs propres attentes. Lorsque l'énoncé comporte un mot inconnu, c'est-à-dire une suite de phonèmes qui ne correspond à aucun mot du lexique, le module de reconnaissance a le choix entre plusieurs solutions : soit ramener à tout prix vers l'un des mots du lexique, quitte à ce que les prononciations diffèrent sévèrement, soit essayer de transcrire la suite de phonèmes avec une orthographe plausible compte tenu de la langue. Si ces deux solutions peuvent être acceptables pour une dictée vocale, la seconde étant par exemple parfaitement adaptée à la transcription de noms propres inconnus du système, ce n'est pas le cas en DHM : non seulement le module de reconnaissance doit indiquer à l'aide d'une étiquette spécifique qu'il s'agit d'un mot inconnu, mais il doit de plus transmettre un code décrivant la prononciation de ce mot, de manière à ce que le système puisse l'ajouter à son vocabulaire et le prononcer à son tour, ne serait-ce que pour demander à l'utilisateur de quoi il s'agit. Pour bien faire, chaque mot reconnu est affecté d'un score de confiance, et l'analyseur syntaxique ou sémantique exploite ces scores de confiance et ses propres préférences pour retrouver (plutôt que de se voir imposer) la transcription la plus vraisemblable de l'énoncé.

Un aspect supplémentaire ayant des conséquences sur la nature du résultat transmis aux autres modules du système de DHM réside dans la prosodie. Qu'il s'agisse du rôle du module de reconnaissance ou d'un autre module, spécifique, il est utile que la transcription écrite de l'énoncé s'accompagne d'un codage, d'une transcription de la

prosodie. Nous le verrons dans les chapitres 5, 6 et 7, la prosodie permet en effet d'aider à l'analyse sémantique (en apportant des indices de focalisation), à la résolution des références quand un geste est utilisé conjointement à une expression référentielle, et à l'identification des actes de langage, en apportant un contour intonatif qui permet de privilégier une hypothèse par rapport à une autre. On attend ainsi plusieurs indications de la part du module d'analyse prosodique : un repérage des accents de focalisation, un découpage temporel de l'énoncé, mot par mot, de manière à apparier mots et gestes en dialogue multimodal, et un codage des principales caractéristiques de l'intonation. Des analyses plus fines, avec par exemple la détection des périodes, nécessite des indications supplémentaires, mais relèvent pour l'instant plus de l'analyse *a posteriori* de corpus oraux que d'analyse en temps réel pour le DHM. C'est d'ailleurs le reproche que l'on peut faire à beaucoup de systèmes de l'état de l'art : ils n'exploitent pas la prosodie, alors que c'est une composante essentielle de la langue orale. Des initiatives telles que celle de (Edlund *et al.*, 2005), qui présente Nailon, un système d'analyse automatique de la prosodie capable de détecter en temps réel plusieurs caractéristiques prosodiques d'un énoncé en DHM, sont importantes.

Un dernier aspect où le module de reconnaissance automatique de la parole a un rôle à jouer est la gestion des tours de parole. Les systèmes de DHM se sont longtemps cantonnés à un fonctionnement alterné des interventions, le système ne coupant jamais la parole à l'utilisateur et ne commençant à parler qu'après la fin de l'énoncé de celui-ci. Plus que cela, nous avons vu (paragraphe 1.1.2) avec le bouton ou la pédale *push-to-talk* qu'il revenait à l'utilisateur d'indiquer à la machine le début et la fin de son intervention. On est maintenant en droit d'attendre d'un système de DHM qu'il laisse l'utilisateur s'exprimer à tout moment, sans contrainte, et donc que ce soit à la machine de détecter le début et la fin des interventions. C'est justement l'une des fonctions du système Nailon, qui se sert par exemple des indices prosodiques de fréquence fondamentale et de rythme pour détecter automatiquement la fin d'une intervention de l'utilisateur.

1.2.2. *Les capacités d'analyse et de raisonnement*

Une fois les signaux captés en entrée du système et transcrits dans une représentation appropriée, de nombreuses analyses et raisonnements vont être lancés de manière à ce que le système comprenne le sens de l'énoncé de l'utilisateur, l'intention de celui-ci, et par conséquent la réponse à lui donner. Les analyses relèvent de la compréhension automatique du langage naturel, donc du TAL, et regroupent les facettes suivantes : l'identification des mots (analyse lexicale) de manière à retrouver leurs sens (sémantique lexicale), stockés dans le système selon un formalisme bien choisi ; l'identification de la structure de la phrase et des fonctions grammaticales des composants identifiés (analyse syntaxique) ; la construction de la sémantique de la phrase en combinant les sens des mots et en suivant les relations syntaxiques (sémantique propositionnelle) ; l'attribution de référents aux pronoms de première et deuxième

personnes, aux expressions référentielles en général, et aux anaphores en particulier (analyse pragmatique parfois appelée « du premier degré ») ; l'identification de l'implicite et des contenus attachés à celui de l'énoncé (pragmatique « du second degré ») ; la détermination des actes de langage afin que le système puisse comprendre la nature de l'intervention de l'utilisateur (pragmatique « du troisième degré »). Au-delà de la simple transcription du « sens littéral » d'un énoncé, on entre ici dans le domaine de la détermination de son « sens en contexte ». Comme pour le TAL, les méthodes et algorithmes implémentés ont évolué pour l'ensemble de ces analyses. Là où, à une époque, les approches symboliques issues de l'IA étaient les seules présentes, on voit maintenant des approches statistiques, et ce à tous les niveaux de la liste ci-dessus. Ces approches ont montré leur efficacité dans beaucoup de domaines du TAL, et ont parfois remplacé complètement les approches symboliques. Dans le domaine du DHM, c'est l'hybridation d'approches symboliques et statistiques qui donne les résultats les plus prometteurs.

Avec comme point de départ une représentation sémantique fidèle à l'énoncé, on arrive alors à une représentation enrichie par le ou les messages, implicites ou explicites, que cet énoncé véhicule. C'est cette représentation enrichie que le système va confronter, en interne, aux représentations manipulées précédemment au fur et à mesure du dialogue. C'est aussi à l'aide des informations qu'elle contient que le système va pouvoir abstraire la structure du dialogue et la comparer aux structures envisagées pour la tâche. Cette approche a été imaginée dès les années 1980 (Reichman, 1985) mais son implémentation informatique dans le cadre d'un système de DHM réel n'a pu se faire que bien plus tard, et est toujours d'actualité. Le système peut aussi procéder à une évaluation de la satisfaction de la tâche, identifier les manques, et décider ainsi quelle va être sa prochaine intervention. Tout ceci relève du raisonnement qu'il met en œuvre afin de traiter l'énoncé de l'utilisateur de la façon la plus pertinente possible, compte tenu de ce qui a déjà été fait, de l'apport que représente cet énoncé, et de ce qu'il reste à accomplir pour satisfaire la tâche. Plus qu'en linguistique et en pragmatique, on se place ici dans l'IA moderne : les thématiques abordées sont celles de la représentation des connaissances, et notamment en suivant des formalismes issus de la logique, de manière à autoriser des déductions automatiques, et celles des systèmes experts et de la décision multicritère. En fait, comme pour les analyses évoquées rapidement dans le paragraphe précédent, on assiste à une hybridation de plusieurs approches. A titre d'exemple, les approches fondées sur le pouvoir d'expression d'une logique bien choisie, et sur l'adéquation de celle-ci au langage naturel, ont exploré différents types de logiques : logiques propositionnelles, logiques modales, logiques temporelles, logiques de description, logiques hybrides.

1.2.3. *Les types de réaction du système et leur manifestation*

Une fois que le système a décidé quelle action effectuer en réaction à l'énoncé de l'utilisateur, il reste à matérialiser cette action. Dans le cas d'un système uniquement

écrit ou uniquement oral, il s'agit de générer un énoncé en langage naturel. La génération automatique est un domaine de recherche à part entière (Reiter et Dale, 2000) qui inclut plusieurs facettes telles que la construction de la phrase et la détermination des expressions référentielles. On retrouve les mêmes problématiques que celles impliquées dans la compréhension automatique, mais inversées. Si certaines ressources linguistiques sont communes, les méthodes et algorithmes en génération sont spécifiques, et ne consistent pas en un simple renversement de leurs équivalents en compréhension. Dans le cas d'un système oral, la dernière étape réalisée par le système est la synthèse vocale, c'est-à-dire la prononciation de l'énoncé retenu. On y retrouve les préoccupations de la prosodie : pour faire vrai, l'énoncé doit être prononcé avec une intonation, un rythme et éventuellement une focalisation qui soient parfaitement en lien avec l'intention de communication du système.

Dans le cas d'un système multimodal, par exemple lorsqu'un avatar représente graphiquement la machine, le problème inclut la génération automatique de gestes, et leur synchronisation temporelle avec les mots du message verbal généré. Quand le geste est possible, les problèmes de génération, et notamment celui de la détermination des expressions référentielles, prennent une autre dimension : parole et geste sont complémentaires, et le système doit choisir quelle part du message affecter à l'une et à l'autre. De plus, la conception graphique de l'avatar lui-même est un domaine qui pose des questions importantes autour du réalisme de l'apparence physique, du regard de l'avatar, de ses mouvements : mouvements des yeux, des sourcils, des lèvres quand un message est verbalisé, de la tête en cas d'acquiescement, et plus généralement du corps. Les indications données à l'utilisateur à travers ces mouvements jouent un rôle dans la communication homme-machine, et il est essentiel que les divers mouvements impliqués vers un même but, par exemple ceux du regard et des sourcils qui indiquent le degré d'attention de l'avatar, soient correctement synchronisés. Par ailleurs, les émotions transmises par le biais des gestes effectués constituent aussi un domaine de recherche à part entière, qui nécessite des études sur les typologies des émotions, sur leur pertinence en DHM, et sur la façon de les rendre, non seulement visuellement, mais aussi par la parole. Enfin, dans le cas d'un système incluant une IHM et manipulant une grande quantité de données, le problème inclut également celui de la présentation graphique d'informations (cas des Immps cités plus haut). Il se peut par ailleurs que l'IHM produise de son côté des sons (*earcons*). La génération de sons et celle de parole doivent alors être réalisées de manière pertinente, c'est-à-dire sans superposition risquant de gêner la perception de l'utilisateur.

1.3. Les enjeux actuels

Une machine peut-elle penser ? Le test de Turing et la quête du Graal de la machine dialoguante fascinent toujours autant, mais les limites des systèmes de DHM actuels ne se posent plus dans ces termes. De manière plus pragmatique, elles se posent en termes de limites dans les capacités à modéliser et à traiter automatiquement la langue

naturelle, dans les capacités à représenter et à raisonner sur des représentations logiques, dans les capacités de traitement et d'intégration de signaux divers et variés par la machine. Les systèmes commercialisés le montrent tous les jours : un système de DHM ne fonctionne bien que dans un cadre applicatif délimité, c'est-à-dire dans un cadre suffisamment restreint pour que le maximum de possibilités d'interaction aient été imaginées en amont, lors de la phase de conception. Contrairement à ce que des tentatives comme Eliza ont pu le faire croire, rien n'est magique, et ce quelques soient les technologies mises en œuvre, qu'elles soient symboliques, statistiques, qu'elles impliquent de l'apprentissage automatique ou non. Tout doit être anticipé, et cela représente une quantité de travail à la mesure des capacités envisagées pour le système.

L'enjeu principal du DHM reste, comme à l'époque de (Pierrel, 1987), l'élaboration pluridisciplinaire de systèmes compréhensifs permettant à l'utilisateur de s'exprimer spontanément comme il le fait avec un interlocuteur humain, et ceci pour une multitude d'applications afin de proposer des systèmes accessibles à tous, dans toutes les situations de la vie courante. Plus précisément, nous allons développer quatre ensembles d'enjeux : les enjeux théoriques, les enjeux qui concernent l'éventail des capacités attendues pour un système, les enjeux techniques qui concernent la conception de systèmes, et les enjeux techniques qui cherchent à faciliter le développement informatique.

1.3.1. *Adapter et intégrer des théories existantes*

Selon (Cole, 1998, p. 191), les avancées récentes n'ont pas inclu le développement de nouvelles théories mais ont porté sur des extensions et des intégrations de théories existantes. On trouve ainsi beaucoup d'approches hybrides, qui exploitent le pouvoir d'expression de plusieurs théories existantes. Ce constat, que nous avions fait plus haut sur les rapprochements entre les approches symboliques et les approches statistiques, est toujours d'actualité. En linguistique, nous avons mentionné les analyses prosodique, lexicale, syntaxique, sémantique et pragmatique. Or ce qui a longtemps été considéré comme une succession d'analyses réalisées en cascade est maintenant appréhendé d'une tout autre façon : une partie des résultats de l'analyse prosodique sert à l'analyse sémantique, une partie des résultats de l'analyse syntaxique sert à l'analyse pragmatique, et celle-ci n'est d'ailleurs pas monolithique mais implique plusieurs facettes quasiment indépendantes les unes des autres. Un enjeu consiste ainsi à revoir complètement les découpages classiques en niveaux d'analyse du langage naturel, et à mieux intégrer les analyses qui ont des buts communs. En DHM, les objectifs constituent une liste qui dépend du système visé mais qui inclut au minimum :

– la détection de la fin de l'énoncé de l'utilisateur ;

– la représentation du sens de celui-ci sous une forme logique, ou, du moins, sous la forme d'une structure de données qui soit manipulable par les algorithmes envisagés ;

– la résolution des références aux objets gérés par l'application ;

– l'identification des contenus implicites véhiculés par l'énoncé, sans y être pour autant explicités ;

– la mise à jour de l'historique du dialogue.

Chaque objectif de cette liste est atteint à l'aide de la collaboration de plusieurs analyses. Par exemple, pour détecter automatiquement la fin de l'énoncé de l'utilisateur, on a besoin d'une analyse prosodique, qui indique quand le contour intonatif descend et apporte alors un indice, et on a besoin d'une analyse syntaxique, qui indique si la suite de mots capturés jusqu'à présent constitue ou non une phrase grammaticale, avec ou sans besoin de mot supplémentaire. En fonction de la personnalité du système, notamment de sa tendance à couper la parole de l'utilisateur, on peut même imaginer qu'une analyse sémantique apporte un argument supplémentaire, dès qu'un résultat sémantique est obtenu. Si l'on reste dans un fonctionnement en cascade des analyses, ce type de mécanisme est impossible. Un enjeu consiste donc à explorer les implémentations d'analyses collaboratives. Si on lance la première analyse à la fin de l'énoncé de l'utilisateur, alors on perd toute possibilité d'interaction en temps réel du système. Un enjeu consiste ainsi à permettre l'exécution des analyses à tout moment, quasiment à chaque mot prononcé par l'utilisateur. Si l'on considère un module comme une boîte noire qui fournit un résultat en une seule fois et dans une seule structure de données, alors l'analyse prosodique ne doit pas se matérialiser en un seul module, mais en plusieurs : un pour la détermination du contour intonatif, un pour la détection des proéminences, un pour le rythme, etc. Un découpage modulaire qui suit le découpage en niveaux d'analyse linguistique ne se justifie donc plus, et l'application des théories linguistiques au DHM constituent toujours des enjeux de recherche. En dialogue multimodal, l'intégration de théories est encore plus cruciale : les aspects gestuels sont liés aux aspects prosodiques, aux aspects ergonomiques, etc. Comme nous allons le voir dans le chapitre 2, les collaborations entre disciplines sont essentielles.

Enfin, pour terminer cette liste d'enjeux théoriques, soulignons l'importance de la méthodologie, avec la nécessité de réaliser des expérimentations, et la nécessité de constituer et d'exploiter des corpus de référence pour le DHM. Cet enjeu lié aux ressources est crucial non seulement pour l'étude de dialogues portant sur une tâche donnée, mais aussi pour l'exécution d'algorithmes d'apprentissage automatique, ou encore pour la constitution de données tels que des lexiques, des grammaires et des modèles de langage, pour le dialogue oral comme pour le dialogue multimodal. Ici aussi, un enjeu réside dans une meilleure intégration de ces ressources. A titre d'exemple, le projet Ozone dont nous avons parlé a permis de commencer à réfléchir au concept de métagrammaire (ou métamodèle), avec comme but l'instanciation à partir d'une base commune d'une grammaire linguistique et d'un modèle de langage statistique.

1.3.2. *Diversifier les capacités des systèmes*

Les enjeux techniques liés aux capacités d'un système de DHM regroupent les enjeux du TAL, de l'IA, des ACA, des SQR, des IHM, et bien d'autres. D'une manière générale, tous les composants dont nous avons parlé peuvent faire l'objet d'améliorations, avec un plus grand éventail de phénomènes pris en compte et une plus grande finesse dans les traitements. (Cole, 1998) met en avant quelques aspects linguistiques comme l'exploration de la nature des segments de discours et des relations de discours, ainsi que le besoin de mécanismes supplémentaires pour gérer des phénomènes-clés tels que la mise en avant d'une information dans un message linguistique. Tous ces enjeux relèvent d'un même but : augmenter la couverture, la fluidité et ainsi le réalisme du dialogue. Pour caricaturer, le but est peut-être d'obtenir un système de dialogue, voire de multilogue (Knott et Vlugter, 2008), naturel en langage naturel, qui soit multimodal, multilingue, multitâche, multirôle, *multithread*, multi-utilisateur et bien sûr capable d'apprentissage...

La question du « réalisme » est une vaste question, qui passe déjà par la rapidité : un système qui met dix secondes à répondre n'a aucune chance d'atteindre le réalisme. Si ce critère est mesurable, il n'en est pas de même d'autres critères : comment mesurer le réalisme d'une voix synthétique, d'une construction de phrase, des gestes d'un ACA ? Le rejet par certains utilisateurs d'une voix artificielle repose parfois sur de petits détails difficiles à mesurer, par exemple un défaut minime dans le rythme d'élocution. La perception de ces défauts minimes peut provoquer un malaise, peut déranger. Le domaine de la robotique, ou encore celui des images de synthèse, utilisent le terme de « vallée dérangeante » pour décrire ce type de phénomène. Le problème est que l'on cherche à se rapprocher de l'humain (du réel pour les images de synthèse), mais qu'il reste un léger écart entre ce que l'on obtient et ce que l'on vise. Or cet écart, aussi minime soit-il, suffit à être perceptible et à déranger. Pour contrecarrer, certains concepteurs rendent l'écart explicite et oublient l'objectif de se rapprocher de l'humain. C'est ainsi que certains jouets mécaniques qui ont l'apparence d'animaux ne sont pas dotés de fourrure. En DHM, c'est par exemple ainsi que le service *web* Ananova prend l'apparence d'une belle jeune femme... aux cheveux verts (Harris, 2005, p. 341).

Enfin, un enjeu essentiel quant aux capacités d'un système de DHM est sa robustesse, c'est-à-dire à la fois sa capacité à gérer ses propres insuffisances, au niveau des analyses linguistiques par exemple, ses propres manques et erreurs, et sa capacité à toujours rebondir, à faire progresser le dialogue coûte que coûte, en s'aidant ou non de la tâche à résoudre. Cela implique de concevoir des modules capables de fonctionner avec des entrées incomplètes, et d'avoir des stratégies pour gérer les problèmes. Cela implique également de prévoir, dès les premières phases de conception, des tests et des paramétrages avec des données réelles, des conditions réelles et non des conditions contrôlées de laboratoire.

1.3.3. *Rationaliser la conception*

Au niveau de la réalisation, les enjeux méthodologiques et techniques sont multiples. Une fois la liste des capacités de compréhension et de génération déterminée, il s'agit de les instancier et de les organiser en modules, composants ou agents dans une architecture, de spécifier les langages d'interaction entre ces éléments, les méthodes d'évaluation, de construction des ressources nécessaires, d'intégration. L'enjeu principal est ici la rationalisation de l'ingénierie des architectures (voir chapitre 4), et d'une manière générale la rationalisation des flux de production, comme dans tout domaine technique professionnel. (Harris, 2004) fait ainsi de son chapitre 9 une description très précise d'une équipe de conception, avec les différents métiers qui interviennent : chef d'équipe dialogue ; « architecte » de l'interaction ; lexicographe, en charge des aspects liés aux corpus ; « scénariste », en charge de l'anticipation des types de dialogues attendus, mais aussi de la définition de la personnalité du système et de ses réactions possibles ; « ingénieur qualité » ; sans oublier les experts en ergonomie, en technique, ainsi qu'un expert du domaine couvert par la tâche. La tâche, et plus généralement le contexte du dialogue, peut nécessiter une intégration avec un autre domaine de recherche. Un premier exemple est celui de la robotique, où l'on commence à voir des systèmes intégrant des capacités propres à la robotique et des capacités de DHM, préférentiellement multimodal (Gorostiza et Salichs, 2011). Un deuxième exemple est celui des IHM quand on cherche à les doter de la parole, tout en conservant d'une part les possibilités de manipulation directe de l'IHM, d'autre part les avantages de l'IHM en termes d'ergonomie, de plasticité : adaptation à l'utilisateur, au terminal, à l'environnement.

1.3.4. *Faciliter le développement informatique*

Au niveau du développement des systèmes, les enjeux techniques relèvent de la facilitation des processus de développement. Un premier pas dans ce sens est la multiplication des boîtes à outils et des ateliers de génie logiciel dédiés au DHM. VoiceXML est un exemple basique, mais il existe beaucoup d'autres plates-formes dédiées par exemple à l'aide à la conception de systèmes de dialogue multimodaux (López-Cózar Delgado et Araki, 2005). Un deuxième pas serait la mise en place d'une librairie proposant un panel riche et performant d'outils de TAL et de gestionnaires de dialogue. C'est un enjeu important, initié avec des tentatives telles que OpenNLP de Apache pour quelques aspects de TAL à l'écrit. Une librairie « OpenDial » serait probablement utile, et permettrait de cibler les efforts ailleurs que vers les composants communs à tous les systèmes. Enfin, un troisième pas dans le même sens serait la matérialisation de tout un ensemble de services liés à la reconnaissance vocale, à la synthèse vocale, aux analyses prosodiques, syntaxiques et sémantiques, dans une couche logicielle de type *middleware*, ou, mieux, dans une carte d'extension d'ordinateur, à l'instar des cartes graphiques pour la visualisation en 3D. Cet enjeu, s'il se réalise un jour, permettrait de disposer de facilités exceptionnelles pour développer un système : tout

processus effectué de manière *hardware* plutôt que *software* gagne énormément en rapidité, et ce serait ouvrir véritablement la porte aux systèmes utilisables en temps réel. Bien entendu, ce n'est pas un enjeu simple, et, si l'on fait la comparaison avec la 3D pour laquelle la carte graphique sert beaucoup plus lors de la conception que lors du rendu final qui nécessite des processus trop spécifiques et trop fins, on pourrait imaginer qu'une « carte dialogue », dans un premier temps, accélère et simplifie la conception de systèmes, sans pour autant procéder au développement complet.

1.4. Bilan

La quête de la machine capable de comprendre le langage humain et de répondre à son utilisateur aussi bien que le ferait un interlocuteur humain dure depuis plus d'un demi-siècle. Les difficultés posées par le traitement automatique des langues n'a pas permis jusqu'à ce jour d'aboutir à des systèmes véritablement compréhensifs. En revanche, nous assistons à une diversification des modalités de communication et des facettes de l'interaction homme-machine. En s'appuyant sur les étapes théoriques, méthodologiques et techniques qui ont marqué l'histoire du dialogue homme-machine, ce premier chapitre fait le point sur les capacités envisageables pour un système de dialogue homme-machine, et sur les limites et les enjeux scientifiques actuels dans ce domaine.

Chapitre 2

Les disciplines du dialogue homme-machine

Le « dialogue » désigne les différentes formes d'entretien entre deux personnes, et caractérise un certain usage du langage. (Clark, 1996, p. 23) avance six propositions portant sur l'usage du langage : il a des buts sociaux ; c'est une action conjointe ; il implique le sens voulu par le locuteur et l'interprétation faite par l'interlocuteur ; il se matérialise avant tout par la conversation de face à face ; il implique souvent plus d'une activité ; son étude est une science à la fois cognitive et sociale. Ces propositions nous permettent de commencer à identifier les disciplines impliquées dans l'étude du dialogue : linguistique et pragmatique bien sûr, mais aussi sciences sociales et sciences cognitives. Ce sont des aspects que nous allons explorer dans ce chapitre, en gardant comme objectif les possibilités d'application au DHM des études en question.

Nous avons mentionné à la fin du chapitre précédent une liste de métiers pouvant constituer une équipe de travail pour la réalisation d'un système de DHM. Cette liste, issue de (Harris, 2004), met surtout en avant des métiers informatiques. On peut tout autant imaginer qu'elle incorpore des linguistes, des sociologues et des psychologues. On peut imaginer qu'elle inclut également, ou surtout, des spécialistes des interfaces entre ces disciplines, notamment des chercheurs qui se situent entre la linguistique et l'informatique, et qui aident au développement de systèmes en extrayant puis formalisant les théories et modèles linguistiques qui leur semblent applicables au DHM. L'application d'une théorie linguistique au DHM peut nécessiter sa simplification, ou du moins sa transformation de manière à obtenir une théorie linguistique légèrement différente mais directement implémentable et donc testable. C'est ce qui a été fait pour la plupart des systèmes cités dans le chapitre précédent, notamment ceux partant de la théorie des actes de langage (Austin, 1962 ; Searle, 1969) et de la structure hiérarchique du dialogue (Roulet *et al.*, 1985). D'une manière générale, la démarche qui vise à formaliser l'interprétation du langage naturel et la gestion du dialogue est la

suivante : elle commence avec l'étude linguistique de phénomènes de communication (écrite et orale), leur analyse méthodique, leur classification et leur caractérisation, c'est-à-dire l'identification de leurs traits pertinents. Cette phase fait appel à la linguistique en général et à la sociologie, qui apporte des explications supplémentaires quant au fonctionnement des tours de parole ou des formules de politesse. Vient ensuite la phase de modélisation, qui consiste à identifier des règles permettant de prendre en compte le maximum de phénomènes. Cette phase fait appel à la linguistique formelle, au TAL, et à l'informatique dès que l'on passe à la formalisation et à l'implémentation. Elle s'accompagne d'hypothèses, afin de déterminer par exemple une règle non explicitée par les théories, et emprunte alors à la psycholinguistique des protocoles expérimentaux pour tester ces hypothèses. Enfin, la gestion du dialogue et la génération de messages à destination de l'utilisateur impliquent des prises de décision pour lesquelles la psychologie et les sciences cognitives s'avèrent précieuses. C'est par exemple la prise en compte des concepts de charge cognitive ou de mémoire de travail qui permet au système de s'adapter aux capacités de son interlocuteur humain.

Ce chapitre, que l'on peut appréhender comme une liste des domaines qu'il est utile de connaître avant de se lancer dans celui du DHM, fait une synthèse sur les aspects cognitifs (section 2.1), sur les aspects linguistiques (section 2.2), puis sur les aspects informatiques (section 2.3), ces derniers devenant très nombreux au fur et à mesure que les capacités de traitement des systèmes s'élargissent.

2.1. Aspects cognitifs

Un système de DHM est un exemple de « système cognitif », c'est-à-dire qu'il se caractérise par le fait qu'il traite des données qui lui sont extérieures, qu'il dispose de connaissances, et qu'il a un comportement fondé à la fois sur ces connaissances et sur les données traitées. Selon la façon de modéliser un système cognitif, on peut suivre la voie du cognitivisme structural, qui met en avant des structures et des mécanismes de fonctionnement de ces structures, ou celle du cognitivisme computationnel, qui met en avant le traitement d'un flux informationnel et qui tend à assimiler les systèmes cognitifs que sont les humains avec les systèmes cognitifs artificiels.

Les sciences cognitives regroupent les disciplines qui étudient les systèmes cognitifs, naturels ou artificiels. De fait, ce sont aussi les disciplines qui s'intéressent au langage, en tant que faculté humaine. La linguistique en fait bien entendu partie : elle aborde le langage à travers l'étude des langues, en tant que systèmes de signes, et comporte une facette sémantique qui s'intéresse aux sens et qui se rapproche ainsi de la sémiotique, dont l'objet plus spécifique est le sens des signes. La psychologie en fait partie également : elle étudie le comportement de l'humain et aborde le langage à travers les opérations de pensée impliquées dans l'activité du langage (psychologie cognitive), l'acquisition par l'enfant (psychologie développementale), ou encore par les influences des autres individus (psychologie sociale). La sociologie étudie les

comportements des groupes et des sociétés, et aborde le langage par l'étude des interactions, comme nous l'avons évoqué avec l'approche de l'analyse conversationnelle, et par l'étude des représentations collectives véhiculées par le langage. Les neurosciences, que l'on peut rapprocher de la psychologie clinique et de la psychopathologie, étudient l'anatomie et le fonctionnement du cerveau, et abordent le langage à travers notamment ses pathologies. La psycholinguistique, branche de la psychologie cognitive liée bien sûr à la linguistique, et aussi de manière historique à la théorie de l'information avec les notions d'information, de code et de message, étudie les activités psychologiques par lesquelles un sujet acquiert et met en œuvre le système de la langue, à travers notamment des expérimentations en laboratoire. La neuropsycholinguistique réunit les préoccupations de la psycholinguistique et de la neurolinguistique (branche des neurosciences portant sur la compréhension, la production et l'acquisition du langage) et se focalise sur l'étude de l'activation des zones du cerveau.

La philosophie, avec ses facettes que sont la philosophie du langage et plus récemment la philosophie cognitive, contribue aux différents points de vue, dont celui de l'histoire des idées, dont celui de la logique sous-jacente au langage, et apporte un éclairage épistémologique. A l'image de la philosophie cognitive, l'ajout du qualificatif « cognitif » permet de préciser, voire de délimiter, un nouveau domaine scientifique dans une discipline établie. C'est le cas par exemple de l'ergonomie cognitive, qui désigne l'étude des interactions entre un utilisateur humain et un dispositif, qui fait ainsi intervenir les notions de charge mentale ou de facteurs humains, et qui intervient dans la conception d'IHM et de systèmes de DHM, notamment pour la génération automatique et la présentation d'information multimédia. C'est le cas également de la linguistique cognitive qui, au départ, apporte à la linguistique la notion de plausibilité cognitive, et s'est maintenant constituée en champ disciplinaire autour de questions portant sur les liens entre langage et pensée, sur la nature des connaissances constituant la faculté de langage, ou sur la modélisation informatique de ces connaissances.

Par ailleurs, l'IA, dont nous avons déjà parlé, est historiquement l'une des disciplines des sciences cognitives, après les tentatives qu'ont été la première cybernétique (étude générale du fonctionnement de l'esprit, impliquant aussi bien des psychologues et des anthropologues que des mathématiciens et des logiciens), la seconde cybernétique (étude de l'évolution des systèmes cognitifs et de leur auto-organisation), puis la théorie des systèmes qui a émergé avec la robotique et les systèmes experts. Son objectif est de développer les systèmes artificiels, et elle aborde le langage par la modélisation et la simulation informatique (Garbay et Kayser, 2011). Même si l'on parle maintenant plus volontiers de l'intelligence ambiante, c'est bien sûr la voie du DHM. (Guyomard *et al.*, 1993-2006) montre que l'IA apporte un point de vue complémentaire aux approches linguistiques : alors que celles-ci supposent que la structure du dialogue concerne les énoncés eux-mêmes, c'est-à-dire leurs formes et leurs contenus linguistiques, l'IA, en s'associant à la logique et à la philosophie du langage, apporte les notions de planification, de représentation, de raisonnement et d'acte de langage pour expliquer la structure du dialogue. Elle a ainsi contribué à définir une approche

par plans du DHM, appelée aussi approche différentielle. Dans cette section, nous explorons les points de vue des sciences cognitives et leurs applications possibles en DHM autour des capacités de perception, de représentation et d'apprentissage.

2.1.1. *Perception, attention, mémoire*

Le système cognitif humain est pourvu de mécanismes perceptifs, qui permettent de distinguer, et ainsi d'extraire, les objets de leur environnement. Ces mécanismes sont rapides, automatiques, fiables. Ils se caractérisent par certains traits remarquables comme l'invariance : dans la perception visuelle, un objet est toujours reconnu, même s'il subit une rotation ou un changement d'échelle (Gaonac'h, 2006). La perception auditive et la perception du langage parlé impliquent, elles aussi, des mécanismes spécifiques, intéressants d'un point de vue cognitif, comme par exemple la capacité à sélectionner une source sonore dans un environnement bruité. Autre exemple, les caractéristiques essentielles de la perception du langage parlé sont l'intensité (voix forte ou douce), la hauteur (voix grave ou aiguë) et le timbre, en tant qu'ensemble de traits comme la texture qui conduit à l'identification du locuteur. Perception visuelle et perception auditive ont fait l'objet d'un nombre impressionnant d'expérimentations et de modélisations en sciences cognitives. La théorie de la Gestalt, le modèle ACT (*Adaptive Control of Thought*) de J. Anderson et ses dérivés, ou encore des modèles moins généralistes comme ceux portant sur la perception de la parole et l'accès lexical (modèles de Forster, de Morton, de Marslen-Wilson), ont été présentés maintes fois dans la littérature, voir (Gaonac'h, 2006). Leurs applications au DHM sont cependant très peu nombreuses, et c'est sur ce point que nous allons nous focaliser.

Le système cognitif qu'est un système de DHM n'a pas vocation à être doté des mêmes capacités que l'humain. D'une part, il n'est pas nécessaire de reproduire la complexité des mécanismes intervenant par exemple dans la reconnaissance d'un mot, quand une technique informatique adaptée, la reconnaissance de la parole, aboutit au même résultat à sa façon. D'autre part, reproduire les erreurs de perception, les faiblesses et les limitations de l'humain n'a pas grand intérêt : le manque de performance des analyses automatiques s'en charge très bien. L'incapacité dans laquelle nous sommes de regarder deux scènes visuelles en même temps, ou d'écouter un message oral à l'oreille droite et un autre message à l'oreille gauche, ne sert pas dans la définition des capacités de traitement d'un système de DHM. En revanche, ce sont des résultats qu'un système peut prendre en compte quand il produit des messages en sortie : ces messages sont à destination d'un humain, ils ne devraient donc pas impliquer deux scènes visuelles simultanées ni deux messages oraux simultanés. C'est ce type de contraintes que nous explorerons dans le chapitre 9 sous le nom de « facteurs humains ». Les capacités de perception à implémenter varient d'un système à l'autre. S'il est utile de doter un robot d'une perception visuelle de manière à ce qu'il puisse se déplacer dans un environnement, la question ne se pose pas de la même façon pour un système comme Shrdlu ou Ozone. Dans le premier cas en effet, le robot ne connaît pas

à l'avance les objets placés dans l'environnement : il doit donc les reconnaître, sous peine de ne pas pouvoir avancer. On exploite alors des techniques de « vision artificielle ». Dans le second cas, la scène visuelle partagée est gérée par le système : c'est lui qui affiche les objets, donc il les connaît et il sait comment ils sont placés. Autrement dit, les mécanismes de distinction d'un objet dans son environnement n'ont pas à intervenir dans Shrdlu et Ozone. Ce n'est pas pour autant que l'ensemble des mécanismes de perception visuelle doit être ignoré. Au contraire, savoir où et comment les objets sont placés dans la scène affichée à l'écran ne veut pas dire savoir comment l'utilisateur va percevoir la scène. L'un des objets peut être visuellement saillant pour un humain, sur des critères comme la taille, la couleur, la forme, critères qui vont provoquer cet effet de saillance de manière prévisible. Si l'on dote le système de capacités à détecter automatiquement cette saillance, ou encore de capacités à détecter automatiquement des groupes perceptifs d'objets, alors on augmente les capacités de compréhension du système. Celui-ci va pouvoir appréhender la scène non seulement selon son point de vue de machine, avec les positions absolues de chaque objet, mais aussi selon un point de vue plus proche de la cognition humaine (Landragin, 2004). L'intérêt est qu'alors le système peut comprendre et compenser une perception biaisée ou erronée de la part de l'utilisateur.

Il en est de même avec les phénomènes d'attention. Les sciences cognitives ont clarifié les différentes formes et fonctions de l'attention (Gaonac'h, 2006, p. 137) : maintien attentionnel (réagir à la nouveauté et maintenir sa vigilance) ; sélection attentionnelle (filtrer certaines informations) ; partage attentionnel (gérer des activités concomitantes) ; contrôle attentionnel (contrôler le déroulement des actions, c'est-à-dire préparer, alterner, superviser). Doter un système de DHM de telles capacités et limites n'a de sens que si l'on cherche à reproduire la faillibilité humaine. Si l'on cherche au contraire à concevoir un système compréhensif, performant et capable d'exploiter tout indice lui permettant de résoudre la tâche pour laquelle il est conçu, rien n'interdit que l'intégralité des informations captées ne soient traitées, sans considération de sélection ou de partage de l'« attention » du système. En revanche, ici aussi, la caractérisation de l'attention humaine est un résultat que le système peut exploiter quand il produit un message à l'intention de l'utilisateur.

La mémorisation et l'oubli posent des questions différentes. Dans les modèles issus des sciences cognitives, plusieurs types de mémoires sont distingués. Au niveau de la perception immédiate et des premières représentations mentales, on parle de mémoire à très court terme ou de mémoire à court terme. C'est à ce niveau que joue l'« empan mnésique », seuil qui limite le nombre d'éléments stockés dans cette mémoire, avec comme conséquence l'incapacité pour un humain de restituer un plus grand nombre d'éléments. Au niveau des premières activités mentales, on parle de mémoire à plus long terme et on distingue (entre autres) la mémoire sémantique, qui stocke des connaissances générales, verbalisables, de la mémoire épisodique, qui stocke des événements particuliers, qui ont eu lieu à un moment donné et qui ont marqué l'esprit. En DHM, prendre en compte l'empan mnésique relève d'une stratégie de

connaissance des limites de l'utilisateur afin de ne pas le surcharger d'informations, par exemple lors de la production de messages. On retrouve donc la même application que précédemment. Cependant, les mémoires sémantique et épisodique peuvent également apporter des idées lors de la conception du système lui-même : c'est par exemple un moyen de penser à sauvegarder tout au long du dialogue un type d'information particulier, comme les événements survenus ; c'est aussi un moyen de séparer les différentes sources de connaissances dans l'architecture du système, et de spécifier les processus d'encodage et de récupération d'informations en mémoire. Plus que cela, la notion d'oubli peut trouver un intérêt : au bout d'un certain temps ou d'un certain nombre de tours de parole, on peut considérer qu'une information transmise et non encore récupérée est oubliée, ou du moins est étiquetée en tant que telle, de manière à ce que le système, si besoin, incite l'utilisateur à la fournir à nouveau.

2.1.2. *Représentation, raisonnement*

Une fois que les mots ont été perçus et reconnus, le système cognitif humain dispose de mécanismes pour les traiter, c'est-à-dire identifier leurs sens, représenter ceux-ci mentalement, associer les sens les uns avec les autres, jusqu'à la construction des sens des phrases et des énoncés et la représentation des connaissances véhiculées. Les sciences cognitives ont proposé plusieurs modèles pour chacune de ces questions. Selon les auteurs – encore une fois, voir les présentations d'ouvrages tels que (Gaonac'h, 2006) – la représentation du sens d'un mot se réduit à un ensemble fini de caractéristiques élémentaires, éventuellement relié à un représentant prototypique de la classe que le mot désigne, ou encore à une liste d'implications, de contenus propositionnels, de procédures qui s'appliquent en contexte, provoquant des effets de sens dans certaines conditions seulement. Selon les approches, associer les sens des mots les uns aux autres peut alors faire intervenir un réseau de relations entre représentations, ou des modèles plus complexes basés sur des structures de traits ou des propositions. Certains auteurs comme G. Fauconnier considèrent le langage et son usage comme la construction mentale d'espaces et d'éléments dans ces espaces, avec des relations entre espaces. Le terme même de « représentation mentale » est employé dans le modèle de (Reboul et Moeschler, 1998, chapitre 6), la théorie des représentations mentales, qui décrit avec une approche relevant de la pragmatique cognitive un ensemble d'éléments, de relations et d'opérations sur des structures à entrées multiples modélisant les représentations mentales, pour la résolution des références. La compréhension du langage fait l'objet de nombreux travaux en sciences cognitives et en linguistique. Comme nous le verrons dans la section 2.2, les modèles viennent surtout de la linguistique, et l'apport de disciplines telles que la psychologie cognitive repose dans des expérimentations, dans des modèles qui rendent compte des limites humaines lors de tâches contrôlées telles que la lecture de textes, ou encore dans des jugements de plausibilité cognitive sur les modèles construits en linguistique. Les sciences cognitives ont contribué à souligner l'importance du contexte, et à clarifier la nature des connaissances qui participent à la construction de représentations : connaissances générales

sur le monde, sur les individus, les systèmes physiques, les objets, leurs catégories, leurs propriétés ; connaissances sur les actions, avec leurs conditions, prérequis et résultats ; connaissances sur les langues ; formats ou patrons prêts à être utilisés pour construire une représentation, en fournissant un cadre structuré, avec des modèles tels que ceux des scripts et des plans.

Les mêmes questions se posent lors de la conception d'un système cognitif artificiel, et les modèles issus des sciences cognitives donnent des pistes intéressantes pour les implémentations informatiques. D'une part, les structures que sont les représentations mentales et les différentes sources de connaissances sont presque directement implémentables : des structures de données informatiques classiques correspondent aux ensembles finis de traits et aux relations entre structures. La gestion de propositions et d'implications est moins évidente, mais reste à la portée de l'informatique. D'autre part, les sciences cognitives elles-mêmes s'informent sur les possibilités d'implémentation, en tant que moyen de procéder à des tests. C'est le cas de certaines approches explorant les notions de script ou de plan pour déterminer la succession d'actions permettant de réaliser une tâche, et pour exploiter les structures résultantes lors de la compréhension d'un texte ou la gestion d'un dialogue.

Une fois que les connaissances sont représentées mentalement, le système cognitif humain dispose de mécanismes pour raisonner sur ces connaissances et construire par inférence de nouvelles données. La manière d'inférer à partir de prémisses un nouveau contenu caractérise le type de raisonnement impliqué. Selon les approches, on distingue ainsi le raisonnement par déduction, dont la conclusion est aussi certaine que les prémisses, à la manière de « X est Y. Tout Y est Z. Donc X est Z » ; le raisonnement par induction, qui consiste à former des représentations générales à partir de faits particuliers, et peut ainsi provoquer des erreurs, comme quand, à force de rencontrer des chats gris, on induit que tous les chats sont gris ; le raisonnement par analogie, qui consiste à exploiter une ressemblance considérée comme non fortuite entre deux connaissances. On peut aussi mentionner le raisonnement par abduction, qui consiste à supprimer les solutions improbables. Tous ces types de manipulation de connaissances s'implémentent, que ce soit en exploitant un moteur d'inférence existant, voire constitutif d'un langage informatique dans le cas célèbre de Prolog, ou en construisant des mécanismes d'inférence spécifiques à un système cognitif artificiel particulier. Le choix des types d'inférence à prendre en compte dans la réalisation d'un système de DHM a des conséquences sur la manière d'interpréter un énoncé de l'utilisateur et sur la décision que le système va prendre selon la nature des inférences réalisées.

Le système cognitif humain prend des décisions sur la base des connaissances représentées mentalement, mais aussi en fonction de ses rapports avec les autres humains : chacune de nos décisions répond à un calcul en termes de croyances, de désirs, d'intentions. Déterminer la nature et le rôle de ces états mentaux est aussi l'une des tâches des sciences cognitives, et fait aussi l'objet d'une application pour le DHM. Un système cognitif humain est capable de raisonner sur la base de ses propres états

mentaux, et est capable d'attribuer des états mentaux à ses interlocuteurs. C'est, entre autres approches, celle de J. Searle sur l'intentionalité, ou encore le principe de la stratégie de l'interprète développée par D. Dennett, et de la théorie de l'esprit qui lui est liée (Reboul et Moeschler, 1998, chapitre 9). En informatique, de telles théories cognitives ont amené la spécification d'agents communicants rationnels, définis par l'ensemble des états mentaux formalisés. Dans le modèle BDI évoqué au paragraphe 1.1.2, trois états mentaux sont modélisés :

– les croyances, qui correspondent à un degré de confiance par rapport à une donnée, cette donnée devenant une connaissance quand la croyance est vraie ;

– les désirs, qui correspondent à l'ensemble des possibilités s'offrant au système cognitif compte tenu de ses croyances et connaissances ;

– les intentions, qui constituent un sous-ensemble de désirs : ceux qui ont été retenus et qui mènent à une action.

Ce modèle a inspiré de nombreux chercheurs et a conduit à des extensions, par exemple le modèle BOID qui ajoute une prise en compte des obligations du locuteur (Broersen *et al.*, 2005). Enfin, pour terminer ce tour rapide et souvent trop schématique des mécanismes cognitifs, soulignons que la dérivation de nombreuses inférences peut amener à une multiplication des croyances et des connaissances, et que, face à cette combinatoire excessive, des notions comme la « pertinence » permettent de faire un tri et de ne retenir que les données les plus importantes, significatives, ou pertinentes d'après le terme retenu par la théorie de la pertinence (Sperber et Wilson, 1995). De tels critères œuvrent dans les systèmes cognitifs humains, et permettent également de contrôler les mécanismes de raisonnement dans les systèmes artificiels. Certains auteurs ajoutent une capacité plus générale de « métacognition », c'est-à-dire un mécanisme cognitif portant sur les mécanismes cognitifs présentés ci-dessus et permettant au système, humain ou artificiel (Sabah, 1997), d'avoir une idée des processus qu'il est en train de mettre en œuvre, de manière par exemple à optimiser ces processus.

2.1.3. *Apprentissage*

L'apprentissage humain peut se faire explicitement, par l'éducation, ou implicitement, par l'association, l'analogie, ou encore l'action et l'exploration (Gaonac'h, 2006). Quand il s'agit de connaissances, le but est de mémoriser les nouvelles connaissances, de manière à les acquérir, à créer ou supprimer des liens entre connaissances déjà acquises, à créer ou modifier des catégories, ou encore à construire de nouveaux scripts ou plans. L'apprentissage concerne donc tous les domaines mentionnés ci-dessus, à partir du moment où des connaissances sont représentées mentalement. Dans le dialogue humain, les sciences cognitives ont décrit plusieurs situations faisant intervenir un apprentissage. La langue étant le vecteur privilégié de toute transmission de connaissances, le dialogue entre un expert et un apprenant constitue un premier corpus

d'étude. Selon la tournure du dialogue, on distingue l'apprentissage actif, pendant lequel l'apprenant pose des questions et, en quelque sorte, oriente le dialogue selon ses désirs, et l'apprentissage passif, pendant lequel c'est l'expert qui oriente le dialogue et transmet ses connaissances. La langue étant elle-même l'objet d'un apprentissage, le dialogue entre un adulte et un enfant constitue un second corpus d'étude, ainsi par ailleurs que le dialogue entre un expert et un apprenant souhaitant acquérir une seconde langue. (Clark, 2009) propose ainsi une liste de phénomènes caractérisant le dialogue entre un adulte et un enfant. Tout d'abord, l'enfant a tendance à mal prononcer les mots, ce qui entraîne l'adulte à sur-prononcer. L'enfant fait des erreurs morphologiques, par exemple dans la conjugaison des verbes, des erreurs lexicales et des erreurs syntaxiques. Celles-ci amènent l'adulte à corriger, à la fin de l'intervention de l'enfant ou en lui coupant la parole. Au niveau de la structure du dialogue, le constat principal est que l'enfant ne suit pas une structure similaire à celle observable dans un dialogue entre deux adultes : souvent, l'enfant est centré sur lui-même et a tendance à suivre sa propre ligne de pensée plutôt que d'interagir avec l'adulte. Par ailleurs, l'adulte privilégie les questions fermées, dont la réponse est oui ou non, aux questions ouvertes, dont la réponse nécessite d'exprimer une connaissance. Il produit des répétitions et reformulations, de manière à augmenter les chances de transmission de l'information. Il s'adapte aux mots que l'enfant connaît, c'est-à-dire qu'il s'aligne d'un point de vue lexical, et il suit un ordre chronologique ou en tout cas logique, par exemple quand il raconte une histoire.

Ces situations d'apprentissage et leurs caractéristiques peuvent inspirer les concepteurs de systèmes de DHM, sans pour autant que le système ou l'utilisateur soit considéré comme un apprenant. Dans les situations où le système doit donner une explication à l'utilisateur, ne serait-ce que pour le faire patienter lors de l'exécution d'une requête, donner les informations de manière chronologique peut guider la génération, de même que produire des questions fermées. Si la machine est capable d'« apprentissage à la volée », c'est-à-dire d'enrichir ses compétences linguistiques pendant et par le dialogue, on peut prévoir des traitements spécifiques pour les énoncés de l'utilisateur à visée corrective ou à visée évaluative des progrès effectués.

Cependant, apprentissage humain et apprentissage automatique (ou apprentissage artificiel) sont très différents, une machine n'ayant pas besoin de passer par tous les stades d'un enfant. Dans le cas de l'apprentissage de la langue, si on décide de doter la machine de cette capacité, alors l'apprentissage n'a pas à porter sur toutes les dimensions de la langue : la prononciation et la morphologie posent par exemple beaucoup moins de problèmes que la sémantique lexicale ou la syntaxe. Le DHM n'est que peu concerné par les erreurs « simples », et il y a des chances pour que, à partir du moment où les règles morphologiques ont été correctement définies, aucun apprentissage ne soit plus jamais nécessaire sur cet aspect. Outre le cas un peu particulier de l'apprentissage à la volée de compétences linguistiques, la plupart des systèmes concernés apprennent de manière implicite. Par exemple, si un utilisateur présente un comportement négatif qui traduit son mécontentement de la machine, celle-ci a

tout intérêt à retenir les conditions qui ont déclenché ce comportement de manière à éviter de le reproduire. Plus un comportement négatif apparaît dans des situations similaires, plus la machine confirme sa prudence par rapport aux conditions. C'est le principe de l'« apprentissage par renforcement », forme d'apprentissage basée sur l'expérience, et c'est l'un des enjeux actuellement explorés du DHM : les capacités d'apprentissage permettent de contourner les difficultés que représentent l'anticipation et la programmation de toutes les situations possibles. En poussant ce principe à l'extrême, on arrive au principe de l'« apprentissage supervisé », c'est-à-dire contrôlé par un opérateur, qui répond aux questions que se pose le système et lui permet ainsi d'apprendre de manière fiable. Une telle procédure intervient avant l'utilisation réelle du système, c'est-à-dire pendant la phase de conception. Il s'agit d'un apprentissage *a priori*, qui peut se dérouler en suivant un dialogue tel que prévu avec un utilisateur, ou plus directement, par intervention de l'opérateur dans chacun des modules concernés. Les données peuvent alors être fournies pas à pas, comme en dialogue, ou d'un seul coup, en s'aidant d'un corpus d'apprentissage, auquel cas on parle d'apprentissage hors ligne. Dans tous les cas, le but est d'entraîner le système pour qu'il affine son comportement, ou tout simplement pour qu'il optimise ses performances (sans rien apprendre de nouveau), avant d'être soumis au public.

D'un point de vue technique, l'apprentissage automatique prend plusieurs formes selon les types de données et d'algorithmes concernés, voir chapitre 6 de (Garbay et Kayser, 2011). Une distinction historique met d'un côté l'apprentissage symbolique, issu de l'IA, qui se limite à des données symboliques ou du moins discrétisées, et l'apprentissage numérique, lié aux statistiques. Dans les années 1990, avec l'avènement de l'ère numérique dont nous avons parlé au paragraphe 1.1.2, l'apprentissage numérique prend le pas sur l'apprentissage symbolique, en exploitant la force de calcul et les corpus de grande taille. A l'heure actuelle, comme l'écrit I. Tellier dans la préface de (Tellier et Steedman, 2009), le travail consiste d'abord à formuler le problème nécessitant un apprentissage, soit en catégorisation, soit en étiquetage (ou annotation), puis à choisir la technique adaptée, sachant que les machines à vecteurs supports (SVM, *Support Vector Machines*) sont performantes pour la catégorisation, et que les champs aléatoires conditionnels (CRF, *Conditional Random Fields*), avec les modèles de Markov cachés (HMM, *Hidden Markov Models*), sont performants pour l'étiquetage.

Au niveau du DHM, on retrouve l'ensemble des approches mentionnées, et ce pour une multitude de modules. Certains aspects comme la reconnaissance automatique de la parole ont depuis longtemps fait appel aux statistiques et à l'apprentissage numérique. Ce que l'on voit plus récemment, c'est un essor des approches hybrides qui concilient apprentissage symbolique et apprentissage numérique, et ce pour l'ensemble des modules d'un système de DHM. Les modules relevant du TAL suivent de plus en plus de telles approches (Tellier et Steedman, 2009), et les modules relevant

de la pragmatique également. Ainsi, la résolution de la référence aux objets fait maintenant appel aux statistiques (Funakoshi *et al.*, 2012), de même que les modèles de représentation sémantique et discursive (Stone et Lascarides, 2010), le terrain commun et le processus d'ancrage (Rossignol *et al.*, 2010), la reconnaissance des intentions du locuteur (de Ruiter et Cummins, 2012), l'attribution d'états mentaux, la gestion du dialogue et la génération automatique de réponses (Rieser et Lemon, 2011). On le voit, il n'est plus question de faire l'impasse sur l'apprentissage automatique en DHM.

2.2. Aspects linguistiques

Un système de DHM manipule la langue, il est donc largement concerné par la linguistique en général, et notamment par la linguistique automatique, ou TAL, mais pas seulement : les avancées récentes de la linguistique de corpus et, ce qui va avec, de la linguistique outillée, ont changé la manière d'analyser des dialogues humains. Nous avons défini le mot « corpus » en tant que collection de matériaux linguistiques attestés. Plus précisément, un corpus se constitue en suivant des règles strictes, de manière à obtenir un échantillon du langage sélectionné selon des critères explicites. Selon le type de dialogue, il peut s'agir d'échantillons sonores ou écrits. Pour le dialogue oral, étudier des enregistrements n'est jamais très pratique et on enrichit souvent le corpus d'une transcription, c'est-à-dire d'une approximation textuelle des énoncés prononcés. Le texte qui en résulte est parfois difficile à comprendre sans transcription de la prosodie, par exemple des pauses, et on est ainsi amené à ajouter des codes ou des annotations, c'est-à-dire des données affectées à des unités textuelles, par exemple aux mots. On obtient alors un corpus annoté, qui contient à la fois les phrases et des observations sur leur énonciation. Quand on procède à des analyses morphologiques et syntaxiques, il est désormais commun d'annoter le corpus avec un jeu d'étiquettes morphosyntaxiques et avec des arbres syntaxiques, ou, plus simplement, des relations entre mots et groupes de mots. Au final, un corpus peut prendre la forme d'une base de données informatique, avec de nombreux champs et des possibilités multiples d'exploration et d'interrogation. Car c'est là tout l'intérêt des corpus : constituer un stock d'analyses de manière à réaliser des calculs de fréquences, des « statistiques descriptives » plus complexes comme des analyses factorielles de correspondances, des recherches de corrélations entre texte et annotations, ou encore des « statistiques inférentielles » comme l'analyse de variance, le but étant alors de tenter de généraliser les résultats obtenus avec un corpus particulier. Cette digression sur la linguistique de corpus nous permet d'ajouter à la liste des disciplines impliquées en DHM les statistiques, en tant qu'outil pour analyser les corpus de dialogue.

2.2.1. *Niveaux d'analyse de la langue*

Tous les domaines de la linguistique, toutes les écoles peuvent inspirer le DHM : à partir du moment où un modèle est proposé, il est tentant d'essayer de l'appliquer

au DHM et de lui faire subir le test de l'implémentation informatique. Ainsi, les domaines de la phonologie, de la morphologie, de la syntaxe, de la prosodie, de la sémantique, et le domaine plus ou moins lié à la linguistique qu'est la pragmatique, fournissent des modèles qui ont un intérêt potentiel en DHM. De fait, certains de ces domaines ont donné lieu à beaucoup plus d'applications que d'autres, et d'autres domaines comme l'étude de l'oral, l'étude des dialogues avec les approches de l'analyse conversationnelle et de l'analyse du discours ont suscité beaucoup d'intérêt, de par leur objet d'étude, bien sûr, mais aussi parce que les échanges pluridisciplinaires ont été fructueux, avec des retombées de part et d'autre. Avant de développer (chapitres 5, 6 et 7) l'ensemble des aspects linguistiques et pragmatiques intervenant dans la compréhension automatique, cette section vise à donner quelques repères préalables concernant l'étude de l'oral et celle du dialogue.

La langue orale que l'on peut observer en dialogue spontané ne ressemble pas à la langue écrite de la littérature ou de la presse écrite : les hésitations et répétitions y sont fréquentes, de même que les accumulations de termes visant à trouver le bon mot, par essais et erreurs. L'énoncé oral se construit en temps réel, alors que l'auteur d'un écrit a eu tout le temps qu'il voulait pour trouver ses mots et peaufiner ses phrases. Ces différences conduisent à deux points de vue par rapport à la syntaxe : soit oral et écrit se caractérisent chacun par une syntaxe spécifique, soit oral et écrit partagent une même syntaxe, avec des aménagements, par exemple des relâchements de contraintes, pour l'oral. Les deux points de vue s'appliquent au DHM. Dans le premier cas, on étudie la syntaxe de l'oral pour spécifier le module syntaxique du système. Dans le second cas, on ajoute au module syntaxique basé sur l'écrit un ensemble de mécanismes additionnels de détection et de réparation des spécificités de l'oral. Par ailleurs, les mêmes mots n'apparaissent pas avec les mêmes fréquences à l'oral et à l'écrit, ce qui entraîne des conséquences sur les modèles de langage utiles à la reconnaissance automatique de la parole. A titre d'exemple, (Blanche-Benveniste, 2010) remarque que les adjectifs constituent 25 % des mots à l'écrit et seulement 2 % à l'oral, ou encore que le mot « dont » est utilisé avec quelques dizaines de verbes à l'écrit, mais pratiquement qu'avec le verbe « parler » à l'oral. Pour le DHM, on peut ainsi dresser une liste prédéfinie des emplois de « dont », ce qui simplifie d'autant la préparation des ressources linguistiques, du moins sur ce point. D'une manière générale, (Blanche-Benveniste, 2010, p. 11) montre que si les fréquences (et les usages) diffèrent, c'est bien un même cadre syntaxique qui permet de rendre compte de l'oral et de l'écrit. La principale différence réside en fait dans la morphologie : l'écrit implique une correction orthographique riche en règles et exceptions bien connues, alors que l'oral met en œuvre une morphologie spécifique, focalisée par exemple sur la prononciation des liaisons. Comme pour la syntaxe, ce constat a des conséquences importantes sur la conception de systèmes de DHM. Dans certains systèmes oraux, le module de reconnaissance de la parole renvoie une transcription écrite des énoncés de l'utilisateur, transcription alors traitée par un analyseur morphosyntaxique. Ce dernier repose sur la morphologie de l'écrit, et celle de l'oral est tout simplement ignorée. Dans la plupart des cas, elle

est prise en compte dans la transcription elle-même, puisque la prononciation a aidé à déterminer cette transcription. Mais dans d'autres cas, la transcription écrite introduit une ambiguïté, comme avec le mot « plus », qui à l'écrit peut signifier « davantage » ou « plus du tout », mais n'est jamais ambigu à l'oral, les deux sens se prononçant différemment. Idéalement, soit on lance l'analyse morphosyntaxique directement sur l'oral, soit on accompagne la transcription d'annotations décrivant la prononciation, soit on implémente des allers-retours entre modules, avec dans notre cas un module d'analyse qui détecte l'ambiguïté et demande au module de reconnaissance de la lever, sans pour autant recommencer une transcription et une analyse.

La langue en dialogue est tributaire des tours de parole et des aspects interactifs. Une phrase à l'écrit n'est pas interrompue, sauf effet volontaire, alors que couper son interlocuteur arrive en dialogue, qu'il s'agisse de dialogue écrit ou oral. Par ailleurs, le dialogue se caractérise par des énoncés d'ouverture et de fermeture, c'est-à-dire des moments-clés où les messages suivent des codes prédéfinis. Ce sont les salutations, les adieux, les formules de politesse, les remerciements ou encore les excuses. Enfin, un dialogue s'étudie en tant que succession pertinente d'énoncés, c'est-à-dire qu'il y a des raisons valables pour qu'un énoncé, par exemple « voici les trajets possibles » suive l'énoncé précédent, « je voudrais aller à Paris ». On peut appeler ces raisons cohérence, continuité thématique, ou acte réactif suite à un acte initiatif. Dans tous les cas, un dialogue s'analyse en tant qu'ensemble d'énoncés, ce que certains appellent discours. (Cole, 1998, p. 198) rappelle que les recherches sur le DHM ont suivi historiquement deux voies tracées par les recherches sur le dialogue humain : d'un côté la voie de l'analyse du discours, à partir de la théorie des actes de langage (Searle, 1969), qui voit le dialogue comme une coopération rationnelle et fait ainsi des liens entre un acte et le précédent (voir paragraphe 1.1.2), et de l'autre côté la voie de l'analyse conversationnelle, qui appréhende le dialogue comme une interaction sociale avec des phénomènes d'organisation des tours de parole, de changement abrupt du topique, de disfluence (Sacks *et al.*, 1974). La voie de l'analyse du discours a mené à des théories computationnelles des actes de langage, et la voie de l'analyse conversationnelle a permis de mieux gérer les tours de parole en DHM. Par exemple, (Sacks *et al.*, 1974) montre que les tours de parole sont réglementés, du moins en langue américaine, par des règles qui s'appliquent à des moments précis, appelés TRP (*transition-relevance place*), c'est-à-dire des moments où la structure du langage permet une interruption. Les critères peuvent être prosodiques, comme la présence d'une pause, aussi bien que syntaxiques ou sémantiques : on peut considérer qu'une phrase est potentiellement terminée quand les actants nécessaires au prédicat ont été exprimés, alors même que le locuteur souhaite peut-être apporter une précision, par exemple un complément différé, une répétition ou une reformulation. Cet exemple est important en DHM car il apporte des paramètres pour le module chargé de détecter la fin des interventions de l'utilisateur, module intervenant dans les systèmes capables de gérer un dialogue en temps réel. Dans un même ordre d'idée, on peut tester la possibilité de couper la parole de l'utilisateur. Deux arguments tempèrent cette initiative : d'une part, même si les

interlocuteurs ont théoriquement la possibilité de parler en même temps, (Levinson, 1983) montre que la totalité des zones de superposition n'atteint pas 5 % dans les dialogues humains qu'il a étudiés ; d'autre part, le DHM ne doit pas forcément se calquer sur le dialogue humain.

Ce dernier point fait l'objet d'un débat d'une portée plus générale que la gestion des tours de parole. Nous avons mis en avant une approche visant à autoriser l'utilisateur à s'exprimer dans sa langue de tous les jours, de manière spontanée. Cette approche, qui tend au dialogue naturel en langage naturel, repose plus ou moins sur le postulat que le DHM peut se calquer sur le dialogue humain. De fait, tout système de DHM a ses limites, et, rapidement, l'utilisateur s'aperçoit que la machine présente des limites de compréhension. Il adapte ainsi son comportement et ses énoncés, et contribue alors à un dialogue qui prend une tournure différente de ce qu'elle aurait été avec un interlocuteur humain. Selon les capacités du système de DHM, cette adaptation peut être modérée, similaire à l'adaptation que requiert tout interlocuteur, ou peut être importante, au point de considérer que la machine est un interlocuteur très particulier. C'est le point de vue de (Jönsson et Dählback, 1988), article au titre évocateur : parler à une machine n'est pas la même chose que parler à votre meilleur ami (paragraphe 3.2.1). C'est aussi le point de vue de (Fraser et Gilbert, 1991) qui propose la méthodologie du Magicien d'Oz pour faire du type de communication entre l'humain et la machine un sujet d'expérience (paragraphe 3.2.3).

2.2.2. *Traitements automatiques*

Les facettes du TAL liées à la compréhension automatique (Jurafsky et Martin, 2009) et à la génération automatique (Reiter et Dale, 2000) sont bien entendu impliquées dans le DHM, qu'il s'agisse des aspects lexicaux, syntaxiques, sémantiques et pragmatiques. Ainsi, les dictionnaires et ressources électroniques, les lexiques sémantiques, ou encore les grammaires lexicalisées apportent des contenus et des méthodes pour le stockage des connaissances lexicales (Mariani *et al.*, 2000). Les formalismes syntaxiques, notamment quand ils ont été conçus avec des préoccupations computationnelles, permettent de concevoir des analyseurs efficaces, performants, et ouverts à des connaissances supplémentaires (Abeillé, 2007). Les modèles sémantiques, depuis les graphes conceptuels (Sowa, 1984) jusqu'aux formalismes de la sémantique formelle et de la sémantique du discours (Kamp et Reyle, 1993) permettent de combiner les contenus des mots et des énoncés (Enjalbert, 2005). La pragmatique, avec la résolution de la référence, dans un contexte uniquement linguistique ou multimodal (Pineda et Garza, 2000), avec l'identification de l'implicite, comme les présuppositions (Van Deemter et Kibble, 2002) et celle des actes de langage (Traum, 2000), a fait également l'objet de nombreux travaux de TAL directement applicables en DHM.

D'autres domaines du TAL interviennent de manière plus ponctuelle, ou pour certains types de dialogue. La résolution des anaphores, qui est une facette de la résolution des références, constitue aussi un domaine à part entière, avec ses propres algorithmes et campagnes d'évaluation (Mitkov, 2002). Un système de DHM n'a généralement pas besoin de mettre en œuvre les algorithmes les plus complexes, dans la mesure où, en dialogue, une anaphore reprend préférentiellement un antécédent récent, pour ne pas dire immédiatement accessible. Néanmoins, l'intégration d'algorithmes qui ont fait leurs preuves dans leurs domaines respectifs reste bien entendu bénéfique, si les ressources de la machine le permettent. Autre domaine du TAL, l'identification des chaînes de coréférence peut s'appliquer en DHM et apporter ainsi un éclairage complémentaire, par exemple sur la façon dont l'utilisateur introduit un nouveau référent et y réfère par la suite, le système pouvant reproduire cette façon de faire en génération. L'identification de relations de discours (Asher et Lascarides, 2003), ou encore la détection des entités nommées constituent, elles aussi, des applications du TAL utiles au DHM, notamment pour le dialogue en domaine ouvert.

Le TAL n'a pas résolu tous les problèmes et ses limites se retrouvent en DHM. C'est le cas de la couverture et de la finesse de la langue, qu'il n'arrive pas à capter avec efficacité et fiabilité. C'est le cas de l'identification des ambiguïtés, de la gestion des mots inconnus, de l'identification d'un usage non littéral de la langue : deuxième degré, ironie, sarcasme, exagération, questions rhétoriques (Clark, 1996, p. 353). Pour ces questions, les systèmes de DHM mettent en place des procédures locales, dépendantes de la tâche, qui fonctionnent sans apporter de solution générale.

2.3. Aspects informatiques

Les aspects informatiques impliqués dans l'élaboration d'un système de DHM sont multiples, comme l'a montré l'historique du domaine avec, entre autres, le traitement de signal audio, le traitement de gestes captés *via* un dispositif spécifique, la capture de visage et plus généralement la vision artificielle, le TAL, la représentation des connaissances, les moteurs d'inférence, l'apprentissage automatique, la synthèse de parole, la présentation d'information multimédia, ou encore l'annotation et l'étude statistique de corpus. Tous ces objectifs nécessitent des compétences théoriques, qu'elles soient linguistiques, psychologiques, ergonomiques, logiques, mathématiques, et mettent en œuvre des structures de données et des algorithmes spécifiques. Dans le chapitre 4, nous explorerons encore un autre domaine de l'informatique, celui des architectures logicielles et de l'ingénierie dirigée par les modèles. On le voit, l'éventail des compétences techniques nécessaires à l'implémentation d'un système complet est extrêmement large. Comme dans la section précédente, nous présentons très rapidement ici quelques repères préalables sur deux aspects choisis parmi d'autres : les ressources électroniques et les IHM plastiques.

2.3.1. *Structures de données, ressources électroniques*

Le temps de la base de données de 14 Ko du système Parry est bien loin. A l'heure actuelle, on cherche à maximiser les données enregistrées dans le système, de manière à pouvoir fournir le maximum de matière aux algorithmes d'analyse. Les réalisations médiatisées d'IBM comme Watson le montrent : ce sont les données et la force brute de calcul qui permettent d'obtenir des systèmes fonctionnels. C'est en tout cas l'une des façons, car la réflexion sur les données reste essentielle. En TAL et d'une manière générale avec l'essor du *web* sémantique, des efforts sont entrepris depuis des années pour améliorer la qualité des données. Cela passe par l'utilisation de langages standardisés, par l'ajout aux données de métadonnées qui en décrivent le contenu de manière sobre et universelle, ou encore par des ontologies, ensembles structurés de termes et de concepts communs à un même domaine ou une même tâche. En DHM, on est longtemps restés avec un fonctionnement fermé, c'est-à-dire des données internes au système et non connectées au *web*. Contrairement aux ressources disponibles sur le *web*, qui évoluent en permanence et peuvent atteindre des tailles démesurées, ces données cloisonnées ont pour avantage d'être accessibles très rapidement, en temps réel, comme l'impose le DHM spontané. Elles nécessitent cependant d'être spécifiées *a priori*, et un enjeu des systèmes de DHM actuels, et notamment des systèmes en domaine ouvert, est d'exploiter en temps réel les ressources du *web*.

2.3.2. *Interfaces homme-machine, interfaces plastiques et ergonomie*

Les ACA, les robots de compagnie et les IHM à commande vocale explorent de nouvelles façons d'interagir avec une machine, qui sont fondées sur la langue mais pas seulement. La transmission visuelle ou sonore d'émotions joue un rôle important dans les ACA, par exemple. Quant aux IHM sur lesquelles on a greffé un module de commande vocale, elles restent utilisables avec les moyens classiques que sont les menus, boutons, curseurs et autres métaphores graphiques. Interaction vocale et interaction classique peuvent ainsi se mélanger et venir compliquer la gestion de l'interaction et du dialogue. Par exemple, est-ce qu'une commande d'annulation porte sur la dernière action effectuée, ou sur la dernière action effectuée avec le même moyen d'interaction ? C'est ce type de question qui se pose avec l'intégration de la commande vocale à nos téléphones portables et assistants personnels. Par ailleurs, et c'est aussi un domaine de recherche à l'œuvre actuellement, le passage d'un ordinateur de bureau vers un ordinateur portable en passant par un assistant personnel pose des questions sur l'adaptation de l'IHM aux contraintes de taille et d'utilisation du matériel sur lequel elle s'exécute. Pour que l'utilisateur puisse changer de matériel de manière transparente, il est nécessaire de concevoir des « interfaces plastiques », c'est-à-dire adaptables, en temps réel, sans arrêter le déroulement de l'application. Le problème prend une ampleur particulière quand l'interaction vocale, voire multimodale, est impliquée (Kolski, 2010, chapitre 9). Là aussi, il s'agit d'un enjeu pour les systèmes de DHM à venir.

2.4. Bilan

Pour concevoir un système capable d'interagir de manière réaliste avec un utilisateur humain, une approche consiste à s'inspirer du fonctionnement humain, donc de la linguistique pour ce qui concerne l'utilisation du langage, de la psychologie cognitive pour ce qui concerne les aspects de perception ou d'attention, et d'une manière générale des sciences cognitives. Que l'on essaie de simuler un mécanisme communicatif humain, de le simplifier pour spécifier un module informatique capable de s'en approcher, ou de s'en inspirer, ce sont plusieurs disciplines scientifiques qui sont impliquées. Ce deuxième chapitre détaille quelques-unes de ces contributions, en terminant par l'ensemble des domaines informatiques concernés par la réalisation d'un système de dialogue, domaines qui sont de plus en plus nombreux.

Chapitre 3

Les étapes de réalisation d'un système de dialogue

Une fois que le but du système a été défini, par exemple renseigner l'utilisateur sur des horaires de trains, et que les moyens d'interaction, vocale ou multimodale, ont été choisis, le ou les concepteurs se retrouvent face à une multitude de questions concernant d'une part les phénomènes que le système doit absolument traiter, et d'autre part les étapes de réalisation : identification des capacités du système, décomposition de celui-ci en modules, détermination et construction des ressources nécessaires à chacun des modules, et bien sûr développement informatique. Un système de DHM est un logiciel comme un autre, et il suit en cela les étapes de conception classiques que sont les spécifications, par lesquelles on définit ce que le système devrait faire, le développement, pendant lesquel on définit comment les spécifications sont rendues concrètes, puis les évaluations, de manière à s'assurer que le système obtenu est utilisable et correspond bien aux spécifications (McTear, 2004). Ces étapes n'ont pas forcément lieu dans un ordre linéaire, sans possibilité de revenir en arrière. On peut ainsi développer un premier système, l'évaluer, puis, en fonction des résultats obtenus, développer un second système plus performant. (Jurafsky et Martin, 2009, p. 872) mettent ainsi en avant trois phases principales, avec d'éventuels allers-retours : une phase initiale consistant à étudier l'utilisateur, par exemple à l'aide d'interviews, et à étudier la tâche, par exemple à l'aide de systèmes de DHM similaires ou de dialogues humains, deuxièmement une phase d'implémentation, qui peut porter sur des simulateurs et des prototypes plutôt que directement sur le système visé, et troisièmement des évaluations, en particulier des tests itératifs avec des utilisateurs. Selon que le cycle concerne un système réel, un prototype ou un simulateur, les tests peuvent prendre diverses formes, surtout si l'on ajoute la possibilité d'impliquer une simulation informatique d'un utilisateur plutôt qu'un utilisateur humain.

Le choix entre les différentes possibilités évoquées ci-dessus repose sur des critères méthodologiques et techniques. Ce sont ces aspects que nous présentons dans ce chapitre, avec une première section qui illustre quelques cas d'étude représentatifs (section 3.1), et une deuxième section qui détaille les méthodologies impliquées dans les principales étapes de conception (section 3.2).

3.1. Comparaison de quelques déroulements de réalisations

Les scénarios qui suivent sont purement indicatifs : même s'ils s'inspirent d'expériences réelles, ils tendent à les caricaturer un peu et servent surtout à illustrer les concepts méthodologiques.

3.1.1. *Un scénario correspondant aux années 1980*

Dans les années 1980, un système de DHM peut encore être réalisé par une personne unique, qui doit donc avoir de solides compétences en informatique. Le concepteur que nous imaginons ici cherche à tester un modèle de dialogue théorique, qui constitue par exemple la proposition innovante de son travail de thèse.

Pour concevoir son système, il définit une architecture linéaire, avec des modules exécutés en cascade, dans l'ordre suivant : reconnaissance automatique de la parole, analyses linguistiques, résolution des références, détection des actes de langage, génération automatique, synthèse de parole. Pour mettre en œuvre ce système, il travaille module après module, dans l'ordre de traitement de la machine. Il exploite un module de reconnaissance de la parole existant, le paramètre avec un vocabulaire correspondant à la tâche visée et l'entraîne avec sa propre voix pendant quelques heures, de manière à optimiser les performances. Pour fonctionner, ce module nécessite l'appui d'une pédale pendant la prononciation de l'énoncé. Une fois l'énoncé terminé, ce qui est indiqué par le relâchement de la pédale, le module commence à traiter le signal enregistré. Il produit une transcription écrite de l'énoncé prononcé, qui entre dans un module d'analyse linguistique. Cette analyse, exploitant par exemple une grammaire d'unification, regroupe les aspects lexicaux, syntaxiques et sémantiques : dans une telle grammaire, le lexique, les règles de grammaire et les représentations des phrases exploitent le même formalisme, celui des structures de traits. C'est ainsi un même mécanisme d'unification de structures de traits qui permet de procéder aux différentes étapes de la compréhension automatique. Le concepteur écrit lui-même les ressources dont le module de compréhension a besoin. La représentation sémantique obtenue est enrichie par le module de résolution des références, qui utilise une base de données comprenant l'ensemble des objets de l'application, puis la représentation complète alors obtenue sert de base à l'identification de l'acte de langage caractérisant l'énoncé. On obtient une représentation constituée d'une valeur illocutoire (ordre ou question, par exemple) sur un contenu sémantique. Cette représentation entre dans le module

dédié à la gestion du dialogue, qui implémente le modèle du concepteur : en fonction de la valeur illocutoire de l'énoncé, le système choisit celle de sa réponse. En fonction du contenu sémantique et de la tâche à résoudre, il définit celui de sa réponse. La génération est réalisée de manière simple, en exploitant des patrons, c'est-à-dire des phrases à trous, qui sont prédéfinis selon des catégories correspondant aux différentes valeurs illocutoires possibles. Pour chacun des modules cités, le concepteur développe son propre algorithme local à partir des données qu'il a sous la main : sa définition de la tâche, du vocabulaire et des phénomènes linguistiques associés. Pour la synthèse vocale comme pour la reconnaissance automatique, il fait appel à un logiciel existant. Tous les modules s'exécutent sur le même ordinateur. Les modules de reconnaissance de la parole et de synthèse vocale sont gourmands en mémoire et en temps de calcul de la machine. Il se passe ainsi plusieurs secondes entre la fin de l'énoncé de l'utilisateur et le début de la réponse du système.

Le concepteur peut alors tester son système en reprenant le déroulement imaginé pour quelques dialogues typiques sur la tâche considérée. De manière manifeste, les résultats sont mauvais. Il reprend donc le paramétrage, voire le développement, de chacun des modules. Pour affiner les capacités de compréhension, il réécrit une partie des ressources. Au bout d'un moment, il relance des tests, en jouant le rôle de l'utilisateur comme lors du développement. Aucune métrique d'évaluation n'est utilisée. Intuitivement, mais aussi parce qu'après tant d'efforts ce serait dommage de jeter le système (d'autant plus que rien n'est réutilisable pour un autre système), les résultats semblent meilleurs. Le concepteur s'aperçoit néanmoins que la totalité des mots définis dans le lexique n'est pas exploitée : au contraire, le système n'arrive à comprendre et à traiter qu'un petit sous-ensemble du lexique. Qu'importe, il est possible d'entretenir un dialogue avec lui, certes simpliste mais qui a le mérite d'exister...

3.1.2. *Un scénario correspondant aux années 2000*

Dans les années 2000, plusieurs concepteurs s'associent, dans le cadre d'un projet international ou d'une activité de laboratoire pérenne et soutenue par des moyens humains constants. Après une phase de spécification commune, ils se répartissent les travaux d'expérimentations, de développement et d'évaluation. Une première étape d'expérimentations consiste à faire passer des sujets, qui ne sont pas les concepteurs, devant une simulation du système tel qu'il est imaginé, selon le principe du Magicien d'Oz. Cette étape, qui se termine avec des interviews des sujets, peut se compléter par des études de corpus de dialogues humains, pour aider à définir les aspects essentiels et les limites du comportement du système, et donc du Magicien d'Oz, pour la tâche envisagée.

Le développement commence par la définition en commun de l'architecture du système, et par la spécification du comportement des modules composant cette architecture : types d'entrées, types de sorties, langage de communication entre modules.

Chacun de ces aspects fait largement appel à l'état de l'art, en exploitant par exemple un langage de communication mis au point dans un autre projet. C'est en suivant de telles contraintes que l'on favorise à la fois les possibilités d'intégration de modules existants et la réutilisabilité des composants créés pour l'occasion. Les concepteurs procèdent alors au développement, c'est-à-dire à la construction des ressources et à la programmation des algorithmes. Pour cela, l'ordre suivi n'est pas celui des traitements effectués par le système : on commence au contraire par le cœur du système, avec la gestion du dialogue et les aspects pragmatiques. En simulant les entrées, on affine alors le comportement du système, puis on précise les capacités de compréhension nécessaires à ce comportement. On peut alors implémenter les modules procédant aux analyses linguistiques. Les ressources sont récupérées en partie d'autres projets et de ressources disponibles, libres ou payantes, afin de garder le maximum d'efforts pour les algorithmes, ou plutôt pour leur paramétrage compte tenu de la tâche.

L'étape suivante consiste à tester et à évaluer l'embryon de système déjà implémenté, en simulant les entrées. Ensuite, de même que les capacités de gestion de dialogue se sont matérialisées en paramètres pour la conception des modules de compréhension (et de génération, celle-ci suivant un développement parallèle), on passe maintenant à la matérialisation de la langue correctement traitée en paramètres pour la reconnaissance de la parole et les traitements de bas niveau. Un système de reconnaissance automatique existant est exploité, avec des performances bien meilleures que celui utilisé dans le scénario précédent, et des possibilités d'interaction accrues. Comme pour chaque phase de développement, des tests et des évaluations sont immédiatement réalisés, avec la reconnaissance seule et avec l'ensemble du système, de manière à affiner la couverture des ressources et la paramétrisation des processus. Les évaluations consistent surtout dans la mesure de taux de rappel et de précision, méthode devenue classique pour l'ensemble des systèmes de TAL (Chaudiron, 2004), adaptée à l'évaluation séparée des modules, mais pas du système dans sa globalité.

Enfin, le système obtenu est soumis à des tests utilisateurs, c'est-à-dire que l'on fait à nouveau appel à des sujets, qui ne sont ni les concepteurs ni les sujets du Magicien d'Oz initial, pour recueillir leurs avis. Ceux-ci sont analysés par les concepteurs, qui décident alors des améliorations à apporter, et procèdent éventuellement à un nouveau cycle de développement : amélioration du cœur du système, puis, conséquemment, des modules de compréhension et de génération, en terminant par la reconnaissance et la synthèse de parole.

Ce scénario compense les points faibles du scénario vu au paragraphe 3.1.1. Cependant, même avec une architecture rationnelle, l'essentiel des processus est exécuté en cascade, et des aspects essentiels de la langue orale tels que la prosodie sont ignorés. C'est typiquement là que l'on assiste aux problèmes décrits précédemment sur les conséquences de la non-prise en compte de la morphologie de l'oral.

3.1.3. *Un scénario actuel*

Un scénario de conception actuel devrait répondre à tous les points faibles identifiés précédemment, et aussi à l'ensemble des enjeux de la section 1.3. Sans revenir sur ceux-ci, notons que le développement d'un système peut maintenant faire appel non plus à des modules existants, mais directement à une boîte à outils qui donne un cadre général, totalement paramétrable. Comme tout cadre, celui-ci peut s'avérer pénalisant à cause des limites qu'il impose. Dans ce cas, le développement de modules complémentaires reste indispensable, et fait l'objet d'expérimentations préalables, de tests et d'évaluations *via* des métriques plus ciblées que les calculs de rappel et de précision. Le système de DHM lui-même peut désormais être évalué globalement, avec des méthodes qui calculent le degré de satisfaction de la tâche, sur des indices comme le nombre de tours de parole pour aboutir au but défini initialement. Ce sont des aspects sur lesquels nous reviendrons dans le chapitre 10. Par ailleurs, l'ensemble des étapes de développement se rationalise, c'est-à-dire que rien n'est fait en aveugle ou au hasard. Par exemple, des corpus sont constitués pour chacun des modules identifiés, et ces corpus sont eux-mêmes scindés en plusieurs parties : une partie pour l'entraînement (l'apprentissage automatique est intégré aux étapes qui en ont le plus besoin), une partie pour les tests. Autre exemple, le Magicien d'Oz est réalisé avec des contraintes et précautions supplémentaires. C'est ce type de contraintes que nous allons maintenant détailler sur quelques étapes de conception emblématiques.

3.2. Description des principales étapes de réalisation

3.2.1. *La spécification de la tâche et des rôles du système*

Un système de DHM sert généralement à aider l'utilisateur pour une tâche donnée, qu'il s'agisse de réservation de billets de train (domaine fermé) ou de renseignements généraux (domaine ouvert). Il lui revient donc le rôle de gérer le dialogue dans une voie qui aboutisse rapidement à la satisfaction de la tâche.

Par rapport à une IHM ou un site *web* classique, un système de DHM laisse l'utilisateur libre de s'exprimer comme il le souhaite, et engage avec lui un dialogue naturel en langage naturel, sans directives trop strictes. Il lui revient donc le rôle de gérer le dialogue avec les spécificités du langage et du dialogue humain.

Ces deux rôles sont-ils conciliables ? Les études de dialogues humains ont identifié un certain nombre de caractères pour rendre compte des aspects naturels du dialogue, sept pour J. Sinclair, étendus à neuf par (Warren, 2006), ou encore dix selon (Clark, 1996). Ces caractères mettent en avant le fait que le dialogue est interactif, coopératif, cohérent, au point parfois d'en être prévisible, et que son succès repose sur les deux interlocuteurs. Les critères de (Clark, 1996, p. 9) incluent notamment les suivants :

coprésence des participants dans un même environnement physique, visibilité, audibilité, instantanéité de l'interaction, évanescence (les énoncés sont fugaces, de même que les actions des interlocuteurs), ou encore simultanéité et aspects temps réel de l'interaction (les interlocuteurs peuvent traiter et produire en même temps). Ces critères apportent de grands principes sur le dialogue naturel, à la fois sur les conditions de l'interaction et sur le déroulement du dialogue. En soulignant notamment les aspects coopératifs, un système de DHM est ainsi vu comme un partenaire plutôt qu'un outil, et on en déduit que la résolution de la tâche est une activité conjointe aux deux interlocuteurs et non un contrôle de la part du système.

Or une tâche telle que la réservation d'un billet de train obéit à des principes structurés. Le système a besoin de connaître un ensemble précis d'informations : gare de départ, gare d'arrivée, jour, horaires, première ou deuxième classe, préférences diverses. Plus que cela, l'ordre dans lequel ces informations sont données par l'utilisateur obéit à certains principes. Comme le montre (Luzzati, 1995, p. 91) : « une demande d'horaire qui n'indique pas la gare d'arrivée et la gare de départ est inconcevable, alors qu'elle est envisageable sans référence temporelle : chacun sait qu'on doit toujours accéder à la ligne de chemin de fer avant de pouvoir chercher l'heure d'un train ». Le déroulement du dialogue est donc fortement contraint (Kolski, 2010). De plus, les études du corpus SNCF montrent que le vocabulaire est relativement limité, les structures des phrases également, autrement dit que la prédominance de la tâche influe sur le langage naturel.

La question principale est donc la suivante : quand il y a une tâche, est-ce que la résolution de la tâche doit prendre le pas sur la spontanéité du dialogue ? Soit on considère que le dialogue est en premier lieu finalisé et on répond positivement à cette question, quitte à ce qu'il manque de cohérence (efficacité avant tout) ; soit on considère que le dialogue vise prioritairement à maintenir une communication agréable avec l'utilisateur, et on accepte alors qu'il prenne trois ou quatre fois plus de tours de parole pour arriver au même résultat. Les deux choix sont acceptables, mais ne doivent pas être évalués de la même façon : si c'est la rapidité à satisfaire la tâche qui dirige l'évaluation, le système finalisé va forcément arriver en première place.

La question posée prend une dimension particulière dès lors que l'on interroge la notion de spontanéité. Il n'y a pas forcément un clivage entre résolution de la tâche et dialogue naturel en langage naturel. Le caractère naturel du dialogue ne se juge pas sur des analyses *a posteriori* de la couverture lexicale et de la diversité des phénomènes linguistiques et pragmatiques, mais se juge d'une part sur le « ressenti de l'utilisateur », avec les réponses qu'il donne au cours d'une interview portant sur la facilité de dialoguer avec le système et sur le degré de satisfaction apporté par les énoncés de celui-ci, et, d'autre part, sur des analyses *a posteriori* vérifiant que les réactions du système sont bien pertinentes par rapport aux énoncés de l'utilisateur. Un utilisateur peut tout à fait être satisfait de son dialogue avec le système, même si la tâche s'est résolue en plus de temps que prévu. Comme le montrent (Sperber et Wilson, 1995)

puis (Reboul et Moeschler, 1998), la pertinence est au cœur du dialogue naturel en langage naturel.

Cette question entre résolution de la tâche et dialogue naturel est aussi liée au rôle d'interlocuteur de la machine. Comme nous l'avons vu à la fin du paragraphe 2.2.1, parler à une machine n'est pas la même chose que parler avec son meilleur ami. A moins d'être induit en erreur et de croire, comme cela est possible au téléphone, que son interlocuteur est humain, l'utilisateur sait qu'il parle à une machine, ce qui peut entraîner un comportement particulier de sa part. Les expérimentations de type Magicien d'Oz et les tests d'utilisation en fin de conception sont révélateurs sur ce point. (Luzzati, 1995) montre, suite à un Magicien d'Oz sur une tâche de réservation de billets de train, que les dialogues vont à l'essentiel avec des structures simples : à chaque acte initiatif correspond un acte réactif, le lexique est essentiellement celui de la tâche, donc réduit, et il n'y a pas de digressions particulières. Les utilisateurs n'argumentent pas leurs demandes, et n'ont pas de face à défendre. Ils s'abstiennent de tout commentaire qui pourrait expliquer ce qu'ils sont en train de dire. Au niveau des références, ils s'abstiennent de toute référence à eux-mêmes ou à la machine. Autrement dit, le dialogue reste naturel mais se restreint de lui-même à ce qui permet la satisfaction de la tâche. Dans de telles conditions, on peut considérer que l'on est bien dans un dialogue naturel en langage naturel : du fait qu'il parle (ou croit parler) à un système, l'utilisateur réduit l'éventail de ses énoncés, mais ce n'est pas au prix de la spontanéité puisqu'il le fait sans y être contraint. A l'extrême, ce mécanisme peut amener à un fonctionnement caricatural, observé par exemple lors d'un Magicien d'Oz décrit dans (Landragin, 2004), où l'utilisateur se limite à seulement deux ou trois phrases : celles qui ont fonctionné au tout début du dialogue, et dans lesquelles l'utilisateur a pris confiance. C'est ce type de comportement que l'on retrouve avec un vrai système de DHM, et qui amène à s'interroger à nouveau sur le caractère naturel : si l'utilisateur se contraint lui-même à ce point, c'est que le fait de parler avec une machine le perturbe. Il n'est donc pas en mode naturel. En fait, et les expérimentations réalisées dans le projet Miamm (Landragin, 2004) le montrent : certains utilisateurs se contraignent sans qu'on le leur demande, et d'autres utilisateurs se placent d'emblée dans un dialogue naturel en langage naturel, sans aucune méfiance vis-à-vis du système. En revanche, et c'est là l'un des enjeux essentiels des concepteurs, le système doit tenir le dialogue : si, au contraire, il se met à montrer son incompréhension, l'utilisateur le plus confiant va vite changer de comportement.

3.2.2. *La spécification des phénomènes couverts*

Imaginer les capacités de compréhension et de dialogue d'un système consiste souvent en une boucle constituée de plusieurs étapes : réflexions sur le comportement attendu du système, d'où spécification de la nature de l'interaction ; spécification de l'étendue des capacités du système ; simulation du futur système et constitution de ce fait d'un corpus de dialogue ; étude du corpus, c'est-à-dire analyse des phénomènes

attendus et nouveaux ; réflexions sur la prise en compte des nouveaux phénomènes, et ainsi retour à la première étape. Une fois que cette boucle est stabilisée, on passe alors à la conception d'un modèle de traitement qui tienne compte du maximum de phénomènes, on implémente ce modèle, et on le teste pour en identifier les points faibles. Avec l'implémentation réalisée, on peut alors revenir à l'étape d'expérimentation et reprendre la boucle, en exploitant un système réel et non une simulation.

Lors de l'étape de spécification de la nature de l'interaction, se pose la question des modalités de communication possibles entre l'humain et la machine, et des dispositifs de capture et de production impliqués. Si le dialogue a lieu par téléphone, la seule modalité est orale, en entrée et en sortie du système. Si le dialogue a lieu en face à face, c'est-à-dire sur un ordinateur avec éventuellement un avatar affiché à l'écran, les entrées peuvent se faire à l'écrit, à l'oral ou de manière multimodale, avec par exemple des gestes effectués sur écran tactile ou à la souris. Par cohérence et pour ne pas perturber l'utilisateur en utilisant une modalité différente de la sienne, les sorties peuvent alors se faire selon les mêmes modalités, ou du moins des modalités équivalentes, les gestes effectués par le système se matérisalisant par un affichage à l'écran et non *via* un dispositif spécifique, sauf dans le cas d'un système de robotique.

L'étape de spécification de l'étendue des capacité du système fait intervenir trois méthodes complémentaires : l'imagination de situations de dialogue ; la réalisation d'expérimentations telles que des simulations du système ; l'analyse de corpus de dialogues pour en déduire des phénomènes et des situations à prendre en compte. Pour chacun de ces trois moyens, on peut ajouter une phase consistant à étendre les phénomènes identifiés par un ensemble de phénomènes similaires, c'est-à-dire à dériver des observations de nouvelles idées. Cette opération de « dérivation » (« il a dit ça, alors il aurait pu dire ça, donc il faut tenir compte des deux ») permet d'aboutir à un ensemble de phénomènes de taille et de couverture plus satisfaisantes.

3.2.3. *La réalisation d'expérimentations et les études de corpus*

Les expérimentations de type Magicien d'Oz que nous avons évoquées et les études de corpus de dialogue sont complémentaires. Ces deux méthodes permettent d'utiliser des dialogues concrets comme bases pour la conception d'un système. Toute expérimentation, si elle est enregistrée, constitue par ailleurs un corpus qui peut être étudié de la même façon que les corpus issus d'enregistrements de dialogues humains, de dialogues homme-machine (sans qu'il y ait simulation ni tromperie), de dialogues homme-homme médiatisés (même remarque), et, pourquoi pas, de dialogues machine-machine comme quand on fait interagir entre eux deux systèmes de DHM.

Le Magicien d'Oz (WOz, pour *Wizard of Oz*, du nom du personnage de F. Baum, ou encore Pnambic, pour *Pay No Attention to the Man BehInd the Curtain*) consiste donc à simuler un système de DHM par un humain, le magicien ou compère, pour

observer le comportement d'un sujet face à cette simulation (Fraser et Gilbert, 1991). Le sujet croit qu'il parle ou qu'il écrit à une machine, mais celle-ci est reliée à un autre ordinateur contrôlé par le magicien, qui peut répondre par écrit ou oralement, en générant ses propres énoncés ou en choissant parmi des énoncés ou des patrons (à trous) prédéfinis. Si la simulation est correctement réalisée, le sujet ne voit pas la supercherie et adopte le comportement qu'il aurait face à une machine, ce qui, déjà, permet d'analyser ce type de comportement. Ensuite, les dialogues ainsi enregistrés permettent aux concepteurs de détecter les situations problématiques, les cas où les réactions du magicien ont perturbé le sujet, et bien sûr les cas d'incompréhension. En se focalisant sur les modules incriminés, les concepteurs peuvent alors améliorer leur système, le rendre plus robuste. Pour détecter de manière fiable les moments où le sujet est perturbé, de même que pour détecter les moments d'incompréhension ou tout simplement d'hésitation, des techniques supplémentaires peuvent être mises en œuvre en même temps que l'enregistrement des dialogues : on peut par exemple faire un suivi du visage du sujet, de manière à capturer ses émotions faciales, ou encore utiliser un oculomètre pour détecter les mouvements de son regard et en déduire des observations sur son attention, et, dans le cas d'une scène visuelle partagée, sur les objets regardés, par exemple au moment de résoudre une référence.

Pour qu'un Magicien d'Oz soit exploitable, il faut néanmoins que les conditions expérimentales soient bien définies. Or il existe presque autant de façons de faire un Magicien d'Oz que de systèmes : le magicien peut être l'un des concepteurs, ce qui lui permet de produire un comportement proche du système visé, mais il peut aussi être un deuxième sujet de l'expérimentation, qui ignore le but de celle-ci et se contente d'appliquer un ensemble de règles visant à simuler le système (c'est le principe du double aveugle, ou du *ghost in the machine*). Par ailleurs, les messages du sujet et du magicien passant par un système informatique, celui-ci peut intégrer un traitement particulier, de manière à pimenter l'expérimentation. Dans (Rieser et Lemon, 2011), du bruit est ainsi ajouté dans les énoncés du sujet avant de les transmettre au magicien, ce qui perturbe le dialogue et permet d'augmenter la fréquence des demandes de clarification. Le choix s'est porté sur la suppression d'un mot de manière aléatoire, mais on peut tout à fait imaginer le remplacement d'un mot par un autre ou d'autres heuristiques locales, et ce non seulement pour les énoncés du sujet, mais aussi pour ceux du magicien. Bien entendu, c'est plus facile à réaliser pour du dialogue écrit que pour du dialogue oral. Dans le cadre de la méthodologie très complexe mise en œuvre dans (Rieser et Lemon, 2011), l'objectif du Magicien d'Oz est quadruple : observer des situations de dialogue ; constituer un corpus d'étude et modéliser ainsi un modèle d'apprentissage automatique ; constituer un corpus d'apprentissage ; contribuer à la spécification d'un modèle d'utilisateur, c'est-à-dire d'un programme informatique dont le but est de simuler le comportement d'un utilisateur du système. Chaque étape et chaque objectif s'accompagnent de précautions, d'évaluations, et de confrontations avec d'autres méthodes comme de l'apprentissage supervisé. D'une manière générale, les consignes données au magicien peuvent être plus ou moins précises. Si elles

le laissent libre de réagir comme il le souhaite aux énoncés du sujet, le risque est d'obtenir un dialogue plus fluide et plus robuste qu'un DHM. Ce sont les principales critiques adressées à ce type d'expérimentation. (Cole, 1998, p. 199) affirme ainsi que les Magiciens d'Oz s'accompagnent souvent d'un trop grand optimisme concernant les performances du système visé, ce qui conduit à la conception de systèmes peu robustes. (Rosset, 2008) ajoute que quand il s'agit de choisir parmi un ensemble de possibilités prédéfinies, un magicien est forcément plus lent qu'un système, ce qui dégrade l'aspect naturel du dialogue et empêche l'exploitation des enregistrements ainsi réalisés. (Denis, 2008, p. 90), qui s'intéresse aux cas d'incompréhension, montre que dans les dialogues obtenus par Magicien d'Oz, les incompréhensions ne durent pas autant qu'en DHM : même lorsqu'il simule une incompréhension, le compère parvient toujours à rétablir un dialogue compréhensible dans l'énoncé qui suit la détection par le sujet. A moins d'entraîner le compère sur cet aspect précis, le Magicien d'Oz ne permet pas de spécifier des stratégies de réparation fiables pour le DHM.

Pour qu'un corpus soit exploitable de manière pertinente, il faut tenir compte des conditions dans lesquelles il a été obtenu, c'est-à-dire dans quel type de situation de dialogue il s'est déroulé, et, s'il s'agit d'extraits, selon quels critères les extraits ont été sélectionnés. Dans tous les cas, un corpus ne reflète jamais les possibilités de dialogue naturel en langage naturel : il n'est pas complet, il ne constitue qu'un « réservoir de phénomènes ». En fonction des conditions et des critères de sélection, on peut considérer ce réservoir comme plus ou moins représentatif. Plus un corpus est considéré comme représentatif d'une certaine situation de dialogue, plus on peut l'exploiter pour concevoir un système chargé de communiquer dans la même situation. Quand le corpus est de taille suffisante, cette exploitation peut comprendre des analyses de fréquence : on implémente en priorité les phénomènes les plus fréquents dans le corpus, ou du moins on cherche à en optimiser le traitement. La fréquence d'apparition d'un phénomène n'a cependant rien à voir avec l'importance du phénomène en question : des énoncés tels que « alerte ! » ou « panne ! », qui peuvent intervenir dans la commande de systèmes dans des milieux dangereux ou en gestion du contrôle aérien, sont très peu fréquents et ne doivent pas pour autant être négligés par le module de compréhension.

Par ailleurs, un corpus peut servir de référence durable pour le domaine du DHM, voire du TAL. C'est le cas quand un soin particulier est donné aux transcriptions et au codage des aspects extralinguistiques : codage de la prosodie, des gestes du visage et des mains, et d'une manière générale de la vidéo montrant l'utilisateur en pleine énonciation. Le système Amitiés, par exemple, a impliqué l'annotation multiniveaux d'un corpus comprenant, outre la transcription de la parole, l'identifiant des locuteurs, le marquage des zones de superposition de parole, ainsi que des annotations sémantiques, dialogiques, thématiques et d'émotions (Hardy *et al.*, 2006). Les efforts fournis sont tels que, comme pour le corpus SNCF utilisé par (Luzzati, 1995), plusieurs systèmes de DHM peuvent les exploiter.

3.2.4. *La spécification des processus de traitement*

Les données déterminent quasiment d'elles-mêmes les processus qu'il faut mettre en œuvre pour les traiter. Plus les phénomènes retenus sont diversifiés, plus les traitements vont devoir être approfondis. Plus il apparaît d'ambiguïtés, plus les traitements vont devoir reposer sur des analyses linguistiques, multimodales et dialogiques fines ainsi que sur une gestion appropriée du contexte. Cette section va nous amener à illustrer les difficultés méthodologiques qui apparaissent alors, en prenant comme exemple un concepteur que nous imaginons seul face au système qu'il est en train de concevoir, un peu comme au paragraphe 3.1.1. Notre concepteur détermine l'architecture de son système, développe ou réutilise chacun des modules identifiés, chacune des ressources nécessaires, et, face aux difficultés rencontrées, opère un certain nombre de simplifications. Car, effectivement, c'est en implémentant que des difficultés apparaissent et que les ambitions initiales retombent quelque peu.

Considérons par exemple que les phénomènes initiaux incluent des anaphores, essentiellement des anaphores pronominales mais aussi quelques exemples d'anaphores associatives. Le module de résolution des références, qui regroupe la résolution des références directes, des références déictiques (et donc la fusion multimodale des gestes de désignation avec les expressions référentielles linguistiques), des déictiques de personne et des anaphores, doit donc intégrer un résolveur d'anaphores pronominales et associatives. Le concepteur se tourne ainsi dans un premier temps vers les résolveurs existants (Mitkov, 2002). Compte tenu du contexte d'implémentation (phénomènes spécifiques de la langue française, lexique considéré, éventail de phénomènes, formats d'entrée et de sortie), il se rend vite compte qu'adapter, paramétrer et intégrer dans son architecture un résolveur existant s'avère délicat. Il décide donc d'implémenter son propre résolveur directement dans le module et avec les paramètres pertinents. Or, cette tâche n'étant que l'un des aspects de l'un des nombreux modules du système, il en est réduit à éliminer certains phénomènes, par exemple l'ensemble des anaphores associatives, de manière à réaliser un résolveur opérationnel dans des temps et avec des moyens raisonnables. C'est dommage (et difficile à admettre), mais ça arrive.

Comme autre exemple de simplification relatif au dialogue multimodal, citons le traitement des références multimodales combinées, par exemple un seul geste déictique lié à deux ou trois expressions référentielles, ou encore plusieurs gestes liés à la même expression référentielle (ou à plusieurs expressions référentielles mais sans correspondance de un à une). Concevoir un module de fusion multimodale capable d'identifier ces situations est très délicat, et devient vite chronophage avec les inévitables problèmes techniques tels que la détection du début et de la fin d'un énoncé multimodal, la capture de trajectoires gestuelles et la gestion de la synchronisation temporelle. Dans le but d'obtenir un système capable de temps de réaction proches des temps de réponse humains, le traitement bas niveau des entrées doit être très rapide et doit éviter une gestion complexe en fin d'énoncé de toutes les hypothèses d'appariements gestes – expressions référentielles. Ainsi, et c'est le cas dans beaucoup de systèmes,

on oublie ce type de situations et on se focalise sur les références déictiques mettant en jeu un seul geste et une seule expression référentielle, ce qui pose déjà bien assez de problèmes (voir chapitre 6). Au final, le système ne traite qu'un sous-ensemble des phénomènes identifiés au départ, mais au moins il fonctionne. Bien entendu, dans le cas d'une conception d'un système par une équipe complète de développeurs, ou dans le cas de la réutilisation d'un système existant, les problèmes ne se posent pas de la même façon. Dans ces deux cas, la spécification des processus de traitement est directement liée à la définition de l'architecture du système (voir chapitre 4).

3.2.5. *L'écriture des ressources et le développement*

C'est au cours de l'implémentation que les véritables problèmes se posent, comme souvent en TAL. En programmant l'algorithme principal d'un module, on s'aperçoit que l'on manque de ressources, que réaliser le traitement est plus complexe que prévu, et qu'il est nécessaire de réduire le nombre des phénomènes à traiter. On peut aussi se rendre compte qu'il manque une donnée en entrée du module, par exemple un aspect prosodique que l'on avait négligé mais qui s'avère être un paramètre important à un moment donné. On constate que la sortie du module ne sera pas aussi complète ou précise que prévu. On s'aperçoit que l'exécution des algorithmes est plus lente que prévu (qu'espéré). En fin de compte, l'implémentation ne pose aucune surprise qu'à des concepteurs qui n'en sont pas à leur premier système de DHM. Dans tous les cas, clarté des spécifications, vérification des ressources disponibles, échanges entre plusieurs spécialistes sont des solutions classiques, peu originales mais essentielles à la conception de systèmes.

Y a-t-il un ordre dans le développement des modules ? *A priori*, si les entrées et sorties de chaque module ont été correctement définies et restent stables, l'ordre est indifférent et le développement peut être réparti entre plusieurs personnes. En pratique, et comme nous l'avons montré avec les scénarios de la section 3.1, mieux vaut commencer par le noyau du système, donc par le gestionnaire du dialogue, et finir avec les modules périphériques tels que la reconnaissance de la parole et la synthèse vocale. C'est une précaution qui permettra plus de tolérance aux petites erreurs de spécifications et aux petites surprises parsemant le développement.

Existe-t-il une liste indicative de ressources ? Tout dépend bien entendu des capacités de traitement envisagées pour le système, et aussi des possibilités de récupération en ligne ou hors ligne de ressources existantes. Avec ce que nous avons déjà présenté et quelques éléments supplémentaires qui s'en déduisent et que nous décrirons plus loin, voici une liste indicative qui montre la diversité des modèles impliqués dans un système de DHM :

– modèles du domaine : base de données des objets instanciés, éventuellement visibles et référables, par exemple les cubes et pyramides de Shrdlu, et le modèle physique décrivant les comportements et évolutions possibles de ces objets ;

– modèles de l'IHM : dans le cas d'un DHM avec affichage d'une interface complémentaire à l'écran, modèle de fonctionnement de celle-ci (actions-réactions et règles de priorité des commandes IHM sur les commandes vocales) ;

– modèles conceptuels : ontologie, graphe ou arbre des concepts concernés, avec à la fois les objets et les actions qui peuvent s'effectuer sur ces objets ;

– modèles de la tâche : suites d'actions conduisant à la satisfaction de la tâche, conditions, arbres de décision ;

– modèles acoustico-phonétiques pour la reconnaissance de la parole, ainsi que pour la synthèse (ce ne sont pas les mêmes modèles, mais ils peuvent avoir une partie commune) ;

– modèles prosodiques : contours intonatifs, règles de mise en relief, pour l'analyse aussi bien que pour la génération (même remarque) ;

– modèles linguistiques symboliques : lexiques, grammaires, structures sémantiques, types de formes logiques, règles pour l'enrichissement de ces formes logiques, mécanismes de déduction ou d'induction, etc. ;

– modèles linguistiques statistiques : modèles de langage pour aider la reconnaissance de la parole, statistiques sur les constructions d'énoncés, sur la détection d'actes de langage, sur les successions d'énoncés, etc. ;

– modèles gestuels : règles de production et formes de geste pour la désignation d'objet, pour la prise de parole, pour le rendu d'émotions, applicables aussi bien en analyse qu'en génération dans le cas de l'affichage d'un avatar ;

– modèles liés à des modalités particulières : base de formes de visages pour un module de suivi et d'analyse de visage (modèles d'émotions, par exemple), modèles pour la lecture sur les lèvres, modèles pour la reconnaissance de l'écriture ;

– modèles cognitifs : limites humaines à prendre en compte dans la génération de messages, charge cognitive, gestion de l'attention et de la saillance ;

– modèles de dialogue : structures possibles, conventions générales, formules de politesse et réactions associées, patrons d'ouverture et de clôture de dialogue, règles de passage de tour de parole ;

– modèles de l'interaction : au-delà du dialogue en langage naturel, modèles de gestion de l'interaction orale, visuelle, multimodale, ou encore *via* l'IHM.

3.2.6. *L'évaluation et le passage à l'échelle*

Dans un cycle de développement tel que l'on en voit en IHM, un « cycle en V » par exemple (Grislin et Kolski, 1996), les tests et les évaluations n'ont pas lieu uniquement en fin de conception, mais lors de plusieurs étapes : le développement des modules conduit à des tests unitaires, puis l'intégration des modules dans l'architecture globale conduit à des tests d'intégration, qui amènent alors à des tests du système global, ces derniers permettant de vérifier que les spécifications fonctionnelles ont été suivies.

Enfin, des tests d'acceptation sont mis en œuvre, de manière à valider l'analyse des besoins. Nous explorerons les différents types de tests envisageables en DHM dans le chapitre 10, et nous nous focalisons ici sur deux aspects méthodologiques particuliers : le passage à l'échelle et la réimplémentation.

Le « passage à l'échelle » est le passage d'un prototype de laboratoire correctement évalué à un véritable système opérationnel, utilisable par le public. Les enjeux principaux sont les suivants :

– passage de conditions de laboratoire contrôlées à des conditions réelles, variables et non maîtrisables ;

– passage de ressources et de données limitées et très fiables à des données réelles, de grande taille et avec des possibilités d'erreurs, de présence de bruit, de non-homogénéité ;

– passage d'une exécution ponctuelle du système pour une durée à chaque fois relativement courte, à un fonctionnement continu impliquant un minimum de redémarrages ;

– passage d'un mode d'utilisation autorisant une certaine tolérance à l'erreur et au dysfonctionnement, à une utilisation ne supportant aucune tolérance au dysfonctionnement et une tolérance très limitée à l'erreur ;

– passage d'un mode d'utilisation impliquant un utilisateur à la fois cadré et de bonne volonté (souvent expert), à un mode impliquant un utilisateur exigeant, parfois malveillant, prêt à jouer avec le système, voire à tenter de le bloquer par tous les moyens possibles. Il n'y a rien de plus éprouvant pour le concepteur que de voir un tel utilisateur, que l'on nomme utilisateur final, torturer son système…

La notion de « réimplémentation » intervient après l'évaluation. Elle consiste à modifier le code du système pour que celui-ci traite les quelques problèmes identifiés comme les plus fréquents ou les plus pénalisants. Le but est de minimiser cette phase de mise à jour du code, c'est-à-dire de tout faire, et ce dès le début de la conception, pour qu'elle soit la plus aisée et la plus rapide possible. (Rieser et Lemon, 2011, p. 16) souligne qu'à l'heure actuelle, aucune méthode rationnelle n'existe en DHM pour transformer ainsi des résultats d'évaluation en code. C'est pourtant le cas dans d'autres domaines de l'informatique, par l'exemple l'infographie et l'image de synthèse (notion de rééclairage après calcul complet d'un rendu), et c'est là encore un enjeu méthodologique essentiel pour les années à venir.

3.3. Bilan

Lorsque l'on se lance dans la réalisation d'un système de dialogue, on ne commence pas à programmer directement en laissant les problèmes apparaître au fur et à mesure. Au contraire, anticiper s'avère essentiel et de nombreuses étapes de travail

ont lieu avant tout développement informatique : définition d'une tâche qui correspond aux objectifs que le système doit remplir ; observation et analyse de dialogues humains portant sur une telle tâche ; détermination des phénomènes linguistiques et interactionnels que le système doit traiter ; réalisation d'expérimentations ; spécification de l'architecture logicielle du système ; identification et recueil des ressources nécessaires à son fonctionnement ; etc. Ce troisième chapitre donne des exemples de scénarios de réalisation incluant ces étapes avant de se concrétiser en développement, test et évaluation.

Chapitre 4

Des architectures pour des systèmes réutilisables

Le mot « architecture » a plusieurs sens, même dans le cadre du DHM. Il peut s'agir de l'architecture conceptuelle, c'est-à-dire de l'ensemble des composants logiciels (modules) et des modes de communications entre ces composants. Il peut s'agir de l'architecture logicielle, c'est-à-dire de la matérialisation de l'architecture conceptuelle sous la forme d'une solution informatique, par exemple un système multi-agents.

Par ailleurs, l'architecture peut concerner l'organisation du système lors de son exécution, reflétant ainsi son fonctionnement. On parle alors d'architecture *run-time*, architecture conceptuelle décrivant comment le système final est constitué. Il s'agit premièrement du schéma des modules, avec la spécification pour chacun d'eux des entrées, des sorties et des traitements effectués, et deuxièmement du schéma des communications entre modules. L'architecture peut concerner aussi non plus l'exécution mais la création du système, c'est-à-dire non plus le système lui-même, mais le système qui a permis de le créer. On parle dans ce cas, qui peut ne pas intervenir dans la conception, d'architecture *design-time*. Plus précisément, il s'agit de l'architecture conceptuelle du système de développement qui crée le système de DHM. C'est essentiellement un ensemble de contraintes séquentielles ou parallèles, qui constitue une chaîne logicielle pour aider le développement en dérivant automatiquement certaines ressources à partir d'autres ressources, et en générant automatiquement des ressources, voire des modules, à partir de modèles. Un atelier de génie logiciel est la matérialisation logicielle d'une architecture *design-time*.

Une fois ces définitions posées, plusieurs constats peuvent être faits par rapport aux réalisations passées et actuelles en DHM. Premier constat : l'architecture conceptuelle

run-time est probablement la caractéristique la plus importante d'un système. C'est ce qui permet de le décrire, d'en montrer les caractéristiques, c'est en quelque sorte une plaquette qui permet de mettre en avant les aspects innovants du système et de montrer qu'il ne s'agit pas d'une amélioration rapide d'un ancien système. Les conséquences sont multiples et conduisent aux deux autres constats. Deuxième constat : malgré les volontés de réutilisation et d'exploitation de ressources existantes, chaque système définit sa propre architecture *run-time*. Celle-ci est directement liée aux capacités de traitement, de compréhension et de génération, et comme tous les systèmes ne fonctionnent pas avec les mêmes entrées et sorties, il est logique que les architectures diffèrent. Néanmoins, les techniques actuelles permettent beaucoup plus de souplesse qu'auparavant quant au fonctionnement d'architectures partiellement implémentées, et on devrait pouvoir se reposer plus facilement sur des architectures de référence. Il n'en reste pas moins que la proposition d'un nouveau système s'accompagne souvent d'une nouvelle architecture, avec l'effet de plaquette et l'effet d'innovation mentionnés plus haut. Troisième constat : alors que la spécification d'une architecture devrait faciliter les collaborations entre chercheurs, c'est au contraire parfois une source de problèmes. Chacun y va de sa propre proposition, et concilier les approches peut s'avérer long et délicat. Plus que cela, il arrive que certains aient tendance à vouloir inclure les propositions des autres en tant que modules dans leur propre architecture. Ce qui ressort de ces problèmes, c'est qu'il manque une architecture générique de référence, fiable et applicable à tout système. L'essor des architectures *design-time*, s'il se confirme, permettra peut-être de donner un tour nouveau à ce verrou.

C'est ce que nous allons aborder dans ce chapitre, avec une section sur les architectures *run-time* (section 4.1), et une autre sur les architectures *design-time* (section 4.2).

4.1. Architectures *run-time*

4.1.1. *Une liste de modules et de ressources*

(López-Cózar Delgado et Araki, 2005, p. 5) montrent une architecture *run-time* très complète, qui intègre notamment des modules pour le traitement de modalités spécifiques comme la lecture sur les lèvres ou la capture du visage de l'utilisateur. D'une manière générale, tous les livres cités dans l'introduction présentent des architectures. Ce sont généralement des schémas composés de boîtes (modules), qui sont étiquetées (processus), avec des flèches entre certaines boîtes (séquences de traitement), elles-mêmes pourvues d'étiquettes (types de données échangées).

Dans les années 1980, les architectures suivent souvent l'ordre des traitements que nous avons déjà évoqué : reconnaissance de la parole, analyse syntaxique, analyse sémantique, gestion du dialogue, génération automatique puis synthèse vocale. Les variantes concernent surtout la façon de gérer les données. (Pierrel, 1987) présente

ainsi un ensemble de données statiques et de données dynamiques. Les premières regroupent un sous-ensemble des modèles dont nous avons fait une liste à la fin du chapitre précédent. Les secondes se réduisent à l'historique du dialogue et au modèle de l'utilisateur, tourné essentiellement vers des paramètres utiles à la reconnaissance vocale : modèles acoustiques individuels, paramètres sur la façon de prononcer les liaisons ou sur les contours prosodiques. Pour certaines tâches, le modèle de l'utilisateur inclut par ailleurs les droits d'accès et de contrôle des objets de l'application.

A l'heure actuelle, on retrouve dans la plupart des systèmes des équivalents de ces modules, avec en plus des modules dédiés à une modalité spécifique ou à la gestion d'un dispositif particulier. Mais, comme l'écrit (Cole, 1998, p. 198), le gestionnaire du dialogue est toujours au cœur du système. C'est lui qui gère l'historique du dialogue, qui enregistre au fur et à mesure du dialogue le déroulement des tours de parole, les énoncés prononcés, leurs caractéristiques linguistiques, notamment à la fois les expressions référentielles utilisées et les référents mentionnés, afin de pouvoir retrouver les informations nécessaires à la résolution d'une nouvelle référence, d'une anaphore ou d'une ellipse nominale ou verbale. Par ailleurs, l'historique stocke également l'état de la tâche, l'étape atteinte dans la stratégie de dialogue en cours, ou encore une description des succès et des échecs de la communication.

Plus qu'une ressource affectée à un module comme c'était le cas dans les années 1980, l'historique du dialogue peut désormais constituer un module à part entière, avec des procédures d'accès et de stockage. En effet, n'importe quel processus, par exemple l'analyse syntaxique, peut théoriquement y faire appel. Nos remarques au paragraphe 2.1.1 sur l'application de l'oubli en DHM vont aussi dans ce sens. Ce principe d'accès et de stockage à n'importe quel moment se généralise dans les systèmes actuels, d'autant plus si l'on cherche à se rapprocher d'un fonctionnement en temps réel, c'est-à-dire avec des analyses effectuées en cours d'énonciation par l'utilisateur. Ainsi, le module chargé de capter le signal audio stocke en temps réel le signal dans une ressource « énoncé », et, toujours en temps réel, les modules de reconnaissance de la parole et d'analyse prosodique mettent à jour cette ressource en ajoutant une ou plusieurs hypothèses de transcription, alors même que l'énoncé n'est pas terminé. En temps réel également, le module d'analyse syntaxique lit ces hypothèses de transcription et indique par une étiquette dédiée chaque moment où la phrase peut être considérée comme autonome. Le module chargé de détecter la fin de l'énoncé pioche lui aussi en temps réel dans cette ressource « énoncé », et indique que le système peut prendre la parole dès que les paramètres prosodiques et syntaxiques le permettent. Les modules sémantiques et pragmatiques se mettent alors à traiter l'énoncé, l'enrichir, et, si le gestionnaire de dialogue l'estime pertinent, prendre la parole peut être décidé, avec comme conséquence potentielle de couper la parole à l'utilisateur si celui-ci, le temps qu'un message soit généré, est toujours en train de parler (sachant que dans ce cas, la reconnaissance de la parole et les analyseurs prosodique et syntaxique continuent leur travail). Des systèmes tels que Nailon (Edlund *et al.*, 2005) ouvrent la voie à ce type de fonctionnement, du moins pour les aspects prosodiques. L'intérêt réside dans

la ressource « énoncé » qui évolue constamment au fur et à mesure des analyses réalisées : loin d'être une donnée figée, elle devient une donnée dynamique, parfois floue ou incomplète, que les modules du système vont exploiter comme ils peuvent, sans les contraintes strictes de complétude et de correction (grammaticale par exemple) imposées trop souvent. On le voit, l'architecture *run-time* correspondante nécessite de solides réflexions sur les ressources et sur les modules impliqués dans la prise de connaissance et dans la mise à jour de ces données.

4.1.2. *Le flux des traitements*

Le flux des traitements effectués par les différents modules de l'architecture peut se faire, comme nous l'avons déjà vu, de manière linéaire, c'est-à-dire en cascade : la sortie d'un module sert directement d'entrée pour le module suivant. D'un point de vue informatique, c'est une contrainte qui peut être à l'origine de l'implémentation réalisée de l'architecture *run-time* ainsi définie. C'est justement la différence entre architecture conceptuelle et sa matérialisation en architecture logicielle : une architecture logicielle linéaire n'a de sens que si l'architecture conceptuelle est elle-même linéaire.

Une architecture logicielle utilisée aux débuts du DHM est celle du tableau noir : les connaissances sont regroupées dans une sorte de base de données qui est accessible à tous les autres modules, un peu comme nous venons de l'illustrer pour les analyses en temps réel. Chacun des modules peut être tributaire de restrictions portant sur les composants de la base de données, et n'en voir qu'une partie seulement. Dans ce cas, l'architecture logicielle comprend un superviseur qui détermine à chaque instant les connaissances à activer. Par rapport à l'implémentation linéaire, celle du tableau noir apporte plus de souplesse et permet à plusieurs modules de collaborer. Bien entendu, encore faut-il que les processus gérés par les modules permettent ce type de collaboration : si l'on garde des modules implémentés de la même façon que dans le modèle linéaire, les processus ne peuvent s'exécuter qu'en cascade, que les données qui leur sont nécessaires soient toutes accessibles dans le tableau noir ou qu'elles transitent de module en module. C'est là aussi une différence entre architecture conceptuelle et architecture logicielle : le tableau noir est utile dès lors que l'architecture conceptuelle prévoit des modules qui partagent les mêmes données. L'agenda du système Gus en est un exemple. Dans des implémentations modernes, il est à noter que le tableau noir actuel, du moins pour les systèmes de DHM connectés, peut être… le *web* (surtout depuis le succès de l'informatique dématérialisée).

(Sabah, 1997) propose au milieu des années 1990 une architecture logicielle intéressante, appelée le carnet d'esquisses. Dans cette implémentation, chaque module est capable d'évaluer ce qu'il produit compte tenu des entrées qui lui ont été fournies. Ainsi, quand le module d'analyse syntaxique reçoit une transcription d'un énoncé, il procède à l'analyse syntaxique, produit ainsi une esquisse, et calcule un score de contentement par rapport à cette esquisse. Ce score de contentement est transmis au

module précédent, plus exactement au module qui a produit la donnée exploitée en entrée. Selon le score, celui-ci peut refaire son travail et produire une nouvelle transcription, la transmettre, et encourager ainsi le module syntaxique à produire une nouvelle esquisse, espérée meilleure. Il s'agit en fait d'une extension du tableau noir avec des boucles de rétroaction. Pour que les modules ne produisent pas toujours les mêmes analyses, du bruit est introduit à chaque étape. D'autres méthodes sont tout à fait envisageables pour remplacer ce bruit qui peut manquer de pertinence : enlever un paramètre utilisé par l'analyse, ou modifier l'importance d'un paramètre, par exemple prosodique.

L'architecture logicielle qui a le plus de succès est probablement le système multi-agents. A chaque module est associé un agent chargé des interactions avec tout le reste de l'architecture. Cet agent gère ainsi les entrées et sorties, et peut d'ailleurs procéder à des processus similaires à celui du carnet d'esquisses. Avec une telle matérialisation, tout problème se résout par l'interaction convergente des différents agents. Le système Trips, successeur du système Trains (Allen *et al.*, 1995), est implémenté sous la forme d'un système multi-agents, avec en gros les agents suivants : interprétation, gestion du dialogue, génération, contexte du discours, contexte référentiel, tâche, qui communiquent tous les uns avec les autres. Les possibilités d'échanges sont ainsi très nombreuses, et c'est l'interaction entre agents qui permet aux données dynamiques de se stabiliser et de produire un résultat satisfaisant.

Cependant, la nécessité de gérer le dialogue selon plusieurs niveaux de priorité, notamment – dans la lignée de notre exemple avec la ressource « énoncé » – un niveau temps réel très proche de la gestion des tours de parole, et un niveau plus réfléchi correspondant aux stratégies de dialogue, a entraîné l'essor d'architectures multicouches (*N-tiers*), avec de nouvelles contraintes et de nouvelles spécifications de systèmes. (Rosset, 2008, p. 83) mentionne un ensemble d'approches de ce type, avec l'exemple typique de l'architecture à deux couches pour gérer de manière simultanée et asynchrone les comportements à court terme comme la prise de parole et les comportements à long terme comme la planification de la tâche et du dialogue, mais aussi des exemples d'architectures à trois couches, capables de gérer de manière séparée plusieurs aspects de la communication, par exemple avec un ACA. Des architectures multicouches similaires sont par ailleurs exploitées depuis longtemps dans le domaine des IHM, avec une gestion de l'interaction qui sépare plusieurs logiques : la logique de persistance qui concerne les données durables, la logique d'application qui concerne la gestion de la tâche, la logique d'interaction qui concerne la gestion des actions de l'utilisateur, et la logique de présentation des données qui gère l'affichage en temps réel, de manière à ne pas présenter des données obsolètes ou qui viennent de faire l'objet d'une action de la part de l'utilisateur. En rationalisant la conception, ces architectures permettent une adaptation à différents terminaux et à différents contextes de travail. La généralisation de cette approche à une architecture pour les systèmes interactifs au sens large est récente. Elle permet d'intégrer modèles de conception logicielle et modèles de communication homme-machine. (Lard *et al.*, 2007) propose

ainsi une approche hybride destinée à spécifier une architecture pour des systèmes qui concilient IHM et DHM.

4.1.3. *Le langage d'interaction entre modules*

Quelle que soit l'architecture retenue, des données sont échangées. Un format s'avère ainsi nécessaire pour encapsuler ces données de manière standardisée, afin que chaque module de l'architecture arrive à décoder et à exploiter ce qui peut l'intéresser. Comme pour les architectures, il y a peut-être autant de propositions de langages d'interaction que de systèmes, notamment en dialogue multimodal, de par la diversité des modalités possibles et des types de contenus exploités. (Denis, 2008, p. 93) en cite quelques-uns et met en avant le langage Mmil, qui a été utilisé dans les projets Miamm, Ozone, Amigo ou encore dans la campagne d'évaluation Media (Devillers *et al.*, 2004), avec sa propre participation. Ce langage prend la forme d'un format de fichier standardisé, qui permet une représentation des événements communicatifs dans un système de DHM, qu'il s'agisse d'événements externes multimodaux (parole, geste) ou d'événements internes (échanges entre modules). L'intérêt est de pouvoir représenter aussi bien les événements que le contenu des messages échangés. La représentation est ainsi multiniveaux, et elle inclut par exemple une représentation du contenu sémantique d'un énoncé, ou encore son acte de langage. Les contenus restent aussi proches que possible de l'énoncé, sans trop d'interprétation. Le contenu sémantique reste ainsi très proche de la forme de l'énoncé, et n'intègre aucune représentation de l'implicite ni même de formalisation sous forme logique : c'est le choix qui a été fait, les enrichissements étant gérés par les modules concernés et non au niveau de l'architecture globale.

L'un des aspects de Mmil a aussi été l'objet de plusieurs travaux relatifs à la matérialisation d'architectures *run-time*. Il s'agit du rapprochement entre l'interaction homme-machine et l'interaction entre modules, purement informatique. A partir du moment où l'on formalise la communication homme-machine comme des ordres ou des questions portant sur des contenus sémantiques, rien n'empêche de faire de même pour la communication entre deux modules d'une architecture : un module qui a besoin d'une information pour terminer une analyse peut ainsi poser une question à d'autres modules, de même qu'un module qui vient de terminer une analyse peut la communiquer sous la forme d'une assertion. Même si ce n'est pas un enjeu essentiel, dans la mesure où la complexité de la communication homme-machine est bien supérieure, c'est une manière d'appliquer des recherches en pragmatique pour la conception d'un langage d'interaction cohérent.

4.2. Architectures *design-time*

La recherche d'un système de DHM générique, c'est-à-dire en partie paramétrable et réutilisable, peut passer par la spécification d'une architecture *run-time* générique.

Dans ce cas, un soin particulier est apporté à la détermination des modules et sous-modules, au langage d'interaction, et aux procédures de paramétrisation par la tâche. Celle-ci étant spécifique au système, tout ce qui concerne la tâche est géré de manière indépendante. Les objets de la tâche sont regroupés dans une base de données qui devient l'un des paramètres du modèle du domaine, le lexique spécifique de la tâche devient l'un des paramètres du module lexical, et ainsi de suite. Si la tâche autorise des énoncés avec une syntaxe spécifique, avec par exemple la disparition de certaines prépositions et de certains déterminants comme on le voit dans les langages spécialisés, alors ces structures syntaxiques particulières deviennent également l'un des paramètres du module syntaxique. On sépare ainsi ce qui est spécifique de la tâche (paramètres) de ce qui est commun à tout système (connaissances et processus).

Une autre façon de résoudre le problème de la généricité est de mettre des efforts dans la conception d'un environnement de développement de systèmes de DHM. Un système n'est plus réalisé de manière autonome, mais devient le produit d'une sorte d'usine à systèmes. C'est cet environnement de développement, ou boîte à outils, ou atelier de génie logiciel, qui va permettre d'importer les spécificités de la tâche et de lancer sur cette base la création du système visé. On entre alors dans le domaine de l'architecture *design-time*, c'est-à-dire de la conception de l'environnement de développement plutôt que du système de DHM lui-même.

4.2.1. *Boîtes à outils et ateliers de génie logiciel*

Un environnement de développement peut consister en une boîte à outils ou en un atelier de génie logiciel. Les boîtes à outils fournissent des interfaces de programmation aux développeurs de systèmes, c'est-à-dire des capacités techniques pour la mise en œuvre de techniques utilisées en DHM. Les ateliers de conception logicielle sont identiques dans leur utilisation aux boîtes à outils, mais ils intègrent souvent un modèle d'architecture en plus de proposer des techniques. Ils proposent également une plate-forme pour le support à l'exécution du code développé. Plus évolués, ils peuvent également émettre des recommandations basées sur le modèle qu'ils implémentent, en rapport avec des standards de développement logiciel.

(McTear, 2004) met l'accent sur l'utilisation de boîtes à outils et d'ateliers de génie logiciel. D'une manière générale, la plupart des grands laboratoires de recherche en DHM et des grandes sociétés informatiques proposent leur propre environnement de développement, basé sur leurs expériences de réalisation de systèmes. Entre autres exemples, on peut citer le fameux VoiceXML, la boîte à outils de CSLU ou encore celle de Carnegie Mellon, qui font régulièrement l'objet de tutoriels dans des conférences ou des écoles d'été. Le constat souvent réalisé est que les boîtes à outils disponibles aident au développement de systèmes simples, mais sont loin de permettre la conception de systèmes de dialogue naturel en langage naturel. Pour atteindre ce but, il faudrait que l'ensemble des aspects TAL soient implémentés pour plusieurs langues,

et cela relève encore des enjeux du TAL et du DHM. En revanche, les outils proposés facilitent le développement de systèmes et permettent de se lancer dans la conception d'un système de DHM sans avoir à tout imaginer à partir d'une feuille blanche comme c'était le cas aux paragraphes 3.1.1 et 3.1.2.

Plus récemment, les efforts entrepris dans le domaine de l'ingénierie dirigée par les modèles (MDE, *Model-Driven Engineering*), et dans sa facette qui concerne les architectures, à savoir les architectures dirigées par les modèles (MDA, *Model-Driven Architectures*), sont venus aux oreilles des chercheurs en communication homme-machine, en commençant par les spécialistes des IHM. Ceux-ci ont ainsi entamé des réflexions sur la remise à plat du cycle de conception d'une IHM, pour que des modèles et métamodèles soient pris en compte lors des toutes premières phases de conception, les dernières étapes consistant dans la génération automatique de l'IHM par l'environnement de développement. De telles réflexions ont débuté en DHM, et leur poursuite constitue un enjeu de taille pour le domaine. Plus précisément, le concepteur spécifie des modèles conceptuels et des processus de transformation de modèles (de modèles de haut niveau vers des modèles de plus bas niveau). L'atelier de génie logiciel génère alors des modèles, puis, après une ou plusieurs phases de génération et de dérivation, finit par produire du code exécutable. A l'heure actuelle, cette démarche très logicielle ne prend pas suffisamment en compte les besoins des utilisateurs et ne s'adapte pas facilement à des processus aussi complexes que ceux de la compréhension automatique du langage naturel. Il faut avouer que la définition des modèles et des métamodèles pose de très nombreux problèmes, surtout si l'on veut tenir compte des spécificités du langage naturel et du dialogue naturel, voire du dialogue multimodal : il faut définir un langage de représentation commun à tous les modèles, de manière à faciliter la gestion de ceux-ci par l'environnement de développement. Or les modèles acoustiques, lexicaux, syntaxiques, pragmatiques, les modèles décrivant la structure d'un dialogue, ou encore les modèles pour le geste de désignation, font intervenir des connaissances très disparates, et pour lesquelles manquent des standards de représentation interconnectés. De plus, les processus de reconnaissance de la parole, d'analyse syntaxique, de résolution de la référence, d'attribution d'états mentaux, etc., font eux aussi l'objet de représentations dans des modèles dédiés. Or, si des descriptions et des formalismes existent, il reste un effort important à faire pour aboutir à des représentations exploitables dans le cadre d'une approche dirigée par les modèles.

4.2.2. Middleware *pour l'interaction homme-machine*

Une alternative, qui rejoint un peu les enjeux de type « OpenDial » décrits à la fin du premier chapitre, réside dans la conception d'un *middleware* adapté aux besoins du DHM. Un *middleware* est un composant logiciel d'interconnexion, qui consiste en un ensemble de services permettant à de multiples processus de s'exécuter sur une ou plusieurs machines et d'interagir à travers un réseau. Plus proche de nos préoccupations,

un *middleware* est aussi et surtout un composant logiciel qui constitue une couche de conversion et de traduction entre deux processus. A l'origine, le *middleware* est essentiellement un *middleware* système, c'est-à-dire une couche de conversion insérée entre un flux de données produit par une machine spécifique et le système d'exploitation de la machine. On distingue ensuite les *middlewares* explicites (couche intermédiaire entre le système d'exploitation et n'importe quelle application exécutée dessus) et les *middlewares* implicites (processus de médiation ou d'interprétation entre une application métier et l'application de présentation qui lui est associée). Le domaine des IHM a amené le développement de plusieurs *middlewares* implicites. (Lard *et al.*, 2007) propose un *middleware* implicite, basé sur une architecture multicouches, pour faciliter le développement de systèmes d'interaction homme-machine, en incluant quelques aspects très simplifiés de DHM. Le principe est l'ajout d'une couche dédiée à l'interaction homme-machine dans l'architecture multicouches de départ. Cette couche est implémentée en tant que *middleware* d'interaction, et fournit des services génériques pour l'interaction homme-machine, de même qu'une abstraction des spécificités des applications et des contextes d'utilisation. Cette proposition se focalise sur les aspects techniques liés à l'architecture générale, et n'aborde quasiment pas les problèmes de TAL et de dialogue naturel en langage naturel. C'est cependant une voie qui constitue un enjeu pour la facilitation des processus de développement de systèmes de DHM.

4.2.3. *Enjeux*

Nous l'avons vu, les pistes entamées vers la réalisation et la diffusion d'architectures *design-time* sont encore balbutiantes. Elles permettent plus de dresser une liste des enjeux pour les années à venir que de faire un bilan des avancées réellement effectuées. De fait, la conception de beaucoup de systèmes actuels fait l'impasse sur le côté *design-time*, la spécification d'une architecture *run-time* fédérant la majorité des efforts.

Dans les sections précédentes, nous avons donc vu qu'un premier enjeu consistait en l'élaboration et la diffusion d'environnements de développement mieux adaptés au DHM, c'est-à-dire aux aspects TAL dans toute leur complexité et ce jusqu'aux techniques de gestion d'un dialogue naturel. Un deuxième enjeu consiste à appliquer au DHM la démarche de l'ingénierie dirigée par les modèles, en tenant compte là aussi de la diversité des aspects impliqués, et donc en implémentant une très grande variété de modèles, qui reprenne un peu la liste mentionnée au paragraphe 3.2.5, en ajoutant tous les aspects techniques, et notamment les aspects d'adaptation du système (interfaces plastiques). Un troisième enjeu réside dans la spécification d'un *middleware* d'interaction plus élaboré que celui présenté dans (Lard *et al.*, 2007), avec là aussi une prise en compte des aspects TAL dans toute leur variété et complexité. Cet enjeu peut d'ailleurs se multiplier si l'on considère qu'un tel *middleware* serait utile pour chacun des domaines du TAL, pour celui des ACA, ou encore pour celui de l'IA, avec un ensemble de services dédiés aux différents types de raisonnement logique.

Enfin, un enjeu souvent mis en avant en IA est la conception de systèmes capables d'évaluer et de modifier leurs propres algorithmes. Avec les principes de dérivation de modèles et d'architectures dirigées par les modèles, ce vieux rêve commence à devenir envisageable. En effet, un système de DHM obtenu par dérivation de modèles peut, au cours de son exécution, mettre à jour certains des modèles sur lesquels il repose. S'il est doté de capacités d'apprentissage, il peut par exemple vouloir enrichir un modèle avec des connaissances nouvelles. S'il est doté de calculs de scores de confiance ou de pertinence, il peut vouloir paramétrer les données d'un modèle en exploitant ces scores, de manière à renforcer l'impact de certains paramètres par rapport à d'autres. Théoriquement, plus le système de DHM est utilisé, plus il est susceptible de remettre en cause les modèles sur lesquels il a été construit. On peut alors imaginer une phase où ce système de DHM décide lui-même de se mettre à jour en relançant tout le processus de dérivation.

4.3. Bilan

Un système de dialogue finalisé est conçu pour aider l'utilisateur dans l'accomplissement d'une tâche donnée. Tout système est ainsi optimisé pour une tâche particulière et s'avère *a priori* peu performant pour une autre tâche. Or, compte tenu de l'important effort de conception, se pose la question de la réutilisation de composants d'un système à l'autre, et, au-delà, de la possibilité de concevoir des composants génériques, exploitables quelle que soit la tâche car paramétrable de manière souple par celle-ci. Ce quatrième chapitre explore cette question de généricité à travers des exemples d'architectures conceptuelles et logicielles, avec d'une part la notion d'architecture fonctionnelle, à l'exécution du système, et d'autre part la notion d'architecture à la conception, avec par exemple le développement dirigé par des modèles, dont le principe est de partir d'un ensemble de modèles décrivant les ressources et les processus, et d'en dériver automatiquement les modules qui constitueront le système final.

DEUXIÈME PARTIE

Les traitements des entrées

Chapitre 5

Analyses et représentations sémantiques

Cette deuxième partie, centrée sur le traitement des énoncés de l'utilisateur en entrée du système, commence avec ce que l'on peut considérer comme l'essentiel des capacités de compréhension d'un système : la détermination du sens des énoncés. Le but est d'obtenir une représentation informatique de ce sens, à partir du signal audio en entrée et de la représentation du contexte en interne au système. Le but est d'obtenir une représentation qui ne fasse pas intervenir d'aspects pragmatiques comme la résolution de la référence ou la détermination de l'acte de langage (voir chapitres 6 et 7). L'appel au contexte est cependant indispensable, parce qu'un énoncé en dialogue peut prendre la suite de l'énoncé précédent, par exemple en ajoutant un complément circonstanciel qui ne s'interprète qu'à l'aide de cet énoncé précédent, ou peut reprendre un terme qui avait fait l'objet d'une désambiguïsation lexicale, ou encore peut présenter une ellipse, c'est-à-dire l'omission d'un constituant, comme un verbe ou un nom, du fait d'un emploi préalable de ce constituant. La détermination du sens complet de « je souhaite le plus court » après avoir parlé de chemin pour aller à Paris peut ainsi intégrer ce concept de chemin, de même que celle de « et combien par l'autre chemin ? » après « combien de temps ça prend par ce premier chemin ? » peut intégrer « prendre du temps », ou du moins le fait que la question porte sur une durée et non une distance. Dans l'exemple de l'introduction, en considérant qu'il s'agit de dialogue oral, on peut aussi considérer que, du fait du contexte, l'énoncé « combien de temps par ce chemin qui semble être le plus court ? » comporte bien une ellipse et ne correspond pas à une phrase avec le verbe partir (« combien de temps part ce chemin… »). Les exemples sont extrêmement nombreux et divers, et c'est ce qui fait la difficulté de ce processus. Notre approche ici n'est pas de proposer un modèle de détermination du sens (il en existe déjà des dizaines), mais de donner une idée des phénomènes linguistiques qui apparaissent en DHM et des processus de traitement à envisager pour les traiter.

La première section de ce chapitre vise ainsi à décrire quelques phénomènes intervenant en dialogue oral et multimodal (section 5.1), et les deux autres à donner un aperçu des traitements informatiques impliqués dans la détermination du sens, avec un ensemble de processus qui analysent les caractéristiques de l'énoncé (section 5.2) et un ensemble de processus qui enrichissent celles-ci avec des éléments complémentaires pour aboutir à une représentation qui soit exploitable pour la gestion du dialogue (section 5.3).

5.1. La langue dans le dialogue et dans le dialogue homme-machine

5.1.1. *Caractéristiques principales du langage naturel*

Un énoncé est composé de mots, qui ont chacun une ou plusieurs significations. Dans une situation donnée, certaines significations ne sont pas possibles et on cherche à identifier le sens que prend le mot en contexte. Plusieurs mots se regroupent pour former un constituant, caractérisé par une fonction dans la phrase (sujet, object direct) et par un rôle actanciel (agent, patient). Certains de ces constituants réfèrent à des objets particuliers du contexte. Même si sens et référence sont reliés, ce point particulier sera l'objet du chapitre 6. Comme les autres mots pleins, le verbe de la phrase a un sens, et c'est avant tout ce sens qui permet de relier entre eux les constituants de la phrase, et d'aboutir à la détermination du sens de l'énoncé. En particulier avec notre exemple de trains, mais aussi d'une manière générale, le sens peut faire intervenir des notions comme le temps et l'espace, et dans ce cas la compréhension de l'énoncé fait intervenir des connaissances sur la temporalité (notion de date, recouvrement d'horaires, durée d'un trajet, repérage d'un événement par rapport à un autre) et la sémantique de l'espace (notion de lieu, de déplacement). Les caractéristiques du langage naturel sont ainsi multiples, et nous nous focaliserons sur les aspects suivants, avec à chaque fois quelques pistes pour leur prise en compte dans des systèmes de DHM : polysémie, métonymie, métaphore, sémantique verbale, implicite, ambiguïté, et structure informationnelle.

La confusion entre « par » et « part » qui peut intervenir en dialogue oral montre un exemple de deux homophones, mots qui se prononcent de la même façon tout en ayant des significations différentes. Or la langue a ceci de particulier qu'un même mot peut avoir plusieurs significations, phénomène appelé polysémie. De fait, tous les mots de la langue ou presque sont polysémiques. Le billet de train dont nous parlions peut désigner le bout de papier ou le droit à la place assise qu'il matérialise, et si l'on se limite au mot « billet », sans prendre en compte le mot composé « billet de train », le nombre de sens possibles dépasse la dizaine. La polysémie provient des relations métonymiques ou métaphoriques (voir ci-dessous), des restrictions de sens, des extensions de sens (« minute » désigne une portion de temps très précise, mais aussi, par extension, un instant de courte durée), ou encore de phénomènes liés à la différenciation d'un trait qui conduit à tenir compte de deux notions. Pour le TAL

et le DHM en domaine ouvert, un enjeu consiste à décrire ces transformations (ou à exploiter une voie parallèle à base de statistiques sur des cooccurrences) de manière à déduire automatiquement les sens polysémiques possibles d'un terme de départ. Pour le DHM en domaine fermé, on se contentera de multiplier les entrées du lexique, en retenant les sens les plus pertinents compte tenu de la tâche.

La langue a aussi ceci de particulier qu'un mot peut prendre la place d'un autre. La métonymie intervient ainsi quand un contenant prend la place d'un contenu (« réserver un train » pour une place dans un train), une partie prend la place d'un ensemble (cas de la synecdoque), une qualité prend la place d'une personne (« les premières classes font du bruit »), un instrument prend la place d'un agent, une action prend la place d'un agent, du lieu où elle se déroule, de ses propres effets, etc. Pour le DHM, on retrouve les deux techniques précédentes : soit on multiplie les entrées du lexique et les structures de phrases possibles, ce qui entraîne une multiplication des possibilités qui reste gérable dans le cadre d'une tâche précise, soit on ajoute des règles autorisant les remplacements métonymiques et permettant une analyse sémantique fine.

Phénomène proche, la métaphore consiste à donner à un mot le sens d'un autre, par analogie. C'est ainsi que les boutons et les menus des IHM sont appelés métaphores. En continuant avec notre dialogue de réservation de billet de train, c'est ainsi que « je ne veux pas d'un escargot, je prendrai le train qui passe par Meudon » est difficile à interpréter pour un système si, du fait de la tâche, « escargot » n'appartient pas au lexique. Là encore, un enjeu pour le TAL ou le DHM en domaine ouvert est de retrouver le trait d'« escargot » qui permet de faire un rapprochement avec un train, et donc de comprendre que l'utilisateur ne souhaite pas un train trop lent. Dans le cas d'un DHM en domaine fermé, une stratégie de repli consiste tout simplement à ignorer la métaphore incompréhensible, et à ne retenir que la deuxième partie de l'énoncé pour déterminer la réaction du système.

Les phénomènes mentionnés jusqu'à présent restaient au niveau du mot. Or certains mots tels que « réservation » ont une caractéristique supplémentaire : ils permettent de faire des liens avec d'autres composants de l'énoncé. Une réservation est faite par quelqu'un, l'agent, et concerne quelque chose, l'objet. L'identification de ces deux actants est nécessaire pour interpréter correctement. On dit que « réservation » a une valence de deux, de même que le verbe « réserver ». La sémantique verbale repose ainsi sur cette notion de valence, qui permet de donner un sens aux prépositions utilisées dans l'énoncé et de mettre en rapport ses différents constituants. Dans « je réserve un billet », la sémantique du verbe « réserver » permet ainsi de déterminer l'agent, « je », donc l'utilisateur, et l'objet, un billet. Même chose pour « j'annule ma réservation du retour », à la fois pour le verbe « annuler », de valence deux, et pour le nom « réservation », avec le possessif qui permet de vérifier l'agent, et le complément du nom « du retour » qui permet d'identifier l'objet. En DHM, l'identification du verbe et de sa sémantique peut ainsi être un point d'entrée pour la compréhension automatique. La prise en compte de la valence ne permet cependant pas de rattacher

tous les composants de la phrase. Ainsi, dans « j'annule définitivement ma réservation du retour » et « j'annule ma réservation du retour parce que je vais rester à Paris », un complément de verbe non nécessaire vient s'ajouter, pour donner des indications (pertinentes ou pas pour le système) sur les conditions d'application du procès dénoté par le verbe. Pour le TAL comme pour le DHM, il s'agit de répertorier les différents verbes avec leurs caractéristiques, et, lors de l'analyse, d'identifier les actants exprimés ainsi que ceux qui ne le sont pas et qui relèvent alors d'une ellipse ou d'un usage particulier. Des dictionnaires sont disponibles pour le premier point, et des stratégies d'interprétation doivent être mises en œuvre pour le second.

Le fait que certains paramètres peuvent ne pas être exprimés est à la source d'un phénomène essentiel de la langue : l'implicite. Tout ce qui n'est pas exprimé mais véhiculé par l'énoncé relève de l'implicite, qui regroupe ainsi des phénomènes très variés. Parmi ces phénomènes se trouvent les ellipses, les sous-entendus (« je ne veux pas d'un escargot » peut sous-entendre « arrêtez de me proposer des trains trop lents »), les présupposés et autres inférences qui relèvent d'une analyse pragmatique complexe et non d'une analyse sémantique. L'identification de l'implicite est un défi pour le TAL et le DHM. Outre le cas des ellipses qui reste modélisable avec des appels à l'historique du dialogue, l'implicite repose sur des hypothèses difficiles à formuler. Dans le cadre d'une tâche précise, les inférences sont plus faciles à réaliser qu'en domaine ouvert car elles s'orientent préférentiellement vers la résolution de la tâche. En domaine ouvert, des approches comme la théorie de la pertinence (Sperber et Wilson, 1995) donnent un cadre intéressant mais difficile à implémenter.

Avec ou sans implicite, un énoncé en langage naturel se caractérise souvent par des phénomènes d'ambiguïté, c'est-à-dire par la possibilité d'obtenir plusieurs interprétations alternatives. Cette multiplicité peut provenir d'un terme ou d'une référence qui laisse un choix entre plusieurs possibilités, même après exploitation du contexte, comme « je veux partir vers six heures » qui peut vouloir dire le matin ou le soir. Elle peut aussi provenir de la structure de la phrase, qui, à cause d'une ambiguïté sur la portée d'une préposition, ne permet pas d'affecter les bons composants à un nom ou à un verbe. Dans « je vais prendre le billet du train de Meudon à Paris », s'agit-il de réserver un aller Meudon – Paris, d'aller chercher un billet à Paris, ou encore (peut-être moins probable) d'aller chercher un billet depuis un train Meudon – Paris ? L'enjeu pour le TAL et le DHM est d'identifier les termes et les structures susceptibles d'engendrer des ambiguïtés, sur la base par exemple d'un inventaire de type grammaire (Fuchs, 2000), c'est-à-dire d'un ensemble de règles portant sur des mots précis et sur l'ordre des mots dans la phrase.

Comme nous le verrons en section 5.2, un ensemble de règles de cette sorte peut aussi servir pour l'analyse syntaxique de la phrase, en phase ou séparément de l'analyse sémantique. Les éléments mentionnés ci-dessus permettent dejà de se faire une idée sur l'étendue et la complexité des traitements. Un dernier aspect sur lequel nous voulons insister ici est la « structure informationnelle ». Cette notion décrit le fait que

certains actants sont mis en avant par rapport aux autres selon des mécanismes qui vont de la structure syntaxique, en incluant l'ordre des mots et l'utilisation de constructions particulières, jusqu'à la prosodie. On peut par exemple faire une distinction entre l'information déjà connue (le support) et l'information donnée par l'énoncé (son apport), entre le topique (ce dont l'énoncé parle) et le commentaire (ce qui en est dit), ou encore entre ce qui est focalisé (le focus, qui reçoit un accent de proéminence prosodique) et ce qui ne l'est pas. Ces dichotomies conduisent à hiérarchiser les constituants de l'énoncé, cette hiérarchisation intervenant comme un point de vue par rapport au contenu sémantique. C'est ainsi que l'on fait la différence entre « je veux *un* billet pour Paris en première classe », « je veux un billet *pour Paris* en première classe » et « je veux un billet pour Paris en *première* classe » (l'italique indiquant l'accent), après, par exemple, le même énoncé sans accent et une erreur de la part du système. Par ailleurs, la structure informationnelle prend aussi sens au-delà des frontières de l'énoncé, et on peut ainsi ajouter aux analyses sémantiques une caractérisation de la progression topicale, qui décrit comment les topiques s'enchaînent dans le dialogue, et quels rapports s'établissent entre les uns et les autres. En DHM, une identification de la structure informationnelle et de la progression topicale permet de mieux gérer la manière dont l'énoncé courant s'insère dans l'historique du dialogue, donc de mieux gérer le dialogue et de parvenir plus efficacement à la satisfaction de la tâche, par exemple en revenant sur le topique principal si le dialogue s'en écarte trop.

5.1.2. *Langue orale et langue écrite*

L'oral et l'écrit diffèrent non seulement par leurs moyens expressifs, par la prosodie avec l'accent de proéminence dont nous venons de voir un exemple, mais aussi par l'utilisation qu'ils font de la syntaxe et par leur morphologie : nombres différents de graphèmes et de phonèmes, marques de pluriel différentes, importance de la prononciation des liaisons, voir paragraphe 2.2.1. Différents registres de langue peuvent coexister à l'oral. Des énoncés tels que « est-ce que je pourrais avoir un train pour Paris ? » peuvent alterner avec « un train pour Paris, c'est possible ? », ce qui donne au final une multitude de possibilités (Cohen *et al.*, 2004). L'oral ferme les yeux plus facilement que l'écrit sur certaines « erreurs », comme les concordances de temps ou l'accord avec l'auxiliaire avoir. L'oral se caractérise par des phénomènes de bruit (hésitation, interjection), de distorsion (répétition, précision sans annulation, correction avec annulation), de fragmentation (reprise après interruption, juxtaposition, dislocation, bribe) et d'ellipse, phénomènes qui peuvent apparaître à l'écrit mais y sont beaucoup moins fréquents, et qui entraînent des structures syntaxiques diversifiées : « le train pour Paris, il part à quelle heure ? », « la nuit, le train pour Grenoble, il s'arrête où ? », « c'est le chemin le plus court, celui qui passe par Meudon ? », ou encore « j'en voudrais deux, de billets ». Les conséquences pour le DHM sont tout simplement la multiplication des structures de phrases possibles.

L'oral donne une importance particulière à la « macrosyntaxe » (Blanche-Benveniste, 2010), c'est-à-dire à une organisation grammaticale qui, contrairement à la syntaxe, n'est pas fondée sur les catégories grammaticales : les unités en sont les énoncés, avec comme composants le noyau (au centre), le préfixe (ce qui est placé avant le noyau et correspond en partie au topique), le suffixe (ce qui est placé après le noyau) et éventuellement le postfixe (ce qui est placé après un suffixe). La macrosyntaxe permet notamment, et c'est aussi son intérêt en DHM, de rendre compte du lien entre des constructions successives qui ne sont ni coordonnées ni subordonnées comme elles le seraient peut-être à l'écrit, mais qui constituent un ensemble, de par le critère prosodique qu'est la période intonative. Celle-ci correspond à une segmentation de l'énoncé sur un ensemble de critères incluant la distribution des pauses et les variations de hauteur (mélodie). Des logiciels comme Analor (Analyse de l'oral) savent les détecter automatiquement, mais sur corpus, *a posteriori* et non en temps réel. L'exploitation automatique de la macrosyntaxe en DHM reste encore un enjeu.

Enfin, l'oral est bien entendu le domaine de la prosodie, qui regroupe les accentuations, c'est-à-dire la mise en relief d'une unité par rapport aux autres (comme on l'a vu avec l'accent de proéminence), le rythme (distribution des pauses, vitesse d'élocution et sa variation dans un même énoncé ou entre plusieurs énoncés), et l'intonation, courbe mélodique qui permet de donner à l'énoncé une valeur illocutoire (ordre, question, assertion). (Rossi, 1999) propose une description de la prosodie en langue française, dont certains aspects peuvent être appliqués au DHM. L'enjeu consiste à mettre en œuvre une analyse en temps réel capable de détecter les groupes rythmiques qui constituent les unités, de détecter les accentuations, sur des critères plus ou moins corrélés d'intensité et de durée, portant sur des unités ou sur l'énoncé complet, d'identifier les périodes intonatives pour aider à l'analyse macrosyntaxique, et de décrire la courbe mélodique des groupes rythmiques, pour annoter ceux-ci en valeurs illocutoires potentielles et en types de modalité, une modalité étant une modification du contenu énoncé en le présentant comme nécessaire, possible, probable, etc. (paragraphe 5.3.1).

5.1.3. *Langue et dialogue spontané*

Dans un dialogue, du fait que deux interlocuteurs se parlent l'un à l'autre, des termes d'adresse peuvent survenir : « *toi*, montre-moi les trains pour Paris », « hé, *machine*, tu en mets du temps ! ». Ces termes qui servent à désigner son interlocuteur, avant tout dans le but d'attirer son attention, ne posent pas de problème particulier en DHM, surtout quand il n'y a comme ici aucune ambiguïté sur l'identité de l'interlocuteur, ce qui ne serait pas le cas dans un polylogue. Il faut juste prévoir un traitement pour, soit les ignorer dans la structure syntaxique de l'énoncé, soit en abstraire un qualificatif positif ou négatif et réagir éventuellement en fonction.

D'une manière générale, la façon dont l'utilisateur s'adresse à la machine peut prendre une multitude de formes, du fait de l'absence de statut social de celle-ci :

« donne-moi un billet de train pour Paris » (forme impérative, plus facile à utiliser avec une machine qu'avec un interlocuteur humain), « réserver un aller pour Paris » (forme infinitive, spécifique au DHM), « billet pour Paris, s'il vous plaît » (forme elliptique), « nous allons réserver un aller pour Paris », « alors on me donne un billet pour Paris », etc. Par ailleurs, les énoncés ne sont pas toujours des phrases complètes – voir le chapitre 10 de (Cohen *et al.*, 2004) – ce qui pose en DHM plusieurs problèmes : d'une part au niveau de la détection de la fin de l'énoncé, seul le critère prosodique étant pertinent dans ce cas de figure, d'autre part au niveau de l'analyse syntaxique qui, si elle aboutit à quelque chose, renvoie une structure incomplète, éventuellement fausse quand l'analyseur n'est pas adapté à ce type de phénomène. Des adaptations sont ainsi nécessaires, comme le montre par exemple (Vilnat, 2005). Enfin, un énoncé peut être incomplet au point de ne s'interpréter qu'en lien avec l'énoncé précédent. Ce sont les réponses à des questions et l'ensemble des « énoncés non phrastiques » (NSU, *non sentencial utterances*), tels que « d'accord », « merci », « désolé », etc., étudiés notamment par (Ginzburg, 2012).

5.1.4. *Langue et gestes conversationnels*

Dans un dialogue humain de face à face, les gestes jouent un rôle à la fois dans l'organisation de l'interaction, dans la transmission d'émotions ou de modalités, et dans la désignation des éléments de la scène visuelle partagée (Kendon, 2004).

Langue et geste sont complémentaires dans la communication spontanée (Landragin, 2006), ce qui conduit à favoriser, quand les moyens techniques le permettent, le DHM multimodal au DHM oral. Ce sont ainsi des gestes qui permettent à l'utilisateur de prendre la parole, de la garder, de signaler qu'il comprend bien ce que le système est en train de lui dire. Ces gestes synchronisateurs s'interprètent sans l'aide de l'énoncé prononcé simultanément. En DHM, un dispositif de capture comme une caméra et des algorithmes de reconnaissances de forme sont nécessaires. Ce sont aussi des gestes qui transmettent une bonne partie de la composante émotionnelle du message : gestes expressifs comme les mimiques faciales, gestes paraverbaux comme les mouvements qui rythment les paroles, qui appuient certains mots. Ce sont encore des gestes qui permettent à l'utilisateur de donner des informations sur les référents de l'énoncé, qu'il s'agisse d'un pointage (geste de désignation ou déictique) vers le référent en question, quand il est présent dans la scène visuelle partagée, ou d'un geste illustratif qui indique une taille, une forme, une action. Enfin, ce sont aussi des gestes qui peuvent porter un acte de dialogue, comme un geste d'interrogation (yeux écarquillés, sourcils levés) ou un geste de citation proche du discours rapporté, comme peut le faire un pointage vers son interlocuteur pour signifier « comme tu l'as dit tout à l'heure » (Clark, 1996, p. 258).

Dans un même ordre d'idée, et même si un geste n'est pas forcément présent, la langue en dialogue peut comporter des « termes déictiques », c'est-à-dire des termes

ou expressions qui renvoient à la situation de communication, avec les trois catégories que sont les déictiques de personne, qui réfèrent aux interlocuteurs à l'aide de formes regroupant les pronoms de personne, les déterminants possessifs et certains noms, les déictiques spatiaux tels que « ici » ou « là », et les déictiques temporels tels que « maintenant », « demain », « tout à l'heure » ou « dans trois jours ». D'autres mots ont aussi une fonction déictique, comme les verbes présentatifs « voici » et « voilà ». Pour tous les cas de déixis, un système de DHM doit faire le lien entre les déictiques prononcés et le contexte situationnel. Ce lien repose donc sur les interlocuteurs, sur les objets de la scène, et sur des repères spatiaux et temporels qui seront autant de paramètres dans la représentation sémantique et pragmatique de l'énoncé.

5.2. Les traitements informatiques : du signal à la représentation du sens

La compréhension automatique d'un énoncé commence par la reconnaissance vocale et l'analyse prosodique, et se termine avec l'enrichissement d'une représentation sémantique par des inférences réalisées essentiellement à partir du contenu de l'énoncé. Cette représentation sera ensuite analysée avec une approche pragmatique qui permettra de l'exploiter pour la gestion du dialogue. Les processus impliqués dans la construction de cette représentation sémantique se déduisent du contenu de la section précédente : processus d'analyse lexicale, de détection de polysémie, de métonymie, de métaphore, etc. Sans refaire une telle liste, cette section a pour but de présenter les processus généraux, en mentionnant des solutions méthodologiques et techniques, et en soulignant les enjeux les plus cruciaux.

5.2.1. *Analyses syntaxiques*

L'analyse syntaxique consiste à mettre en évidence et à représenter dans une donnée informatique la structure d'une phrase, avec ses constituants que sont le verbe, le sujet, le complément d'objet direct, etc. Pour ce faire, elle nécessite une liste des mots de la langue, avec, pour chaque entrée, la catégorie (verbe, nom commun) et les propriétés morphologiques (genre, nombre, personne). Ces dernières permettent de gérer la morphologie en même temps que se déroule l'analyse syntaxique. Suite aux travaux de R. Montague (Muskens, 1996), la compréhension automatique, particulièrement pour l'écrit, a pris une voie qui consiste à réaliser une analyse syntaxique de la phrase avant d'en faire une analyse sémantique, celle-ci conduisant à la détermination de la forme logique ou des formes logiques décrivant les sens possibles. Une analyse pragmatique tenant compte des aspects contextuels enrichit alors cette représentation sémantique pour obtenir une forme propositionnelle, résultat de l'ensemble du processus de compréhension automatique.

Pour le traitement de l'oral spontané, ce processus qui passe par une analyse syntaxique globale de la phrase n'est pas toujours adapté : comme nous l'avons vu plus

haut, un énoncé en dialogue peut correspondre à une phrase incomplète, et d'autres mécanismes doivent être mis en œuvre pour que le système ne bloque pas à cause d'une analyse syntaxique impossible à réaliser. On en vient à implémenter des analyseurs syntaxiques capables de produire des analyses partielles, c'est-à-dire capables de gérer la sous-spécification par exemple d'un actant, ou des analyses locales, qui produisent un résultat avec les données disponibles, même si elles sont parcellaires.

Analyse globale et analyse locale ou partielle, ainsi qu'analyse macrosyntaxique, se complètent ainsi dans le but d'exploiter au mieux les données en entrée. Le principe de l'analyse locale permet notamment de modérer l'importance de la syntaxe dans le processus de compréhension, et de mettre en avant l'analyse sémantique, qui, en quelque sorte, dirige les opérations et fait appel à des analyses syntaxiques locales quand cela est nécessaire. Ce type de mécanisme est plus adapté aux caractéristiques de l'oral comme les phénomènes de distorsion et de fragmentation, ou tout simplement à la présence fréquente de questions, type de phrase qui apparaît très peu souvent dans les textes écrits et n'est ainsi pas bien pris en compte par les analyseurs conçus et entraînés sur l'écrit. Il peut de plus s'avérer plus robuste face aux éventuelles erreurs de la reconnaissance de la parole. C'est particulièrement le cas dans le DHM en domaine fermé : pour une tâche de réservation de train, on connaît quasiment à l'avance l'ensemble des éléments que peut comporter une requête de l'utilisateur, or ces éléments peuvent ainsi aider l'analyse, selon une approche descendante, et en complément de l'approche ascendante amorcée avec la reconnaissance de la parole.

Il existe ainsi une multitude de façons d'implémenter une analyse syntaxique, de même qu'il existe une multitude de formalismes de représentation. En fonction de l'énoncé de l'utilisateur, une implémentation peut s'avérer plus performante qu'une autre à certains moments, pour certains phénomènes. Si l'on dispose des moyens informatiques nécessaires, on peut envisager comme (Vilnat, 2005) de mettre en œuvre plusieurs algorithmes, c'est-à-dire de procéder à une « multi-analyse syntaxique ». Le principe est d'extraire une analyse fiable à partir des résultats de plusieurs analyseurs, chaque résultat pouvant s'accompagner d'un score de contentement, comme dans l'approche du carnet d'esquisses vue au paragraphe 4.1.2. Le constat à la base de cette approche a trait au succès des traitements qui combinent des sorties multiples, que ce soit au niveau de la combinaison d'hypothèses faites par le module de reconnaissance de la parole ou la combinaison d'hypothèses d'étiquetage morphosyntaxique (Vilnat, 2005, p. 20). Si des scores de contentement ne sont pas calculables, une stratégie consiste à privilégier les résultats produits le plus grand nombre de fois par l'ensemble des analyseurs mis en œuvre. Dans ce cas, il est utile de maximiser le nombre d'analyseurs, sachant que les contraintes du DHM en temps réel risquent de freiner cette approche : même si l'analyse syntaxique n'est pas la plus gourmande en mémoire et en temps de calcul, les précautions sont toujours de mise.

5.2.2. *Ressources sémantiques et conceptuelles*

La liste des mots de la langue avec leur catégorie et propriétés morphologiques constitue la ressource principale nécessaire à l'analyse syntaxique. D'autres ressources peuvent aussi être exploitées, comme des données statistiques sur les successions de mots et structures de phrases possibles. Pour aller plus loin et aborder le sens, d'autres données sont nécessaires. Il s'agit tout d'abord des sens des mots, ou du moins d'une représentation formelle, par exemple une structure de traits, qui range dans des cases des aspects tels que les types d'objets (inanimés concrets, abstraits, animés, individus humains), les types de propriétés (gradables, non gradables), d'événements (actions et procès, qui sont délimités dans le temps), et d'états (non délimités dans le temps). Les dictionnaires de langue ne sont malheureusement pas exploitables pour remplir de telles structures, car ils ne sont pas assez structurés et expriment les sens en langage naturel, ce qui rend le problème circulaire : il faut interpréter automatiquement les définitions d'un dictionnaire pour obtenir des ressources utiles à l'interprétation automatique.

Plusieurs approches sont exploitées, si possible de manière complémentaire, pour construire ce qui s'appelle un lexique sémantique. L'approche componentielle vise à spécifier pour chaque mot l'ensemble des traits sémantiques décrivant le sens du mot, et permet ainsi la constitution d'un lexique relativement proche dans son fonctionnement d'un dictionnaire. En DHM, c'est une approche tout à fait réaliste à partir du moment où l'on se place en domaine fermé, c'est-à-dire avec une liste de mots de taille limitée. L'approche des bases de données lexicales à la WordNet permet d'obtenir une liste organisée de mots, sans forcément d'équivalents de définitions (on constitue un thésaurus plutôt qu'un dictionnaire), mais avec de nombreuses relations orientées entre mots : hyponymie (« est une sorte de », par exemple de « TGV » à « train »), méronymie (« est une partie de », par exemple de « wagon » à « train »), troponymie (« X, c'est Y d'une certaine manière », une des sortes d'implication entre deux verbes), antonymie (contraire), synonymie, etc. Ces relations, notamment parce qu'elles sont orientées, permettent plus de possibilités d'analyse qu'une approche purement componentielle : on peut ainsi modéliser le fait qu'un trajet « long » est « non court », mais qu'un trajet « non long » n'est pas forcément « court ». Enfin, l'approche des lexiques dérivés, construits à partir de corpus ou de ressources telles que celles obtenues par les approches précédentes mais en suivant les principes d'une théorie lexicale particulière, permettent d'obtenir des structures de données plus riches. Un exemple célèbre et plusieurs fois exploité en DHM, dans le cadre d'une tâche délimitée, est celui des graphes conceptuels (Sowa, 1984). Il s'agit d'un formalisme qui permet d'aller plus loin dans la représentation des connaissances, avec des relations multiples entre concepts, et qui conduit à la notion d'« ontologie », en tant que modèle de données représentatif d'un ensemble de concepts dans un domaine, permettant de raisonner sur les objets relevant de ce domaine. Un système de DHM portant sur la réservation de train requiert non seulement des ressources sur la langue, par exemple un lexique sémantique, mais aussi des ressources sur le monde des trains, des transports et des réservations, avec

l'ontologie correspondante, que celle-ci soit représentée par un graphe conceptuel ou par un ensemble de structures de traits.

Par ailleurs, certains formalismes syntaxiques exploitent des données lexicales telles que celles constituant un lexique sémantique, ce qui conduit à mettre en œuvre des grammaires lexicalisées (Abeillé, 2007), et ainsi à mélanger lexique, syntaxe et sémantique lexicale. Un enjeu du DHM réside dans la constitution et l'exploitation de ces lexiques syntaxiques sémantiques, notamment pour le DHM en domaine ouvert. Nous citions plus haut l'initiative WordNet, et nous pouvons citer ici, en tant que l'une des pistes actuellement creusées pour la linguistique et le TAL en général, l'initiative FrameNet, qui consiste à répertorier des constructions de phrases, en les reliant les unes aux autres dès qu'elles décrivent un sens similaire. C'est en quelque sorte un lexique syntaxique sémantique orienté vers l'identification automatique des rôles actanciels, mais pas seulement, et c'est ce type d'initiative que le DHM peut exploiter.

5.2.3. *Analyses sémantiques*

Analyser sémantiquement un énoncé veut tout d'abord dire déterminer les sens de chacun des mots pleins utilisés, ce que permet un lexique sémantique, et veut aussi dire construire le sens de la phrase. Ceci peut se faire à partir de la structure identifiée par l'analyse syntaxique : on suit les relations entre composants, ce qui permet de construire les relations sémantiques. L'analyse exploite alors un lexique syntaxique sémantique, et se fonde par exemple sur les traits sémantiques et les fonctions grammaticales pour déterminer les rôles actanciels des différents composants de la phase. Car c'est là l'un des rôles fondamentaux d'une analyse sémantique, avec notamment les travaux de C. Fillmore qui met en avant cet aspect dans les grammaires de cas, suite au constat que la structure syntaxique n'est pas suffisante pour rendre compte des liens entre un verbe et ses actants, voir d'une manière générale (Enjalbert, 2005). Un autre rôle est, en confrontant les traits sémantiques des éléments en présence, de choisir parmi les sens possibles d'un mot, c'est-à-dire de résoudre les cas de polysémie et autres aspects de la langue mentionnés au paragraphe 5.1.1.

Pour le DHM en domaine fermé, l'analyse sémantique peut quasiment s'arrêter à cette étape, dans la mesure où la détermination de la sémantique verbale et l'identification des actants permet de deviner le contenu sémantique de l'énoncé, en tout cas dans les cas les plus simples tels que « je voudrais aller à Paris » ou « je réserve un aller ». Pour le DHM en domaine ouvert ou pour les systèmes qui cherchent une compréhension fine des phénomènes linguistiques, un autre rôle de l'analyse sémantique est d'exprimer le sens sous forme logique, afin de pouvoir lancer des calculs d'inférences sur cette forme logique et celles déjà enregistrées au fur et à mesure du dialogue, et de modéliser ainsi une partie de l'implicite. Une méthode consiste à suivre les principes des logiques mathématiques et à procéder à un calcul des prédicats, c'est-à-dire à la formalisation du contenu de l'énoncé avec des variables, des

relations, des prédicats, des connecteurs logiques (conjonction, disjonction, implication) et des quantificateurs (universel, comme dans « tous les trains comportent des premières classes », ou existentiel, comme dans « un train vient de tomber en panne près de la gare de Palaiseau »). L'enjeu devient alors de formaliser le langage naturel avec les contraintes de la logique, ce qui pose d'innombrables questions et problèmes. C'est ainsi que les logiques modales, temporelles, et les logiques hybrides peuvent s'appliquer au DHM.

Au-delà des frontières de l'énoncé, c'est aussi dans cette voie que des théories comme celle du changement de contexte (FCS, *File Change Semantics*) et la théorie de la représentation du discours (DRT, *Discourse Representation Theory*) vont proposer des cadres formels pour l'interprétation d'énoncés (Kadmon, 2001). La DRT notamment, qui décrit comment sont construites des structures de représentation du discours (Kamp et Reyle, 1993), fait l'objet de nombreuses extensions et adaptations pour des implémentations informatiques, par exemple DRL, *Discourse Representation Language*, voir (Kadmon, 2001). Pour le DHM, une extension importante est celle de la SDRT, *Segmented* DRT (Asher et Lascarides, 2003), qui passe en revue toutes les facettes du dialogue, et des implémentations ont été réalisées notamment dans certains SQR, ou encore dans le projet Verbmobil, voir (Cole, 1998). D'autres extensions prennent en compte les phénomènes de sous-spécification présents à l'oral, comme la CDRT, *Compositional* DRT (Muskens, 1996) ou la UDRT, *Underspecified* DRT. Par ailleurs, la DRT a même droit à une extension pour le dialogue multimodal, avec l'intégration des aspects linguistiques et des aspects visuels et gestuels dans un même cadre formel. Il s'agit de la MDRT, *Multimodal* DRT (Pineda et Garza, 2000).

Une voie totalement différente est le recours à des probabilités, de manière à obtenir une grammaire sémantique probabiliste. C'est ce qui est fait dans le système Tina (Seneff, 1995). D'autres approches exploitent les modèles de Markov cachés, en ajoutant une structure hiérarchique pour combiner les avantages d'une grammaire sémantique et ceux des statistiques. (Jurafsky et Martin, 2009, p. 859) présentent ainsi les modèles de compréhension cachée (HUM, *Hidden Understanding Model*).

Les analyses sémantiques peuvent donc se faire de multiples façons, et, comme toujours, peut combiner plusieurs approches. Deux remarques terminent cette section : premièrement, comme nous l'avons déjà souligné avec l'importance de l'analyse syntaxique locale ou partielle par rapport à une analyse syntaxique globale et complète, l'analyse sémantique peut elle aussi rester incomplète. Comme l'écrit (Enjalbert, 2005, p. 303), « il faut absolument abandonner l'idée de compréhension "complète", incompatible avec la variété extrême des problèmes à traiter. Les objectifs de compréhension doivent être ramenés à une tâche particulière, qui va orienter, limiter l'analyse, et fournir des informations complémentaires. D'ailleurs n'est-ce pas ainsi que fonctionne le lecteur humain, captant certaines informations dans un texte, en fonction de son intérêt, de ses objectifs de lecture – quitte à revenir sur une lecture plus complète ? ». Deuxièmement, déterminer le sens d'un énoncé n'est pas un problème

uniquement sémantique : c'est bien entendu aussi un problème pragmatique, avec la détermination des référents (les variables dans la forme logique) et l'identification des actes de langage. Cette dernière permet de comprendre les énoncés elliptiques et les énoncés non phrastiques. L'analyse pragmatique est ainsi utile à l'analyse sémantique, de même que l'analyse sémantique est l'un des paramètres permettant d'identifier les caractéristiques pragmatiques d'un énoncé. Analyse sémantique et analyse pragmatique vont ainsi de pair (Cole, 1998, p. 189).

5.3. L'enrichissement de la représentation du sens

Une fois que l'on dispose d'une représentation du contenu sémantique explicite de l'énoncé, on peut aller plus loin dans l'analyse et intégrer quelques aspects sémantico-pragmatiques relevant de la connotation, de la mise ou relief, ou des inférences telles que les « explicitations » de la théorie de la pertinence, c'est-à-dire les hypothèses communiquées explicitement par l'énoncé et qui indiquent par exemple l'attitude propositionnelle comme l'intention ou la croyance (Sperber et Wilson, 1995). On peut également enrichir la forme logique obtenue par l'analyse sémantique en intégrant un calcul des « implicitations », c'est-à-dire les hypothèses communiquées non explicitement telles que les sous-entendus. Ceux-ci peuvent être dérivés à partir de l'énoncé seul, qu'il soit linguistique ou multimodal, ou à partir du contexte. Dans cette section, nous nous limitons à ce qui peut être dérivé de l'énoncé seul.

5.3.1. *Au niveau de l'énoncé linguistique*

L'enrichissement des analyses décrites dans les sections précédentes consiste en l'ajout d'indications et de contenus propositionnels à la représentation sémantique obtenue. Selon les théories, ces indications et contenus peuvent être très variables, et nous retiendrons ici les suivants : connotations, modalités, ironie, saillance, focus, présuppositions et sous-entendus.

L'ajout de « connotations », c'est-à-dire d'éléments de sens qui s'ajoutent au sens littéral, se fait en suivant les liens d'hyponymie, de synonymie, etc., et en identifiant des traits sémantiques pertinents liés aux éléments de l'énoncé, dans les cas où de tels traits apportent des paramètres utiles à la compréhension automatique. Pour le DHM, l'enjeu consiste à ne pas générer automatiquement trop de connotations, et pas n'importe quand, mais à produire des informations qui permettent de combler des trous dans le processus de compréhension, par exemple pour interpréter correctement les métaphores et les comparaisons.

Les « modalités » sont des façons de modifier le contenu sémantique en exprimant l'attitude de l'utilisateur par rapport au contenu de son énoncé. Le terme n'a donc rien à voir avec celui lié à la multimodalité de la communication. Selon le type d'attitude,

on parle de modalité épistémique (degré de croyance de ce qui est dit, comme dans « ce chemin semble être le plus court »), de modalité aléthique (vérité ou possibilité de réalisation de ce qui est dit), de modalité déontique (obligation, interdiction, permission), de modalité intersubjective (conseil, reproche), de modalité appréciative, etc. En DHM, la présence de verbes comme « sembler », « vouloir » ou « pouvoir », ainsi que d'adverbes comme « peut-être », « probablement » ou « apparemment », est un indice essentiel à exploiter pour identifier automatiquement une modalité et l'utiliser pour enrichir la représentation sémantique.

L'ironie est l'exemple typique où le contenu sémantique de l'énoncé ne correspond pas du tout au contenu véhiculé. C'est un phénomène qui se rapproche de la litote, et qui peut être difficile à détecter, même par un interlocuteur humain. Comme toujours en DHM, il est souhaitable d'arriver à détecter ce type de comportement de la part de l'utilisateur, de manière à essayer de ramener le dialogue dans une voie plus neutre, mais on peut considérer que ce n'est pas forcément une priorité, surtout en domaine fermé.

La prise en compte de la « saillance », le profilage selon l'approche de la linguistique cognitive (Langacker, 1987, p. 39), permet de faire une hiérarchisation entre les différents éléments de l'énoncé, par exemple le nombre de billets, la destination et le type de place dans « je veux un billet pour Paris en première classe » selon l'accentuation, mais aussi entre des formes de phrase légèrement différentes, qui conduisent toutes au même contenu sémantique : « c'est un billet que je veux pour Paris en première classe », « c'est pour Paris que je veux un billet en première classe », etc. Dans ces énoncés, le présentatif « c'est … que » a pour effet de mettre un élément en saillance, et cette mise en saillance doit faire partie de la description du sens de l'énoncé. L'enjeu pour le DHM est de mettre en œuvre un modèle décrivant l'ensemble des facteurs de saillance, et à calculer pour chaque énoncé les saillances relatives de chacun des éléments, de manière à les hiérarchiser et à tenir compte de cette hiérarchisation lors de la détermination de la réaction du système.

Dans un même ordre d'idée, les approches fondées sur le focus, c'est-à-dire sur la saillance avec un point de vue avant tout prosodique, point de vue qui traduit l'importance de ce facteur dans la langue anglaise, explorent les facteurs permettant d'identifier le focus dans un énoncé, ainsi que les mécanismes par lesquels l'effet de focus s'étend d'un mot à un segment linguistique plus large (problème de l'association avec le focus). Certaines approches comme celle de (Beaver et Clark, 2008) considèrent qu'un élément focalisé soulève un ensemble d'alternatives qui sont calculées de manière compositionnelle en tant que sens focalisé ou alternatif, c'est-à-dire en tant qu'ensemble de contenus propositionnels. L'enjeu pour le DHM est d'identifier automatiquement ces contenus alternatifs, qui s'ajoutent au contenu sémantique issu des analyses plus classiques.

Pour continuer dans cette voie consistant à déterminer des contenus propositionnels supplémentaires, un aspect essentiel de l'interface entre sémantique et pragmatique consiste dans l'identification des « présuppositions ». Il s'agit de contenus implicites, qui se déduisent du contenu de l'énoncé, en tant que suppositions préalables : « bonjour, je voudrais changer une réservation pour un départ à huit heures » présuppose que le locuteur a déjà effectué une réservation, et que cette réservation concernait un départ à un horaire différent. De manière plus floue, on peut aussi supposer que le changement ne concerne que l'horaire, et pas la gare de départ ni celle de destination. Dans tous les cas, un système de DHM a tout intérêt à identifier ces contenus implicites, pour être à même de réagir avec pertinence : retrouver la réservation précédente (et, s'il faut un nom ou un numéro pour cela, demander celui-ci à l'utilisateur), vérifier les paramètres et éventuellement demander confirmation à l'utilisateur que seul l'horaire est bien à modifier.

Enfin, l'identification des sous-entendus repose cette fois sur une interprétation fine d'indices linguistiques ou sur des informations contextuelles plus larges que celles inclues dans l'énoncé. Dans « seul le train de sept heures s'arrête à Valence », un sous-entendu possible peut ainsi, du fait de l'usage de « seul », être « on pouvait s'attendre à ce que d'autres trains s'arrêtent à Valence ». Plusieurs règles ou lois permettent de déterminer des sous-entendus : la loi d'informativité, qui s'applique dans cet exemple avec « seul », la loi d'exhaustivité, qui permet d'inférer « certains trains ne s'arrêtent pas à Valence » à partir de « certains trains s'arrêtent à Valence », la loi de litote, qui permet d'inférer « tous les trains pour Palaiseau sont peu agréables » à partir de « certains trains pour Palaiseau sont peu agréables », ou encore la loi de négation, d'inversion argumentative, etc. D'une manière générale, les sous-entendus peuvent englober des phénomènes très disparates, à l'image de « je crois que je vais prendre l'avion », qui, énoncé après un long dialogue cherchant à réserver des billets de train, peut sous-entendre « aucune de vos propositions ne me convient », « vous n'êtes pas efficace », ou encore « j'en ai marre, j'arrête de parler avec vous ». Pour le DHM, l'enjeu est déjà d'identifier les sous-entendus les plus proches du contenu sémantique, de manière à tenir compte de phénomènes linguistiques comme ceux provoqués par l'emploi de « seul » ou de « certains ». Pour le DHM en domaine fermé, les sous-entendus plus flous peuvent se déduire de paramètres liés à la tâche, par exemple de la difficulté à satisfaire l'utilisateur compte tenu du nombre de propositions refusées par lui.

5.3.2. *Au niveau de l'énoncé multimodal*

En dialogue multimodal, la représentation sémantique de l'énoncé oral se confronte avec d'autres aspects sémantiques, notamment ceux portés par les gestes, les mimiques, et l'attitude générale de l'utilisateur, telle qu'elle est capturée et identifiée par les modules de détection des émotions et de détection des gestes conversationnels.

A titre d'exemples, un geste de découragement, une mimique faciale proche d'une grimace de dégoût, ou encore un ensemble d'indices vocaux et gestuels traduisant un énervement manifeste de l'utilisateur (rapidité des mouvements, du débit de parole, de son intensité) sont autant d'indices sémantiques qui peuvent apporter un éclairage important sur le contenu sémantique de l'énoncé oral. Deux cas principaux se présentent : soit les indications apportées par les autres modalités sont compatibles avec l'énoncé oral, par exemple quand les marques d'énervement vont de pair avec « ça ne va pas, je veux un aller pour Paris et pas pour Lyon », et dans ce cas ces indications permettent de préciser les états mentaux de l'utilisateur (ce qui aide à la détermination de la réaction du système), soit ces indications ne sont pas compatibles avec l'énoncé oral, et on est alors probablement en présence d'un comportement ironique, d'une litote, ou en tout cas d'un fort sous-entendu qu'il s'agit de décrypter, et de clarifier en posant par exemple une question explicite à l'utilisateur.

Cette confrontation de contenus sémantiques issus de modalités différentes s'appelle « fusion d'informations multimodales », ou fusion multimodale, et c'est un aspect sur lequel nous allons revenir dans le chapitre 7 pour la confrontation des actes de dialogue, et tout de suite dans le chapitre 6 pour la résolution des références.

5.4. Bilan

En dialogue oral spontané, un énoncé de l'utilisateur en entrée du système se caractérise par ses propriétés prosodiques, lexicales, syntaxiques et sémantiques. Toutes ces caractéristiques linguistiques posent des problèmes d'identification et de traitement automatique, et amènent à mettre en œuvre des techniques dédiées, en fonction de l'éventail des phénomènes retenus. Ce cinquième chapitre fait le point sur le traitement des entrées et montre comment aboutir à des représentations internes au système opérationnelles. L'accent est mis sur la reconstruction du sens explicite et implicite de l'énoncé, de manière à ce que le système raisonne sur une représentation sémantique enrichie qui soit la plus proche possible de celle correspondant à l'intention de l'utilisateur.

Chapitre 6

La résolution des références

Ce processus, de même que ceux que nous étudierons dans les prochains chapitres, requiert des arguments en entrée, qui peuvent faire l'objet de prétraitements, et retourne un résultat en sortie, résultat qui nécessite des post-traitements. On est ici en présence des signaux capturés en entrée, de préférence après une première étape d'interprétation qui consiste en une ou plusieurs transcriptions écrites de l'énoncé oral (plusieurs dans les cas d'ambiguïté), avec des indications prosodiques, et des représentations simplifiées pour les gestes. Pour que la multimodalité soit correctement gérée, ces transcriptions et représentations s'accompagnent de repères temporels, par exemple les dates de début et de fin de chaque mot prononcé, de chaque accentuation prosodique et de chaque trajectoire gestuelle. Ces aspects constituent un premier ensemble d'arguments pour la résolution des références. Un deuxième ensemble regroupe les résultats des analyses lexicales, syntaxiques et sémantiques telles que présentés dans le chapitre précédent. Enfin, un troisième ensemble d'arguments consiste en l'historique du dialogue, au cas où une référence ferait appel à une référence préalable, ce qui est le cas des anaphores pronominales et associatives. Comme pour l'énoncé en cours de traitement, l'historique regroupe deux types de représentations : les représentations issues des analyses linguistiques, et les représentations sous forme de liste chaînées de mots avec repères temporels et prosodiques. Les premières sont essentielles pour interpréter « le » dans « mets-le dans la boîte » qui fait suite à « prends un cube vert », et les secondes pour interpréter « ce que j'ai pris pour une pyramide » ou « ce que j'ai appelé forme bizarre », autrement dit dans les cas où l'expression référentielle contient des propriétés qui ne sont pas celles du référent et qui ne sont donc pas accessibles à partir de celui-ci.

Le résultat de la résolution de la référence est une mise à jour du résultat des analyses linguistiques, dans lequel les variables encore libres sont désormais affectées à des référents, de préférence les bons, c'est-à-dire ceux qui correspondent à l'intention de l'utilisateur. En cas d'ambiguïté, plusieurs représentations alternatives sont produites. En cas d'impossibilité à affecter un référent, une représentation sous-spécifiée est produite et transmise aux modules chargés des analyses pragmatiques. Ainsi, dans l'exemple de l'introduction, l'énoncé U2 présente une référence, « ce chemin qui semble être le plus court », qui est accompagnée d'un geste de pointage, et qui réfère ainsi de manière multimodale à l'un des chemins affichés à l'écran. Les repères temporels du côté de l'énoncé oral et ceux du côté du geste permettent de vérifier s'il y a bien synchronisation temporelle au moment (approximatif) de l'acte de référence. La prosodie peut apporter également un argument dans ce sens, par exemple si « ce » est légèrement accentué du fait de la production simultanée d'un geste. Le geste apporte l'identité du référent, l'énoncé oral apporte la représentation sémantique qui en a été faite et dans laquelle l'expression référentielle est restée sous la forme de variable. Le processus de fusion des informations multimodales permet alors de résoudre la référence, c'est-à-dire d'affecter l'identifiant du référent à l'expression référentielle.

C'est ce processus que nous allons détailler dans ce chapitre, avec premièrement les références à des objets comme c'est le cas pour des trajets de train (section 6.1), deuxièmement les références à des actions (section 6.2), et troisièmement le cas particulier des références qui vont chercher un référent dans l'historique du dialogue, avec les phénomènes d'anaphores et de coréférences (section 6.3).

6.1. Résolution des références à des objets

La référence a fait l'objet de très nombreux travaux, issus de la philosophie du langage, de la logique et de la linguistique (Abbott, 2010), travaux qui ont mis en avant la notion de référence, les catégories d'expressions référentielles (démonstrative pour « ce chemin »), et tout un ensemble de phénomènes caractéristiques de la langue, comme le mode de présentation d'un référent ou la distinction entre usage attributif d'une expression (« le train pour Palaiseau, quel qu'il soit, s'arrête toujours à Villebon ») et usage référentiel (« le train pour Palaiseau a huit minutes de retard »), qui implique un référent précis (Charolles, 2002). En DHM, le problème revient toujours à construire un lien entre une forme linguistique et un élément de la base de données gérée par la tâche, qu'il s'agisse d'un train en particulier ou d'un type de train, voire de la classe générique de tous les trains, telle qu'elle est définie dans le modèle conceptuel. Il arrive également que l'expression référentielle, par exemple « un cube » dans « prends un cube et mets-le sur la pyramide de gauche », laisse le système choisir entre plusieurs possibilités, l'utilisateur imposant que le référent appartienne à un ensemble bien précis et concret, mais sans choisir parmi les alternatives : n'importe laquelle conviendra.

D'une manière générale, résoudre les références à des objets se fait en exploitant les propriétés mentionnées et la détermination de l'expression référentielle. « La pyramide verte » indique ainsi trois critères de recherche dans l'ensemble des objets accessibles au moment de l'énonciation : l'objet doit être unique, il doit avoir une forme de pyramide (catégorie) et une couleur verte (modifieur). Si la base de données du système comprend plusieurs objets qui vérifient ces propriétés de forme et de couleur, on est face à une ambiguïté. Si elle n'en comprend aucun, résoudre la référence ne peut pas se faire. Si elle comprend un seul objet qui vérifie les deux propriétés, alors cet objet est considéré comme le référent et le processus de compréhension peut passer à l'étape suivante.

Il se peut cependant que les termes utilisés ne correspondent pas tout à fait aux propriétés telles qu'elles apparaissent dans la base de données, ce qui pourrait expliquer les cas où aucun référent n'est identifié. Parcourir le modèle conceptuel des propriétés est alors nécessaire, de manière à vérifier si un objet d'une forme proche ou d'une couleur proche existe, objet qui pourrait être un candidat pertinent. C'est le genre de situation qui peut arriver en DHM, l'utilisateur n'ayant pas forcément le vocabulaire exact ayant servi à construire le modèle conceptuel et la base de données des objets de l'application. C'est aussi le genre de situation qui arrive en dialogue humain, et qui montre d'ailleurs tout l'intérêt d'un dialogue : les interlocuteurs peuvent communiquer pour se mettre d'accord sur les termes les plus adéquats compte tenu du référent. Cette convergence vers un terme commun a été étudiée notamment dans (Brennan et Clark, 1996), avec la notion d'alignement lexical : les interlocuteurs alignent leur comportement linguistique, notamment au niveau des termes utilisés. Un système de DHM peut donc faire la même chose, ce qui veut dire employer le même terme que l'utilisateur, et proposer un autre terme quand il ne comprend pas celui-ci.

En dehors des propriétés qui apparaissent sous la forme de mots pleins, la référence fait intervenir une multitude de formulations, par exemple la relative de « ce chemin qui apparaît à gauche de l'écran ». L'enjeu en DHM est de comprendre le sens de cette relative restrictive, pour en déduire un critère de recherche dans la base des objets, en l'occurrence une propriété de disposition spatiale qui peut se calculer avec les coordonnées de l'objet tel qu'il apparaît sur l'écran. En revanche, dans « ce chemin qui semble être le plus court », la relative n'est pas restrictive et n'a en fait rien à voir avec l'identification du référent, pour laquelle l'expression « ce chemin » est nécessaire et suffisante. Enfin, notons l'importance toute particulière du déterminant : avec « la pyramide verte », le déterminant défini au singulier se traduit par un critère de recherche bien précis : trouver l'unique pyramide verte dans un ensemble d'objets qui *a priori* comporte des pyramides et des objets d'autres formes, ainsi que des objets verts et des objets d'autres couleurs. Le démonstratif de « ce chemin » a un tout autre rôle : il signale que le référent est saillant dans la situation de communication, soit parce qu'il vient d'être mentionné, auquel cas on est en présence d'une anaphore comme dans « le chemin qui va de Paris à Meudon semble être le plus court » suivi de « je choisis ce chemin », soit parce qu'il est désigné simultanément par un geste, ce

qui est le cas dans l'énoncé U2. C'est cette analyse fine des déterminants qui va nous permettre d'implémenter une résolution performante des références aux objets.

6.1.1. *Le modèle des domaines de référence multimodaux*

Suite à une intuition de (Corblin, 1995) et à un ensemble de travaux effectués à Nancy (Landragin, 2004, p. 107), la notion de domaine de référence a montré son intérêt pour la résolution de la référence, dans un contexte linguistique comme multimodal (Landragin, 2006). L'idée est que l'identification des référents passe systématiquement par l'identification d'un sous-ensemble contextuel auquel ils appartiennent. Ce sous-ensemble, qui ne s'étend pas à l'intégralité du contexte mais correspond par exemple à un espace attentionnel, est appelé domaine de référence. Il permet de justifier l'emploi du défini singulier, comme dans « la pyramide verte », même dans les cas où le contexte comprend plusieurs pyramides : s'il existe un espace attentionnel préalablement délimité au cours du dialogue, et que cet espace attentionnel comprend une seule pyramide verte, alors il y a des chances que le défini singulier ne relève pas d'une erreur de l'utilisateur mais bien d'une interprétation localisée dans le domaine de référence qu'est l'espace attentionnel en question.

Par rapport à la théorie de la représentation du discours et à son extension qu'est la MDRT pour le dialogue multimodal (Pineda et Garza, 2000), le modèle des domaines de référence multimodaux procède à un traitement plus fin de la focalisation à un sous-ensemble contextuel. Par rapport à l'approche des domaines de quantification (suite aux travaux de R. Montague), les domaines de référence durent sur plusieurs énoncés et tiennent compte des mécanismes de restriction et d'élargissement contextuels au fur et à mesure que le dialogue progresse d'un point de vue référentiel. Nous en parlons ici pour la résolution de la référence, mais ils sont aussi utilisés pour la génération automatique d'expressions référentielles (Denis, 2011). Ils ont par ailleurs été appliqués à des phénomènes plus larges que celui de la référence, par exemple la gestion du dialogue (Grisvard, 2000), en lien avec la théorie des représentations mentales qui en est aussi à l'origine, voir chapitre 6 de (Reboul et Moeschler, 1998). Parmi les travaux proches se trouvent ceux de (Beun et Cremers, 1998) sur les espaces focaux, ceux de (Wright, 1990) sur les domaines référentiels, ou encore, pour une extension du même principe au discours, ceux de (Luperfoy, 1992) sur les chevilles du discours.

Un exemple typique d'utilisation des domaines de référence autour de l'exemple de l'introduction est mis en avant en remplaçant U4 par « et combien de temps avec l'autre chemin ? » : « l'autre chemin » ne s'interprète correctement que dans un domaine de référence qui comprend deux chemins, avec l'un des deux déjà focalisé. Or « voici les chemins possibles » avait eu comme effet de construire un domaine de référence comportant deux chemins, et « ce chemin » avait focalisé l'un des deux. Le deuxième chemin est donc parfaitement accessible. De même, en remplaçant U4 par « je voudrais voir les trajets directs » : « les trajets directs » ne s'interprète pas dans

l'ensemble de tous les trajets imaginables, mais dans le domaine de référence défini par l'énoncé précédent, « voici les trajets possibles ». Dans ce domaine de référence qui ne comprend que des trajets vers Paris, le modifieur « direct » permet d'extraire l'ensemble des trajets directs vers Paris, ce qui correspond bien au référent voulu par l'utilisateur. Par ailleurs, si l'on remplace U4 par « je voudrais voir les trajets pour Marseille », le domaine de référence actif, qui comprend les trajets pour Paris et donc aucun trajet pour Marseille, conduit à l'identification d'aucun référent, et donc à une réponse négative de la part de l'utilisateur, telle que « il n'y en a pas », ou, en exploitant le fait que le domaine de référence a été construit par rapport à la destination de Paris, « il n'y en a pas, il n'y a que des allers pour Paris ». Cette réponse privilégie l'interprétation courante dans le domaine de référence actif, ce qui est l'un des fonctionnements de ce modèle. Un autre système aurait pu effacer les trajets pour Paris et afficher ceux pour Marseille en énonçant « voici les trajets possibles » comme en S2, mais cela revient à ignorer le dialogue tel qu'il s'est déroulé jusqu'à présent, et à reprendre la requête à zéro, ce qui ne va pas dans le sens d'un dialogue naturel. Si le système choisit de considérer cet énoncé U4 comme une nouvelle requête avec la construction d'un nouveau domaine de référence, alors il y a une sorte de rupture référentielle dans le dialogue, et cette rupture peut faire l'objet d'une demande de confirmation, telle que « pour Marseille ? », ou d'une matérialisation de la rupture, telle que « on abandonne Paris pour Marseille, donc », énoncée au moment même où les trajets pour Paris disparaissent de l'écran au profit de ceux pour Marseille.

6.1.2. *Analyse de la scène visuelle*

Une première source pour la détermination d'espaces focaux ou de sous-ensembles contextuels qui fassent l'objet de domaines de référence est la perception visuelle. Nous l'avons vu au paragraphe 2.1.1, le système de DHM connaît la nature et l'emplacement spatial de tous les objets affichés sur la scène visuelle, donc, dans notre exemple, des trajets de train qui ont été mis en valeur graphiquement, et, dans le cas d'une tâche comme celle de Shrdlu, des formes géométriques qui constituent le micromonde physique de la tâche. Dans ce contexte, l'utilisateur peut se focaliser sur un sous-ensemble, par exemple celui des formes placées à gauche, ou celui de toutes les formes creuses. Ce sous-ensemble visuel se détermine à l'aide de critères tels que ceux avancés par la théorie de la Gestalt : proximité spatiale entre objets, similarité, continuité, etc., voir une formalisation hiérarchique dans (Landragin, 2004) et (Landragin, 2006). Il ne devient un domaine de référence qu'à partir du moment où l'utilisateur exprime une référence à un groupe perceptif (le groupe de formes à gauche) ou à un élément isolé spatialement ou de par ses propriétés intrinsèques comme la taille et la couleur. Ce domaine de référence permet alors de rendre compte de phénomènes de focalisation attentionnelle, en autorisant par exemple l'interprétation de « le cube vert », non plus comme l'unique objet de la scène visuelle vérifiant la propriété d'être de forme cubique et celle d'être de couleur verte, mais comme l'unique objet du domaine de référence visuel ayant ces propriétés. Dans le cas où la scène comporte un

autre cube vert, non placé dans l'espace attentionnel, ce mécanisme permet d'éviter une réaction du système telle que « je ne comprends pas de quel cube vert il s'agit », mais, au contraire, de doter celui-ci de capacités à modéliser l'attention et à résoudre les références de manière pertinente.

Un phénomène important pouvant intervenir dans ce cadre est celui de saillance visuelle, qui permet de rendre compte de l'attraction de l'attention de l'utilisateur par un objet en particulier, quand celui-ci se distingue des autres objets visibles par des propriétés spécifiques : mise en avant spatiale, plus grande taille, couleur différente. En plus de capacités à détecter automatiquement les groupes perceptifs, un système de DHM reposant sur une scène visuelle affichée à l'écran a ainsi tout intérêt à détecter automatiquement les objets visuellement saillants, comme nous l'avons vu au paragraphe 2.1.1. Au niveau des domaines de référence, la saillance d'un objet ne contribue pas à construire un nouveau domaine potentiel, mais à focaliser l'un des éléments d'un domaine de référence construit selon les groupes perceptifs. C'est ainsi que le pronom exophorique, par exemple « il » sans aucun antécédent linguistique possible, peut être interprété comme référant à l'objet le plus saillant dans le domaine de référence courant. Ici encore, le modèle des domaines de référence multimodaux permet ce niveau approfondi de compréhension de la part du système.

6.1.3. *Analyse des gestes de désignation*

Dans le cadre d'une interaction multimodale avec écran tactile, l'utilisateur peut effectuer des gestes, par exemple des pointages ou des entourages, de manière à référer aux objets affichés. Si la trajectoire gestuelle se superpose parfaitement aux objets visés, le système de DHM peut résoudre la référence sans trop de difficultés. Si la trajectoire est approximative, et par exemple passe à côté ou recouvre involontairement un objet qui ne fait pas partie des référents intentionnels, alors le système doit faire face à des cas d'indécision. C'est là que la notion de domaine de référence peut apporter un éclairage utile.

D'une manière générale, la capture d'un geste peut mener à la détection d'une ambiguïté sur l'intention à l'origine du geste : un même geste, avec la même forme ou la même trajectoire, peut découler de plusieurs intentions. Un mouvement de la main capté par une caméra peut par exemple correspondre à un geste paraverbal qui appuie un mot en particulier mais ne réfère pas, ou peut désigner un objet précis et réaliser ainsi une référence. La présence d'une expression référentielle dans l'énoncé linguistique, ainsi que des techniques d'apprentissage appliquées à la reconnaissance des gestes paraverbaux, permettent de lever ce type d'ambiguïté. Une fois que le système est sûr que le geste effectué est un geste déictique, l'analyse de la trajectoire gestuelle peut elle-même conduire à la détection d'une ambiguïté. Dans le cadre d'une interaction sur écran tactile, un exemple consiste en un geste qui entoure trois objets, mais

qui déborde également sur un quatrième et se termine à portée immédiate d'un cinquième. Sur la base d'une analyse structurelle de cette trajectoire d'entourage, c'est-à-dire d'une détection des aspects remarquables de la trajectoire tels que les points d'inflexion, les croisements, les zones à courbure constante ou encore les zones de fermeture (Bellalem et Romary, 1996), d'une analyse de la scène visuelle en termes de groupes perceptifs (paragraphe 6.1.2), et éventuellement de calculs d'indices géométriques comme des taux de recouvrement ou des distances relatives, le système peut alors écarter le quatrième objet, par exemple parce qu'il ne fait pas partie du même groupe perceptif que les autres objets concernés, et parce que la trajectoire gestuelle présente un léger mouvement d'évitement au moment où elle déborde sur ce quatrième objet. En revanche, il peut décider de garder le cinquième objet en tant que candidat potentiel, dans la mesure où ce cinquième objet fait partie du même groupe perceptif que les trois objets clairement entourés. Si l'historique du dialogue comporte un domaine de référence qui sépare cet objet des trois entourés, la décision aura été l'inverse. Quoi qu'il en soit, nous voici maintenant avec deux hypothèses : une portant sur trois objets, l'autre sur quatre. C'est ce qui constitue une préanalyse du geste en contexte visuel, et c'est cette préanalyse qui va se confronter avec l'analyse sémantique de l'énoncé oral simultané : en gros, soit l'expression référentielle indique un nombre (« ces trois objets », « ces quatre formes », « cet objet, cet objet et cet objet ») et l'ambiguïté est levée, soit ce n'est pas le cas. L'ambiguïté est alors confirmée et le système doit décider entre choisir l'une des alternatives ou poser une question à l'utilisateur (Landragin, 2006).

D'autres ambiguïtés et analyses sont envisageables. Dans le cadre d'un DHM s'appuyant sur une IHM, tout geste est ainsi *a priori* ambigu entre un geste conversationnel, à destination du système de dialogue, et un geste de manipulation directe, à destination de l'IHM. Au niveau des analyses, d'autres approches consistent par exemple à ce que le module en charge des gestes ne propose pas d'hypothèses en cas d'ambiguïté (Martin *et al.*, 2006), ce qui peut conduire le gestionnaire de dialogue, dans le cas où le module linguistique n'arrive pas à trouver le référent par lui-même, à décider d'une réaction sans disposer d'hypothèses (et donc à poser une question sur l'identité du référent). L'approche de (Kopp *et al.*, 2008) consiste à mettre en œuvre, pour la génération automatique de geste mais l'idée est applicable en compréhension automatique, un formulateur gestuel qui raisonne sur la base d'un ensemble de traits avec des valeurs traduisant la localisation, le sens de la trajectoire, la direction de chacun des doigts (quand la configuration de la main est détectée par une caméra ou un gant de désignation), la direction de la paume, ou encore la forme générale de la main. D'une manière générale, les processus à mettre en œuvre sont tributaires des modalités captées en entrée : là où la détection par caméra nécessite de nombreux paramètres afin par exemple de reconstruire le sens d'un geste iconique (ou d'un geste en langue des signes), l'utilisation d'un écran tactile réduit toute l'interaction gestuelle à la capture d'une simple trajectoire, comme dans l'interaction classique avec la souris.

6.1.4. *Résolution de la référence en fonction de la détermination*

L'analyse du contexte visuel et celle des gestes se fait en parallèle des analyses linguistiques. Comme nous l'avons vu dans le chapitre précédent, celles-ci collaborent entre elles pour aboutir à une représentation formelle du sens de l'énoncé, qui tient compte des caractéristiques prosodiques, lexicales, syntaxiques et sémantiques de celui-ci. Dans « combien de temps avec ce chemin qui semble être le plus court ? », la prosodie indique par exemple une légère accentuation du démonstratif de l'expression référentielle « ce chemin », la syntaxe et la sémantique concluent avec la prosodie que « qui semble être le plus court » n'est pas une relative restrictive qui pourrait servir à identifier le référent, et le lexique sémantique permet de faire un lien entre « chemin » et le concept de voyage décrit dans le modèle conceptuel de l'application, qui s'avère bien compatible avec le fait de durer un certain temps, sujet de la question principale de l'énoncé. L'expression référentielle « ce chemin » n'est cependant pas complètement analysée. Notamment, aucun référent ne lui est encore affecté. Pour cela, le modèle des domaines de référence multimodaux procède à la détermination d'un domaine de référence sous-spécifié qui traduit les contraintes linguistiques portées par l'expression référentielle. Ces contraintes sont tout d'abord celles des mots utilisés, donc de la catégorie et des modifieurs en tant que filtres pour chercher le référent parmi les objets accessibles. Dans certains cas comme dans « colorie la pyramide en rouge » ou « supprime ce fichier », la sémantique du verbe et la sémantique de la phrase apportent des filtres supplémentaires : le fait de ne pas être rouge pour « la pyramide » et le fait d'être supprimable pour « ce fichier ». Une autre contrainte est celle portée par la détermination. Selon le fonctionnement du démonstratif, du défini et de l'indéfini, les critères de recherche du référent vont être différents. Ainsi, le démonstratif « ce N » impose la focalisation du référent, soit de manière préalable par une mention au même référent, soit de manière simultanée par un geste de désignation. De son côté, le défini « le N » fonctionne par extraction du seul N dans un domaine de référence. Comme l'écrit (Corblin, 1995, p. 51), « le N » consiste toujours à opposer, pour en prédiquer quelque chose, un N précédemment mentionné aux autres entités. Il oppose nécessairement l'élément qui est un N dans le domaine de référence, aux éléments qui ne sont pas des N. Enfin, l'indéfini fonctionne en sélectionnant un élément quelconque d'un ensemble. Ces trois cas sont loin de couvrir l'ensemble des expressions référentielles que la langue permet (les pluriels posent leurs propres mécanismes, de même que les pronoms personnels ou les noms propres), mais ils illustrent les trois mécanismes principaux impliqués par la résolution de la référence : exploitation d'une mise en relief, extraction, sélection.

A ce stade, le système de DHM dispose donc d'un domaine de référence sous-spécifié portant les contraintes linguistiques de l'expression référentielle, et c'est ce domaine sous-spécifié qu'il va tenter d'apparier avec les domaines de référence apportés par le contexte visuel, par l'analyse du geste si un geste est effectué, et par l'historique du dialogue. Un des rôles de celui-ci consiste en effet à sauvegarder les domaines de référence successifs, de manière à rendre compte des phénomènes d'élargissement

contextuel, de restriction contextuelle, d'anaphore, et aussi d'altérité, avec les expressions en « autre » telles que « les autres trains » ou « l'autre pyramide ». Selon la tâche, l'appariement peut se tester dans un ordre précis, en privilégiant par exemple la perception visuelle à l'historique du dialogue et en s'arrêtant dès qu'un résultat complet est obtenu, ou peut consister à expliciter toutes les possibilités, de manière à laisser le gestionnaire de dialogue décider quelle alternative choisir en cas d'ambiguïté. Avec notre exemple U2 impliquant l'expression référentielle « ce chemin » et un geste de désignation vers l'un des chemins affichés à l'écran, nous sommes dans le cas de figure le plus simple qui soit : le domaine de référence sous-spécifié impose une focalisation existante dans un domaine qui regroupe les différents trajets pour aller à Paris, le geste apporte une hypothèse de trajet désigné, et l'appariement conduit à considérer que la focalisation porte sur cette hypothèse, et donc à la résolution de la référence. Dans d'autres cas plus complexes, il peut être nécessaire de déterminer le type d'accès aux référents, avec une analyse fine des combinaisons des types d'accès et des types de déterminants (Landragin, 2006). Dans tous les cas, un formalisme tel que les structures de traits permet d'implémenter un tel modèle, l'appariement s'opérant par l'opération d'unification. L'enjeu pour le DHM réside donc surtout dans la détermination de tous les types de référence, afin d'écrire le module chargé de déduire des formes linguistiques les contraintes formalisées dans les domaines de référence, contraintes qui vont orienter l'unification des structures de traits.

Par ailleurs, les énoncés qui comportent plus d'une référence peuvent poser des problèmes relatifs à la fusion multimodale. Dans un exemple tel que « ce trajet est-il plus long que ceux-ci et ceux-ci ? », trois expressions référentielles peuvent faire l'objet d'un geste de désignation, voire de plusieurs dans le cas de « ceux-ci ». Si le système reçoit cinq gestes de désignation, une analyse approfondie de la synchronisation temporelle et des possibilités de correspondances entre gestes et expressions s'avère nécessaire, afin de déterminer quels gestes s'associent avec quelles expressions. La seule contrainte due à l'usage naturel de la langue et du geste est que l'ordre de succession des gestes suit l'ordre de succession des expressions. Dans des cas extrêmes observés pour des tâches incitant à multiplier les références (Landragin, 2004, p. 45), la combinatoire peut devenir telle que des heuristiques sont nécessaires. Ces phénomènes nous conduisent en particulier à distinguer plusieurs niveaux de fusion multimodale. Là où beaucoup d'approches focalisées sur les signaux procèdent à un appariement des gestes et des expressions uniquement sur le paramètre de la synchronisation temporelle, c'est-à-dire en opérant une fusion multimodale physique, d'autres approches comme celle des domaines de référence mettent en avant un autre niveau de fusion multimodale : le niveau sémantique (Martin *et al.*, 2006 ; López-Cózar Delgado et Araki, 2005). Le chapitre 7 présentera un troisième niveau, pragmatique, relatif aux actes de dialogue.

La référence peut s'appliquer à bien d'autres entités que des objets concrets comme des pyramides ou des trajets. Dans l'exemple classique « mets ça ici » (Bolt, 1980),

une première référence multimodale concerne effectivement un objet concret, désigné linguistiquement par « ça » (c'est-à-dire l'expression référentielle la plus vague qui soit en français), mais concerne aussi un lieu, avec « ici ». Cette deuxième référence s'accompagne nécessairement d'un geste et constitue donc un cas de référence multimodale. La résolution de cette référence pose d'autres problèmes que ceux que nous avons vus jusqu'à présent. La nature du référent est indiquée par « ici », mais la détermination exacte du référent dépend de plusieurs paramètres : nature de l'action, ici un positionnement ; nature de l'objet à placer, notamment son gabarit (mettre un clou n'a rien à voir avec mettre de la moquette, selon l'exemple de L. Romary dans le cadre d'une tâche d'aménagement d'intérieur) ; nature des objets déjà présents dans le lieu indiqué (voir la section suivante pour ce qui concerne les paramètres liés à l'action). Par ailleurs, la référence peut s'appliquer à des objets abstraits, qui peuvent être des concepts connus de la tâche, par exemple « retard » ou « prix », comme dans « un retard serait inacceptable » ou « quel est le prix de ce trajet ? », ou qui peuvent être des états, des actions ou des procès, comme dans « être en retard serait inacceptable » ou « combien coûte ce trajet ? ». D'une manière générale, tout mot plein peut référer. Dans tous les cas, le procédé de résolution de la référence peut s'inspirer de celui détaillé pour les objets concrets, ou de celui que nous allons voir maintenant pour les actions.

6.2. Résolution des références à des actions

L'exemple « mets ça ici », outre deux références multimodales, comporte une référence à une action, portée par le verbe à l'impératif. Or, selon la tâche et les possibilités qu'elle offre, faire le lien entre ce mot « mets » et l'une des actions exécutables par l'application n'est pas forcément évident. Car c'est bien là le cœur du problème : étant donné un énoncé, à quelle fonction de l'application fait-il référence ? Avec l'exemple d'un logiciel de dessin à commande vocale, « mets ça ici » peut vouloir déclencher une action de déplacement d'un objet, action qui *a priori* est plutôt liée au verbe « déplacer » qu'au verbe « mettre », celui-ci pouvant par exemple déclencher la création d'un objet (« mets un cube ici »). Résoudre la référence fait ainsi intervenir la sémantique du verbe employé, sa valence (paragraphe 6.2.1), mais aussi les objets concernés et, d'une manière générale, la tâche en cours (paragraphe 6.2.2).

6.2.1. *Référence aux actions et sémantique verbale*

Le modèle de tâche comprend la liste des actions que l'application est capable d'exécuter. Résoudre la référence aux actions consiste donc à se ramener à l'un des éléments de cette liste. Pour de tels éléments de l'application, (Duermael, 1994) utilise le terme d'opérateur, et le définit comme un modèle d'action, constitué de préconditions, de post-conditions et d'un corps. Le corps correspond à une fonction de

l'application. Les préconditions, indispensables, permettent de s'assurer de l'applicabilité de cette fonction, en vérifiant par exemple que les objets concernés sont bien compatibles avec la fonction. Les post-conditions servent à la simulation de l'exécution de la fonction, juste avant de réaliser celle-ci : le but est de simuler les effets avec des représentations éphémères des objets et des connaissances, de manière à voir ce que ces objets et connaissances subissent, et ce qu'il en sort. Cette anticipation permet d'une part de détecter des problèmes peu prévisibles, par exemple collatéraux, et importants, comme la suppression d'un objet. Si le système estime cela pertinent, il peut alors prévenir l'utilisateur d'une telle conséquence et lui demander confirmation. L'anticipation permet d'autre part de mettre en œuvre une gestion dynamique des actions, avec l'implémentation d'une fonction d'annulation, ce qui n'est pas forcément simple dans les cas de suppression ou de modification importante d'objets.

Une fonction de l'application nécessite des paramètres. Elle s'applique souvent à des objets, selon des propriétés précises. Une fonction de déplacement d'un objet nécessite ainsi de connaître l'objet concerné et le lieu envisagé, en plus des préconditions sur le fait que l'objet doit être déplaçable et que le lieu de destination peut bien être occupé par lui. L'exécution de la fonction nécessite donc ces deux paramètres. Dans le cas le plus simple, l'énoncé de l'utilisateur comprend un verbe dont la sémantique est clairement liée à celle de l'opérateur intentionnel, et dont la valence correspond au nombre de paramètres requis. C'est le cas avec « déplace ça ici » : comme de plus les deux paramètres n'ont pas la même nature, la résolution de la référence à l'action se fait très simplement. Dans d'autres cas, par exemple dans « je déplace ça ici » ou dans « déplace ça », la résolution de la référence nécessite soit d'ignorer un paramètre qui en fait n'en est pas un mais seulement un moyen d'expression neutre vis-à-vis de l'exécution de la tâche, soit de détecter l'absence d'un paramètre, ce qui entraîne le système à poser une question sur celui-ci. C'est bien entendu le cas le plus fréquent dans les dialogues de réservation de billets de train (et c'est aussi dans ce but que les rôles actanciels sont identifiés lors de l'analyse sémantique), avec l'exemple U1 en tête : « je voudrais aller à Paris » n'indique ni la gare de départ, ni l'horaire de départ, qui sont des paramètres nécessaires. Le système peut considérer que l'utilisateur, avec cet énoncé, ne fait qu'amorcer une requête et que la complétion de celle-ci va faire l'objet du dialogue. Il peut alors « planifier » les demandes de précision, et commencer comme on l'a vu au paragraphe 3.2.1 par la gare de départ. Il peut aussi tenter de déduire les paramètres manquants en faisant appel à l'historique du dialogue (si une gare de départ a été mentionnée à un moment donné, c'est peut-être un candidat pertinent), au contexte situationnel (la gare de départ correspond à l'emplacement du terminal utilisé pour dialoguer) ou au bon sens (quelqu'un qui cherche à acheter un billet de train peut vouloir partir de suite, en tout cas c'est un choix que le système peut proposer avant même de poser des questions).

Comme toujours avec le langage, des ambiguïtés peuvent survenir et compliquer la résolution de la référence. Dans un système comme Shrdlu, c'est par exemple le cas quand on veut mettre un objet sur un autre ou emboîter deux objets : un énoncé qui

fait référence à une action nécessitant deux objets comme paramètres peut conduire à deux interprétations, la bonne et celle où les deux objets sont inversés. Pour décider, une analyse des prépositions employées et des rôles actanciels est déterminante. Plus compliqué, un énoncé comme « réserve-moi un train pour Paris tout de suite » peut entraîner une ambiguïté sur l'horaire de départ : est-ce le plus tôt possible (dès qu'un train part) ou le complément « tout de suite » concerne-t-il uniquement l'ordre de réservation ? Une phrase peut comprendre des compléments optionnels, au sens de la valence verbale, ainsi que des éléments intermédiaires, non prévisibles compte tenu du verbe utilisé mais seulement en parcourant ses hyperonymes, des éléments accessoires, non prévisibles compte tenu du verbe car correspondant à des circonstants, ou encore des éléments extrapériphériques comme des modifieurs logiques ou discursifs, par exemple « comme vous savez ». Une tâche pour le DHM est d'exploiter ces éléments linguistiques pour mieux gérer les conditions et paramètres d'exécution des fonctions de l'application. Au-delà des aspects linguistiques, il arrive aussi que la résolution de la référence aux actions fasse intervenir des aspects purement applicatifs. Si l'on considère par exemple que la tâche implémente une fonction de suppression d'objet, qui ne fonctionne qu'avec un seul paramètre, donc un objet unique, alors un énoncé tel que « supprime ces objets » aboutit soit à un message d'erreur de la part du système (« il me faut un seul objet. Quel objet dois-je supprimer ? »), soit à la mise en œuvre d'une succession d'exécutions de la fonction de suppression. Cette dernière solution n'est pas forcément pertinente, par exemple si la suppression d'un objet entraîne des conséquences sur les autres objets.

Pour résoudre de tels exemples, la résolution de la référence peut faire intervenir un modèle temporel. Le but est de prendre en compte les contraintes temporelles intervenant dans l'exécution d'une fonction, ce qui permet de modéliser de manière fine des actions séquentielles, ainsi que les interactions entre l'exécution des actions et l'évolution parallèle du monde des objets. Avec l'exemple de la réservation d'un billet de train, un tel modèle temporel est plus que nécessaire à partir du moment où plusieurs terminaux permettent à plusieurs utilisateurs de réserver des billets pour les mêmes trains (Kolski, 2010). L'intérêt d'un modèle temporel réside aussi dans une meilleure exploitation des caractéristiques linguistiques des verbes : classes sémantiques, classes aspectuelles (inchoatif ou non inchoatif selon le début hypothétique de l'action, terminatif ou non terminatif selon la fin de celle-ci), rôles du participe passé ou encore des prépositions. Sur certains de ces points, les systèmes de DHM actuels ont encore à progresser.

6.2.2. *Analyse de l'énoncé « mets ça ici »*

L'exemple « mets ça ici », qui a fait les débuts du dialogue multimodal (Bolt, 1980), peut être décortiqué de manière plus approfondie quand on tient compte des cas de figures suivants, tous dans le cadre de l'utilisation par commande vocale d'un logiciel de dessin :

– la référence « ça » désigne un objet statique, qui fait partie d'une palette graphique et ne doit donc pas être déplacé. Dans ce cas, « mettre » réfère à une action de création à l'identique et non d'un déplacement ;

– la référence « ça » désigne un objet qui n'est pas rangé à la bonne place (c'est-à-dire là où se trouvent les autres exemplaires de la même classe d'objets), ou qui n'est pas dans la bonne configuration ou orientation. Dans ce cas, « mets ça ici » est probablement plus qu'un déplacement : il s'agit peut-être aussi d'une rotation ou d'un rangement selon les paramètres des objets déjà rangés ;

– la référence « ça » dépend du résultat de la résolution de la référence « ici » : le lieu désigné par « ici » est par exemple un lieu de rangement pour des objets d'une certaine catégorie, et le geste accompagnant « ça » est potentiellement ambigu entre plusieurs objets de catégories différentes ;

– la référence « ici » dépend du résultat de la résolution de la référence « ça » : c'est la distinction entre « mets de la moquette ici » en désignant un point d'une pièce, et « mets un clou ici » avec le même geste ;

– la référence « ici » dépend de connaissances spécifiques, par exemple quand « ça » désigne une prise électrique : le geste accompagnant « ici », même s'il est effectué avec précision, peut ne pas désigner le lieu exact pour la prise, celui-ci devant suivre des normes quant à la hauteur ou la distance par rapport à une autre prise, normes qui deviennent prioritaires dans le placement et entraînent une réinterprétation du geste en tant que désignation approximative ;

– les deux références peuvent être produites en même temps qu'un seul geste, qui décrit éventuellement une trajectoire (déplacement), ou peut s'interpréter uniquement avec les deux extrémités (disparition puis réapparition). L'ambiguïté peut être importante, par exemple quand l'action de déplacement laisse une trace visible à l'écran, trace qui devient elle-même partie du dessin en cours. Dans l'hypothèse du déplacement, une ambiguïté supplémentaire peut intervenir si la tâche implémente deux types de déplacement : l'un sans effet sur les objets présents sur le chemin parcouru, l'autre conduisant à l'écartement de tout obstacle.

Ces exemples montrent l'importance du modèle des actions et de la sous-spécification du langage naturel. Pour les résoudre, il est utile de mettre en œuvre un processus de résolution de la référence en plusieurs étapes, comprenant :

– l'analyse du contexte visuel (analyse des groupes perceptifs et des différences entre le « ça » et les objets déjà placés en « ici ») ;

– l'analyse des trajectoires gestuelles (présence d'une phase d'évitement, par exemple) ;

– les analyses linguistiques (sémantique verbale, rôles actanciels, aspects temporels) ;

– la confrontation des trois analyses ainsi effectuées pour résoudre de manière parallèle les références aux objets et les références aux actions (fusion multimodale en tenant compte des contraintes de chacune des modalités) ;

– la confrontation des analyses pragmatiques, ce qui fera l'objet du chapitre 7.

6.3. Gestion des anaphores et des coréférences

Nous avons mentionné l'un des rôles de l'historique du dialogue, à savoir retenir les objets référencés ainsi que les expressions utilisées pour ce faire, au fur et à mesure du dialogue. C'est ainsi qu'il est possible de résoudre les anaphores et d'identifier les coréférences. La résolution des anaphores est un processus beaucoup étudié en linguistique et en TAL (Mitkov, 2002), qui consiste à faire le lien entre une expression anaphorique et son antécédent. Ainsi, dans « prends un cube et mets-*le* dans la boîte », « le » est une expression anaphorique, c'est-à-dire qu'elle ne peut pas s'interpréter dans le contexte visuel immédiat mais nécessite d'explorer le cotexte linguistique, afin de trouver un référent déjà mentionné qui est ainsi repris. La recherche de l'antécédent aboutit à identifier « un cube » et à construire une relation anaphorique entre « le » et « un cube ».

Dans un système de DHM, ce processus nécessite plusieurs étapes. Il s'agit tout d'abord d'identifier les expressions vraiment anaphoriques, et de les distinguer de celles qui sont des références telles qu'on les a vues tout au long de ce chapitre. Pour cela, la forme linguistique est essentielle, les pronoms de troisième personne favorisant clairement une interprétation anaphorique, alors qu'une expression telle que « le cube » ou « le vert » peut référer aussi bien directement qu'anaphoriquement : « prends un cube rouge et un cube vert, mets le rouge dans la boîte et le vert par-dessus ». L'impossibilité à résoudre la référence directe est aussi un indice : si plusieurs référents sont possibles, peut-être qu'il s'agit d'une anaphore. Une deuxième étape consiste à regarder le genre, le nombre, éventuellement la catégorie si l'expression anaphorique en contient une, de manière à faire la liste des antécédents possibles. Quand plusieurs antécédents sont identifiés, il s'agit alors de faire un choix parmi eux. Les critères sur lesquels repose ce choix sont la proximité, par exemple en nombre de mots, entre l'antécédent et l'anaphore, la saillance du référent correspondant à l'antécédent, ou encore les fonctions grammaticales : si la fonction de l'antécédent est la même que celle de l'anaphore, il y a parallèle syntaxique, et c'est un argument pour privilégier cet antécédent plutôt qu'un autre. En TAL comme en DHM, ce processus peut être implémenté en faisant appel à des statistiques, ou encore à de l'apprentissage automatique, de manière à pondérer l'importance de chacun des paramètres de résolution en fonction de tests sur corpus. En DHM, l'antécédent peut appartenir à un énoncé antérieur, et l'identité du locuteur n'est pas une contrainte : l'utilisateur peut très bien reprendre anaphoriquement une référence faite par le système et inversement.

Jusqu'à présent, nos exemples d'anaphores sont également des coréférences, c'est-à-dire que l'antécédent et l'expression anaphorique désignent le même référent : une fois que la relation entre les deux est identifiée, l'attribution d'un référent à l'expression anaphorique consiste à reprendre le référent déjà attribué à l'antécédent. Or l'anaphore a ceci de particulier qu'elle peut être associative, c'est-à-dire exploiter un lien conceptuel entre deux référents différents. C'est ainsi que « donne-moi un billet pour Paris. *Le prix* ne doit pas dépasser vingt euros » ou « dessine un triangle. Colorie *un côté* en rouge » font intervenir à chaque fois deux référents liés entre eux, la référence au second se comprenant grâce à la mention du premier, par une relation d'anaphore associative. L'anaphore n'est alors pas coréférente : les deux référents ne sont pas identiques, et chacun nécessite son propre processus de résolution de la référence.

Comme pour la référence, l'anaphore et la coréférence peuvent porter aussi bien sur des objets concrets que sur des objets abstraits, et en particulier sur des événements. Dans « la réservation n'a pas fonctionné, je n'ai reçu aucun billet », il y a un lien entre le fonctionnement de la réservation et le fait de recevoir un billet. Dans « j'ai réservé un aller pour Paris avec des préférences d'horaire et de place. Je recommence avec un aller pour Lyon », le verbe « recommencer » ne se comprend qu'à l'aide de l'antécédent qui explicite une réservation. Enfin, un exemple de coréférence événementielle dans le cadre de notre tâche favorite est le suivant : « je réserve un billet pour Paris. Je souhaite un aller ». Dans tous ces exemples, le système de DHM doit faire face à deux phrases, ou deux propositions, qui décrivent toutes les deux un événement, et qui sont liées l'une à l'autre. Le lien dépend de la nature des événements et de leur représentation dans un modèle conceptuel. Ainsi, recevoir un billet peut être considéré comme la dernière étape constitutive d'une réservation. Les enjeux pour le DHM sont ici multiples : il s'agit premièrement de déterminer le lien anaphorique ou coréférentiel, deuxièmement de relier les contenus sémantiques des deux phrases en fonction de ce lien, troisièmement d'inférer un contenu sémantique qui pourrait recouvrir les deux phrases, ou, si ce n'est pas possible, d'expliciter le type de relation de discours qui opère entre les deux. Ce sont des aspects essentiels à la compréhension, qui permettent d'appréhender avec efficacité la cohérence d'un dialogue, mais qui font intervenir de nombreuses connaissances et qui sont délicats à implémenter. Dans le cas du premier exemple, la deuxième phrase « je n'ai reçu aucun billet » exprime la cause du constat fait dans la première phrase. Pour identifier cette relation de discours, il faut comprendre non seulement le lien entre les deux événements, mais il faut aussi comprendre que l'utilisateur exprime son problème, avec une description argumentée. Le système peut ainsi réfuter le lien en répondant « la réservation a fonctionné, mais le billet vous a été envoyé hier et n'arrivera que demain ». Dans le cas du deuxième exemple, la sémantique du verbe « recommencer » permet au système d'entamer une nouvelle réservation en reprenant l'ensemble des paramètres de l'ancienne, à l'exception de la destination qui est explicitée. Dans le cas du troisième exemple, les deux événements sont tout simplement les mêmes (mais encore faut-il le comprendre), ce

qui permet au système de définir une requête avec l'ensemble des paramètres, ceux indiqués dans la première phrase et ceux indiqués dans la seconde.

On le voit, la référence est bien une question pragmatique, qui va au-delà de la simple identification d'un référent : avec la notion de domaine de référence, avec l'exploitation de l'historique du dialogue, avec les liens qui sont faits entre les différentes modalités, avec les notions de coréférence et de cohérence, elle apparaît comme un mécanisme complexe qui contribue à donner de la consistance au dialogue.

6.4. Bilan

Les liens entre un énoncé linguistique et le monde de la tâche à accomplir se font par des références : d'une part des références aux objets accessibles et manipulables dans ce monde, ce qui peut se faire à l'aide de mots et d'expressions bien choisies, ou à l'aide de la combinaison d'un geste de désignation avec une expression linguistique adéquate (dans le cas des systèmes multimodaux) ; d'autre part des références aux diverses actions réalisables dans le monde de la tâche. Ce sixième chapitre déroule les processus de résolution des références aux objets et aux actions, et montre comment plusieurs énoncés peuvent être reliés entre eux par des relations d'anaphore et de coréférence, relations qu'un système se doit d'identifier pour être à même de gérer la cohérence d'un dialogue.

Chapitre 7

La reconnaissance des actes de dialogue

Les énoncés « réserve-moi un aller pour Paris », « quelle est la durée de ce trajet ? » et « je n'arrive pas à communiquer avec toi » diffèrent de par leur contenu sémantique, mais aussi de par l'acte de langage qu'ils effectuent : le premier est un ordre (« dire de »), exprimé à l'impératif, et requiert du système qu'il exécute l'ordre ; le second est une question (« demander »), exprimée à l'interrogatif, et requiert une réponse de la part du système ; le troisième est une assertion (« dire que »), exprimée sous une forme déclarative, et requiert que le système prenne en compte ce qui est dit pour en tirer des conclusions, quelles qu'elles soient. La nature et les mécanismes d'identification de ces actes de langage, catégorisés ici selon le point de vue de la théorie de la pertinence (Sperber et Wilson, 1995), constituent un volet de la pragmatique (dite du troisième degré, voir paragraphe 1.2.2), qui s'avère essentiel à la gestion du dialogue : c'est en comprenant quel acte de langage réalise l'utilisateur qu'un système de DHM peut déterminer sa propre réaction. Du moins, et c'est ce que nous allons voir dans ce chapitre et dans le suivant qui lui est lié, c'est l'un des paramètres qui permet au système de décider comment continuer le dialogue.

Comme pour le modèle des actions décrit dans le chapitre précédent, l'accomplissement d'un acte de langage fait intervenir des préconditions et des post-conditions, et son identification nécessite un certain nombre de paramètres. En compréhension automatique, le processus chargé de l'identification des actes de langage nécessite ainsi des arguments en entrée, qui peuvent faire l'objet de pré-traitements, et retourne un résultat en sortie. En entrée, on a ici besoin de la représentation sémantique obtenue par le chapitre 5, avec les indications prosodiques, et notamment celles qui concernent le contour intonatif de l'énoncé, et avec les résultats de la résolution des références obtenus par le chapitre 6. On a aussi besoin de l'historique du dialogue, avec les représentations sémantiques et pragmatiques calculées pour les énoncés précédents, en

incluant l'identification des actes de langage qui a été faite. Enfin, dans le cas du dialogue multimodal, on a également besoin d'une représentation des gestes effectués et d'une manière générale des contenus portés par les modalités traitées, afin de leur affecter éventuellement un acte de langage, qui, du fait de la nature de ces modalités, est appelé « acte de dialogue » plutôt qu'« acte de langage ». Nous le verrons, un geste peut en effet exprimer un ordre, une question ou une assertion.

Le résultat de la reconnaissance des actes de dialogue est l'affectation d'une ou de plusieurs étiquettes sur les contenus sémantiques, ces étiquettes décrivant les actes de dialogue réalisés. Comme toujours, plusieurs hypothèses alternatives peuvent être produites en cas d'ambiguïté, et une représentation sous-spécifiée en cas d'impossibilité à reconnaître un acte particulier. Les représentations pragmatiques ainsi obtenues constituent le paramètre principal pour la gestion du dialogue et la détermination de la réaction du système, et viennent mettre à jour l'historique du dialogue.

Après une première section visant à décrire la nature des actes de dialogue (section 7.1), ce chapitre présente quelques méthodes utilisées en DHM pour leur identification automatique (section 7.2), et notamment les processus mis en œuvre en dialogue multimodal, avec le processus de fusion multimodale au niveau des actes de dialogue (section 7.3).

7.1. Nature des actes de dialogue

7.1.1. *Définitions et phénomènes*

(Austin, 1962) distingue pour chaque énoncé un acte locutoire, qui correspond à la production de l'énoncé, un acte illocutoire, celui de demander, d'ordonner, etc., et un acte perlocutoire qui correspond à la production, souvent intentionnelle, de certains effets sur les croyances et le comportement de l'interlocuteur. Le terme d'acte de langage sert à décrire les actes illocutoires. (Searle, 1969) étend les catégorisations de J. Austin et caractérise cinq types d'actes principaux sur la base de critères tels que la condition de sincérité ou la direction d'ajustement de l'acte, c'est-à-dire est-ce qu'il agit sur le monde où est-ce l'inverse : les actes assertifs, les actes directifs, dont le but est de faire faire quelque chose à l'interlocuteur et dans lesquels on trouve la question et l'ordre, les actes commissifs qui engagent le locuteur à une action future, les actes expressifs et les actes déclaratifs. Cette théorie fait l'objet de très nombreuses variantes et adaptations, par exemple celle de (Clark, 1996) qui considère les types d'actes suivants : assertion, ordre, requête, question fermée, promesse, offre, remerciement, compliment, salutation, adieu. D'une manière générale, les différentes démarches et les différents critères qui conduisent à déterminer une liste d'actes plutôt qu'une autre sont présentées par exemple dans (Traum, 2000).

(Sperber et Wilson, 1995), dans le cadre de leur approche cognitive et pragmatique du dialogue, proposent d'abstraire les catégories en « dire de » (que l'on appellera

ordre), « demander » (question) et « dire que » (assertion), qui se focalisent sur ce qui doit être identifié pour que l'énoncé soit interprété. Ces trois types d'acte de langage ne se basent pas sur la syntaxe de l'énoncé, comme le tout début du chapitre pouvait le faire croire, n'impliquent pas des conditions telles que celles de J. Searle, et, du moins pour un humain, s'identifient à l'aide de tests linguistiques simples : l'ajout de « s'il te plaît » permet de tester l'ordre, l'ajout au début de l'énoncé de « dis-moi » permet de tester la question, et l'ajout d'« après tout » permet de tester l'assertion. La syntaxe apporte un indice, sans qu'une structure syntaxique soit pour autant liée à un type d'acte : « puis-je avoir un aller pour Paris ? », sous sa forme de question, satisfait surtout le test du « s'il te plaît » caractérisant l'ordre. La prosodie apporte également un indice, avec par exemple un contour intonatif montant (ou plutôt constant, en tout cas pas descendant) qui permet d'interpréter « vous avez des billets pour Paris » comme une question plutôt qu'une assertion. Pour ce dernier exemple, l'acte de question, surtout si la prosodie est peu marquée, peut cependant ne pas être très manifeste, en tout cas moins que « avez-vous des billets pour Paris ? ». Suite à cette remarque, on peut faire comme (Kerbrat-Orecchioni, 2012) une distinction entre la valeur illocutoire et la force illocutoire : dans les deux exemples, la valeur est celle de question, alors que la force est plutôt faible dans le cas de la forme assertive et plutôt forte dans le cas de la forme interrogative. Cela permet de mieux caractériser l'acte de langage de l'énoncé, et ainsi d'y réagir de manière pertinente.

Au-delà d'actes de langage, le dialogue met en œuvre des actes liés au déroulement et aux modalités de la communication. Nous avons vu que la possibilité d'actes gestuels conduit à parler d'« actes de dialogue ». Ce terme sert également à désigner des actes qui se comprennent dans un contexte de dialogue, c'est-à-dire en tenant compte des énoncés précédents. Ainsi, compte tenu de son contenu sémantique restreint, un énoncé tel que « oui » se voit attribuer un acte de langage d'assertion, mais se modélise de manière plus précise par un acte de dialogue de type accusé de réception ou réponse à une question quand on considère que l'énoncé précédent est soit « je veux une place en première classe », soit « reste-t-il une place pour Paris dans le prochain train ? ». Intégrer ainsi des aspects dialogiques dans la notion d'acte peut poser des problèmes, et certains auteurs refusent cette vision des choses, considérant que les liens entre énoncés sont réalisés à un autre niveau d'analyse, qui concerne la gestion du dialogue (voir chapitre 8). Il n'en reste pas moins qu'un acte de langage se comprend dans un dialogue, ce que (Grisvard, 2000, p. 102) montre avec l'exemple « tu effaces la séquence », qui, précédé de « qu'est-ce qui se passe si j'appuie sur OK ? », s'interprète comme une assertion, alors que, précédé de « bon alors, qu'est-ce que je fais ? », s'interprète comme un ordre. C'est pour ce type d'exemples que le recours à l'historique du dialogue s'avère indispensable. Enfin, les actes de dialogue peuvent être considérés comme constituant une catégorie d'« actes conversationnels » (Jurafsky et Martin, 2009), avec la catégorie des actes de tour de parole, des actes d'ancrage (accusé de réception, demande d'accusé de réception, réparation, demande de réparation, continuation, annulation) et des actes d'argumentation (élaborer, clarifier,

contrer, etc.). Ce point de vue plus global peut encore s'élargir en intégrant la possibilité d'actes conjoints, c'est-à-dire réalisés de manière coopérative par le locuteur et l'interlocuteur, comme lorsqu'un énoncé du premier complète celui du second.

Au final, un énoncé comme « combien ce temps avec ce chemin qui semble être le plus court ? » peut regrouper à lui seul plusieurs actes conversationnels : des actes explicites comme la question portant sur un temps de parcours et le commentaire sur le fait qu'il semble être le plus court (paragraphe 7.1.2), ou comme le fait tout simple de prendre la parole suite à une action du système de dialogue, mais aussi des actes tacites comme celui correspondant à un ancrage de l'énoncé « voici les trajets possibles » : c'est parce que l'utilisateur a bien compris cet énoncé qu'il peut attribuer aux éléments graphiques qui apparaissent le statut d'alternatives répondant à sa requête initiale. Par ailleurs, l'acte de bonne réception de l'énoncé précédent est aussi un acte tacite qui intervient ici. On en vient alors à la notion d'« acte multifonctionnel » (Bunt, 2011), ce que nous verrons plus loin sous le nom d'acte composite (paragraphe 7.1.3).

7.1.2. *Le problème des actes indirects*

Dans l'exemple de l'introduction repris ci-dessus, « qui semble être le plus court » est une relative qui peut s'interpréter comme la proposition « ce chemin semble être le plus court ». Avec sa forme assertive, ce commentaire de la part de l'utilisateur peut sembler peu important, en tout cas pas vraiment adressé au système mais relevant plutôt d'une pensée exprimée comme ça, au passage. Le seul test linguistique qui fonctionne avec cette proposition est celui de l'assertion : « après tout, ce chemin semble être le plus court ». Pourtant, avec ce commentaire, l'utilisateur peut vouloir faire réagir le système, à la manière d'une question : « ce chemin semble être le plus court, n'est-ce pas ? ». Le phénomène décrit ici est celui de l'acte de langage « indirect ». Il a été beaucoup étudié d'un point de vue théorique (Searle, 1969 ; chapitre 5 de Levinson, 1983 ; Moeschler, 1985) et aussi d'un point de vue plus formel (Searle et Vanderveken, 1985 ; Asher et Lascarides, 2001). Certains auteurs considèrent que l'acte de langage effectué par l'énoncé est bien une assertion, et que c'est notre façon d'interpréter qui en fait une question. D'autres considèrent qu'une forme peut prendre la place d'une autre par convention. C'est le cas notamment avec l'ordre, qui, sous sa forme impérative, peut sembler un peu brutale, au point de lui préférer la question (qui en devient un acte indirect conventionnel). D'autres encore, et c'est notamment le cas de (Asher et Lascarides, 2001), mettent en perspective l'énoncé courant avec le précédent ou le suivant, de manière à expliciter la relation de discours qui existe entre les deux et qui peut faire pencher l'analyse vers l'acte indirect. Plus précisément, la relation de discours devient elle-même un acte, et la question à forme d'assertion devient un acte complexe, ce qui ne pose pas de problème pour un système de DHM déjà habitué à gérer des ambiguïtés dans tous les sens. Quelle que soit la voie suivie pour l'interprétation, l'important pour le déroulement du DHM est que le système comprenne qu'il peut réagir en répondant au commentaire, c'est-à-dire en confirmant ou infirmant qu'il

s'agit du chemin le plus court. Si le système ne détecte pas cette possibilité d'interprétation « indirecte », alors il ne peut répondre qu'à la question posée par la proposition principale, « combien de temps avec ce chemin ? ». Ce n'est pas forcément grave, mais cela diminue grandement les capacités de compréhension et surtout de coopération.

Un autre exemple typique du phénomène d'acte indirect est la question qui cache un ordre, comme « peux-tu m'écouter ? » dans notre tâche de renseignement ferroviaire. Avec les tests linguistiques vus au début du paragraphe 7.1.1, on voit que celui de l'ordre et celui de la question fonctionnent bien avec cet exemple : « peux-tu m'écouter, s'il te plaît ? » et « dis-moi, peux-tu m'écouter ? ». L'hypothèse de la question simple n'est pas tenable très longtemps : le système a tout à fait les capacités d'écouter l'utilisateur, donc répondre « oui, je peux » n'apporterait pas grand chose au dialogue. Sauf cas particulier, comme quand l'utilisateur vient à peine d'aborder le système et qu'il n'a pas encore constaté les capacités de compréhension de celui-ci, ce ne sont *a priori* pas les capacités du système qui sont en jeu. La question « peux-tu m'écouter ? » pourrait aussi porter sur les conditions de la situation de communication : en environnement bruité, l'utilisateur peut croire que le système ne va peut-être pas l'entendre (« peux-tu m'entendre ? » serait d'ailleurs plus pertinent dans ce cas). Plus vraisemblablement, c'est une injonction que l'utilisateur est en train de faire, voire un sous-entendu tel que « fais un peu attention à ce que je dis » ou « je te dis que je veux aller à Paris, pas que je cherche une promotion pour mes prochaines vacances ». Si les sous-entendus sont difficilement identifiables, l'intention d'ordre est plus facile à inférer, et on considère donc cet énoncé comme un exemple d'acte indirect, pouvant amener à une réaction du système telle que « je vous écoute » ou « redites-moi votre requête ».

Par rapport à l'exemple « ce chemin semble être le plus court », il est cependant ici possible de répondre en même temps à la question, par « oui, je vous écoute » ou « oui, redites-vous votre requête ». Autrement dit, on peut considérer que la question et l'ordre s'effectuent en même temps, ce qui constitue un acte « composite ».

7.1.3. *Le problème des actes composites*

Outre les cas où un énoncé réalise à la fois un acte d'accusé de réception de l'énoncé précédent et l'acte de langage porté par la forme linguistique, le cas des actes composites est intéressant dès qu'un tour de parole comporte plusieurs actes de langage. D'une certaine manière, c'est le cas dans « bonjour, je voudrais aller à Paris » et dans « d'accord, je réserve un billet », dans la mesure où ces deux énoncés comportent en fait chacun deux segments de discours, un acte de langage étant attribué à chaque segment. La détermination des actes de langage comprend ainsi une segmentation préalable des énoncés. C'est surtout le cas de « combien de temps avec ce chemin qui semble être le plus court ? » qui est intéressant et qui relève de l'acte composite, avec une composante « demander », qui est directe et concerne un temps de parcours,

et une autre composante « demander », qui est indirecte et prend la forme de l'assertion « ce chemin semble être le plus court ». Selon les auteurs, la première composante, c'est-à-dire celle qui occupe la place principale du point de vue de la construction de la phrase, est appelée « acte directeur » ou « acte de première intention », alors que la seconde composante est appelée « acte subordonné » ou « acte de seconde intention ». Pour ce qui concerne l'acte indirect, la composante assertive, c'est-à-dire celle qui correspond à la forme linguistique, est appelée « acte de surface » ou « acte secondaire », alors que la composante interrogative, c'est-à-dire celle qui est proche de l'intention communicative ou en tout cas de l'acte auquel le système doit réagir, est appelée « acte profond » ou « acte primaire ». Autre exemple du foisonnement de termes, (Kerbrat-Orecchioni, 2012) distingue de son côté la valeur patente (littérale, explicite, primitive) de la valeur latente (indirecte, implicite, dérivée), avec plusieurs cas où la véritable intention est celle de la valeur latente.

Nous avons ici analysé « combien de temps avec ce chemin qui semble être le plus court ? » comme un seul segment discursif, affecté donc d'un acte composite à deux facettes. Une autre possibilité aurait consisté à considérer deux segments, sur des critères moins syntaxiques (il y a une seule phrase) que fonctionnels ou communicatifs, voir les segments fonctionnels dans (Bunt, 2011). Le résultat est le même avec les deux approches : le système se trouve face à deux actes, et il peut réagir à l'un aussi bien qu'à l'autre, voire aux deux. L'avantage de la solution considérant l'énoncé comme un segment unique est qu'elle met en avant un acte par rapport à l'autre, ce qui est plus délicat à faire si l'on considère deux segments autonomes. L'acte direct qui correspond à la question de la principale est en effet mis en avant par rapport à celui du commentaire, du fait justement d'être lié à une principale et non à une subordonnée ou une relative. C'est donc un critère surtout syntaxique qui prime ici. Il existe cependant des situations où un seul segment mène clairement à l'identification d'un acte composite. C'est le cas de l'énoncé « que pensez-vous de huit heures ? » en réponse à la question « avez-vous un aller pour Paris pour demain matin ? ». Il s'agit d'une question du système, qui porte sur un horaire, en réaction à une demande de l'utilisateur. Or le fait qu'un horaire soit mentionné prouve que le système a compris la demande : non seulement il répond par l'affirmative (si la réponse avait été négative, cet énoncé n'aurait pas pu survenir), mais en plus il demande une précision pour valider la réservation. Nous avons donc bien ici un acte composite, ou, si l'on reprend le point de vue de (Asher et Lascarides, 2001), un acte complexe incluant une relation d'élaboration entre les deux énoncés.

Sans remettre en question la classification en trois actes de langage principaux que fait la théorie de la pertinence, ce type d'exemple montre que le langage naturel ne rentre jamais dans des catégories totalement compartimentées. L'interprétation proposée pour l'exemple de l'introduction est de plus réductrice. Pour bien faire, il faudrait intégrer les aspects de tours de parole, d'ancrage compte tenu des énoncés précédents, et d'une manière générale de l'ensemble des fonctions communicatives qui interviennent en dialogue, comme la gestion de temps et de la tâche en cours.

7.2. Identification et traitement des actes de dialogue

7.2.1. *Classification et identification des actes*

En DHM, la détermination de la nature et d'une typologie des actes de langage, ainsi que des actes de dialogue, doit faire face à deux objectifs potentiellement contradictoires. Il s'agit d'une part de déterminer une classification précise des types d'actes pour que cela devienne le principal critère de raisonnement pour le système : plus il y a de types d'actes, plus l'identification automatique peut poser problème, mais plus l'interprétation est fine et plus le système peut réagir avec pertinence. D'une manière générale, une classification avec un nombre raisonnable de types est plus pratique à gérer pour le système. Il s'agit d'autre part d'envisager des corpus annotés en actes de dialogue, de manière à exploiter ces corpus pour améliorer les performances d'identification du système, par exemple *via* une phase d'apprentissage automatique. Or constituer un corpus de référence sur ces aspects se fait pour l'instant à la main, et il est délicat de demander à des annotateurs de choisir, pour chaque énoncé, un type d'acte parmi cent. La tâche d'annotation en devient trop pénible et risque d'entraîner une multiplication des erreurs. Par ailleurs, conceptualiser cent types d'actes est loin d'être évident pour un humain, et cela n'aide pas le concepteur de systèmes.

Plusieurs propositions (techniques) de classifications d'actes ont été faites, qu'il s'agisse de la Fipa (*Foundation for Intelligent Physical Agents*), de Damsl (*Dialog Act Markup in Several Layers*), de langages basés sur XML comme KQML ou de la récente norme ISO 24617-2, spécifiée par un groupe de travail réunissant les principaux chercheurs internationaux en DHM. L'intérêt, notamment de cette dernière, est de proposer une classification hiérarchique, qui permet d'appréhender les types d'actes avec différents niveaux de granularité, voir également le tableau récapitulatif de (Harris, 2004, p. 104). C'est peut-être la solution au dilemne du paragraphe précédent : là où un système de DHM peut exploiter tous les niveaux de granularité, un annotateur de corpus se contentera, dans un premier temps, du premier niveau.

L'identification de l'acte de langage, ou de l'acte de dialogue, est un processus qui nécessite les paramètres suivants :

– les mots de l'énoncé eux-mêmes, et leurs propriétés sémantiques, notamment pour le verbe (Rosset *et al.*, 2007) ;

– l'analyse syntaxique de la phrase, avec notamment le mode ;

– l'analyse prosodique de l'énoncé, avec notamment le contour intonatif (Wright-Hastie *et al.*, 2002) ;

– le type de l'acte précédent dans le dialogue, et d'une manière générale toute information issue de l'historique du dialogue qui permet de rattacher l'énoncé courant aux précédents ;

– à la manière d'un modèle de langage et dans le but d'exploiter des méthodes telles que les CRF ou les HMM (voir paragraphe 2.1.3), la succession des actes précédents.

Comme le souligne (Jurafsky et Martin, 2009, p. 880), le module dédié à la reconnaissance des actes peut se diviser en deux parties : une partie chargée des actes généraux, et une partie chargée de la reconnaissance d'actes spécifiques comme les actes correctifs de l'utilisateur suite à une erreur du système. Les deux parties fonctionnent de la même façon, c'est-à-dire, dans les systèmes actuels, selon une tâche d'étiquetage après une phase d'apprentissage automatique, mais avec des modèles différents. Les énoncés correctifs sont en effet plus difficiles à reconnaître que des énoncés habituels, et nécessitent des indications spécifiques telles que la présence de mots comme « non » ou « je n'ai pas », la présence d'une répétition, accompagnée éventuellement d'une articulation exagérée, d'une paraphrase, d'un ajout ou d'une omission de contenu.

7.2.2. *Cas des actes indirects et des actes composites*

Le processus d'identification des actes indirects reprend les mêmes paramètres que ceux présentés dans la section précédente, mais met en avant quatre aspects complémentaires particulièrement importants :

– un répertoire de conventions dialogiques, qui inclut quelques exemples typiques d'actes indirects : si la situation en cours de traitement correspond à l'un de ces exemples, alors le système peut s'appuyer sur la solution préconisée ;

– les préférences du locuteur, c'est-à-dire le modèle de l'utilisateur, si celui-ci a été mis à jour au fur et à mesure du dialogue, notamment dans les cas de détection d'actes indirects (« quand il dit ça, c'est pour faire ça ») ;

– le modèle de la tâche du système, et notamment la liste de ses capacités : c'est en effet un moyen pour identifier les actes indirects tels que « peux-tu m'écouter ? » ;

– des hypothèses sur les états mentaux de l'utilisateur, en suivant par exemple un modèle BDI (voir paragraphe 2.1.2), de manière à ce que le système puisse détecter quand l'utilisateur connaît déjà la réponse à la question qu'il est en train de poser, afin d'interpréter celle-ci comme un acte indirect.

Dans un même ordre d'idée, l'identification des actes composites fait appel à la même liste de paramètres, avec quelques aménagements :

– un ensemble de mots et de constructions linguistiques qui sont souvent utilisés pour exprimer un acte de deuxième intention : épithètes, adverbes évaluatifs, appositions, subordonnées relatives, etc. ;

– un répertoire de conventions dialogiques, qui inclut des exemples d'actes composites avec un ensemble de réactions possibles pour chacun des cas : par exemple, on

réagit à un acte regroupant une question et un commentaire par la réponse à la question et, éventuellement, par la confirmation ou l'infirmation du commentaire (surtout si celui-ci s'avère faux) ;

– les préférences du locuteur, qui peuvent se paraphraser en « quand il dit ça, c'est pour faire ça et ça » ;

– le modèle de la tâche de manière à déterminer l'ordre d'importance des différents actes en présence ;

– des hypothèses sur les états mentaux de l'utilisateur, de manière à favoriser l'acte dont la satisfaction aura le plus d'incidence sur ces états mentaux. On rejoint ici clairement la théorie de la pertinence, avec la notion d'effets contextuels (Sperber et Wilson, 1995).

Dans les deux cas, le processus mis en œuvre dans les systèmes actuels repose encore une fois sur de l'apprentissage automatique. Pour ces actes à la fois explicites et implicites que sont les actes indirects et composites, une technique adaptée est la classification supervisée, avec les étiquettes d'actes comme classes cachées à détecter (Jurafsky et Martin, 2009).

7.3. Traitement des actes de dialogue multimodaux

Nous l'avons vu au paragraphe 5.1.4, certains gestes peuvent porter un acte de dialogue. Des yeux écarquillés, dans le cas d'un système de DHM avec capture par caméra, peuvent être l'équivalent d'une question telle que « qu'avez-vous dit ? ». S'ils s'accompagnent d'un geste de désignation, l'intention communicative peut être quelque chose comme « qu'est-ce que cela ? ». Un geste de pointage peut aussi correspondre à un ordre, de même, dans le cas d'un système avec un écran tactile, qu'un geste en forme de croix sur un objet peut signifier l'ordre de supprimer cet objet. Enfin, et c'est le cas de la majorité des gestes expressifs, un geste peut procéder à une assertion, dont le contenu s'ajoute au contenu de l'énoncé linguistique simultané selon le processus de fusion multimodale au niveau sémantique (voir paragraphe 6.1.4). Bien entendu, si aucun énoncé linguistique n'accompagne le geste, l'interprétation de celui-ci se termine avec la détermination de son seul acte de dialogue. En revanche, en dialogue multimodal, nous faisons face à des exemples où l'énoncé linguistique se voit attribuer un acte de langage, le geste également, qu'il s'agisse d'un « dire de », d'un « demander » ou d'un « dire que » (nous restons ici aussi dans le cadre de la théorie de la pertinence), et où l'interprétation automatique passe par le traitement de ces actes de dialogue multimodaux.

Le processus pour ce faire est celui d'une fusion multimodale à un niveau pragmatique : les deux actes, celui du geste et celui de la parole, sont confrontés et unifiés de manière à obtenir un seul acte qui caractérise l'énoncé multimodal complet. Tout

d'abord, quand les deux actes en présence sont du même type, par exemple deux assertions ou deux questions, la fusion multimodale consiste essentiellement à vérifier la compatibilité des contenus sémantiques. L'exemple des yeux écarquillés, si ce geste intervient en même temps qu'un énoncé oral, ne porte pas de contenu sémantique particulier. La fusion est donc immédiate quand l'énoncé oral est du type « demander ». Même chose pour un geste brusque qui vient illustrer de manière injonctive l'ordre également transmis par un énoncé oral simultané. La fusion peut être moins immédiate avec deux assertions : cette fois, le geste porte un contenu sémantique, par exemple celui d'un quasi-linguistique particulier. Soit ce contenu est compatible avec celui de l'assertion réalisée par la parole et la fusion multimodale aboutit à un seul acte d'assertion avec un contenu sémantique unifié, soit les deux contenus sémantiques ne sont pas compatibles et le système fait face à deux actes différents, autrement dit à un « acte multimodal composite ». Enfin, quand les deux actes en présence sont de types différents, par exemple un geste interrogatif et un énoncé oral de type « dire que », plusieurs cas sont possibles. Soit les contenus sémantiques peuvent fusionner, par exemple quand le geste ne porte pas de sens particulier, et le système de DHM peut alors soulever l'hypothèse d'un « acte multimodal indirect » : l'énoncé linguistique ressemble à une assertion, mais la prise en compte du geste remet en cause cette interprétation et propose celle de l'acte profond « demander ». Si le contenu sémantique s'y prête, l'hypothèse est retenue, le geste jouant alors exactement le même rôle qu'un contour intonatif de question. Soit les contenus sémantiques ne fusionnent pas, et dans ce cas on fait face à deux actes distincts, ou à un acte composite qui comporte une question avec son contenu sémantique et une assertion avec son propre contenu sémantique. On peut alors considérer qu'une hiérarchie opère entre les deux : l'acte linguistique l'emporte sur l'acte gestuel, ne serait-ce que parce qu'un dialogue est avant tout linguistique. Comme pour l'exemple « combien de temps avec ce chemin qui semble être le plus court ? », le système aura à décider de sa réaction compte tenu des trois possibilités qui s'offrent à lui : réagir à l'acte premier (ici l'assertion linguistique), réagir à l'acte second (ici la question gestuelle), ou réagir aux deux actes. Ceci relève de la stratégie de dialogue, et c'est l'objet du chapitre suivant.

7.4. Bilan

Lorsque l'on utilise un système de dialogue homme-machine, c'est souvent moins dans le but de discuter d'égal à égal que de faire faire une tâche à la machine. Un énoncé porte ainsi un acte de requête, d'ordre ou encore de question. Ces actes de dialogue, qui peuvent être indirects et non indiqués explicitement par l'énoncé, doivent être identifiés correctement par le système. Celui-ci peut alors déterminer quel type de réaction adopter et appréhender le dialogue comme une succession structurée d'actes significatifs. Ce septième chapitre montre comment les actes de dialogue sont traités, et donne des exemples d'actes complexes, d'une part au niveau du langage qui s'avère naturellement complexe, d'autre part au niveau des interactions entre langage et gestes communicatifs.

TROISIÈME PARTIE

Le comportement du système et son évaluation

Chapitre 8

Quelques stratégies de dialogue

Le gestionnaire du dialogue est souvent présenté comme le cœur d'un système de DHM. C'est là où arrivent tous les résultats des processus de compréhension automatique, c'est là où les différentes informations se confrontent, y compris celles qui ont été stockées au fur et à mesure du dialogue, c'est là où les décisions sont prises, impliquant éventuellement une phase de résolution de problème ou de recherche d'information dans une base de données, et c'est de là que partent les messages à destination de l'utilisateur, sous une forme abstraite qui est matérialisée ensuite par les modules chargés de la génération automatique et de la synthèse.

Comme pour chacun des chapitres de la deuxième partie de ce livre, il s'agit ici d'un processus qui requiert des arguments en entrée et retourne un résultat en sortie. Un premier ensemble d'arguments regroupe les résultats des analyses prosodiques, lexicales, syntaxiques, sémantiques et pragmatiques, sous la forme d'une représentation du sens de l'énoncé, toutes références résolues, et accompagnée d'une étiquette indiquant l'acte de dialogue réalisé. Un deuxième ensemble d'arguments regroupe les ressources suivantes, gérées indépendamment ou en lien (voire en interne) avec le gestionnaire du dialogue : modèle de la tâche, des objets et du monde dans lequel se déroule la tâche, historique du dialogue, modèle de l'utilisateur. Le résultat visible de la gestion du dialogue est le choix d'une réaction du système. Cette réaction peut être une action dans le monde des objets de l'application, action alors visible à l'écran. Elle peut consister en la présentation visuelle d'une information. Elle peut aussi prendre la forme d'un énoncé linguistique. Dans ce cas, le gestionnaire du dialogue retourne un message sous une forme qui indique le « quoi dire », sans encore entrer dans les détails du « comment le dire », processus pris en charge par le générateur automatique, par le module dédié à l'ACA si le système est représenté visuellement par un avatar, et par le module dédié à la présentation d'un contenu multimodal (voir chapitre 9). Les autres

résultats de la gestion du dialogue sont la mise à jour des ressources citées, notamment de l'historique du dialogue qui peut en particulier comporter une structure permettant de relier les énoncés les uns aux autres, mais aussi des objets et du monde dans lequel se déroule la tâche, par exemple si le système doit modifier l'un de ces objets.

Le but de ce module central pour le système est de prendre des décisions qui d'une part vont dans le sens d'un dialogue naturel avec l'utilisateur (on revient ici au dialogue naturel en langage naturel), d'autre part manifestent une personnalité (*persona*) perceptible par l'utilisateur, et transparaissant aussi bien dans les contenus sémantiques émis que dans la façon de transmettre ceux-ci. L'un des rôles d'un système étant de se montrer coopératif, la modélisation de cet aspect constitue un enjeu particulièrement important, du moins pour les systèmes en domaine fermé tels que celui relatif au renseignement ferroviaire. Nous explorons ainsi la prise en compte des aspects naturels et coopératifs en DHM (section 8.1), puis nous explorons les aspects techniques de la gestion du dialogue, avec quelques approches et enjeux (section 8.2).

8.1. Aspects naturels et coopératifs de la gestion du dialogue

8.1.1. *But commun et coopération*

Un dialogue naturel se définit déjà en tant que discours, c'est-à-dire en tant que suite d'énoncés formant un ensemble cohérent, qui fait sens. Il se définit aussi en tant qu'activité finalisée et qu'activité collaborative (McTear, 2004). Le premier terme met en avant le but commun aux interlocuteurs et, dans le cas de dialogue finalisé, la tâche à résoudre. C'est ce but qui va provoquer une planification, c'est-à-dire la détermination de plans, et ainsi d'énoncés, pour réaliser le but. Quand des obstacles surviennent, c'est parce qu'un élément fait défaut pour réaliser le but. Le second terme met en avant la coopération, c'est-à-dire le principe général qui dit que les interlocuteurs cherchent à s'entraider plutôt qu'à se contrecarrer et à mettre en péril le dialogue. Dialoguer, c'est tout faire pour continuer à dialoguer, tant que le but n'est pas réalisé. Cette notion de coopération, qui peut sembler à première vue assez abstraite, a fait l'objet de très nombreux travaux visant à en déduire des sortes de règles sur lesquelles un gestionnaire de dialogue pourrait s'appuyer.

(Grice, 1975) considère que le dialogue est guidé par un ensemble de maximes, ou grands principes, qui permettent aux interlocuteurs d'interpréter et de produire des énoncés de manière pertinente. Son « principe de coopération » est formulé de la manière suivante : « faites en sorte que votre contribution à la conversation soit, au moment où elle intervient, telle que le requiert l'objectif ou la direction acceptée de l'échange verbal dans lequel vous êtes engagé », et ses maximes sont les suivantes : maximes de quantité (faites que votre contribution soit aussi informative que nécessaire, ne faites pas votre contribution plus informative que nécessaire) ; maximes de qualité (faites que votre contribution soit véridique, ne dites pas ce que croyez faux

ou ce pour quoi vous n'avez pas de preuve) ; maxime de pertinence (parlez à propos) ; maximes de manière (soyez clair, évitez l'ambiguïté, soyez bref, soyez ordonné). Malgré leur côté un peu vague, par exemple pour la maxime de pertinence, ces principes ont eu une influence déterminante sur les travaux qui ont suivi. Leur côté vague a aussi permis de les interpréter de manière large, en incluant un grand nombre de phénomènes. Le modèle hiérarchique de l'école de Genève (Roulet *et al.*, 1985) souligne ainsi deux contraintes issues des maximes de Grice, contraintes qui jouent un rôle dans la détermination de la structure des dialogues : la complétude interactionnelle, c'est-à-dire la tendance à faire progresser le dialogue vers la satisfaction des deux intervenants, et la complétude interactive, c'est-à-dire la tendance, lorsqu'il y a désaccord, à résoudre le conflit. Le modèle genevois a ensuite fait l'objet de diverses versions et extensions, notamment pour intégrer des critères pas seulement linguistiques (Reboul et Moeschler, 1998, p. 87), mais, comme beaucoup d'autres modèles, le point de départ reste toujours un principe de coopération suivi par les interlocuteurs. La maxime de pertinence est totalement reformulée par la théorie de la pertinence, qui en fait la ligne directrice d'une approche très aboutie des inférences réalisées dans le dialogue coopératif (Sperber et Wilson, 1995). Plus récemment, (Allwood *et al.*, 2000) décrit une approche des principaux mécanismes de coopération en dialogue, avec les critères de but commun, de confiance, mais aussi de considération cognitive et de considération éthique, critères qui élargissent le périmètre d'application des maximes de Grice. Les auteurs redéfinissent alors les notions de coordination ou de coopération en fonction de ces quatre critères, avec des degrés variables. Le but commun est ainsi décrit comme un degré de contribution mutuelle à un but partagé, un degré de conscience de ce but partagé, un degré d'accord sur ce but, un degré de dépendance au but, et un degré d'antagonisme impliqué dans le but. Autre exemple, la considération éthique comporte par exemple le fait de ne pas forcer l'autre, de ne pas l'empêcher de suivre ses propres motivations.

L'ensemble de ces critères et de ces travaux permet de se faire une idée générale de la couverture et de la complexité des mécanismes à l'œuvre en dialogue. En contrepartie, ils conduisent au constat qu'implémenter un système de DHM coopératif n'est pas chose aisée, et que les liens entre les théories et les exemples concrets ne sont pas évidents à faire. Certains concepteurs de systèmes redéfinissent leur notion de coopération, avec des préoccupations plus proches d'aspects techniques ou de critères directement implémentables. Un des traits de la coopération dans (Luzzati, 1995, p. 39) relève par exemple de la gestion par le système de ses propres erreurs. La coopération se matérialise avant tout par le choix des réponses du système aux énoncés de l'utilisateur. Si l'on prend comme exemple « combien de temps avec ce chemin ? », une première réponse du système peut être « deux heures ». Dans ce cas, le système, qualifié éventuellement de communicant, se contente de répondre à la question posée, sans rien dire de plus ni de moins que la valeur demandée. Une deuxième réponse possible peut être « deux heures, à cause d'un changement à Versailles ». Ici, la réponse à la question est d'une part évaluée, d'autre part expliquée. Elle est évaluée par

le système qui, sur la base de critères tels que la durée moyenne d'un aller entre Palaiseau et Paris, constate qu'une durée de deux heures est longue, et risque de ce fait de déplaire à l'utilisateur. Elle est alors expliquée par le système, qui cherche et décrit la raison principale de ce temps élevé. Avec une telle réponse, le système peut être considéré comme coopératif. Cependant, il reste dans son rôle de système chargé de satisfaire les requêtes de l'utilisateur. Or, une troisième réponse possible outrepasse ce rôle : en répondant « deux heures à cause d'un changement à Versailles, mais si vous passez par Meudon, vous en aurez pour cinquante minutes », le système fait preuve d'un degré supplémentaire de coopération – on peut alors le qualifier de collaboratif – consistant à proposer à l'utilisateur un changement d'orientation, à savoir essayer un chemin différent de celui initialement choisi. C'est ce changement d'orientation qui, si l'utilisateur l'accepte, contribue à la coconstruction d'un but commun.

8.1.2. *Tours de parole et aspects interactifs*

La tâche de réservation de billets de train peut conduire à des dialogues naturels, comme dans l'exemple de l'introduction, mais aussi à des situations plus proches de la caricature d'un certain type de communication entre l'humain et la machine : « je voudrais aller à Paris », « quel jour ? », « ce serait demain », « quelle heure ? », « disons vers neuf heures », « quelle classe ? », « en première, s'il vous plaît », etc. L'utilisateur peut produire tous les énoncés imaginables, le caractère artificiel relève ici du comportement du système. En ne produisant que des questions, et toujours avec la même forme de phrase, le système a beau faire avancer la tâche, il ne contribue pas du tout aux aspects linguistiques qui font un dialogue naturel en langage naturel, et on ne peut pas dire que son comportement soit coopératif d'un point de vue interactionnel. Un système a donc tout intérêt à ajouter de l'interactivité, ce qui peut se faire en variant les formes des phrases, mais aussi en variant les contenus des interventions, et en ajoutant par exemple un court énoncé chargé de faire des liens entre l'intervention de l'utilisateur et celle du système. (Denis, 2008, p. 67) donne ainsi plusieurs réponses possibles à la requête « je veux aller à Paris » de l'utilisateur :

– « quand désirez-vous partir ? » : réaction sur l'un des paramètres manquants de la requête, c'est-à-dire initiation d'une contribution pertinente ;

– « d'accord. Quand désirez-vous partir ? » : accusé de réception puis réaction par rapport à la requête ;

– « à Paris. Quand désirez-vous partir ? » : répétition du seul paramètre indiqué par l'utilisateur, ce qui permet d'une part de procéder à l'accusé de réception, d'autre part de permettre à l'utilisateur de vérifier que sa demande a bien été comprise, en tout cas sur ce point (si ce n'est pas le cas, l'utilisateur peut ainsi réagir tout de suite) ;

– « vous voulez aller à Paris. Quand désirez-vous partir ? » : répétition démonstrative, qui peut prendre plusieurs formes de paraphrase, de l'énoncé complet de l'utilisateur, ce qui fait office d'accusé de réception, d'accusé de compréhension, et permet ici aussi à l'utilisateur de réagir tout de suite en cas d'erreur.

On reste ici dans des cas où le système prend la parole une fois que l'utilisateur a terminé son énoncé, et réciproquement, autrement dit dans des interactions alternées. Or les travaux de l'analyse conversationnelle ont souligné la diversité dans le dialogue humain des phénomènes d'organisation des tours de parole, d'organisation des séquences ou segments, d'organisation des réparations, etc. (Sacks *et al.*, 1974). Des corpus enregistrés ont montré qu'un dialogue n'était pas seulement une suite ordonnée d'énoncés. Un dialogue entre deux interlocuteurs implique qu'il y a deux canaux de communication opérant de manière simultanée : le canal principal, qui est celui du locuteur conduisant le dialogue à un instant donné, et le canal secondaire (*backchannel*), occupé par exemple par des indices d'écoute fournis par l'interlocuteur. Ces indices peuvent prendre la forme de sons non lexicaux ou d'énoncés courts avec un acte de langage montrant la bonne réception, voire la compréhension, comme « hum », « oui » ou « ah bon ». Ils peuvent aussi prendre la forme d'une complétion de l'énoncé du locuteur, ou encore d'une répétition d'un fragment de celui-ci. Dans tous les cas, comme ils sont brefs et ne constituent pas vraiment un tour de parole, ils n'empêchent pas le locuteur de continuer son intervention. Ils peuvent donc entraîner une superposition des voix. En DHM, certains travaux ont intégré pour le comportement du système la production de tels énoncés (Ward et Tsukahara, 2003 ; Edlund *et al.*, 2005), mais les problèmes techniques sont nombreux : la reconnaissance de la parole fonctionne alors même que le système produit du bruit, ce qui peut provoquer des baisses de performances ; la production d'un énoncé de contrôle peut survenir au moment même où l'utilisateur allait laisser la parole, ce qui peut provoquer un petit instant de flottement, et, d'une manière générale, le moment de production d'un énoncé de contrôle n'est pas forcément le plus pertinent, à cause par exemple d'un léger décalage temporel par rapport aux moments les plus propices que sont les TRP (voir paragraphe 2.2.1). Il n'en reste pas moins que doter le système de la capacité à exploiter le canal secondaire augmente le réalisme du dialogue, de même que le doter de la capacité à occuper le terrain quand l'utilisateur ne dit rien, c'est-à-dire à produire un message de relance quand l'utilisateur ne répond pas, à produire un message de maintien du dialogue quand l'utilisateur ne sait plus quoi dire, et même à produire un message de mise en attente dès qu'un traitement risque de durer plusieurs secondes.

8.1.3. *Interprétation et inférences*

Pour réagir de manière pertinente à un énoncé, il faut bien sûr le comprendre, et comprendre aussi ses implicitations et explicitations (voir section 5.3). Déterminer les contenus sémantiques et les actes de langage est un premier pas, mais il faut aussi faire les bonnes inférences, qui se fondent sur l'énoncé et son contexte et permettent à l'interlocuteur humain de comprendre tout de suite les sous-entendus et autres contenus implicites. (Grice, 1975) a là aussi proposé une terminologie et des principes généraux qui permettent de mieux comprendre les phénomènes dialogiques. En utilisant le terme implicature pour les inférences pragmatiques, il distingue les implicatures conventionnelles, déclenchées par une utilisation conventionnelle de la langue, les

implicatures conversationnelles particulières, déclenchées par la mise en relation de l'énoncé dans le contexte de son énonciation, et les implicatures conversationnelles généralisées, déclenchées de manière contextuelle sans l'aide des éléments linguistiques de l'énoncé. Ces types d'implicatures permettent d'expliquer la compréhension et la coopération dans le dialogue. C'est quand le locuteur semble violer l'une des maximes qu'il faut se pencher sur les implicatures. Calculer une implicature revient à déterminer ce que le locuteur a implicitement supposé, afin de préserver le principe de coopération (Denis, 2008, p. 18).

Par ailleurs, un dialogue ne consiste pas seulement en un échange d'informations, qu'elles soient explicites ou implicites. Il peut aussi s'agir de négocier, de convaincre son interlocuteur, de prouver quelque chose. Cet aspect, il est vrai peu présent en DHM, fait également l'objet de travaux qui conduisent à identifier des actes de langage en quelque sorte enrichis par une mise en perspective. Parmi les trois types que nous avons vus dans le chapitre 7, l'assertion est par exemple peu explicite quant à l'intention sous-jacente. (Baker, 2004) souligne que dans le dialogue de négociation, il ne s'agit pas tant de produire des assertions, que de faire des propositions ou des offres. Une assertion peut apporter un argument, et c'est là aussi une facette du dialogue peu prise en compte dans les systèmes de DHM.

8.1.4. *Dialogue, argumentation et cohérence*

Des chercheurs en linguistique et en pragmatique comme J. Moeschler ont beaucoup étudié le dialogue argumentatif et ont fait des rapprochements avec la théorie des actes de langage, voire la théorie de la pertinence. Ils définissent ainsi une catégorie particulière d'actes de langage, les « actes d'argumentation » qui sont réalisés dès lors qu'un énoncé est destiné à servir une certaine conclusion (Moeschler, 1985, p. 189). La forme linguistique pouvant elle-même comprendre des instructions argumentatives, par exemple les connecteurs argumentatifs « mais », « par contre », « donc », « parce que » ou « décidément », l'interprétation d'un énoncé ajoute une dimension argumentative consistant à identifier l'orientation (sémantique) et le type d'acte (pragmatique).

Dès que plusieurs énoncés sont en jeu, il s'agit en outre de les relier les uns aux autres par des relations argumentatives, qui permettent de structurer le dialogue en détectant comment tel énoncé apporte un argument dans le sens de telle affirmation. Tout ceci ajoute une dimension d'analyse dont nous n'avions pas encore parlé, et qui consiste à ajouter un ensemble d'étiquettes supplémentaires à chaque énoncé, cet ensemble comportant une indication sur l'acte et l'orientation de l'énoncé, et sur ses relations avec les énoncés précédents et suivants. C'est d'ailleurs là que l'on peut constater que les implémentations pour le DHM ne sont pas nombreuses : cette dimension d'analyse, un peu comme celle du modèle hiérarchique de l'école de Genève, s'applique surtout *a posteriori*, donc plutôt sur corpus qu'en temps réel dans un

système effectif. Par ailleurs, nos exemples de réservation de billets n'impliquent pas vraiment de dimension argumentative, et c'est surtout un enjeu pour certains systèmes de DHM, ludiques ou en domaine ouvert.

La gestion d'un dialogue argumentatif est bien décrite d'un point de vue théorique, notamment par (Moeschler, 1985), qui propose un ensemble de stratégies de dialogue telles que : la négociation anticipée, stratégie argumentative visant à anticiper les contre-arguments que l'on pourrait opposer et à les réfuter tout de suite (« vous pourrez croire que c'est plus cher en passant par Versailles, mais non... ») ; la négociation factuelle, stratégie discursive ayant pour objet la mise en accord sur certains faits décisifs pour la poursuite de l'interaction ; la négociation interactionnelle, stratégie visant à imposer une image de soi et de l'autre lors de l'interaction ; la négociation métadiscursive, visant à donner des indications permettant d'interpréter rétroactivement la fonction d'une intervention ; ou encore la négociation méta-interactionnelle, visant à définir les droits et les obligations des interlocuteurs. Pour appliquer correctement ces stratégies, il est nécessaire d'identifier les successions d'actes discursifs, ce que J. Moeschler propose de faire *via* la notion de mouvement discursif, sorte de structure qui réunit plusieurs actes argumentatifs, et, sur la base notamment de leur orientation, permet d'identifier les cas de concession ou de conclusion argumentatives.

Par ailleurs, la gestion d'un dialogue argumentatif fait intervenir des possibilités d'interruption : contrairement à une requête ou à une narration, pour lesquelles l'interlocuteur attend généralement la fin de l'énoncé, une intervention argumentative peut susciter l'interlocuteur à couper la parole au locuteur, de manière à étouffer un argument avant même qu'il ne soit complètement énoncé. C'est ce que montre (Dessalles, 2008, p. 17), en partant du constat – suite à une étude sur corpus – que les interlocuteurs sont souvent capables d'anticiper la nature de l'argument avant même qu'il ne soit totalement exprimé. L'interruption est ainsi une stratégie de dialogue, certes un peu brutale, mais qui peut être prise en compte par un système de DHM. Bien que ces stratégies de dialogue s'appliquent théoriquement à tout type de dialogue, notons toutefois qu'elles restent plus pertinentes dans le cas de dialogues de négociation et d'argumentation que dans le cas de dialogues de commande ou de renseignement.

Un dernier aspect qui a aussi une importance particulière dans les dialogues de ce type est la gestion de la cohérence et de la cohésion. Il s'agit d'identifier les relations entre plusieurs énoncés, un peu comme le font les relations argumentatives. (Prévot, 2004) fait le point sur cette question qui a suscité de nombreux travaux en linguistique, et distingue une cohérence sémantique, surtout spatiale et temporelle, une cohérence implicite, regroupant des aspects liés aux intentions, au propos et à certaines conventions de dialogue, et des indices de surface qui définissent la cohésion : structure informationnelle, ellipses, anaphores, chaînes de coréférence, c'est-à-dire des phénomènes linguistiques qui dépassent les frontières de la phrase. (Moeschler,

1985, p. 190) ajoute la cohérence argumentative, qui caractérise un discours ou dialogue dans lequel les instructions données par les connecteurs argumentatifs sont satisfaites, toute contradiction argumentative étant résolue. Tous ces aspects permettent à un système de DHM de gérer le dialogue en connaissance de cause, c'est-à-dire en exploitant un indicateur de cohérence et de cohésion, et de disposer d'indications supplémentaires pour son choix de réponse, processus que nous allons voir maintenant.

8.1.5. *Choix d'une réponse*

Quand l'utilisateur pose une question au système, le choix de la réponse de celui-ci s'impose : soit il connaît la réponse et la donne, soit il ne la connaît pas et s'en excuse, tente de diriger le dialogue vers une autre voie, ou, éventuellement, demande à l'utilisateur de reformuler sa question. Dans le contexte d'un dialogue construit sur plusieurs interventions, une question peut impliquer beaucoup plus de complexité. (Luzzati, 1995, p. 61) montre que dans le cas de la vente de billets de train, il est fréquent pour un guichetier humain et donc pour un système de produire des réponses « maximalistes », c'est-à-dire des réponses qui comportent systématiquement plus de renseignements qu'il n'en est demandé, pour des raisons d'efficacité, notamment pour prévenir d'éventuelles questions supplémentaires. Cela rejoint un peu le principe de répétition démonstrative vu au paragraphe 8.1.2. Afin de mieux déterminer quand le système peut produire ce type de réponse qui viole l'une des maximes de quantité de Grice, il est utile de caractériser les différents types de questions possibles. (Van Schooten *et al.*, 2007) étudient l'importance des questions successives, et sont amenés à proposer une typologie des questions. Parmi les types de questions, on a déjà vu des exemples de questions fermées, pour lesquelles les réponses possibles sont « oui », « non » ou « je ne sais pas », des exemples de questions ouvertes, pour lesquelles la réponse est une valeur propositionnelle, comme une expression référentielle, une quantité ou une entité nommée, mais nous n'avions pas encore rencontré de questions telles que « comment fonctionne une réservation ? ». Or cette question nécessite une explication, ce qui peut être très complexe pour un système de DHM.

Quand l'utilisateur donne un ordre au système, le choix de la réponse s'impose également : en même temps que le système exécute l'ordre, du moins si toutes les conditions sont remplies (voir paragraphe 6.2.1), il peut produire un énoncé qui, d'une part annonce l'action en cours d'exécution (surtout si celle-ci n'est pas visible), et d'autre part permet à l'utilisateur d'enchaîner, de manière à poursuivre le dialogue naturel. Suite à l'ordre « réserve-moi un billet pour Paris, demain à huit heures en première classe », un énoncé du système tel que « d'accord » ou « voilà, c'est fait » est probablement insuffisant, car il contribue à clore le dialogue. Un énoncé incluant une relance s'avère ainsi plus pertinent, comme « c'est fait, voulez-vous un autre voyage ? », et ne met pas l'utilisateur dans la situation de ne plus savoir quoi dire.

Quand l'utilisateur produit une assertion, comme par exemple « je n'ai pas la carte senior », c'est là que le système doit faire preuve de sa capacité à gérer un dialogue : une assertion apporte une information *a priori* nouvelle pour le système, sinon elle ne sert à rien, et c'est cette nouveauté qui doit déclencher des inférences. Sur la base de ces inférences, et donc sur la base de ce qui a déjà été dit et des connaissances communes aux deux interlocuteurs, le système doit comprendre si l'assertion vient combler un manque qui bloquait une situation, situation que les inférences débloquent et qui vont permettre au système de savoir quoi répondre, ou, au contraire, si l'assertion vient inciter le système à proposer quelque chose, comme « nous allons voir si vous pouvez l'obtenir. Quel âge avez-vous ? ».

Par ailleurs, un énoncé de l'utilisateur, quel que soit son acte de langage, peut amener le système à réagir d'une manière imprévue, par exemple lorsqu'il est incapable de résoudre une ambiguïté sur un référent ou tout simplement de comprendre l'énoncé. Là aussi, plusieurs stratégies de dialogue sont possibles. (Denis, 2008, p. 43) s'intéresse particulièrement à la détection des problèmes et à la robustesse des systèmes, et en vient à proposer des stratégies de dialogue dans les cas où deux problèmes s'additionnent. Ainsi, une vision classique de la gestion d'une ambiguïté consiste à choisir parmi les alternatives, quitte à se tromper, plutôt que d'engager un sous-dialogue de clarification qui risque d'une part de donner une image négative du système, d'autre part de faire diminuer les chances d'arriver rapidement à la satisfaction de la tâche. Rattraper une erreur sur un référent peut en effet s'avérer plus rapide qu'un sous-dialogue de clarification. A cette vision, A. Denis ajoute un phénomène qu'il a observé en corpus et qui n'est que très peu étudié en DHM, à savoir le cas où la demande de clarification elle-même est la source d'une nouvelle incompréhension, ou d'une divergence d'interprétation entre l'utilisateur et le système, ce qui peut conduire à une situation inextricable. L'existence de cette possibilité renforce la pertinence de la stratégie consistant à forcer un choix.

8.2. Aspects techniques de la gestion du dialogue

8.2.1. *Gestion et contrôle du dialogue*

Décomposer en tâches le processus de gestion du dialogue n'est pas chose facile. De nombreuses approches se sont succédées, et il s'avère difficile de les positionner les unes par rapport aux autres compte tenu de la diversité des paramètres exploités et des recouvrements entre processus. D'une manière très générale, la « gestion du dialogue » comporte trois phases plus ou moins imbriquées les unes dans les autres : premièrement le contrôle du dialogue, qui cherche à gérer le processus interactif de manière à déterminer un type de réaction suite à une succession d'énoncés structurée, deuxièmement la modélisation du contexte de dialogue, qui s'intéresse à l'historique du dialogue et aussi à la manière dont les contenus des énoncés sont ancrés, et

troisièmement l'initiative qui ajoute un comportement particulier aux considérations précédentes (McTear, 2004 ; Jokinen et McTear, 2010).

Le contrôle du dialogue, qui fait l'objet de cette section, a été l'objet d'un très grand nombre de travaux (Pierrel, 1987 ; Sabah, 1989 ; Carberry, 1990 ; Luzzati, 1995 ; Traum et Larsson, 2003, etc.). En caricaturant un peu, ce sont tout d'abord des méthodes à états finis qui ont été mises en œuvre : qu'il s'agisse d'automates à états finis ou de « grammaires de dialogue », le principe est de déterminer *a priori* l'ensemble des situations possibles, et les moyens de passer d'une situation à une autre. Ce type d'approche fonctionne très bien pour des dialogues où l'initiative est toujours du même côté, par exemple du côté du système, dans la mesure où ce sont alors les énoncés du système (questions dans les systèmes très directifs, différents actes de langage dans des systèmes plus souples) qui sont affectés aux états, les réponses de l'utilisateur étant prises en compte par les transitions (Jurafsky et Martin, 2009, p. 863). Viennent ensuite les méthodes à base de patrons (à trous), qui permettent de reconnaître des situations sans figer l'initiative d'un côté ou de l'autre.

Suivent alors les méthodes à base d'un état d'information, c'est-à-dire les méthodes qui ajoutent une mémoire, contenant un peu ce que l'on veut, sachant que cette mémoire va aider à déterminer les suites possibles du dialogue. Selon les auteurs, l'état d'information va contenir l'historique du dialogue, le terrain commun, un modèle avec des états mentaux, un modèle de l'utilisateur, etc. L'important, c'est d'exploiter et de mettre à jour des données qui vont se comporter un peu comme des variables globales dans le gestionnaire de dialogue. C'est ainsi que l'approche de la planification va connaître un grand succès, avec les travaux initiaux que sont (Cohen et Perrault, 1979) et (Allen et Perrault, 1980) : le but est de reconnaître et de planifier des plans, les actes de langage étant planifiés au même titre que les actions (les actes de langage du locuteur font partie d'un plan que l'interlocuteur doit découvrir pour répondre de manière pertinente), sur la base d'une modélisation des états mentaux des interlocuteurs. Afin de concilier dialogue naturel et satisfaction de la tâche, on peut distinguer deux types de plans gérés en parallèle : les plans du discours et les plans du domaine. D'une manière générale, les plans autorisent de très nombreuses possibilités en termes de contrôle du dialogue, voir l'article de N. Maudet dans (Gardent et Pierrel, 2002).

Arrivent ensuite les théories de l'action conjointe (voir paragraphe 8.2.2 avec la notion de terrain commun), puis le contrôle du dialogue s'inspirant de la théorie des jeux, chaque énoncé étant un coup joué, pour lequel le locuteur cherche à maximiser son gain (Caelen et Xuereb, 2007). Enfin, plus récemment, on voit apparaître des techniques d'apprentissage automatique pour le contrôle du dialogue, avec des modèles de type MDP (*Markov Decision Process*) ou POMDP (*Partially Observable MDP*), qui étendent celui de l'état d'information en ajoutant un moyen probabiliste de décider de l'action future en fonction de l'état courant (Jurafsky et Martin, 2009, p. 883). Dans un même ordre d'idée, (Singh *et al.*, 2002) utilise l'apprentissage par renforcement

pour obtenir un ensemble de décisions optimales. Les règles de décision sont affinées en procédant à plusieurs milliers d'échanges entre le système et un utilisateur simulé.

On le voit, le contrôle du dialogue peut impliquer de nombreuses techniques. Ajoutons que l'implémentation peut faire intervenir en parallèle d'autres données, comme par exemple une analyse des thèmes abordés, de manière à diriger la réaction du système plutôt vers les thèmes en cours que vers les thèmes abandonnés. C'est l'approche de (Vilnat, 2005), qui fait coopérer trois sous-modules d'analyses pragmatiques distinctes :

– une interprétation « thématique » qui gère la cohérence globale des thèmes abordés durant le dialogue ;

– une analyse « intentionnelle » qui fournit une représentation fonctionnelle du dialogue où les rôles des diverses interventions sont explicitées, voir la présentation des approches fondées sur la structure intentionnelle dans l'article de N. Maudet dans (Gardent et Pierrel, 2002) ;

– une gestion de l'interaction qui permet de réagir aux différents types d'incompréhension en permettant au dialogue de rester efficace.

De son côté, (Rosset, 2008) ajoute des stratégies dépendant de choix ergonomiques et propose un modèle de dialogue construit autour d'un ensemble de phases : acquisition (obtention des informations nécessaires à la satisfaction de la tâche), négociation, navigation, postacceptation (transition vers une négociation, une navigation ou la clôture du dialogue), et métatraitement (repérage et traitement des erreurs), voir le système Arise (Lamel *et al.*, 2003).

8.2.2. *Modélisation de l'historique du dialogue*

La modélisation du contexte du dialogue prend elle aussi une grande variété de formes dans les systèmes théoriques et dans les systèmes implémentés. Un premier aspect consiste à fournir des éléments contextuels pour la compréhension complète d'un énoncé, après les analyses sémantiques et pragmatiques de celui-ci, donc au niveau du rôle de l'énoncé dans le déroulement du dialogue, et par exemple dans la satisfaction de la tâche. Parmi ces éléments, on peut ainsi trouver des hypothèses sur les croyances ou les désirs du locuteur, ce qui permet de mettre en perspective le contenu sémantique de son énoncé. On trouve également toutes les informations relevant de ce qui a déjà été dit au cours du dialogue, donc de l'historique. C'est là un deuxième aspect de la modélisation du contexte du dialogue : formaliser et sauvegarder dans une structure de données les résultats des interprétations contextuelles, ainsi que faire apparaître dans cet historique une structure décrivant le déroulement du dialogue, par exemple la voie suivie pour résoudre la tâche.

Une notion qui a émergé et qui a suscité de nombreuses propositions dans les travaux théoriques sur le dialogue est celle de « terrain commun », c'est-à-dire d'ensemble des informations partagées par le locuteur et son interlocuteur, soit parce qu'elles sont mutuellement manifestes, par exemple parce qu'elles ont été verbalisées, soit parce qu'elles se déduisent de ce qui a été dit : par exemple, quand le système répond « à Paris. Quand désirez-vous partir ? », il est manifeste pour les deux interlocuteurs que l'utilisateur désire aller à Paris et que le système est au courant de ce désir. Le terrain commun devient ce que les interlocuteurs construisent au fur et à mesure du dialogue. C'est ainsi que le langage est vu comme une « action conjointe » (Jurafsky et Martin, 2009). Plus précisément, on peut distinguer un terrain commun communautaire, lorsque les connaissances partagées vont au-delà de l'interaction entre les deux interlocuteurs, et un terrain commun plus personnel, dont les connaissances ne valent que pour les deux interlocuteurs (Denis, 2008, p. 45). Pour faire partie du terrain commun, un énoncé du locuteur doit être ancré par l'interlocuteur. On distingue ainsi le « processus d'ancrage », processus par lequel les interlocuteurs mettent à jour le terrain commun, et le « critère d'ancrage », critère que les interlocuteurs cherchent à atteindre, avec une volonté de croyance mutuelle de compréhension. Plusieurs modèles d'ancrage ont été proposés. Les premiers, qui ne font pas encore ces distinctions, mettent à jour automatiquement le terrain commun, à chaque énoncé. En réaction, (Clark et Schaefer, 1989) propose le modèle des « contributions de discours », dans lequel l'ancrage ne peut se réaliser que lorsque le critère d'ancrage est atteint. D'autres propositions tentent alors de poursuivre cette voie pour arriver à des modèles implémentables, notamment le modèle des actes d'ancrage, avec ses neuf degrés d'ancrage, d'« inconnu » et « non compris » à « accepté » (Traum et Hinkelman, 1992), puis le modèle des croyances faibles qui apporte une modélisation explicite des croyances de compréhension nécessaires à l'ancrage : il considère que les interlocuteurs effectuent des hypothèses sur la compréhension de leur partenaire et que la confirmation de ces hypothèses permet d'atteindre le critère d'ancrage. Un mécanisme de renforcement permet alors de transformer une croyance mutuelle faible en croyance mutuelle, voir (Denis, 2008). Enfin, une tendance plus récente consiste à intégrer aux modélisations de l'ancrage des aspects numériques, avec par exemple le calcul d'un score qui permet de caractériser l'utilité à ancrer une information. Comme souvent, cette approche numérique vient en complément des approches symboliques.

Si l'on reprend maintenant notre exemple fétiche et la distinction des trois actes de langage principaux de la théorie de la pertinence, la réception par le système de l'énoncé « je voudrais aller à Paris » conduit à l'ensemble des raisonnements suivants : l'analyse du sens de l'énoncé conduit à l'identification de l'expression d'une intention. Le système retient donc cette intention, qui va lui permettre de planifier ses prochaines actions. Les rôles du système sont d'informer son utilisateur, ici sur les moyens de parvenir à Paris, et de lui vendre des billets de train. L'énonciation de « je voudrais aller à Paris » par l'utilisateur est très probablement l'expression d'un désir, celui que le système l'aide à aller à Paris. Satisfaire ce désir passe par l'identification, parmi

l'ensemble des moyens possibles pour aller à Paris, de celui qui va satisfaire le plus l'utilisateur. Pour l'instant, le système ne connaît pas les préférences de celui-ci, en revanche il peut déjà faire un tri dans sa base de données des trajets afin de lui proposer les trajets les plus pertinents.

A ce stade, le système décide donc de faire une proposition à l'utilisateur, en incluant plusieurs alternatives de manière à laisser choisir celui-ci. Le système décide également de manifester un accusé de réception, et d'exploiter la multimodalité, c'est-à-dire d'afficher à l'écran quelques trajets menant à Paris (avec une mise en relief des gares parisiennes impliquées), en même temps qu'il prononce une phrase relativement courte, « voici les trajets possibles », phrase qui n'inclue pas les informations sur les trains (trop longues à verbaliser). De cette manière, l'ancrage de l'énoncé de l'utilisateur est effectué par la mise en relief des gares parisiennes, et la prise en compte des états mentaux de l'utilisateur est montrée par l'énoncé oral. On peut considérer que cette réponse est coopérative : ni trop courte, ni trop longue, pertinente compte tenu de l'énoncé de l'utilisateur, impliquant la reconnaissance d'une intention et d'un acte de langage indirect, elle satisfait globalement les critères théoriques décrits dans les sections précédentes. En plus de la génération de cette réponse, le système sauvegarde dans sa modélisation de l'historique du dialogue les états mentaux identifiés, ainsi que cette première matérialisation d'un plan consistant à identifier un moyen d'aller à Paris, afin de vendre le billet de train correspondant.

L'énoncé suivant de l'utilisateur, « combien de temps avec ce chemin qui semble être le plus court ? », pose tous les problèmes que nous avons décrits dans les chapitres de la deuxième partie de ce livre, et soulève ici de nouvelles questions : qu'apporte cet énoncé par rapport aux états mentaux et au plan amorcé ? En interprétant l'énoncé, le système se rend tout d'abord compte que son intervention a été comprise : l'utilisateur a bien vu les alternatives proposées, puisqu'il pose une question sur l'une d'entre elles. Il y a donc ancrage de la réponse du système. Par ailleurs, la question amène à l'identification dans la base de données applicative d'une propriété d'un trajet, et l'évaluation du commentaire amène le système à effectuer une comparaison sur des propriétés de trajets. Cette évaluation est déclenchée parce que le système a déterminé que le commentaire était en fait une question, *via* un acte de langage indirect. Si cette question indirecte s'avère pertinente, il faut pouvoir y répondre, et pour cela il est nécessaire de comparer les durées. A l'issue de cette comparaison, le système sait qu'effectivement, le chemin désigné par l'utilisateur est le plus court. Il peut donc répondre à la question indirecte. Comme celle-ci a été exprimée sous la fome d'un commentaire, le système gère en parallèle la croyance associée, à savoir que l'utilisateur croit que le chemin est probablement le plus court. A ce stade, les états mentaux de l'utilisateur ont été mis à jour, l'historique du dialogue également, et le système fait face à plusieurs faits : il connaît la réponse à la question directe, il connaît la réponse à la question indirecte, et il cherche toujours à satisfaire l'utilisateur en l'incitant à faire un choix, si possible rapidement. C'est ce qui peut conduire le système, dans l'exemple tel qu'il apparaît dans l'introduction, à générer la réponse « vingt minutes ». Cette réponse un peu courte

correspond à la décision de répondre à l'acte premier, rapidement et efficacement, et d'ignorer l'acte second qu'est le commentaire. Ce choix a été fait pour deux raisons : d'une part parce que le système croit que la réponse « vingt minutes » va satisfaire l'utilisateur (c'est une durée courte, donc satisfaisante), d'autre part parce qu'il considère que confirmer la croyance, ici vraie, n'est de ce fait pas indispensable. Enfin, en plus de la génération de la réponse, le système met à jour l'historique du dialogue avec une description de tout ce que l'on vient de voir, et avec une matérialisation de l'avancée du dialogue, notamment du fait que la réponse du système continue à focaliser le dialogue uniquement sur le chemin sur lequel l'utilisateur s'est focalisé. On retrouve ici certains des aspects discutés dans le chapitre 5.

Pour implémenter ce type de processus, nous voyons que les analyses linguistiques et pragmatiques sont essentielles, avec par exemple l'identification des actes de langage indirects, qu'une identification des états mentaux de l'utilisateur apporte des possibilités de raisonnement, et que la gestion d'un historique du dialogue correctement structuré est elle aussi indispensable. A titre d'exemple, (Vilnat, 2005) propose la notion de page de l'historique, qui contient pour chaque intervention l'identifiant du locuteur, les représentations sémantiques et pragmatiques, le topique (ou propos) concerné, le but concerné, l'état de la structure du dialogue, l'état des variables interactionnelles, l'état du plan en cours de développement, etc. On le voit, l'historique est une structure complexe, multiforme et multifonction.

Si l'on reprend les trois actes de langage principaux, on peut considérer les processus et structures suivants :

– « dire que » : le locuteur exprime une assertion dans le but de faire savoir quelque chose au système. Celui-ci met à jour sa base de connaissances qui fait partie du terrain commun et que l'on peut désigner par CG (*Common Ground*). En effet, en fournissant une information au système, le locuteur contribue à rendre cette information mutuellement connue et manifeste (du moins après la prise en compte du processus d'ancrage) ;

– « dire de » : le locuteur exprime un ordre dans le but de faire faire quelque chose au système. Celui-ci met à jour une liste d'actions à effectuer, que l'on peut désigner par TDL (*To Do List*). Il s'agit d'une sorte de pile (ou tas) répertoriant au fur et à mesure du dialogue les choses à faire, en enlevant un item dès que l'action correspondante est effectuée. Gérer ce type de liste permet au système de savoir ce qui lui reste à faire, sans que cela soit associé au traitement de l'énoncé courant, mais, au contraire, en rendant possible l'exécution d'une action plusieurs tours de parole après la requête de l'utilisateur ;

– « demander » : le locuteur exprime une question dans le but de savoir quelque chose de la part du système. Celui-ci met à jour une liste d'interrogations auxquelles il doit répondre, que l'on peut désigner par QUD (*Questions Under Discussion*, ou *Questions Under Debate*), c'est-à-dire une structure similaire au TDL, chargée cette fois de répertorier et de gérer tout au long du dialogue les questions posées.

Selon les approches, seule l'une des trois structures peut être mise à jour lors du traitement d'un acte de langage, ou, au contraire, les mises à jour multiples sont autorisées. Dans ses modèles successifs, (Ginzburg, 2012) a détaillé le fonctionnement de sous-structures et proposé des structures complémentaires, par exemple LM (*Latest Move*), SG (*Shared Ground*), *Facts*, *Pending* (structure éphémère), etc. D'autres approches mettent en avant une structure répertoriant les engagements du locuteur, CS (*Commitment Store*), une structure spécifique aux énoncés saillants, Sal-Utt (*Salient Utterances*), ou encore un sous-ensemble de la structure QUD, chargé d'un type particulier de questions, *Issues*, suite à une distinction entre plusieurs types de questions selon leurs fonctions dans le dialogue (Denis, 2008).

8.2.3. *Gestion du dialogue et gestion de la multimodalité*

L'exemple précédent comportait une initiative de dialogue (voir paragraphe 8.2.1) : quand le système décide de répondre « voici les trajets possibles » en affichant un ensemble d'alternatives sur la scène visuelle, il fait un choix sur ce qu'il va dire et afficher. L'affichage des trajets possibles et de la direction de Paris relève d'un acte de type « dire que », sauf qu'il s'agit d'un « dire visuellement » plutôt que d'un « dire ». L'énonciation de « voici les trajets possibles » relève d'un « dire que ». Le système aurait pu faire bien d'autres choix, par exemple poser une question telle que « est-ce qu'un trajet avec un changement à Meudon vous irait ? », qui permet de faire avancer la tâche en testant l'une des alternatives. L'énoncé se serait alors caractérisé par l'acte de langage « demander ». Le choix d'acte de langage en réaction à un acte de l'utilisateur fait ainsi partie de la gestion du dialogue. Dans le contexte multimodal que nous étudions ici, l'acte choisi est un acte multimodal composite : le système veut indiquer plusieurs choix possibles à l'utilisateur, et il le fait d'une certaine manière qui implique fortement des fonctionnalités de présentation d'information multimédia.

La gestion du dialogue comporte d'autres aspects que notre exemple n'illustre pas, par exemple l'explicitation des conditions qui permettent au système d'abandonner ses buts, de s'engager dans une voie visant à satisfaire un but en particulier, et d'une manière générale d'explorer les interactions entre buts, croyances et intentions. C'est ce que montre (Cohen et Levesque, 1990) avec un exemple d'un robot qui dit qu'il va apporter quelque chose, qui ne le fait pas, et qui explique ensuite qu'il a trouvé autre chose à faire, cet exemple servant à illustrer une théorie de l'action rationnelle. Par ailleurs, la gestion du dialogue inclut également la capacité à gérer une incompréhension (voir paragraphe 8.1.5 avec le même type de préoccupations pour la gestion des ambiguïtés). Le système peut choisir de résoudre l'incompréhension sans faire appel à l'utilisateur (on parle alors de robustesse interne), ou de résoudre l'incompréhension grâce à l'utilisateur, c'est-à-dire en déclenchant un sous-dialogue de clarification (robustesse externe). Comme le montre (Denis, 2008, p. 35), les deux approches sont complémentaires : on ne peut pas se contenter d'une très bonne robustesse interne couplée avec une faible robustesse externe, car le dialogue a aussi pour fonction de

discuter de ce qui ne va pas, et car certains problèmes nécessitent une demande de clarification. Mais on ne peut pas se contenter de l'inverse : un système qui s'appuie systématiquement sur l'utilisateur pour résoudre ses problèmes d'interprétation finit par être agaçant. C'est le cas par exemple de quelques-uns des premiers systèmes dotés de capacités d'apprentissage automatique à la volée, qui, pour être sûrs d'avoir bien intégré un nouveau terme ou tout simplement d'avoir bien compris, posent une question fermée quasiment à chaque tour de parole.

La gestion du dialogue multimodal, et par exemple du dialogue avec un système d'information, c'est-à-dire un système dédié à la présentation d'informations complexes comme des données géographiques, comporte encore d'autres aspects spécifiques à la gestion de la quantité d'informations. Si le système décide de présenter les détails de trente trajets de train, ou de montrer une carte géographique annotée avec des éléments de réponse liés à la requête de l'utilisateur, il est nécessaire de contrôler la manière de transmettre ces informations. Cela peut se faire en planifiant, c'est-à-dire en répartissant la transmission sur plusieurs tours de parole. Cela peut aussi se faire en répartissant l'information sur plusieurs modalités de communication, un peu comme « voici les chemins possibles » mais surtout comme nous le verrons dans le chapitre 9. Sur ce critère de quantité d'informations, proche de la notion de charge cognitive, (Horchani, 2007) présente et modélise trois stratégies de dialogue suite à une requête telle que « je voudrais aller à Paris » :

– l'énumération : dans le cas où le nombre de solutions reste raisonnable (mais encore faut-il décider d'un seuil, surtout que ce seuil dépend de la quantité moyenne d'informations contenue dans une solution), le système présente une liste regroupant toutes les solutions, liste pouvant être verbalisée, affichée, ou partiellement verbalisée et partiellement affichée ;

– la restriction : dans le cas où le nombre de solutions dépasse le seuil raisonnable, le système suggère des critères afin de restreindre l'espace de recherche. Le système peut aussi proposer des réponses conditionnelles. Plus que la transmission d'une réponse, l'acte de dialogue est ici la transmission de conditions pour résoudre un problème ;

– la relaxation : dans le cas où aucune solution n'est trouvée, le système suggère soit des solutions alternatives, soit des critères de recherche alternatifs. Les réponses éventuellement présentées sont des réponses suggestives.

Dans un contexte multimodal, la gestion du dialogue comporte en outre la gestion temporelle des tours de parole, notamment quand la recherche et *a fortiori* la présentation d'une information complexe prend du temps. On a vu qu'il n'était pas forcément pertinent pour le système de couper la parole à l'utilisateur. En revanche, la question peut se poser d'une manière tout à fait différente quand il s'agit non pas de produire un énoncé oral, mais d'amorcer une présentation d'information multimédia. En effet, une action purement visuelle du système peut être envisagée alors même que l'utilisateur est en train de parler, surtout si cette action visuelle apporte une information rapide

qui peut s'avérer efficace pour la suite de l'interaction. Au-delà des aspects liés à la génération de messages multimodaux sur lesquels nous allons revenir dans le chapitre 9, ce sont bien des aspects liés à la gestion du dialogue qui sont en jeu ici.

8.2.4. *Un système de dialogue peut-il mentir ?*

Un dernier aspect sur lequel nous n'avons pas assez insisté est l'importance de la tâche dans la gestion du dialogue. L'exemple de l'introduction va nous permettre d'illustrer quelques comportements du système en fonction de ses propres priorités, et notamment la possibilité de mentir, comportement fascinant s'il en est pour un système de DHM.

Revenons donc sur la gestion du dialogue au moment de la réception de U2, « combien de temps avec ce chemin qui semble être le plus court ? ». Au paragraphe 8.2.2, nous avons supposé que la durée du chemin en question était de vingt minutes, ce que le système évalue comme court, et nous avons supposé que le commentaire était vrai, à savoir que le chemin désigné est bien le plus court dans l'ensemble des solutions identifiées. Supposons maintenant comme au paragraphe 8.1.1 que la durée du trajet est de deux heures, ce que le système peut évaluer comme une durée longue, *a priori* peu satisfaisante compte tenu de la distance correspondante. Ajoutons aux priorités du système celle de vendre un billet de train, priorité qui peut prendre plusieurs matérialisations : faire aboutir le dialogue sur une vente et pas simplement sur un renseignement horaire, proposer des promotions, privilégier la vente de certains billets par rapport à d'autres en fonction de contraintes telles que la date très proche du transport, la durée de validité des billets, etc. En fin de compte, le système peut être amené à produire l'énoncé « deux heures », sans plus de précision, pour l'une des raisons suivantes :

– le système sait que c'est une durée peu satisfaisante, mais c'est le trajet le plus court – autrement dit le commentaire de l'utilisateur est vrai – et cela semble donc être la meilleure solution. Le système n'a pas considéré que confirmer le commentaire était nécessaire, pour les raisons évoquées au paragraphe 8.2.2. Il n'a pas non plus osé dire à l'utilisateur de prendre un taxi, peut-être parce que cela sort de son domaine. . . ;

– le système voit que ce n'est pas le trajet le plus court, mais l'autre trajet possible dure dix minutes de moins et implique un changement supplémentaire. L'un dans l'autre, la croyance de l'utilisateur est fausse mais ceci ne constitue pas un véritable problème. Plutôt que de perdre du temps à expliquer les avantages et inconvénients de chacun des deux trajets possibles, le système préfère répondre à la question directe et ignorer le commentaire ;

– le système voit que ce n'est pas le trajet le plus court, et qu'il existe une autre possibilité qui prend moins d'une heure. Selon des priorités d'ordre pragmatique, il considère cependant que répondre à l'acte premier est beaucoup plus important que répondre à l'acte second, et que produire une réponse courte doit être privilégié avant

tout. Il répond donc en se contentant du strict minimum. Si l'utilisateur n'est pas satisfait, il posera bien une question sur l'autre chemin affiché ;

– le système voit que l'autre possibilité est bien plus rapide, mais aussi bien moins chère : privilégiant la vente du billet le plus rentable, il décide donc de ne rien dire, espérant que l'utilisateur ne posera pas de question supplémentaire et restera focalisé sur ce trajet…

Dans l'un ou l'autre de ces cas de figures, des réponses telles que « deux heures, et c'est bien le chemin le plus court », « deux heures, l'autre chemin étant à peine plus court », « deux heures, mais l'autre chemin est plus court », ou encore « deux heures, à cause d'un changement à Versailles » seraient probablement préférables. Aucun mensonge n'apparaît cependant, ou alors sous la forme d'un mensonge par omission. Il peut en effet sembler délicat d'envisager un système capable de mentir sciemment, en donnant une information fausse, par exemple en répondant « une heure » alors que le trajet dure deux heures : d'une part l'utilisateur peut s'en apercevoir, d'autre part cela nécessiterait de dupliquer la base de données de l'application, de manière à éviter toute contradiction ultérieure. On peut espérer qu'aucun concepteur de système n'en vienne à de telles extrémités.

Pour terminer sur cet exemple, notons qu'il y a un lien étroit entre la gestion des actes de langage composites et la gestion du dialogue : dans le cas où la durée est de deux heures et qu'il ne s'agit pas du chemin le plus court, le système fait face à plusieurs possibilités de réaction, selon les priorités données aux deux actes de langage et à la manière de poursuivre le dialogue :

– quand un utilisateur exprime une croyance et que cette croyance s'avère fausse, le système peut considérer comme prioritaire de rétablir la vérité. Il peut ainsi renverser l'importance donnée linguistiquement à l'acte premier puis à l'acte second, et décider de réagir à l'acte second seul (« non, ce n'est pas le chemin le plus court ») ou aux deux actes, mais en commençant par la réponse à l'acte second (« non, ce n'est pas le chemin le plus court : il prend deux heures »). Dans le cas d'un système coopératif, on peut même ajouter la transmission de l'identifiant du chemin le plus court, c'est-à-dire de la solution qui, elle, rend le commentaire vrai : « non, ce n'est pas le chemin le plus court. Voici le chemin le plus court », avec une mise en valeur à l'écran du trajet correspondant ;

– quelle que soit l'importance donnée à la vérité ou fausseté du commentaire, le système peut considérer que d'un point de vue linguistique, il doit répondre d'abord à l'acte premier puis à l'acte second, ce qui donne « deux heures, mais ce n'est pas le chemin le plus court », ou, avec un comportement coopératif très appréciable, « deux heures, mais le trajet le plus court est celui-ci ».

De nombreuses autres réponses sont possibles, et cette illustration prouve d'une part que générer des réponses en langage naturel dans le but d'augmenter le réalisme

du DHM pose de nombreux problèmes, et d'autre part qu'une identification fine des contenus sémantiques, des actes de langage et des états mentaux des interlocuteurs est nécessaire pour obtenir un comportement adéquat du système, c'est-à-dire compréhensif, pertinent, cohérent et adapté à la tâche en cours de résolution.

8.3. Bilan

La tâche à accomplir est le fil directeur du dialogue : le dialogue progresse quand la tâche progresse. Néanmoins, un dialogue homme-machine réaliste ne doit pas se construire sur cette seule priorité : il doit aussi se préoccuper de la fluidité, de la spontanéité linguistique des échanges. Ce huitième chapitre confronte comportement par rapport à la tâche et comportement linguistique, afin de montrer comment se rapprocher d'un dialogue naturel en langage naturel. Des exemples de stratégies de dialogue illustrent comment un système peut être optimisé dans ce sens, mais aussi comment un système peut être amené à mentir.

Chapitre 9

La gestion de la multimodalité en sortie du système

A chaque fois que le gestionnaire de dialogue décide de produire un message à destination de l'utilisateur, ce qui arrive généralement peu de temps après la fin d'une intervention de celui-ci (mais peut aussi survenir en plein milieu d'un énoncé), un processus de génération automatique se met en œuvre. Pour le dialogue écrit ou oral, c'est le domaine de la génération automatique de textes qui est concerné. Pour le dialogue multimodal, qu'il s'agisse d'un système d'information susceptible d'afficher des données complexes, d'un système gérant un micromonde représenté à l'écran, d'un système doté d'un dispositif de retour d'effort, d'un ACA ou d'un robot capable de produire des gestes tout en parlant, la génération d'un énoncé en langage naturel se couple avec celle d'un geste ou d'un retour visuel. Le processus peut alors impliquer la génération multimodale, c'est-à-dire la production de références multimodales, dans le sens inverse de celui étudié dans le chapitre 6, ainsi que la transmission d'informations multimédia. Pour ce dernier point, le domaine concerné est celui des systèmes de présentation d'informations multimédia, ou IMMPS, *Intelligent MultiMedia Presentation Systems* (Stock et Zancanaro, 2005), domaine de recherche à part entière, un peu comme celui des ACA. La gestion des sorties d'un système de DHM peut ainsi impliquer de nombreux traitements, répartis dans de multiples modules.

Pour appréhender ces processus, on peut faire une distinction entre le « quoi » et le « comment ». Le premier est du ressort du gestionnaire du dialogue (Jurafsky et Martin, 2009). Il intègre un « quoi dire » et éventuellement un « quoi afficher » et un « quoi faire », chacun d'eux incluant un contenu sémantique et, surtout pour le premier, un acte de dialogue. Le second est du ressort de la génération, et c'est lui que nous allons étudier dans ce chapitre. Pour ce qui concerne la génération de textes, les

étapes de traitement successives sont les suivantes : planification du contenu, c'est-à-dire choix de la manière d'agencer entre elles les différentes propositions constituant le contenu sémantique ; aggrégation des phrases, c'est-à-dire affectation des propositions à des phrases et détermination des relations de discours ; lexicalisation, autrement dit le choix des mots ; génération des expressions référentielles, pour l'instant dans un cadre uniquement linguistique qui nécessite de choisir entre référence directe et anaphore ; réalisation linguistique, avec l'application des règles syntaxiques et morphologiques pour obtenir une phrase bien construire (Reiter et Dale, 2000). Dans le cadre du dialogue oral, s'ajoute une phase de détermination de la prosodie et de synthèse vocale, qui peut inclure un rendu oral des émotions, ainsi qu'une gestion des actes de dialogue, avec notamment la génération d'un acte qui matérialise le changement de tour de parole. Si le dialogue implique un ACA, le « comment » peut, lui aussi, se décomposer en plusieurs processus : choix d'un type de comportement physiquement perceptible, compte tenu du « quoi », puis instanciation (ou rendu) de ce comportement, voir chapitre 9 de (Garbay et Kayser, 2011). La gestion d'une tête parlante nécessite en particulier une phase d'animation du visage (lèvres, yeux, mains, corps en général) incluant le rendu visuel d'émotions. Tous ces processus impliquent diverses techniques, depuis l'utilisation de patrons, qu'ils soient syntaxiques, prosodiques, gestuels, animatiques, avec ou sans variables paramétrables, jusqu'à la gestion de phénomènes linguistiques et discursifs comme c'est le cas pour la génération automatique d'énoncés en langage naturel quand elle exploite les principes de la structure informationnelle.

Les liens entre le comportement général du système et l'ensemble de ces traitements est parfois difficile à faire. Le rendu des émotions est un moyen privilégié pour transmettre quelques indications, par exemple sur l'orientation positive ou négative du message. Faire varier cette orientation en fonction des réponses, des incompréhensions, des ambiguïtés, ou tout simplement de l'incapacité du système à répondre à une requête, augmente le réalisme de l'interaction : une orientation systématiquement positive peut énerver l'utilisateur dans les cas d'incompréhension répétée, et, bien entendu, une orientation majoritairement négative ne contribue pas à l'aspect coopératif du dialogue. Au-delà d'une simple orientation positive ou négative (certains systèmes peuvent se fâcher quand ils détectent un comportement malveillant de la part de l'utilisateur), les modèles d'émotions actuels impliquent plusieurs dimensions, chacune se matérialisant potentiellement sur plusieurs modalités : la valence (positive ou négative), l'activation (faible à fort), le degré de contrôle (la peur n'est par exemple pas liée à un sentiment de contrôle de la situation, alors que la colère l'est), et le degré d'imprévu, voir chapitre 3 de (Garbay et Kayser, 2011).

Pour atteindre un réalisme satisfaisant, le comportement du système peut également intégrer certains aspects de la communication spontanée décrits dans les chapitres 3 et 5. Il peut par exemple produire des hésitations et des répétitions, comme un interlocuteur humain (Rosset, 2008, p. 84), ou encore une gestion de la charge cognitive de l'utilisateur, comme nous allons le voir avec les facteurs humains qui font

l'objet de ce chapitre, avec tout d'abord quelques principes généraux pour la conception des modules chargés des sorties du système (section 9.1), en particulier au niveau pragmatique des actes de dialogue (section 9.2), puis la description de quelques processus particulièrement importants dans le dialogue multimodal (section 9.3).

9.1. Méthodologie pour la gestion des sorties

9.1.1. *Principes généraux pour la multimodalité en sortie*

Un présentateur d'informations multimédia a pour rôle de traduire les messages provenant du gestionnaire de dialogue en tenant compte le mieux possible des caractéristiques particulières des informations à présenter (donc à afficher ou à verbaliser), du terminal sur lequel s'effectue le dialogue, de l'environnement physique (dialogue en milieu bruité, dans un avion, sur un terrain d'opération) et de l'utilisateur. Quand l'information est amenée à être répartie sur plusieurs modalités de communication, on parle de « fission multimodale », processus inverse dans son objectif de celui de la fusion multimodale décrit dans le chapitre 6. Le terme « information » regroupe aussi bien les énoncés en langage naturel ou multimodaux que des données issues du modèle de l'application, comme les caractéristiques d'un ensemble de trains. Certaines informations peuvent être affectées d'étiquettes décrivant leur statut compte tenu de la tâche en cours : caractère d'urgence et d'importance (critique, par exemple). D'autres caractéristiques peuvent faire l'objet d'étiquettes, ou de calcul de la part du présentateur afin de tester les possibilités de présentation : caractère discret ou continu, volume, complexité, nombre d'éléments (paragraphe 9.1.2). C'est notamment ce qui permet des gestions totalement différentes d'énoncés en langage naturel et de données telles que des cartes géographiques ou des bases de données d'horaires. De manière un peu schématique, c'est le gestionnaire de dialogue qui décide de :

– « qui » : à qui l'information est destinée ;

– « quoi » : quelles informations sont présentées ;

– « dont » : quelle partie de l'information est mise en valeur ;

– « où » : sur quel ensemble de dispositifs l'information peut être présentée ;

– « quand » : quand et pendant combien de temps dure la présentation.

C'est le présentateur multimédia qui réalise ces décisions, c'est-à-dire qui procède au « comment ». Cela se fait en choisissant le ou les dispositifs à exploiter, en divisant l'information pour déterminer la partie revenant à chacun des dispositifs, en la divisant pour répartir sa présentation dans la durée impartie, en choisissant la manière de mettre en valeur la partie concernée, en gérant éventuellement une interface spécifique à l'affichage, avec par exemple des métaphores graphiques comme des ascenseurs et des boutons de navigation dans l'espace occupé par l'information.

On peut résumer les préoccupations d'un présentateur multimédia en un ensemble de principes généraux, à la manière des maximes de Grice. La conception de systèmes incluant un présentateur multimédia requiert une prise en compte fine des aspects pragmatiques et cognitifs de la communication, et c'est dans ce but que sont énoncés ces principes, qui restent à matérialiser (comme les maximes de Grice) par une théorie telle que la théorie de la pertinence (Sperber et Wilson, 1995). Une première facette concerne la prise en compte des caractéristiques des informations et de leur ancrage dans l'historique du dialogue, ce qui implique, dans le cas d'une communication incluant DHM et IHM, l'historique de l'interaction qui sauvegarde l'ensemble des manipulations directes effectuées sur les objets composant l'IHM. Les premiers principes pour la conception de présentateurs multimédias naturels, adaptatifs et centrés sur l'utilisateur sont ainsi les suivants :

– bien présenter en répartissant de manière pertinente les informations sur les canaux de communication ;

– bien présenter en se préoccupant du rendu et de la valorisation de l'information sur chaque canal de communication ;

– bien présenter en exploitant de manière pertinente le contenu sémantique du message ;

– bien présenter en maintenant une cohérence et une cohésion avec les messages précédents.

Une deuxième facette regroupe la prise en compte des caractéristiques du terminal et de l'environnement physique et situationnel :

– bien présenter en exploitant de manière pertinente les moyens de présentation ;

– bien présenter en exploitant de manière pertinente les conditions de présentation.

On en vient alors à la prise en compte de l'utilisateur, avec ses capacités physiques et cognitives, ses rôles dans la tâche en cours d'exécution, et ses préférences de communication telles qu'elles ont été définies et identifiées au cours de l'interaction :

– bien présenter avec une exploitation fine des attentes de l'utilisateur ;

– bien présenter pour favoriser une perception adéquate du message ;

– bien présenter pour favoriser des réactions adéquates de la part de l'utilisateur.

9.1.2. *Facteurs humains pour la présentation multimédia*

S'adapter aux capacités physiques et cognitives de l'utilisateur relève des facteurs humains (voir le début de la section 2.1). C'est le domaine de la psychologie cognitive, et notamment de l'ergonomie cognitive (Gaonac'h, 2006), avec des préoccupations comme celle de la gestion de la charge cognitive, de la mémoire et de l'attention de l'utilisateur, préoccupations qui s'appliquent utilement au DHM, voir le chapitre 9 de

(Cohen *et al.*, 2004). Cette adaptation complète celle explorée par les IHM plastiques (paragraphe 2.3.2), avec plusieurs facettes caractérisant l'adaptabilité des IHM, et plus généralement du DHM, aux terminaux, aux droits de l'utilisateur (droits d'accès à certaines informations et pas à d'autres), aux rôles de celui-ci (selon son rôle dans la résolution de la tâche, certaines informations sont plus importantes que d'autres) et aux préférences : préférences sur le filtrage des données, sur les modalités à privilégier, ou encore sur les manières de mettre en valeur une partie de l'information. Tous ces processus interviennent à la fois dans la répartition des informations sur les canaux de communication et dans la mise en valeur d'informations particulières.

Pour réaliser ces processus, un premier ensemble de paramètres regroupe les caractéristiques de l'information à transmettre, avec trois catégories principales : le contenu sémantique, les aspects pragmatiques, et l'ancrage dans l'historique du dialogue. Parmi tous les éléments qui constituent le contenu sémantique, on retient comme paramètres essentiels :

– le niveau de criticité, qui peut amener au choix d'une mise en valeur forte ;

– le niveau d'urgence, qui peut bloquer tous les processus en cours pour forcer l'utilisateur à réagir immédiatement ;

– la complexité de l'information : nature de la structure de données, volume et nombre d'éléments impliqués ;

– la constitution de l'information : discrète ou continue, linguistique, chiffrée, répartie sur une, deux ou trois dimensions ;

– l'étendue de l'information, avec par exemple un choix consistant à montrer l'information dans sa totalité, ce qui peut la rendre illisible, et à ajouter une loupe (zoom) sur une partie focalisée, et donc lisible ;

– les contraintes de présentation inhérentes à l'information multimédia elle-même : visuelle pour de la cartographie, sans contraintes pour un message en langage naturel, qui reste aussi bien affichable que verbalisable.

Les aspects pragmatiques regroupent les valeur et force illocutoires vues au paragraphe 7.1.1, c'est-à-dire l'acte de dialogue porté par l'action de présenter, avec plusieurs degrés selon le caractère marqué ou peu marqué de cet acte, mais aussi les valeur et force perlocutoires, qui reflètent les effets réalisés par l'expression de cet acte de dialogue et que nous verrons en section 9.2. Comme nous l'avons vu avec le traitement des entrées et la gestion du dialogue, ces aspects pragmatiques prennent un sens quand ils sont reliés à la structure du dialogue, et ainsi à la modélisation de l'historique multimodal : historique des messages échangés, des données affichées, des actions d'affichage effectuées, de manière à autoriser des mentions ultérieures à ces actions.

Un deuxième ensemble de paramètres regroupe les moyens et les conditions de présentation. Les premiers sont essentiellement les caractéristiques du terminal utilisé, notamment la liste des dispositifs en fonctionnement (haut-parleur, écran tactile) avec leurs limites, la disponibilité de chacun d'entre eux, et tout un ensemble de contraintes quant à la transmission d'informations par leur biais : contraintes de dimensions comme la taille de l'écran, contraintes de temps de traitement, contraintes sur la constitution de l'information compte tenu de la modalité choisie, etc. Pour les seconds, nous proposons d'appliquer à la présentation multimédia les trois fonctions du geste identifiées par (Cadoz, 1994) pour décrire les possibilités d'interaction gestuelle :

– les contraintes épistémiques, liées à la prise de connaissance de l'environnement : prise en compte du niveau de bruit ambiant tel qu'il est capté par le microphone et du niveau de luminosité ambiante, voire des vibrations dans le cas d'un DHM dans un avion, du moins si l'on dispose des capteurs adéquats ;

– les contraintes ergotiques, liées à la transformation de l'environnement : niveau sonore et niveau de luminosité à ne pas dépasser pour ne pas gêner l'environnement, ce qui peut se concevoir facilement dans un bureau de travail ou un terrain d'opération impliquant plusieurs autres systèmes et utilisateurs ;

– les contraintes sémiotiques, liées à l'émission d'informations pour l'environnement, avec la quantité et la qualité du débit de parole, qui peut par exemple s'avérer trop intense compte tenu de l'environnement de communication.

Dans la lignée de ce dernier point, un troisième ensemble de paramètres regroupe les aspects humains relevant de l'adaptation à l'utilisateur. Il s'agit tout d'abord de s'adapter aux capacités physiques de l'utilisateur, en cas de handicap, c'est-à-dire de contraintes sur le fonctionnement des canaux de communication, mais aussi, d'une manière générale, avec les contraintes et les préférences concernant le niveau d'exploitation de ces canaux : canal auditif déjà accaparé, par exemple. Il s'agit ensuite de l'adaptation aux rôles de l'utilisateur et à ses préférences individuelles. Il s'agit enfin des facteurs humains en tant que préférences universelles, avec les facteurs physiologiques, les facteurs linguistiques et les facteurs cognitifs. Les premiers sont liés à la nature des modalités : pour la génération de son, comme une alerte de type bip sonore, le système doit savoir que les aigus sont plus stridents que les graves, que plus l'alerte est forte et plus elle a des chances d'être perçue (mais plus elle est stressante). Pour la génération visuelle, le système peut être amené, comme on l'a vu en interprétation (paragraphe 6.1.2), à implémenter les critères de groupement perceptif issus de la théorie de la Gestalt, ou encore à prendre en compte les observations des théories de la couleur, afin d'exploiter le rouge, c'est-à-dire la couleur perçue le plus rapidement par l'humain, pour les messages urgents. Quelle que soit la modalité, le système peut exploiter les notions de « saillance » et de « prégnance » : un élément saillant, c'est-à-dire qui se distingue des autres éléments par des propriétés singulières, par exemple

le seul élément bleu de la scène visuelle, se perçoit plus facilement ; un élément prégnant, c'est-à-dire qui a fait l'objet de répétitions préalables au point d'imprégner la mémoire de l'utilisateur, se perçoit également plus facilement.

Les facteurs linguistiques relèvent des préférences lexicales, prosodiques, syntaxiques, sémantiques et pragmatiques de l'utilisateur. Le système peut s'adapter à ces préférences en s'alignant sur les usages de l'utilisateur, c'est-à-dire en employant les mêmes termes, dans les mêmes structures de phrases, avec un usage similaire du langage. Au niveau du dialogue naturel en langage naturel, il s'agit aussi d'appliquer les maximes de Grice lors de la détermination du message, de minimiser les risques d'ambiguïtés en anticipant sur les conditions de production de celles-ci (on peut par exemple éviter une anaphore pronominale quand plusieurs antécédents sont possibles), et on peut aller jusqu'à éviter les actes de langage indirects et composites, du moins dans des dialogues simples où la tâche prime. Il s'agit également d'exploiter la structure informationnelle, notamment pour la mise en relief d'une partie de l'information, et d'appliquer des règles simples de cohésion et de cohérence, par exemple une utilisation pertinente des connecteurs, en tenant compte des contenus des messages précédents.

Quant aux facteurs cognitifs, il s'agit de l'ensemble des éléments qui ont fait l'objet de la section 2.1 : prise en compte des limites humaines quant à la mémoire à court terme, aux capacités de représentation mentale, ou encore à la gestion de l'attention de l'utilisateur. Il s'agit par exemple de ne pas faire en sorte que le système tente de capturer l'attention de l'utilisateur dans des directions variées. D'une manière générale, un principe peut consister à exploiter ce qui a déjà fonctionné correctement. Si le système constate que tel message a eu une influence positive et efficace en étant produit visuellement plutôt qu'oralement, il peut décider de l'employer à nouveau avec le même type de production dans des conditions similaires. Au final, les paramètres mentionnés sont gérés de la manière suivante :

– paramètres issus du domaine applicatif, de la tâche et du modèle de l'utilisateur : les niveaux d'urgence et de criticité, les informations autodescriptives (information structurée et quantifiée), et les contraintes et préférences de présentation multimédia spécifiques au type de tâche ou à la tâche elle-même ;

– paramètres calculés par le gestionnaire du dialogue : les valeurs et forces pragmatiques, les étiquettes telles que celles relatives aux émotions, les indications de cohésion et de cohérence, les mises en valeurs, et les contraintes et préférences sur les termes linguistiques et sur la gestion du dialogue ;

– paramètres décidés par le présentateur multimédia sur la base des contraintes des autres niveaux : l'ordonnancement des informations à présenter, par exemple en fonction des niveaux d'urgence, la manière de dissocier une information en plusieurs phases de présentation, la manière de dissocier une information sur les canaux de communication, les degrés de mise en valeur de chaque information, par exemple en fonction de la criticité, ainsi que les manières de mettre en valeur.

9.2. Pragmatique pour la présentation multimédia

9.2.1. *Valeurs et forces illocutoires*

Un aspect important réside dans l'exploitation des valeurs et des forces illocutoires imposées par le gestionnaire du dialogue. En interprétation comme en génération, au contenu sémantique du message s'ajoute comme on l'a vu une valeur illocutoire désignant l'acte de dialogue réalisé par l'énonciation (ou la présentation), valeur qui peut être un « dire que », un « dire de », un « demander » ou une combinaison de plusieurs de ces actes (acte composite), et qui est liée à une intention sous-jacente. A cette valeur on donne ici de l'importance à la force, c'est-à-dire au degré d'intensité avec lequel la valeur doit être transmise. La manière de présenter, par exemple une alerte, dépend de la valeur illocutoire : si l'on veut juste informer, *via* un « dire que », on peut présenter d'une manière neutre, en tout cas très diffèrente de la manière choisie pour un « dire de » qui revient à donner l'ordre à l'utilisateur de prendre en compte la signification de l'alerte, selon un mode de fonctionnement correspondant à un « inciter à agir ». Le système de dialogue peut avoir besoin d'une confirmation de la réception du message. On peut ainsi distinguer un acte qui consiste à informer, sans indication particulière sur la suite de l'interaction, d'un acte qui consiste à informer en demandant un accusé de réception, comme le font par exemple les boîtes de dialogue incluant les boutons « OK » et « annuler » que l'utilisateur doit cliquer s'il veut poursuivre sa tâche. Pour ces deux exemples, la notion d'acte s'avère utile, montrant par là-même que la génération automatique peut procéder à des stratégies similaires à celles de l'interprétation automatique. L'exemple de l'alerte peut ainsi prendre la forme d'un acte composite avec un « dire que » explicitant la nature du problème, et un « dire de » donnant à l'utilisateur l'ordre de réagir. Quant à l'exemple avec accusé de réception, il combine un « dire que » portant l'information à présenter et un « demander » portant sur l'accusé de réception : « OK » représente une réponse positive à cette question, « annuler » une réponse négative.

9.2.2. *Valeurs et forces perlocutoires*

Ces exemples qui incitent ou obligent l'utilisateur à réagir d'une certaine manière se rapprochent de la notion d'acte perlocutoire, visant à créer certains effets sur les états mentaux de l'utilisateur (paragraphe 7.1.1). Ce que les valeurs illocutoires ne peuvent faire explicitement par le biais d'actes concrets, les valeurs perlocutoires peuvent amener le système, par exemple l'ACA, à manifester un comportement qui contribue à produire chez l'utilisateur un certain effet. Gérer une valeur perlocutoire est complexe, surtout quand elle s'accompagne d'une certaine force. Dans tous les cas, il revient au présentateur multimédia de trouver une forme d'expression du message qui rende correctement le but perlocutoire. Il peut s'agir d'une prosodie particulière, ou d'une expression ou attitude d'un ACA qui traduit une attente, indiquant à l'utilisateur qu'une réaction de sa part est souhaitable. Dans notre exemple, le système peut

répondre « je n'ai plus de place dans le train pour Paris » à la requête de l'utilisateur, ce qui constitue un simple « dire que » sans but perlocutoire autre que celui d'informer. Mais le système peut aussi répondre « une heure avant le départ, je n'ai plus de place dans le train pour Paris », ce qui correspond à la même valeur illocutoire, avec en plus un léger sous-entendu consistant à avertir l'utilisateur qu'il aurait pu s'y prendre plus tôt, et avec en plus une valeur perlocutoire visant à modifier les croyances de l'utilisateur sur la façon de réserver un billet. Dans un même ordre d'idée, « je n'ai plus de place pour Paris, seulement pour Massy-Palaiseau » porte une valeur perlocutoire consistant à tenter de modifier le but initial de l'utilisateur, c'est-à-dire à l'inciter à considérer un but alternatif, mais sans l'exprimer ouvertement comme le ferait « êtes-vous prêt à changer votre destination pour Massy-Palaiseau ? ».

Pour le langage naturel, gérer la valeur perlocutoire se rapproche ainsi de la gestion des inférences et de la planification dans le dialogue. Pour d'autres modalités, comme par exemple dans une IHM, il peut s'agir plus simplement de rendre manifestes les différentes possibilités qui s'offrent à l'utilisateur suite à une action de celui-ci. Dans une IHM, on appuie sur les boutons, on cherche à taper du texte dans les zones qui ressemblent à des champs de saisie, à cliquer sur les cellules d'un tableau : on sait que tout élément affiché a une fonction et on explore cette fonction avec les moyens offerts par l'interaction clavier-souris. En conséquence, le présentateur multimédia qui gère l'IHM en même temps que le DHM doit tenir compte des fonctions des éléments de l'IHM lors des différentes phases de présentation : pour chaque élément impliqué dans la dernière action ou dans l'action en cours, il doit connaître les possibilités d'interaction en entrée, éventuellement les inhiber (bouton grisé, cellule de tableau affichée avec une couleur particulière), et informer si besoin le module de gestion des entrées.

Cette stratégie consistant à anticiper sur les actions futures de l'utilisateur, en fonction d'une valeur perlocutoire particulière, peut avoir des conséquences nombreuses et complexes. Quand le système de DHM pose une question à l'utilisateur et qu'il attend une réponse parmi un ensemble d'alternatives clairement identifiées, quand il incite l'utilisateur à parler d'un nouveau topique, quand il présente des informations dans un ordre précis et incite ainsi l'utilisateur à faire référence à ces informations, il entraîne une réduction des possibilités d'interaction. Le prochain énoncé de l'utilisateur peut être quasiment prédictible. De ce fait, les modules de reconnaissance de la parole et d'analyses linguistiques ont tout intérêt à être prévenus. Compte tenu des difficultés de la reconnaissance de la parole en domaine ouvert, plusieurs modèles de langage peuvent être impliqués : un modèle généraliste sert au début du dialogue, mais un modèle orienté sur les chiffres, dates et valeurs peut être exploité lorsque l'utilisateur doit répondre à une question du système portant sur de telles données. De même, suite à la présentation d'informations dans un ordre significatif, l'utilisateur peut être amené à utiliser des références mentionnelles telles que « le premier », « le second » et « les deux derniers », ou encore des expressions quantifiées en « chaque… » ou « tous les… ». Le présentateur multimédia doit ainsi être conscient qu'il met en évidence un

ordonnancement, surtout quand cet ordonnancement n'était pas explicité par le gestionnaire de dialogue (c'est tout simplement le cas quand on affiche des informations de gauche à droite), de même qu'il doit être conscient quand il agglomère des informations et peut ainsi inciter à des références à un groupe plutôt qu'à des éléments individuels. L'intérêt est qu'il prévienne non seulement le module de reconnaissance de la parole pour adapter le modèle de langage, mais aussi le module d'analyse sémantique et celui de résolution de la référence, pour leur indiquer la logique qui sous-tend l'ordonnancement ou l'agglomération. Lors de la compréhension de l'énoncé de l'utilisateur, le système a alors tous les éléments pour identifier le bon référent.

9.3. Processus de traitement

9.3.1. *Répartition de l'information sur les canaux de communication*

Nous le voyons avec l'étendue des facteurs humains et la complexité des aspects pragmatiques en génération, la présentation d'informations multimédia, si elle est implémentée en suivant un modèle approfondi, implique de nombreux paramètres et des processus complexes, souvent interdépendants. Parmi ceux-ci, un processus essentiel pour le dialogue multimodal consiste à répartir l'information à transmettre sur les canaux de communication. Une première étape dans ce processus consiste à prendre en compte les contraintes, une deuxième étape à prendre en compte les préférences, et une troisième étape, qui peut éventuellement entraîner de nouvelles décisions, consiste à faire les liens entre les composantes de l'information ainsi réparties, c'est-à-dire à expliciter des liens entre les modalités.

La première étape privilégie les contraintes. Il s'agit de prendre en compte les contraintes inhérentes à l'information (modalité visuelle pour une carte géographique), les contraintes liées au terminal (s'il ne possède pas de haut-parleur, tous les énoncés en langage naturel devront être affichés sous une forme écrite, avec les contraintes morphologiques et orthographiques que cela représente), les contraintes liées à l'environnement de présentation (un fort bruit ambiant peut entraîner l'abandon de la modalité vocale), ainsi que les contraintes liées aux capacités et aux rôles de l'utilisateur.

L'étape suivante consiste à prendre en compte un ensemble de règles portant sur :

– l'urgence ou la criticité du message : si l'un des deux paramètres est à un niveau élevé, il est peut-être nécessaire d'exploiter tous les canaux de communication ;

– le contenu du message : choix du canal le mieux adapté à la constitution de l'information, ou encore exploitation simultanée de plusieurs canaux si l'information est d'une grande complexité ;

– l'acte de communication : un seul canal favorisé pour un acte simple, deux canaux favorisés pour un acte composite ;

– l'historique de l'interaction : utilisation privilégiée d'un canal déjà exploité ;

– les préférences de l'utilisateur : utilisation d'un seul canal si c'est une préférence exprimée ;

– les facteurs humains : répartition dans le temps de l'affichage d'une information de très grande taille, pour laquelle la prise de connaissance nécessite une attention maintenue ; prise en compte de l'attention de l'utilisateur si le système est capable de la détecter (en cas d'attention soutenue, le système n'a pas besoin de faire des efforts particuliers pour répartir l'information de manière explicite).

La troisième étape, qui consiste à expliciter un lien entre les modalités, rejoint la notion de déixis (voir paragraphe 5.1.4), avec des termes tels que le verbe présentatif « voici ». Quand une partie de l'information est affichée et que l'autre partie est verbalisée, il se peut que l'utilisateur ne fasse pas spontanément le lien entre les deux parties. Le danger est alors qu'il considère que le système lui adresse deux messages distincts : les deux parties de l'information peuvent être de nature différente et ne pas dépendre l'une de l'autre, contrairement par exemple à la vidéo où la bande son et l'image sont immédiatement perçues comme les deux parties d'une même information, de par leur synchronisation temporelle. Il peut donc être utile que le système ajoute au contenu sémantique à transmettre une indication qui permette à l'utilisateur de faire le lien. Cette indication est portée par une modalité et rappelle l'existence de l'autre. Ce peut être une icône visuelle qui indique que le système est en train de parler. Ce peut être une expression référentielle démonstrative, qui s'appuie sur le contexte visuel, ou encore un énoncé tel que « sur la carte géographique, vous pouvez voir les trains qui vont à Paris » ou « le train avec changement à Meudon est celui qui clignote ». La génération de tels énoncés nécessite plusieurs étapes de traitement, par exemple :

– le gestionnaire de dialogue produit la requête de présentation « rendre manifeste le trajet Meudon-Paris à l'utilisateur » ;

– le présentateur multimédia choisit une réalisation à la fois visuelle (affichage du trajet sur la carte géographique, avec clignotement pour que le trajet se distingue bien des autres trajets déjà affichés) et auditive (acte de langage « dire que » pour informer l'utilisateur de l'identité du trajet ainsi affiché), avec la génération d'une déixis pour que l'utilisateur rapproche les deux réalisations ;

– le présentateur multimédia demande au générateur de langage naturel de matérialiser la déixis multimodale, en lui indiquant la nature de l'affichage retenu ;

– le générateur de langage naturel choisit d'effectuer une référence multimodale, « celui qui clignote », et de construire une phrase simple autour de cette référence, « le trajet Meudon-Paris est celui qui clignote » ;

– le présentateur multimédia envoie de manière synchrone cet énoncé au module de synthèse de parole et effectue le clignotement.

Une autre possibilité est le choix de la phrase présentative « voici le trajet Meudon-Paris ». Cette phrase est plus courte et plus efficace, mais le lien entre « voici » et le clignotement est moins explicite que dans « celui qui clignote ».

9.3.2. *Gestion de la redondance et de la fission multimodales*

Si le présentateur multimédia souhaite accentuer la présentation d'un objet comme le trajet Meudon-Paris, il peut décider d'afficher un message textuel « Meudon-Paris » sur le trajet en question, en même temps qu'il génère l'énoncé oral « voici le trajet Meudon-Paris ». Dans ce cas, le choix relève de la « redondance ». L'intérêt est que si l'un des canaux de communication ne fonctionne pas bien, l'autre permet de compenser. Par ailleurs, plus on émet l'information, plus il y a de chances pour que l'utilisateur la perçoive. De même, plus on répère l'information, plus il y a de chances pour qu'elle imprègne l'utilisateur, comme on l'a vu avec la notion de prégnance. Cependant, la redondance n'est pas forcément un avantage systématique : trop de messages n'incitent pas l'utilisateur à maintenir son attention sur le problème en cours. Comme l'on dit souvent, trop d'informations tue l'information. De plus, trop de messages augmentent le temps de traitement et donc de réaction, et, s'ils sont mal gérés, peuvent conduire l'utilisateur à ne pas faire le lien entre les différentes matérialisations de la même information. En conclusion, il faudrait n'exploiter la redondance que dans les situations où il est manifeste qu'il s'agit de redondance. Par ailleurs, il vaut mieux s'abstenir d'utiliser la redondance dans un même canal de communication, par exemple en couplant la génération d'un énoncé oral avec celle d'un son, toutes les deux occupant le canal audio, qui s'en trouve ainsi surchargé.

Gestion de la redondance et répartition de l'information sur les canaux de communication relève d'une même problématique, celle de la fission multimodale. Au niveau des signaux, audio ou visuels, il s'agit de diriger l'information vers le bon canal de communication, selon la nature de celle-ci. Présenter une vidéo implique ainsi de transmettre la bande son dans le canal audio et l'image dans le canal visuel, ce qui consiste en une fission multimodale dirigée par des contraintes. Au niveau de la sémantique, il s'agit de dissocier le contenu de l'information sur plusieurs modalités, afin de mieux gérer sa complexité en obtenant des messages monomodaux simplifiés. Un exemple relevant d'une fission multimodale dirigée par les préférences consiste à afficher la partie de l'information qui nécessite une attention maintenue, et à verbaliser la partie qui ne fait que capter l'attention sélective. Au niveau de la pragmatique, il s'agit de dissocier l'acte de dialogue du message sur plusieurs modalités, afin d'obtenir des actes de dialogue plus simples, comme c'est le cas quand on dissocie les composantes d'un acte composite pour en faire des actes simples. Là aussi, il s'agit d'une fission multimodale, ce qui nous permet d'avancer que les trois niveaux identifiés pour la fusion multimodale ont des équivalents en génération.

9.3.3. *Génération d'expressions référentielles*

L'exemple « celui qui clignote » fait intervenir plusieurs phénomènes de référence mentionnés dans le chapitre 6 : la mention d'une propriété, ici sous la forme d'une relative, qui singularise l'objet et permet de l'identifier, ainsi que l'emploi d'un pronom à la fois démonstratif et anaphorique, démonstratif parce qu'il réfère de manière démonstrative à un objet focalisé par ailleurs, anaphorique parce qu'il reprend d'un antécédent sa tête nominale, à savoir « trajet ». L'ensemble de ces phénomènes donnent une idée des processus (Mellish *et al.*, 2006) qu'un générateur d'expressions référentielles doit gérer. Une expression référentielle doit permettre à l'utilisateur d'identifier le référent de manière non ambiguë. Pour cela, un algorithme désormais classique, l'algorithme incrémental (Reiter et Dale, 2000), repère les propriétés qui permettent de distinguer l'objet ciblé des autres objets, en privilégiant certaines propriétés pour certains types d'objets, suivant en cela des préférences presque universelles. Cet algorithme a fait l'objet de nombreuses extensions et adaptations, pour prendre en compte tout type de propriétés, y compris des relations spatiales entre plusieurs objets, pour gérer un critère de cohérence entre les propriétés utilisés pour référer de manière groupée à plusieurs référents, ou encore pour tenir compte de points de repères contextuels, notamment quand les propriétés choisies sont vagues (« trajet long ») et gradables (Kopp *et al.*, 2008 ; Krahmer et Van Deemter, 2012).

Les enjeux linguistiques pour les travaux à venir sont une meilleure gestion de la saillance, de la redondance, des ensembles de référents, des référents autres que des objets, une meilleure exploitation des propriétés vagues, la gestion de relations entre plus que deux objets, et un meilleur contrôle de la production involontaire d'ambiguïtés. Dans un cadre dialogique, il s'agit aussi d'intégrer aux processus la possibilité de coconstruire une référence avec l'interlocuteur (Krahmer et Van Deemter, 2012). Des enjeux plus psychologiques résident, comme pour la génération automatique en général, dans des collaborations avec des psycholinguistes, *via* par exemple des expérimentations visant à caractériser la notion de saillance, et dans la prise en compte des facteurs humains, autrement dit dans l'application des principes développés ci-dessus à la référence. D'un point de vue technique, les algorithmes actuels combinent représentation de connaissances à l'aide de logiques adaptées, recherche dans des graphes, satisfaction de contraintes, modélisation du contexte. Ils sont de plus en plus complexes, et, comme souvent en TAL et en DHM, explorent actuellement les combinaisons d'approches symboliques et d'approches basées sur des données de corpus.

9.3.4. *Valorisation d'une partie de l'information et synthèse*

Au bout de la chaîne de traitement des sorties se trouvent la valorisation d'une partie de l'information, visuelle, en langage naturel ou de toute autre façon, et les processus de matérialisation, notamment la synthèse d'ACA et la synthèse vocale.

Pour mettre en valeur une partie de l'information, le processus reprend le même principe que celui de la répartition de l'information sur les canaux de communication : on commence par prendre en compte les contraintes, puis par gérer un ensemble de règles modélisant les préférences. Les contraintes sont les mêmes que précédemment, par exemple un niveau sonore à ne pas dépasser malgré l'intention de mettre en valeur, ce qui peut se traduire par une accentuation prosodique. Les règles reposent sur le contenu du message, l'acte de dialogue, avec par exemple une intensité particulière pour un « dire de », les facteurs humains et les préférences individuelles exprimées par l'utilisateur. La mise en relief proprement dite peut prendre une multitude de formes, surtout pour un message en langage naturel. Elle peut faire intervenir la structure informationnelle, des constructions syntaxiques comme les présentatifs ou les topicalisations, et des procédés prosodiques variés (une pause avant et une pause après l'élément à valoriser, par exemple). Au niveau de la gestion d'un ACA, des rendus spécifiques sont mis en œuvre. A propos de données telles qu'une carte géographique ou un tableau de chiffres, toute exploitation des couleurs, de la taille relative des éléments, des critères de la théorie de la Gestalt est envisageable.

Dans le dialogue naturel en langage naturel, le rendu de la voix du système est un aspect essentiel, qui peut rebuter l'utilisateur comme on l'a vu au paragraphe 1.3.2 avec le problème de la « vallée dérangeante ». De manière générale et un peu schématique, on peut actuellement faire prononcer à un système n'importe quel texte ou énoncé, avec une qualité qui s'approche de ce que pourrait produire un humain ne maîtrisant pas le contenu sémantique. Il manque encore une meilleure prise en compte d'aspects contextuels, par exemple le rendu de nuances révélatrices d'une compréhension fine. C. d'Alessandro dans (Chaudiron, 2004) montre que l'évolution des systèmes de synthèse vocale a suivi le chemin suivant : systèmes de type 1, capables de restituer des messages préenregistrés ; systèmes de type 2, gérant des messages simples fabriqués à partir d'un vocabulaire fixe et à l'aide de méthodes de concaténation ; systèmes de type 3, capables de procéder à une véritable synthèse à partir de texte ; et systèmes de type 4, à composante visuelle. Le paradoxe de la synthèse vocale utilisée en DHM est que l'on voit souvent ce processus comme un module exécuté en bout de chaîne, c'est-à-dire peu concerné par les processus d'analyses prosodiques et linguistiques, alors que les réalisations actuelles montrent que pour bien prononcer un énoncé, le système doit en comprendre le sens et maîtriser la prosodie au point de pouvoir, si besoin, s'aligner sur celle de l'utilisateur. Les phénomènes d'accentuation, de proéminence prosodique, d'intonation, de rythme, font partie des préoccupations de la synthèse vocale, voir le chapitre 11 de (Cohen *et al.*, 2004). Nous avons souligné à quel point la phase du « comment le dire » était importante en dialogue, pour des raisons notamment de cohésion et de cohérence. Or, une fois que la suite de mots constituant le message en langage naturel est décidée, c'est surtout la prosodie qui va permettre de contrôler la synthèse. (Theune, 2002) montre par exemple qu'une spécification approfondie des directives prosodiques est essentielle, et propose un modèle de génération, pas forcément orienté pour le dialogue, dans lequel plusieurs processus

se succèdent pour enrichir le texte à prononcer en texte annoté, celui-ci servant à la synthèse de parole.

L'alignement prend une importance particulière dans ce contexte, suite notamment à des travaux sur l'alignement prosodique, mais aussi lexical (Brennan et Clark, 1996) et d'une manière générale selon toutes les dimensions du langage (Pickering et Garrod, 2004). Plus récemment, (Branigan *et al.*, 2010) souligne à quel point l'alignement intervient en DHM, mais aussi à quel point il diffère de l'alignement en dialogue humain : l'alignement avec une machine est vu comme une stratégie pour que la communication fonctionne. De fait, l'alignement est considéré comme l'un des nombreux enjeux majeurs du DHM.

9.4. Bilan

En plus de retours vocaux, le système peut manifester ses réactions et réponses à l'aide de retours visuels, voire gestuels s'il se matérialise par un avatar affiché à l'écran. La tâche peut également impliquer la présentation d'informations à l'utilisateur. La production de tous ces messages en sortie du système fait intervenir de nombreux choix de la part de celui-ci : quelle information mettre en avant ? Quelle partie de l'information rendre visuellement ? Quelle partie rendre oralement ? Ce neuvième chapitre fait le point sur les paramètres nécessaires à de tels choix, et sur les techniques de génération automatique de messages linguistiques et multimodaux. On y retrouve des préoccupations linguistiques abordées lors du traitement des entrées, mais aussi des aspects cognitifs tels que la prise en compte de facteurs humains pour la production de messages adaptés.

Chapitre 10

L'évaluation de systèmes de dialogue multimodaux

Qu'elle intervienne en toute fin de conception, sur le système final, ou au cours même de la conception, sur des prototypes ou des modules de système, l'évaluation a pour rôle de mesurer les performances, de comparer ces performances à celles de systèmes existants, et d'identifier les points forts et les points faibles. Si les moyens le permettent, ces derniers peuvent entraîner la reprise d'une phase de conception, de manière à améliorer le système. Souvent décriée, peut-être parce qu'elle appuie là où ça fait mal, l'évaluation peut aussi apporter un regard précieux et des méthodes exploitables pour la conception et l'implémentation. A propos de l'évaluation des algorithmes de génération automatique d'expressions référentielles, (Krahmer et Van Deemter, 2012) notent que les premiers travaux n'étaient pas clairs sur les paramètres utilisés dans les algorithmes, et que c'est lorsque des évaluations ont commencé à être réalisées que les chercheurs ont été obligés de dévoiler leurs cartes, c'est-à-dire de décrire leurs paramètres favoris. Ce sont aussi les campagnes d'évaluation qui contribuent à réunir les chercheurs autour des mêmes problématiques, et ainsi à participer à la dynamique générale. En revanche, elles s'accompagnent encore de nombreuses contraintes qui peuvent rebuter certains. Par exemple, comparer plusieurs systèmes nécessite de projeter les résultats de chacun d'entre eux vers un formalisme commun, qui puisse autoriser les comparaisons. Or cette projection peut s'avérer très coûteuse en temps, alors qu'elle n'apporte rien au système lui-même.

Dans ce chapitre, nous ne présentons pas les méthodes d'évaluation spécifiques à chacun des modules d'un système de dialogue, telles que les méthodes d'évaluation d'un moteur de reconnaissance de la parole, celles concernant un résolveur d'anaphore, ou encore la qualité perçue d'un ACA, voir chapitre 9 de (Garbay et Kayser,

2011), ou de la synthèse de parole, sans oublier l'évaluation des modèles et des processus de dérivation de modèles utilisés dans les architectures *design-time* (voir section 4.2). Chaque domaine, ne serait-ce que celui très général de l'interprétation ou celui de la gestion du dialogue, a ses propres critères, qui renvoient aux techniques utilisées, ce qui sort largement du cadre de ce livre. Nous présentons les caractéristiques d'une démarche visant à évaluer un système de DHM, avec les spécificités de l'interaction orale et de l'interaction multimodale, en tenant compte des aspects objectifs (nombre de tours de parole, durée moyenne des énoncés, nombre de refus) et des aspects subjectifs (interview d'un utilisateur sur son ressenti, remplissage de questionnaires), qui font éventuellement l'objet d'analyses statistiques descriptives et inférentielles, voir section 2.2 et (Jurafsky et Martin, 2009, p. 872).

Nous présentons tout d'abord les méthodes actuellement utilisées, en DHM oral et multimodal aussi bien qu'en IHM (section 10.1), ce qui nous amène à souligner les points faibles de l'évaluation, et à présenter les enjeux pour les années à venir (section 10.2) ainsi que quelques pistes, notamment pour le dialogue multimodal (section 10.3).

10.1. Faisabilité de l'évaluation de systèmes de dialogue

On voit paraître un nombre grandissant d'articles sur l'évaluation des systèmes de dialogue oraux et multimodaux. Des paradigmes d'évaluation sont proposés, de plus en plus larges et complexes, regroupant notamment des métriques, des tests utilisateurs et des méthodes d'analyse de questionnaires remplis par des sujets après leur utilisation du système. Ces efforts sont pertinents et louables, mais ne doivent pas faire oublier plusieurs constats récurrents qui restent particulièrement valables.

Premier constat : contrairement aux systèmes de recherche d'information, de reconnaissance de la parole ou d'analyse syntaxique, les systèmes de DHM restent souvent au niveau de prototypes de recherche difficiles à réaliser et à faire fonctionner correctement, ainsi que très sensibles au comportement des utilisateurs. A part quelques exemples marginaux, ludiques par exemple, il n'existe à l'heure actuelle aucun système fiable commercialisé et utilisé de manière profitable par une population de taille importante. Autrement dit, le passage à l'échelle reste un problème majeur en DHM et les évaluations réalisées s'en tiennent à des prototypes de recherche ou à des systèmes professionnels tellement finalisés (militaires par exemple) qu'ils ne s'adressent qu'à un nombre extrêmement réduit d'utilisateurs. L'évaluation pour le DHM se cantonne donc à un périmètre limité qui, sans remettre en cause son utilité, nuance quelque peu sa portée.

Deuxième constat : il risque d'exister bientôt autant de méthodologies d'évaluation que de systèmes proprement dits. Ce n'est pas un problème en soi, mais cela soulève des interrogations. En particulier, on peut s'interroger sur le bien-fondé d'une

méthode d'évaluation proposée par les concepteurs d'un système dans le but d'évaluer ce seul système, la méthode étant elle-même évaluée par son application au système en question. Cette description peut sembler caricaturale, elle reflète pourtant une certaine réalité, ou en tout cas elle s'en approche. Cette situation est inévitable compte tenu du nombre réduit de systèmes et, face aux avancées de chaque système, de la nécessité de prendre en compte des aspects qui ne sont pas traités par les méthodologies d'évaluation existantes. Ainsi, on en vient à proposer ou à étendre une méthodologie d'évaluation en vue de pouvoir évaluer les avancées d'un nouveau système. Les avancées technologiques rapides ne font qu'augmenter ce phénomène. L'évaluation pour le DHM semble ainsi perpétuellement en retard sur son objectif.

Troisième constat : l'évaluation sert non seulement à améliorer le développement d'un système particulier (en passant par des mesures, des diagnostics et des questionnaires de satisfaction), mais aussi à comparer des systèmes les uns par rapport aux autres. Plusieurs campagnes ont été lancées, et ce qu'il en ressort finalement, c'est qu'il est très difficile de comparer plusieurs systèmes de dialogue, même s'ils ont été réalisés pour des contextes applicatifs comparables, par exemple le renseignement ferroviaire ou hôtelier pour ne citer que ces deux applications largement exploitées. Pour cet aspect, le domaine du DHM semble poser un problème plus délicat que les autres domaines du TAL, et contribue à l'image de fragilité attachée à son évaluation.

Face à ces constats, on peut s'interroger sur la faisabilité de l'évaluation pour le DHM. Dans ce but, cette section passe en revue les principaux problèmes et quelques méthodologies qui nous semblent prometteuses. La question de la faisabilité nous semble constituer un problème de fond qui n'est pas assez discuté et pour lequel nous tentons d'apporter quelques pistes de réflexion. Les critiques que nous venons de porter avec les constats précédents ne nous empêchent pas, dans un second temps, de proposer des pistes pour une meilleure prise en compte de la multimodalité dans des paradigmes existants. La section 10.2 présente ainsi quelques possibilités d'extension à la multimodalité des méthodologies prévues pour l'oral et des peu nombreuses méthodologies déjà tournées vers la multimodalité. L'exemple « mets ça ici » qui nous a servi au paragraphe 6.2.2 nous permet de les illustrer.

10.1.1. *Quelques expériences d'évaluation*

Reprenons comme dans les scénarios de la section 3.1 un personnage générique de concepteur : qu'il s'agisse de la dernière étape de conception ou d'une étape intermédiaire, notre concepteur de système se pose inévitablement le problème de l'évaluation. Compte tenu des simplifications effectuées par rapport aux théories linguistiques et pragmatiques, le système est-il suffisamment performant ? D'une manière générale, qu'il y ait eu des simplifications ou non, le système a-t-il des réactions satisfaisantes ? Chaque module remplit-il bien son rôle ? L'architecture est-elle pertinente pour les

traitements réalisés ? Le comportement du système correspond-il à l'idée que l'on s'en faisait au départ et qui était à la base des indications données au Magicien d'Oz ?

Bien entendu, la première idée qui vient à l'esprit une fois que le système est opérationnel est de procéder à des tests utilisateurs. Souvent, c'est là que le moral du concepteur est mis à rude épreuve : entre les sujets qui ne comprennent pas comment se servir du microphone et du bouton (ou de la pédale) *push-to-talk* associé ; ceux qui n'arrivent pas à utiliser l'écran tactile ; ceux qui ne contrôlent pas leur action sur le matériel et vont jusqu'à le détériorer ; ceux qui refont trois fois chaque geste ou redisent trois fois le même bout de phrase de peur que le système en ait raté une partie ; ceux qui s'expriment tellement spontanément que leurs phrases foumillent d'incises, de subordonnées relatives, d'hésitations et de corrections ; ceux qui sont tellement intimidés devant le système qu'il s'expriment en style télégraphique ; et surtout ceux qui dépassent le cadre applicatif préalablement défini… Il suffit souvent d'un seul mot (qui n'avait pas été imaginé au départ) pour que soit généré le désespérant « je n'ai pas compris… », situation vécue au cours de l'évaluation du projet Ozone, avec le système multimodal de réservation de trains qui intégrait une modélisation complète des trains, des gares, des horaires, mais dont le lexique ne contenait pas le « demain » qu'un des évaluateurs a énoncé.

Face à ce constat, le concepteur procède alors avec ses sujets à une formation sur l'interaction homme-machine, sur ce qu'est le dialogue finalisé spontané, sur la façon dont le système fonctionne, et surtout sur le domaine applicatif et son étendue. A l'issue, un sujet produit directement des énoncés valables et le système fonctionne beaucoup mieux. Si ce n'est pas le cas, une séance d'entraînement (ne comptant pas dans l'évaluation) peut être envisagée. Mais justement, qu'en est-il de l'évaluation ? A force de répéter des exemples de phrases que le système est capable de traiter, les sujets sont forcément conduits à énoncer ces mêmes phrases, et il devient difficile d'évaluer le côté « spontané » du dialogue. Même chose dans le cas où le module de reconnaissance de la parole nécessite un apprentissage : à entraîner ce module sur des phrases de référence répétées par le locuteur, celui-ci en vient à se faire une idée précise des capacités du système et à diriger son comportement en conséquence.

Néanmoins, notre concepteur peut désormais comparer les énoncés des sujets avec les énoncés imaginés au départ. C'est parfois un deuxième coup porté à son moral : en essayant de faire fonctionner le système ou tout simplement en s'en tenant aux exemples fournis, les sujets réduisent parfois leurs énoncés à des phrases très simples, dépourvues de tous les phénomènes prévus au départ et qui étaient sensés faire l'intérêt et le caractère innovant du système. A titre d'exemple, un des sujets enregistrés dans le corpus Magnet'Oz, voir le chapitre 13 de (Van Deemter et Kibble, 2002), n'a produit en une demi-heure quasiment que deux phrases, à savoir « mets ça ici » et « mets ça là », ce qui relève effectivement de la multimodalité attendue, mais s'avère pour le moins restreint ! Notre concepteur qui souhaite évaluer son système en le testant par exemple sur des situations de référence multimodale intéressantes se retrouve donc

frustré. Au point qu'il incite parfois le sujet à produire un énoncé proche de celui qu'il voudrait évaluer (situation vécue au cours de l'enregistrement de Magnet'Oz). A ce stade, les conditions expérimentales ne sont plus un souci et on en vient à reprendre la liste des phénomènes initiaux pour les faire passer un à un dans le système... Si cette méthode permet au concepteur de faire un diagnostic de son système, il est clair qu'elle ne permet pas de produire une évaluation satisfaisante.

Notre concepteur tente alors une autre méthode : obtenir de ses sujets leurs impressions ainsi que toutes sortes d'informations sur leur expérience avec le système. Qu'il s'agisse d'un entretien enregistré ou du remplissage, libre ou dirigé, d'un questionnaire, les informations recueillies sont bien souvent en deçà des attentes du concepteur : informations trop générales (« oui, le système a bien répondu, mais pas toujours », « c'est un beau projet », « c'était amusant ») ; manque de précision (« c'était un peu lent ») ; décalage par rapport à ce qui est attendu (« j'ai utilisé des gestes ») ; hors sujet (longue appréciation de l'apparence des icônes et des objets affichés, critique de la tâche applicative ou des performances du système expert intégré dans le système). Au final, le concepteur se retrouve avec des anecdotes à raconter mais rien qui puisse être exploité proprement pour une évaluation ou un diagnostic de son système.

Pour les systèmes destinés à des spécialistes, c'est-à-dire à des utilisateurs qui connaissent la tâche ou à des utilisateurs qui se servent déjà du même type de système mais dans des versions construites sur des interfaces graphiques classiques, le pire est parfois à venir : ces spécialistes ont leurs habitudes et cherchent à les retrouver ou à les tester face au nouveau système multimodal. La moindre erreur, la moindre incertitude, le moindre écart par rapport à la normalité qu'ils se sont définie sont à leurs yeux intolérables. Ils sont alors d'autant plus susceptibles de remettre en question l'intérêt de la commande vocale ou multimodale. Autrement dit, ce sont les utilisateurs les plus exigeants que l'on puisse imaginer, et obtenir d'eux des éléments d'évaluation est pour le moins risqué. C'est pourtant là que se trouve le principal enjeu du domaine du DHM en tant que domaine de recherche appliquée : avec l'objectif d'apporter à la société la spontanéité de la communication multimodale, il est nécessaire d'aller au-delà des prototypes de recherche et de viser la réalisation de systèmes grand public ou la transformation de systèmes graphiques existants en systèmes multimodaux. Autrement dit, il est nécessaire de viser le passage à l'échelle, avec tous les enjeux qualitatifs et quantitatifs que celui-ci comporte (voir paragraphe 3.2.6).

La prise en compte de tels aspects lors de la réalisation de systèmes de dialogue multimodaux pose de nombreux défis, et une réalisation parfaitement fiable n'est pas pour tout de suite. Or, dans l'absolu, une véritable évaluation n'est pertinente que dans des conditions réelles de fonctionnement. C'est peut-être là que le domaine du DHM se démarque de la plupart des autres domaines du TAL : le passage à l'échelle des systèmes reposant sur l'écrit et mettant en œuvre moins de composants logiciels sensibles n'est pas aussi délicat qu'en dialogue. Certains systèmes, par exemple les moteurs d'indexation et de recherche, sont même conçus très tôt pour traiter de grosses

masses de données dans des conditions réelles d'utilisation. Leur passage à l'échelle en est ainsi facilité, et des évaluations peuvent être entreprises sur des bases stables.

10.1.2. *Méthodologies pour les interfaces homme-machine*

Le passage à l'échelle est également maîtrisé dans le domaine des IHM, qui peut inspirer le DHM dans la mesure où l'interaction entre l'humain et la machine constitue le même socle. Comme le présentent (Kolski, 1993) et (Grislin et Kolski, 1996), une évaluation consiste à « vérifier » et à « valider » le système. Le système est vérifié s'il correspond aux spécifications issues de la définition des besoins, et est validé s'il correspond aux besoins en respectant les contraintes du domaine d'application. L'évaluation ergonomique d'une IHM consiste à s'assurer que l'utilisateur est capable de réaliser sa tâche au moyen du système qui lui est proposé. La notion d'utilité détermine si l'IHM permet à l'utilisateur d'atteindre ses objectifs de travail. Celle d'utilisabilité rend compte de la qualité de l'interaction homme-machine, en termes de facilité d'apprentissage et d'utilisation. Ces notions peuvent tout à fait s'appliquer au DHM, de même que les critères ergonomiques identifiés pour les IHM : l'acceptabilité sociale (des systèmes inacceptables socialement seraient par exemple des systèmes posant des questions indiscrètes aux utilisateurs), l'acceptabilité pratique (contraintes de production, de coût, de fiabilité). C'est aussi dans l'acceptabilité que l'on retrouve l'utilisabilité, avec des critères tels que : facile à apprendre, à mémoriser, peu d'erreurs, capacité de guidage de l'utilisateur, gestion de la charge de travail, etc.

Au-delà de ces critères, le domaine des IHM propose un ensemble de méthodes qui se combinent pour aboutir à une évaluation pertinente : analyse des avis d'utilisateurs représentatifs, analyse des activités d'utilisateurs représentatifs (enregistrement vidéo, observation, utilisation d'un oculomètre, mesures physiologiques), jugement d'experts, grilles d'évaluation parcourant la liste des qualités d'un bon système. Si ces méthodes ne peuvent être utilisées que dans le cas d'un système effectif, d'autres sont envisageables au cours même de la conception du système : jugement d'experts, modélisation théorique de l'interaction (approches analytiques : modèles formels prédictifs, modèles de qualité, modèles logiciels). Enfin, un troisième ensemble de méthodes est dédié à l'évaluation *a priori*, c'est-à-dire dès les phases de spécification, par exemple en prenant en compte l'importance des facteurs humains lors de la conception du système, ou en suivant les principes de l'ingénierie cognitive, notamment celui visant à replacer l'utilisateur au cœur de chacune des étapes de spécification et de conception.

Beaucoup parmi ces méthodes n'ont pas d'équivalent en DHM. C'est le cas par exemple des approches analytiques qui visent à modéliser de manière formelle le comportement d'un utilisateur : si certains comportements face à une IHM peuvent être prévisibles et formalisables, le problème est plus complexe en DHM. Il suffit de se reporter à l'ensemble des chapitres précédents. Certaines méthodes sont en revanche parfaitement exploitables, et sont même matérialisées de manière plus précise pour

le DHM. A propos de la méthode classique du jugement par un expert, (Gibbon *et al.*, 2000) montre par exemple qu'en DHM, au moins trois experts sont nécessaires pour identifier à peu près la moitié des problèmes d'utilisabilité. Plus il y a d'experts impliqués, plus il y a de problèmes identifiés mais plus l'évaluation s'avère coûteuse, en temps comme en moyens humains. Plus récemment, (Kühnel, 2012) présente un ensemble de méthodes, incluant le test par un groupe d'experts, mais aussi la méthode de la visite guidée cognitive (*Cognitive Walkthrought*), qui consiste avant tout en une analyse de la tâche avec décomposition en actions : au moins un expert suit un chemin déterminé comme optimal pour résoudre la tâche, et vérifie à chaque étape que l'étape suivante est accessible pour tout utilisateur novice.

10.1.3. *Méthodologies pour le dialogue oral*

Plus spécifiquement sur le DHM oral, un certain nombre de méthodes ont été proposées (Antoine et Caelen, 1999 ; Devillers *et al.*, 2004 ; Dybkjær *et al.*, 2004 ; Walker, 2005 ; Möller *et al.*, 2007 ; Kühnel, 2012). Elles constituent une sorte de cadre de référence comprenant des recommandations pour mettre en œuvre des tests d'interaction avec des utilisateurs, des méthodes pour analyser automatiquement ou semi-automatiquement les traces d'interaction obtenues, des repères pour déterminer des métriques d'évaluation, ou encore des principes pour constituer et analyser des questionnaires remplis *a posteriori* par les utilisateurs. On retrouve donc quelques-unes des méthodes utilisées en IHM. Chaque évaluateur de système peut ainsi piocher dans ce stock pour déterminer la ou les méthodes qu'il va appliquer. En fait, un seul test semble insuffisant et une véritable évaluation semble devoir grouper plusieurs types de tests. Les campagnes d'évaluation (Evalda/Media : méthodologie d'évaluation de la compréhension hors et en contexte du dialogue), les groupes de travail (groupe MadCow, groupe « compréhension de parole » du GdR I3) et les divers consortiums de projets européens exploitent largement ce principe. Lorsque plusieurs systèmes sont en jeu et que l'évaluation est comparative, des règles de fonctionnement peuvent être définies de manière à mieux contrôler la qualité de l'évaluation. La campagne d'évaluation par défi avec sa gestion croisée des rôles des concepteurs des systèmes en jeu (Antoine, 2003) en est un exemple.

Les principales propositions de méthodologie s'accompagnent chacune d'une idée originale qui vient simplifier la mise en œuvre d'un type de test en lui apportant un moyen d'être opérationnalisé dans un contexte déterminé. Le paradigme du groupe MadCow (Hirschman, 1992) apporte ainsi la notion de gabarit qui caractérise les solutions minimales et maximales à une requête et rend ainsi son évaluation plus rigoureuse. La paradigme Paradise, *Paradigm for Dialogue System Evaluation* (Walker *et al.*, 2001), se focalise sur la maximisation de la satisfaction de l'utilisateur et propose de prendre la satisfaction de la tâche comme référence. Autre exemple d'idée originale, (López-Cózar Delgado *et al.*, 2003) proposent d'évaluer un système en générant

automatiquement des énoncés utilisateurs de test, c'est-à-dire en modélisant le comportement de l'utilisateur, y compris ses erreurs. En France, cette méthode a été reprise dans le paradigme Simdial (Allemandou *et al.*, 2007), dans lequel la simulation déterministe d'un utilisateur permet d'évaluer automatiquement les capacités dialogiques du système, notamment grâce à la notion de phénomène perturbateur, qui, à l'instar du bruit dans les Magiciens d'Oz de (Rieser et Lemon, 2011), permet d'introduire des contestations ou des demandes de reformulation qui vont permettre d'évaluer le comportement général et la robustesse du système. Par ailleurs, la méthodologie DQR, Donnée–Question–Réponse, voir notamment le chapitre de J. Zeiliger *et al*. dans (Mariani *et al.*, 2000), introduit le principe de questionner le système sur le point à évaluer, avec l'avantage de déplacer ainsi l'objet de l'évaluation de la donnée vers la question, et donc ni sur les réponses ou réactions du système (méthode « boîte noire », qui ne nécessite pas d'explorer les structures internes au système, mais qui manque de précision), ni sur les structures sémantiques du système (méthode « boîte transparente », précise et conduisant facilement à un diagnostic, mais qui nécessite de disposer de représentations sémantiques de référence). Encore faut-il que le système soit capable de répondre aux questions Q de DQR. Le paradigme adapté DCR, Demande–Contrôle–Réponse-Résultat-Référence (Antoine et Caelen, 1999), minimise ce problème en remplaçant la question par un contrôle qui est une simplification ou une reformulation de la demande utilisateur initiale. Pour sa part, le paradigme Peace, Paradigme d'Evaluation Automatique de ComprEhension, voir le chapitre de L. Devillers *et al*. dans (Gardent et Pierrel, 2002), apporte l'idée originale de modéliser l'historique du dialogue par une paraphrase, ce qui permet de rester dans le mode « boîte noire » tout en permettant une évaluation de la compréhension en contexte.

10.1.4. *Méthodologies pour le dialogue multimodal*

Dans le contexte du DHM multimodal, les propositions sont loin d'être aussi pertinentes. Le paradigme Promise, *Procedure for Multimodal Interactive System Evaluation* (Beringer *et al.*, 2002), est présenté comme une extension de Paradise à la multimodalité, avec des principes pour affecter des scores aux entrées et sorties multimodales. La proposition reste en fait à un niveau très approximatif, bien en deçà de la variété des phénomènes multimodaux. Les aspects intéressants de l'article concernent le dialogue oral, avec des considérations sur le niveau de complétude de la tâche et le niveau de coopération de l'utilisateur. Les travaux de N.O. Bernsen et L. Dybkjær, qui font pourtant référence dans le milieu du dialogue multimodal, sont plutôt décevants en ce qui concerne l'évaluation. (Bernsen et Dybkjær, 2004) présentent ainsi une méthodologie prévue pour un système, avec une focalisation sur la méthode du questionnaire rempli *a posteriori* par les utilisateurs. La raison donnée est d'ailleurs que les autres méthodes ne sont pas encore bien établies. Malheureusement, les questions du questionnaire restent à un niveau très superficiel pour ce qui concerne la multimodalité : « avez-vous utilisé la souris ou avez-vous pointé sur l'écran ? », « quelles étaient vos impressions en produisant un geste ? », et « auriez-vous aimé en faire plus avec le

geste ? si oui, pour faire quoi ? ». Les réponses qui ont été fournies par les utilisateurs semblent également très pauvres, d'autant plus qu'une des conclusions des auteurs est que les utilisateurs ont préféré parler plutôt qu'exploiter les possibilités multimodales... Pour sa part, (Dybkjær *et al.*, 2004) est plus une revue de méthodologies et de projets qu'une proposition de nouvelle méthodologie pour la multimodalité : le propos reste au niveau de recommandations générales. Dans un autre registre, (Vuurpijl *et al.*, 2004) présente un outil, appelé « μ-eval », pour la transcription des données multimodales et l'évaluation d'un système. Or l'évaluation ne concerne que les tours de dialogue et ne traite pas les phénomènes multimodaux. Enfin, (Walker *et al.*, 2004) se focalise sur les modèles utilisateur et les stratégies de dialogue (oral) mais quasiment pas sur les aspects multimodaux.

D'une manière générale pour l'évaluation des systèmes multimodaux, on ne retrouve donc pas les principes appliqués dans les campagnes d'évaluation des systèmes oraux. Chaque proposition de méthodologie n'est réalisée que dans le cadre d'un seul système, qui est d'ailleurs le système sur lequel la méthodologie est testée (voir notre deuxième constat). Plutôt que de nous focaliser sur ces travaux, nous allons reprendre quelques problèmes identifiés et méthodologies proposées dans le cadre du dialogue oral, afin de vérifier leur pertinence dans le cadre du dialogue multimodal.

10.2. Enjeux pour l'évaluation des systèmes multimodaux

10.2.1. *Evaluation globale ou évaluation segmentée ?*

Une première question qui se pose lors de la mise en place d'une procédure d'évaluation est le choix entre une évaluation globale et une évaluation segmentée (ou évaluation par module). L'évaluation « globale » considère le système de dialogue comme une boîte noire et s'intéresse aux énoncés échangés, c'est-à-dire aux traces d'interaction. L'évaluation porte alors sur la pertinence de la réaction du système par rapport à l'énoncé utilisateur, sur l'avancée de la tâche au fur et à mesure des échanges, mais ne prend en compte ni l'architecture ni les fonctionnalités internes du système. Sa mise en œuvre relève des tests utilisateurs et du passage sur corpus (incluant les suites de test). L'évaluation proprement dite peut consister en une simple analyse subjective des traces d'interaction, ou en l'application de métriques objectives permettant d'aboutir à des résultats chiffrés (Walker *et al.*, 2001). Dans tous les cas, l'application de cette méthode aux systèmes multimodaux pose le problème de la couverture des tests effectués, donc de la couverture des situations de dialogue évaluées et la couverture du corpus de test. Pour valider un système fondé sur une interaction spontanée, il faudrait en effet tester un très grand nombre de situations de manière à rendre compte de la variabilité de l'interaction. Ce problème est déjà présent en dialogue oral, mais la multiplication des paramètres dans les situations de communication multimodale le rend plus prégnant ici. A titre d'exemple, les références aux objets peuvent prendre une variété de formes linguistiques incluant divers types de pronoms et de groupes

nominaux. En DHM multimodal, chacune de ces formes existe et peut de plus s'associer à un geste de désignation. La variété des gestes de désignation est donc à prendre en compte, de même que celle des contextes visuels dans lesquels les gestes ont été produits (voir chapitre 6). La combinatoire est ainsi bien plus grande et la couverture des situations de référence devrait suivre cette augmentation d'échelle.

L'évaluation « segmentée » considère le système comme une boîte transparente et s'intéresse aux fonctionnalités et représentations internes. L'évaluation porte alors sur les entrées et sorties de chaque module. Dans le cadre du dialogue oral, on se limite souvent à la sortie du module sémantique et on compare les représentations sémantiques obtenues à des représentations de référence. L'application de cette méthode aux systèmes multimodaux pose quelques problèmes, certains déjà présents pour l'oral mais rendus plus prégnants, et d'autres spécifiques à l'introduction de la multimodalité. Ainsi, si nous nous focalisons sur l'évaluation de la compréhension multimodale, c'est-à-dire sur la fusion des informations captées en entrée (sachant que l'évaluation de la génération multimodale pose des questions similaires) :

– on peut faire comme pour l'oral et se focaliser sur les représentations sémantiques multimodales obtenues en sortie du module gérant l'analyse sémantique globale, c'est-à-dire du module chargé de la fusion multimodale. Selon le système (par exemple selon que la multimodalité regroupe le langage naturel et le geste conversationnel, ou regroupe au contraire la reconnaissance des émotions sur le visage de l'utilisateur avec la lecture sur ses lèvres et l'analyse du langage naturel), ces représentations sémantiques multimodales sont très variables et couvrent des phénomènes très différents. Le problème majeur qui se pose alors est la détermination de représentations multimodales de référence qui soient communes à plusieurs systèmes. Spécifier des représentations exhaustives est quasiment impossible, surtout que les technologies évoluent et rendent toute spécification rapidement obsolète ;

– on peut considérer au contraire un système multimodal comme un processus de fusion en plus d'un ensemble de systèmes monomodaux, chacun d'eux se caractérisant par un type de représentation sémantique. L'évaluation concerne alors d'une part le processus de fusion, et d'autre part chacun des systèmes monomodaux, avec à chaque fois des représentations de référence. On devra donc spécifier au préalable des représentations sémantiques de référence pour les trajectoires gestuelles, d'autres pour l'interprétation des émotions, etc. L'intérêt de cette méthode est de mieux cibler le diagnostic, car on peut identifier de quelle chaîne de traitement monomodale vient un manque. Ses inconvénients sont bien entendu la multiplicité des représentations de référence indispensables, ainsi que la nécessité d'une méthode d'évaluation spécifique à la fusion multimodale ;

– d'autre part, en poursuivant cette voie, on peut considérer que l'évaluation doit s'appliquer de manière modulaire, c'est-à-dire en utilisant des représentations de référence pour évaluer les sorties de chacun des modules du système. Le problème majeur de cette approche, outre une forte dépendance à l'architecture, est la multiplication

des modules et donc celle des évaluations et des représentations associées : représentations lexicales, syntaxiques et sémantiques du langage naturel, des expressions du visage, des trajectoires gestuelles, etc. Une évaluation modulaire constitue donc une charge importante de travail. Autre problème : en multipliant les évaluations locales, il devient difficile de confronter les mesures pour obtenir une évaluation globale du système. Le diagnostic devient plus précis, mais au détriment d'une mesure simple permettant d'appréhender la qualité du fonctionnement global. D'autre part, et c'est là un point fondamental, les erreurs d'un module peuvent être rattrapées par les performances d'un autre module, sans conséquence sur ce fonctionnement global. Il ne s'agit pas de compenser la mauvaise réalisation d'un module par la réalisation exemplaire d'un autre module, mais de compenser des erreurs inévitables par des procédures de rattrapage pertinentes. L'exemple typique concerne les erreurs de reconnaissance vocale : il est illusoire d'espérer un module de reconnaissance qui soit capable de performances parfaites (100 % de mots correctement reconnus) avec un vocabulaire de grande taille ou même de quelques centaines de mots. En revanche, si les modules sémantiques et pragmatiques sont capables de transformer des informations sémantiques et contextuelles en contraintes sur la reconnaissance, le système pourra retrouver de manière sûre l'énoncé prononcé, même si le moteur de reconnaissance a des performances médiocres. Autrement dit, l'évaluation du module de reconnaissance n'a aucun intérêt, et seule compte l'évaluation de l'interaction entre module de reconnaissance et modules sémantiques et pragmatiques. L'évaluation modulaire n'est donc pas si simple à mettre en œuvre ni si pertinente.

10.2.2. *Faut-il gérer un corpus multimodal ?*

Un deuxième problème qui se pose lors de la mise en place d'une procédure d'évaluation concerne l'exploitation d'un corpus multimodal. Il s'agit ici d'un corpus de test, bien entendu différent du corpus préalablement utilisé pour le recueil de phénomènes ou pour un éventuel apprentissage. On fournit en entrée du système toutes les situations comprises dans le corpus, et on teste soit les sorties du système soit ses représentations sémantiques internes. Si la procédure semble claire, elle pose néanmoins un certain nombre de problèmes liés au codage de la multimodalité.

Au niveau le plus « brut », un corpus multimodal est un enregistrement des signaux captés par le système : signal audio capté par le microphone, signal gestuel capté par l'écran tactile, signal vidéo capté par une ou plusieurs caméras, etc. Si un tel enregistrement est idéal pour une simulation des entrées du système, il ne permet aucune manipulation des données (on pensera à la dérivation d'exemples à partir d'un exemple initial, voir paragraphe 3.2.2), et s'avère difficilement caractérisable en termes de phénomènes. Pour ce faire, l'annotation du corpus, donc le passage du niveau brut à un niveau interprété, est souvent une opération indispensable. Or l'annotation d'un corpus multimodal pose des problèmes liés à la nature des signaux enregistrés. Contrairement à la parole qui peut se transcrire de manière relativement simple et objective

en phrases écrites, le geste et les autres modalités ne peuvent pas se transcrire simplement, comme on l'a vu au paragraphe 6.1.3. Il n'en reste pas moins que ces freins techniques ne doivent pas faire perdre du vue que l'exploitation de corpus reste indispensable en DHM.

10.2.3. *Peut-on comparer plusieurs systèmes multimodaux ?*

Un troisième problème concerne la mise en œuvre d'une « évaluation comparative ». Le principe est de comparer plusieurs systèmes de dialogue ayant des compétences similaires sur le même type d'application. Mais, dans les travaux existants qui se limitent au dialogue oral, la comparaison porte rarement entre un système de dialogue oral et un autre type de système de référence, par exemple un système de dialogue écrit. L'intérêt serait pourtant d'évaluer l'apport de la parole en tant que source d'amélioration de la communication entre l'utilisateur et sa machine, ou source d'amélioration de l'efficacité de gestion de la tâche. La question se pose surtout dans le cadre du dialogue multimodal. Bien souvent, on présente la capacité multimodale comme un plus par rapport à la capacité linguistique : la multimodalité est avancée comme étant plus efficace, plus rapide, plus précise et plus directe, en particulier pour les actions de référence qui permettent un accès direct aux objets (sans passer par le biais de descriptions spatiales complexes et potentiellement ambiguës). Une procédure d'évaluation comparative devrait donc inclure des systèmes oraux en plus des systèmes multimodaux. De plus, et particulièrement dans les milieux professionnels, la multimodalité est aussi présentée comme un plus par rapport à l'interaction graphique classique à base de fenêtres, de menus, d'icônes et de boutons. L'accès aux objets affichés à l'écran est en effet le fondement de l'une comme de l'autre. Une procédure d'évaluation comparative devrait donc inclure des interfaces graphiques en plus des systèmes multimodaux. Les aspects d'efficacité, de rapidité et de précision deviennent autant de mesures permettant de comparer un système multimodal et une interface graphique pour la même tâche (tous les systèmes multimodaux ne sont néanmoins pas réalisés après une première version graphique).

Questionnaire à destination des sujets d'un test utilisateur, évaluation à vocation de diagnostic, évaluation globale, évaluation segmentée, évaluation fondée sur les réactions du système ou par comparaison de ses représentations internes avec des représentations de référence : la majorité des méthodes proposées pour le dialogue oral semble applicable au dialogue multimodal, au prix de quelques précautions de mise en œuvre. Parmi les approches qui nous semblent très délicates à appliquer au multimodal se trouvent l'évaluation comparative et le passage sur corpus. Même si ce dernier reste utile dans un but de test ou d'entraînement, il est pour l'instant difficile d'imaginer une réutilisation de corpus multimodaux « tout trouvés », annotés et pouvant être adaptés à l'évaluation d'un système pour lequel il n'a pas été conçu. On peut cependant espérer un développement futur des corpus multimodaux, surtout si la synergie qui s'organise

autour de la paire corpus-campagne d'évaluation, synergie que l'on peut observer avec Evalda/Media (Devillers *et al.*, 2004), s'étend à la multimodalité.

Finalement, l'impression qui ressort est que le domaine du DHM, et *a fortiori* celui du dialogue multimodal, se prête moins bien que les autres domaines du TAL à l'évaluation. Une méthodologie à la fois précise, fiable, objective, complète, indépendante des systèmes, des modalités et des tâches, reste un graal inaccessible. Les acronymes des méthodologies proposées reflètent eux-mêmes cette impression : Paradise, Peace, Promise, etc. Heureusement, les problèmes soulevés, les questions posées, les aspects traités, même s'ils n'aboutissent pas à une méthodologie d'évaluation unanime, contribuent grandement à l'amélioration des processus de réalisation et de test d'un système de dialogue.

10.3. Eléments méthodologiques

Parmi les aspects méthodologiques que nous avons évoqués, nous en reprenons ici quelques-uns qui constituent autant d'éléments pour une méthodologie d'évaluation des systèmes de dialogue multimodaux. Ces propositions concernent ainsi le niveau d'expertise de l'utilisateur, les questionnaires pour les sujets de tests utilisateurs, la mise en œuvre de questions et de commandes évaluatives pour les systèmes multimodaux (il s'agit d'étudier la faisabilité d'un DQR-DCR multimodal), ainsi que la mise en œuvre de paraphrases d'un historique multimodal (il s'agit d'étudier la faisabilité d'un Peace multimodal). Nous laissons de côté les autres idées proposées dans le cadre du dialogue oral, comme par exemple la simulation du comportement d'un utilisateur avec la génération automatique d'énoncés de test. Le comportement multimodal d'un utilisateur reste en effet mal connu et une telle simulation peut sembler pour l'instant bien délicate. C'est en tout cas un enjeu pour le dialogue multimodal.

10.3.1. *Expertise de l'utilisateur et complexité du système*

Nous nous intéressons ici au « niveau d'expertise » de l'utilisateur en tant qu'information fournie par l'utilisateur et utile au dépouillement d'un questionnaire ainsi qu'à une meilleure analyse de son comportement face au système. Cette préoccupation peut sembler peu compatible avec le dialogue naturel en langage naturel : à partir du moment où le système visé doit permettre un dialogue spontané, tout utilisateur qui sait parler est expert. Dans les faits, les systèmes de DHM ne font que s'approcher du dialogue naturel, et, surtout quand ils intègrent des dispositifs de capture multimodaux, restent avant tout des systèmes informatiques, pour lesquels les avis d'utilisateurs plus ou moins experts sont essentiels. Dans les méthodologies citées aux paragraphes 10.1.3 et 10.1.4, ce niveau d'expertise est souvent réduit à l'opposition entre novice et expert, voir chapitre de L. Devillers *et al*. dans (Gardent et Pierrel, 2002). A la suite du *Danish Dialogue Project*, il est parfois fait mention de deux indicateurs, un

pour le niveau d'expertise vis-à-vis du domaine et l'autre vis-à-vis du système, mais les deux indicateurs restent binaires, entre novice et expert (Dybkjær *et al.*, 2004). Les perspectives données à la fin de ces articles soulignent la nécessité d'une échelle plus fine pour mieux analyser les données fournies par les tests.

Or des travaux classiques en IA proposent depuis longtemps des typologies à plusieurs niveaux, chaque niveau se caractérisant par un type de comportement face à un problème. Dans (Dreyfus et Dreyfus, 1986), ce sont ainsi cinq niveaux d'expertise qui sont détaillés :

– le novice, qui se contente d'appliquer des règles fixes et déterministes ;

– le pratiquant, qui reconnaît des situations et agit en fonction d'expériences passées ;

– le compétent (*competent*), capable de réaliser des plans stratégiques ;

– le très compétent (*proficient*), qui élague inconsciemment des plans n'ayant aucune chance d'aboutir ;

– l'expert, qui se déplace inconsciemment dans un espace de raisonnement.

Fournir ces définitions aux utilisateurs pour qu'ils soient capables de déterminer leur niveau d'expertise est peut-être un peu compliqué et risqué. Une solution consiste à leur poser directement des questions telles que « avez-vous reconnu une situation de communication à laquelle vous êtes habitué ? » et à en déduire pas à pas leur niveau d'expertise.

Dans le cas d'utilisateurs professionnels tels que des militaires connaissant parfaitement leur tâche et ayant une connaissance plus ou moins approfondie du système, le niveau d'expertise est une information fondamentale. On peut imaginer un opérateur qui a le statut d'expert pour un système à commande vocale mais celui de novice pour le système multimodal dédié à la même tâche. L'analogie avec le nombre d'heures de vol d'un pilote est immédiate car elle correspond au même problème : à la place ou en complément d'une étiquette telle que « compétent » ou « expert », il est intéressant de tirer parti du nombre d'heures que l'utilisateur a passé avec le système. Pour un avion comme pour un système de dialogue, ce nombre d'heures inclut implicitement un temps de formation et d'apprentissage, et une seule mesure peut suffire. Si le temps passé à suivre une formation est considéré comme une mesure pertinente, par exemple si les tests utilisateur comportent deux groupes de sujets, l'un ayant suivi une formation rapide et l'autre une formation approfondie, on pourra soit garder deux indicateurs (deux nombres d'heures), soit procéder à un calcul avec des pondérations différentes pour la formation et pour la pratique. Suivant le même principe, on peut faire une distinction entre le temps passé avec le vrai système et celui passé avec un Magicien d'Oz. L'analogie repose alors sur le nombre d'heures qu'un pilote passe dans un simulateur, et le niveau d'expertise inclut une nouvelle pondération. Toujours

selon ce principe, on peut dissocier le temps éventuellement passé à entraîner le moteur de reconnaissance vocale, temps qui a permis à l'utilisateur de se faire une idée précise des phrases opérationnelles. Ou bien on peut considérer que répéter des énoncés préparés constitue une sorte de formation, et donc inclure ce temps dans le temps de formation.

Il est en tout cas possible d'aller au-delà de la dichotomie novice-expert, le tout est de savoir comment une information plus précise peut être exploitée. En effet, disposer de plusieurs niveaux d'expertise permet de comparer plusieurs utilisateurs ayant des niveaux similaires (égaux pour une mesure discrète, dans un même intervalle pour une mesure continue comme avec les calculs précédents), aussi bien que des utilisateurs de niveaux variés. L'évaluation du système de dialogue inclut dans ce cas de nouveaux indicateurs, par exemple :

– des utilisateurs de même niveau ont-ils le même comportement multimodal et ont-ils les mêmes problèmes pour utiliser le système ?

– un utilisateur d'un niveau plus élevé est-il plus efficace, plus rapide et plus précis dans son utilisation de la multimodalité ? couvre-t-il un ensemble plus large de phénomènes ? fait-il appel à plus de fonctionnalités du système ?

Enfin, si l'on en vient à gérer un niveau d'expertise non pas pour le système multimodal dans sa globalité mais seulement pour les aspects multimodaux de l'interaction, un autre problème se pose : celui de la légitimité d'une mesure indépendante du système et donc valable pour tout système multimodal. Ce niveau d'expertise multimodale se baserait sur l'utilisation d'un ou de plusieurs systèmes multimodaux et nécessiterait la prise en compte d'un indicateur de complexité de chaque système. En reprenant l'analogie précédente, cent heures de vol sur un petit avion de tourisme n'équivalent pas à cent heures de vol sur un chasseur, même si les deux types d'avion présentent des points communs dans leur façon d'être pilotés. La pondération du temps par un indice de complexité de chaque système est une solution possible. Cet indice de complexité devrait dépendre des aspects suivants : nombre de mots du vocabulaire du système, nombre de constructions syntaxiques reconnues, nombre de types de gestes traités, nombre d'actes de langage, nombre de dispositifs (terminaux, écrans), nombre de langues détectées, nombre de types d'échanges possibles, éventuellement nombre de primitives de la tâche applicative. On retrouve l'ensemble des fonctionnalités des systèmes multimodaux, et cela rend difficile la mise en œuvre d'un indice absolu. Une solution consiste donc à choisir un système de référence et à calculer, sous la forme d'un ratio, l'écart entre ce système et le système considéré.

10.3.2. *Questionnaires pour les utilisateurs*

Le questionnaire constitue la principale source d'informations subjectives. Il permet d'élargir le champ de l'évaluation et de procéder à un diagnostic, non seulement

du système mais aussi du contenu de la formation des utilisateurs. Nous allons reprendre ici les questionnaires proposés pour le dialogue oral et les compléter en y ajoutant des préoccupations multimodales.

Certaines questions d'ordre général peuvent être reprises des questionnaires de N.O. Bernsen et L. Dybkjær. Il s'agit essentiellement des questions portant sur les impressions de l'utilisateur, sur l'utilité perçue du système, et sur les voies d'améliorations possibles : « était-ce facile ou difficile d'utiliser le système ? pourquoi ? » ; « pouviez-vous comprendre ce qu'il disait ? » ; « arrivait-il à suivre ce que vous vouliez lui dire ? » ; « que pensez-vous de son comportement sur l'écran ? » ; « qu'est-ce qui était bien dans votre interaction avec le système ? » ; « qu'est-ce qui était mauvais dans votre interaction avec le système ? » ; etc. (Bernsen et Dybkjær, 2004). A ces questions nous pouvons rapidement en ajouter d'autres qui portent sur les modalités et sur la multimodalité :

– pouviez-vous comprendre les gestes qu'il produisait ? arrivait-il à tenir compte de vos gestes ? que pensez-vous du fonctionnement du retour d'effort ?

– les messages qu'il produisait étaient-ils correctement émis ? de manière synchrone ? sans incohérence ni manque ? arrivait-il à relier correctement vos gestes avec ce que vous disiez ? comprenait-il vos messages dans leur intégralité ?

Les questions générales du paradigme Paradise peuvent également être étendues rapidement à la multimodalité, avec le même principe consistant à répondre par un indice de satisfaction compris entre 1 et 5 :

– le système était facile à comprendre (valable en oral et en multimodal) ;

– le système comprenait ce que je disais (on remplace « dire » par « produire ») ;

– j'ai obtenu les informations que j'ai demandées (toujours valable) ;

– le rythme des interactions était correct (toujours valable, mais très important en multimodal car inclut les éventuels problèmes de synchronisation temporelle entre les différentes modalités – autrement dit la réponse ne s'interprète pas de la même façon selon que le système est oral ou multimodal) ;

– à chaque moment je savais ce que je devais dire (« produire », donc) ;

– le système expliquait clairement ce qu'il avait compris (toujours valable) ;

– les questions ou suggestions du système m'ont aidé (toujours valable).

Nous pouvons d'autre part proposer quelques exemples de questions plus ou moins ciblées sur les phénomènes multimodaux et reparties en quatre catégories :

1) questions portant sur les conditions d'interaction : vous sentiez-vous libre quand vous parliez ? quand vous produisiez des gestes ? quand vous utilisiez le dispositif à retour d'effort ? le système a-t-il correctement géré les dispositifs à sa disposition ? avez-vous ressenti des contraintes dans leur utilisation ? lesquelles ?

2) questions portant sur le traitement des entrées : le système a-t-il bien compris votre façon de communiquer avec lui ? vous a-t-il paru sensible à vos émotions ? d'une manière générale, le système a-t-il bien pris en compte votre expressivité ?

3) questions portant sur la gestion des sorties : que pensiez-vous de la façon dont les réponses vous étaient présentées (que ce soit oralement, graphiquement, ou les deux) ? l'avatar vous semblait-il naturel ? lorsque l'avatar produisait un geste, cela vous semblait-il pertinent ?

4) questions portant sur les rapports entre la gestion des entrées et celle des sorties : avez-vous senti une cohérence entre vos messages et les messages produits par le système ? l'expressivité vous semble-t-elle meilleure en entrée ou en sortie ?

On le voit, il n'est pas difficile de spécifier un ensemble de questions portant sur la gestion de la multimodalité. Celles-ci nous semblent en tout cas plus précises que celles de (Bernsen et Dybkjær, 2004) et devraient éviter des réponses trop évasives. Elles ne font cependant pas oublier que rien ne remplace l'écriture spontanée d'un texte décrivant les impressions du sujet, en particulier avant que toute question ne soit posée. En effet, une question dirige l'attention du sujet vers le thème mentionné et peut de ce fait introduire un biais par rapport aux impressions initiales de celui-ci. Même l'ordre des questions peut avoir une influence sur les réponses. Spécifier un questionnaire devrait donc se faire sous la direction d'un psychologue, ce qui n'est pas fait dans Paradise, dans Peace ou ici, et ce qui constitue l'un des enjeux de cet aspect méthodologique.

10.3.3. *Extension de DQR et de DCR au dialogue multimodal*

J. Zeiliger *et al*. dans (Mariani *et al.*, 2000) ont retenu une méthodologie de type « boîte noire » et permettant de faire un diagnostic du système, méthodologie qui repose sur des tests génériques pour l'évaluation de la compréhension d'un énoncé isolé. Les aspects contextuels ont été négligés (nous y reviendrons avec Peace), mais c'était le prix à payer pour obtenir une méthodologie simple et bien délimitée. Le principe est de procéder à des évaluations ponctuelles, chacune d'entre elles étant centrée sur un phénomène particulier. Ainsi, dans la matérialisation DQR, l'évaluation ponctuelle prend la forme d'une question Q adressée au système et permet de vérifier sa bonne compréhension de la demande initiale D. Un des exemples donnés concerne la résolution des anaphores, avec la demande, la question et la réponse suivantes :

– D = « vous prenez la rue à droite et vous la suivez sur 300 mètres » (énoncé initial, tel qu'il a été adressé au système dans le but de faire avancer la tâche) ;

– Q = « suivre rue à droite ? » (question adressée au système juste après l'énoncé D et destinée à évaluer la bonne compréhension de D) ;

– R = « oui » (réponse du système montrant que l'anaphore a été bien comprise et rendant l'évaluation positive).

Les auteurs spécifient sept niveaux caractérisant la portée des questions posées. Nous reprenons ici ces niveaux en indiquant à chaque fois comment étendre le paradigme pour pouvoir l'exploiter en dialogue multimodal.

– Niveau 1 = « information explicite ». Il s'agit du repérage d'une information explicitée dans l'énoncé, l'intérêt étant de tester la bonne compréhension de l'énoncé littéral compte tenu de la grande variabilité du langage spontané. Les exemples donnés par les auteurs se contentent de reprendre une partie de l'énoncé et de demander une confirmation de la compréhension de cette partie : D = « vous prenez à droite après les bâtiments blancs aux volets bleus » puis Q = « volets blancs ? » ou « volets bleus ? ». L'extension de ce principe à la multimodalité consiste à poser des questions sur les éléments de l'énoncé multimodal. Avec D = « mets ça ici » + geste en (x_1, y_1) + geste en (x_2, y_2), on peut tester les capacités de capture de la multimodalité en posant les questions Q suivantes : « ça ? » + geste en (x_1, y_1) ; « mettre ici ? » + geste en (x_2, y_2) ; « mettre ça ? » + geste en (x_2, y_2) ; « mettre ça ici » + geste en (x_2, y_2) + geste en (x_1, y_1) ; etc. La procédure peut sembler naïve, mais elle permet de tester de manière simple le bon appariement des gestes avec les expressions référentielles, ce qui constitue un processus non négligeable de la fusion multimodale. Une attention particulière sera donnée à la synchronisation temporelle entre les mots prononcés et les gestes produits. Ainsi, un décalage temporel entre « ça » et le geste dans la question Q pourra conduire selon le système à une réponse positive reflétant sa robustesse pour l'appariement multimodal même quand les conditions de production sont déviantes, ou au contraire à une réponse négative reflétant l'incapacité du système à sortir d'un certain intervalle temporel.

– Niveau 2 = « information implicite ». Ce niveau concerne la résolution des anaphores, des ellipses, des incomplétudes et autres informations implicites mais récupérables aux niveaux syntaxiques et sémantiques. Un exemple fait intervenir : D = « donnez-moi un billet pour Paris et aussi pour Lyon » et Q = « billet pour Lyon ? ». La résolution de la référence étant l'un des principaux aspects de la multimodalité spontanée, un DQR multimodal devra bien entendu en rendre compte. Ainsi, en reprenant comme D la primitive universelle de la multimodalité, « mets ça ici » avec deux gestes de désignation, les questions Q pourront introduire des précisions sur les référents, en partant par exemple de la mention de leur catégorie et en allant jusqu'à donner leur identifiant unique tel qu'il est géré par le système : « mettre cet objet ? » + geste en (x_1, y_1) ; « mettre ce fichier ? » + geste en (x_1, y_1) ; « mettre "submis.tex" ? » (sans geste) ; « mettre obj_{4353} ? » (sans geste) ; etc. La procédure d'évaluation inclut donc la paraphrase en langage naturel d'une référence multimodale. Ce qui reste simple pour le geste déictique l'est beaucoup moins pour les autres types de gestes coverbaux. Imaginons par exemple que « mets ça ici », ou plutôt « déplace ça ici » pour ne pas trop compliquer l'exemple, s'accompagne d'un seul geste qui part de l'objet à déplacer et aboutit au lieu de destination. Selon une première hypothèse qui reprend la présentation du paragraphe 6.2.2, cette trajectoire gestuelle est considérée comme la matérialisation de la nécessaire transition entre la désignation d'objet et la désignation de lieu. Dans ce cas, seules les extrémités de la courbe sont utilisées lors des analyses

sémantiques : le point (x_1, y_1) puis l'objet présent en ce point ou dans un voisinage immédiat sont unifiés avec « ça », et le point (x_2, y_2) est unifié avec « ici ». Autrement dit on revient au cas précédent. Selon une seconde hypothèse, la trajectoire est considérée comme la combinaison de ces deux désignations avec un geste coverbal illustratif apportant une caractéristique de l'action de déplacement, à savoir le chemin (ou points de passage) à suivre. La trajectoire est analysée d'un point de vue temporel (courbe produite de manière régulière, sans point d'arrêt significatif) et d'un point de vue structurel (arc de cercle), avant d'être unifiée à « déplace », c'est-à-dire d'être interprétée comme un chemin de déplacement. Si l'on veut tester cette fonctionnalité du système multimodal, il suffit de poser une question Q supplémentaire : « suivre cette trajectoire ? » ou « déplacer selon ces points de passage ? », en reprenant dans un cas comme dans l'autre le geste complet. Le seul inconvénient reste celui qui vaut pour l'ensemble de la méthodologie DQR, à savoir la nécessité pour le système de traiter de telles questions.

– Niveau 3 = « inférence ». Il s'agit ici de la construction du sens complet de l'énoncé, la difficulté étant l'identification des sous-entendus, identification qui fait appel à des raisonnements de sens commun et à des inférences pragmatiques. Avec D = « je voudrais un aller-retour pour Paris », les auteurs proposent Q = « vouloir billet ? ». Cet aspect est indépendant des modalités de communication, et reste valable dans l'état pour le dialogue multimodal.

– Niveau 4 = « interprétation du type d'acte illocutoire ». On entre ici dans les niveaux de dialogue, avec un premier aspect concernant les actes de langage et la capacité du système à identifier le bon type d'acte, même en cas d'acte de langage indirect. Avec D = « un billet pour Paris », qui peut faire suite à une question ou qui peut correspondre à une demande initiale, la question Q = « est-ce une demande ? » permet d'évaluer l'acte identifié par le système. En multimodal, on retrouve les actes de dialogue, notamment gestuels, discutés dans les chapitres précédents, et on envisage les questions Q suivantes : « ce geste est-il une demande ? » + geste ; « ce geste accompagne-t-il la parole ? » + geste ; « l'énoncé multimodal est-il une demande ? » (pour tester la fusion multimodale au niveau pragmatique) ; etc. Avec cet aspect et ceux détaillés lors des niveaux 1 et 2, nous avons fait le tour des principaux problèmes qui se posent pour le traitement des entrées en dialogue multimodal.

– Niveau 5 = « reconnaissance des intentions ». Il s'agit ici de déterminer les intentions ou les buts sous-jacents aux énoncés de l'utilisateur, donc à un niveau plus profond que le niveau 4. Le principe est d'interroger explicitement les états intentionnels, avec des questions Q telles que : « l'utilisateur sait-il, veut-il… ? ». De tels états intentionnels sont indépendants des modalités de communication, et l'extension de DQR à la multimodalité ne change rien à ce niveau.

– Niveau 6 = « pertinence de la réponse ». L'objet de la question évaluative est ici assez large puisqu'il s'agit de tester la pertinence des réponses du système. Les aspects couverts sont donc *a priori* les capacités linguistiques (et donc multimodales) dont font preuve les réponses, leur adéquation par rapport à l'énoncé initial de l'utilisateur,

par rapport aux connaissances de l'application, par rapport aux moyens de communication, par rapport au profil de l'utilisateur, etc. Dans le chapitre de J. Zeiliger *et al.* de (Mariani *et al.*, 2000), les exemples de questions Q sont les suivants : « cette question est-elle agressive ? » ; « cette question est-elle nécessaire ? » ; « cette proposition est-elle possible à cet instant ? ». Ces exemples interrogent à la fois la forme et le contenu de la réponse. En dialogue multimodal, il faudrait donc ajouter tous les aspects liés à la multimodalité en sortie, c'est-à-dire aux choix que le système a fait lors de la détermination du contenu et de la forme de la réponse multimodale. Ainsi, des exemples possibles pour Q sont : « le choix de la ou des modalités de sortie est-il pertinent ? » ; « le message est-il surchargé ? » ; « le message est-il redondant ? » ; « le message est-il synchronisé ? » ; « les informations présentées sont-elles pertinentes ? ». Ces questions font le tour des principaux problèmes qui se posent en sortie dans le dialogue multimodal. Elles intègrent cependant des aspects métalinguistiques qui ne sont généralement pas implémentés dans le modèle conceptuel et le lexique des systèmes. Plus que le principe d'extension de DQR à la multimodalité, c'est la faisabilité même de ce sixième niveau qui semble irréaliste.

– Niveau 7 = « pertinence de la stratégie ». Ce dernier niveau teste la qualité de la stratégie de dialogue, c'est-à-dire si elle a été efficacement menée et si elle est réussie. En fait, les questions couvrent non seulement la stratégie de dialogue, mais également la stratégie de gestion de la tâche : « le client est-il content ? » ; « y a-t-il trop de questions de confirmation indirectes ? » ; « le mécontentement est-il dû à la stratégie ? ». Peuvent également être interrogés la lenteur, le nombre d'incidences, les raisons d'une rupture. Ces aspects étant indépendants des modalités, rien n'est à ajouter et nous obtenons finalement un DQR multimodal en bonne et due forme.

Comme nous l'avons évoqué au paragraphe 10.1.3, l'autre matérialisation de cette méthodologie est le paradigme DCR (Antoine et Caelen, 1999) qui, en remplaçant la question évaluative par un contrôle C, minimise le problème de la capacité du système à répondre à cette question parfois métalinguistique. Le contrôle consiste en une simplification ou une reformulation de la demande utilisateur initiale. Ainsi, en reprenant quelques-uns des exemples précédents, on fera intervenir les contrôles multimodaux suivants : « mets "submis.tex" en (x_1, y_1) » ; « déplace obj_{4353} de (x_1, y_1) à (x_2, y_2) » ; « déplace obj_{4353} selon les points de passage (x_3, y_3), (x_4, y_4), (x_5, y_5) ». Passer du DQR multimodal à un DCR multimodal nécessite donc la paraphrase de manière simple et non ambiguë des références multimodales, avec la description en langage naturel de coordonnées spatiales. Les autres aspects ne posent pas de problème particulier, en tout cas pas plus de problème que le passage de DQR à DCR.

10.3.4. *Vers d'autres méthodes d'évaluation*

Les principes de Peace, exposés dans l'article de L. Devillers *et al.* de (Gardent et Pierrel, 2002), sont la reformulation de l'historique en une phrase unique, l'utilisation

de cette phrase pour une évaluation contextuelle de l'énoncé courant, et l'exploitation de représentations sémantiques de référence. Nous avons déjà parlé de la difficulté d'appliquer ce dernier principe au dialogue multimodal, et c'est donc l'idée de reformulation de l'historique qu'il s'agit d'étudier ici.

La modélisation de l'historique du dialogue est un problème récurrent en DHM, et s'avère particulièrement complexe en dialogue multimodal (Landragin, 2004). Comme nous l'avons vu dans le chapitre 6, l'historique doit conserver à la fois l'identifiant des référents (pour ressortir ceux-ci lors de l'interprétation d'une anaphore) et les mentions utilisées pour y référer (pour interpréter les références mentionnelles ou métalinguistiques, ainsi que pour interpréter les ellipses, en particulier les ellipses nominales). En multimodal, il en est de même et les formes référentielles multimodales doivent donc être conservées, de même que l'état de la scène visuelle, à chaque étape, conduisant ainsi au moins à un historique linguistique, un historique gestuel et un historique visuel. Une chaîne de référence faisant appel aux modalités utilisées ou aux souvenirs de l'utilisateur est alors interprétable, par exemple : « l'objet que je viens de désigner », « les deux objets groupés un peu plus loin », « celui de gauche », « celui qui était à droite », « le dernier ». Ces expressions référentielles montrent d'elles-mêmes que la paraphrase d'un historique multimodal est une tâche impossible à réaliser, ou alors au prix de simplifications telles que le biais introduit enlèvera toute plausibilité à l'évaluation. En effet, le seul processus de paraphrase automatisable est l'utilisation systématique des identifiants des référents, or cette solution semble plus destructrice en dialogue multimodal qu'en dialogue oral : elle met en effet de côté l'ensemble des aspects multimodaux. Il nous apparaît donc difficile d'appliquer les principes de Peace au dialogue multimodal.

L'évaluation des systèmes de dialogue multimodaux s'avère en fin de compte plus complexe que celle des systèmes oraux (pourtant déjà bien délicate), surtout quand la multimodalité est considérée comme l'association complémentaire du langage naturel et d'autres modalités de communication sur lesquelles s'appuie le langage. Dans ce dernier chapitre nous avons proposé quelques briques méthodologiques, et en particulier une extension des paradigmes DQR et DCR à la multimodalité. Plusieurs aspects restent à étudier pour obtenir une méthodologie couvrant le champ occupé actuellement par le dialogue multimodal. Un aspect concerne une piste qui est explorée actuellement pour simplifier la réalisation de systèmes multimodaux, celle de l'ingénierie dirigée par les modèles (voir section 4.2). D'autre part, quand les systèmes de dialogue multimodaux seront suffisamment nombreux, il nous apparaît utile de revenir sur la méthode d'évaluation par défi. Son principe, que ce soit l'étape de dérivation d'énoncés à partir d'un ensemble d'énoncés initiaux ou l'échange des rôles entre différents concepteurs, nous semble en effet pertinent pour le dialogue multimodal.

10.4. Bilan

L'évaluation en dialogue homme-machine ne se caractérise pas par l'efficacité, l'objectivité et le consensus que l'on observe dans d'autres domaines du traitement automatique des langues. Les systèmes restent conçus pour une tâche donnée, ce qui rend difficile toute évaluation comparative ou normée. De plus, les avancées technologiques rendent vite obsolètes les paradigmes d'évaluation et ont pour conséquence une multiplication de ceux-ci. Ce dixième chapitre fait une synthèse sur les méthodes existantes et propose un ensemble de réflexions autour de l'évaluation de la multimodalité dans les systèmes à forte composante linguistique.

Conclusion

Comme l'écrivait déjà (Luzzati, 1995, p. 6) il y a presque vingt ans, le DHM est toujours « un type nouveau de communication qu'il s'agit d'inventer presque en même temps que les matériels qui la supportent, et dont la nature sera fonction des compétences que l'on parviendra à insuffler à la machine, aussi bien en ce qui concerne les mécanismes de la compréhension que ceux de la génération ou de l'interaction ». Nous l'avons vu, malgré les avancées techniques et par exemple l'utilisation grandissante d'algorithmes d'apprentissage automatique, la quantité de travail nécessaire pour réaliser un système de DHM est fonction des capacités envisagées pour ce système, qu'il s'agisse de capacités de capture de signaux divers et variés, de capacités de TAL, de capacités de raisonnement logique ou de capacités de production et de rendu visuel de messages. Nous avons souligné dans ce livre l'importance d'une méthodologie pluridisciplinaire, qui intègre expérimentations, études de corpus, confrontations de théories pour leur application au DHM, et dont l'une des facettes, l'évaluation, est d'une complexité telle que les efforts de recherche doivent être poursuivis. Nous avons souligné l'intérêt d'une gestion dynamique des tours de parole, l'importance de la prosodie et de la sémantique (plutôt que la syntaxe) dans le processus d'analyse linguistique, ainsi que les rôles centraux des processus de résolution des références, d'identification des actes de dialogue et de planification. A partir d'un exemple à première vue simple d'une tâche de renseignement ferroviaire, nous avons donné un panorama des techniques actuelles et des enjeux pour le dialogue en domaine fermé.

Sans reprendre les enjeux mentionnés dans chacun des chapitres, revenons sur les quatre ensembles d'enjeux détaillés à la section 1.3. Le premier ensemble regroupe les enjeux théoriques, avec l'exploration de théories linguistiques et leur adaptation pour le DHM, adaptation qui peut passer par une remise en cause de certains ancrages historiques tels que le découpage en syntaxe, sémantique et pragmatique. Nous avons souligné l'importance des travaux à l'interface entre deux disciplines, avec l'exemple désormais évident de l'interface entre théories linguistiques et implémentations informatiques. Cette position intermédiaire présente les avantages de contribuer à identifier dans les travaux linguistiques ceux pour lesquels des applications sont pertinentes, et

de contribuer au DHM avec des compétences théoriques et avec une vue d'ensemble parfois enviable. Mais elle s'avère aussi très inconfortable : le chercheur à l'interface ne contribue pas aux théories linguistiques (seulement à leurs applications) et ne produit pas de développement informatique (seulement des préalables). A ce titre, il n'a aucun résultat direct à valoriser, et son apport, qu'il s'agisse d'un modèle formel ou de pistes pour une implémentation, peut être critiqué aisément : seule une implémentation s'accompagne de « preuves » et peut résister à la critique. Avec ce livre, nous espérons avoir contribué à montrer à quel point ces travaux à l'interface restent indispensables.

Le deuxième ensemble d'enjeux porte sur l'éventail des capacités attendues pour un système. Nous avons montré que des capacités de compréhension étendues étaient la base pour des échanges pertinents et pour un dialogue réaliste. Par ailleurs, comme le souligne (Cole, 1998, p. 200) pour les systèmes en domaine fermé, c'est avant tout la robustesse et l'aspect temps réel qui doivent encore être améliorés, ainsi que les capacités du système à diriger (sans que ce soit trop manifeste) l'utilisateur dans une voie qui fonctionne.

Le troisième ensemble regroupe les enjeux méthodologiques et techniques autour de la conception de systèmes. Nous avons donné des exemples de flux de travail complexes, faisant intervenir la mise en œuvre non seulement d'architectures *run-time*, mais aussi, ce qui est moins répandu, d'architectures *design-time*, indispensables pour autoriser une certaine souplesse dans le développement ainsi qu'une certaine réutilisabilité. Par ailleurs, nous avons montré que la quantité de travail nécessaire à l'élaboration d'un système de DHM dépassait celle d'une thèse de doctorat, et, par conséquent, qu'une équipe s'avérait désormais nécessaire. Les éléments de cette équipe se distinguent selon différents métiers, à l'image de ce qui se fait dans d'autres domaines de l'informatique.

Enfin, le dernier ensemble d'enjeux est celui de la facilitation du développement informatique, avec les boîtes à outils, les ateliers de génie logiciel, et, peut-être un jour, les *middlewares* et cartes *hardware* dédiés au TAL, à la compréhension automatique et à la gestion du dialogue. De nos jours, réaliser un système de DHM qui suive l'état de l'art pour la majeure partie de ses fonctionnalités et qui ajoute un aspect innovant est un véritable défi, à moins de travailler dans un environnement qui met à disposition une plate-forme continuellement mise à jour. La généralisation de ce type de plate-forme et des moyens de facilitation que nous avons mentionnés serait une avancée majeure pour le domaine du DHM. Non seulement cela permettrait d'accélérer de manière significative les résultats des recherches, mais, également, de procéder à des évaluations plus fiables, plus comparables qu'elles ne le sont actuellement.

Bibliographie

ABBOTT B., *Reference*, Oxford University Press, Oxford, 2010.

ABEILLÉ A., *Les grammaires d'unification*, Hermès-Lavoisier, Paris, 2007.

ALLEMANDOU J., CHARNAY L., DEVILLERS L., LAUVERGNE M., MARIANI J., « Un paradigme pour évaluer automatiquement des systèmes de dialogue homme-machine en simulant un utilisateur de façon déterministe », *Traitement Automatique des Langues*, 48(1), p. 115-139, 2007.

ALLEN J.F., PERRAULT C.R., « Analysing Intention Utterances », *Artificial Intelligence*, 15, p. 143-178, 1980.

ALLEN J.F., SCHUBERT L.K., FERGUSON G., HEEMAN P., HWANG C.H., KATO T., LIGHT M., MARTIN N., MILLER B., POESIO M., TRAUM D.R., « The TRAINS Project: A Case Study in Defining a Conversational Planning Agent », *Journal of Experimental and Theoretical Artificial Intelligence*, 7(1), p. 7-48, 1995.

ALLWOOD J., TRAUM D.R., JOKINEN K., « Cooperation, Dialogue and Ethics », *International Journal of Human-Computer Studies*, 53, p. 871-914, 2000.

ANTOINE J.Y., Pour une ingénierie des langues plus linguistique, mémoire d'Habilitation à Diriger des Recherches, Université de Bretagne Sud, Vannes, 2003.

ANTOINE J.Y., CAELEN J., « Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat) », *Langues*, 2(2), p. 130-139, 1999.

ASHER N., GILLIES A., « Common Ground, Corrections, and Coordination », *Argumentation*, 17, p. 481-512, 2003.

ASHER N., LASCARIDES A., « Indirect Speech Acts », *Synthese*, 128(1–2), p. 183-228, 2001.

ASHER N., LASCARIDES A., *Logics of Conversation*, Cambridge University Press, Cambridge, 2003.

AUSTIN J., *How to do things with words*, Oxford University Press, Oxford, 1962.

BAKER M.J., Recherches sur l'élaboration de connaissances dans le dialogue, mémoire d'Habilitation à Diriger des Recherches, Université de Nancy 2, 2004.

BEAVER D.I., CLARK B.Z., *Sense and Sensitivity: How Focus Determines Meaning*, Blackwell, Oxford, 2008.

BELLALEM N., ROMARY L., « Structural Analysis of Co-Verbal Deictic Gesture in Multimodal Dialogue Systems », *Progress in Gestural Interaction. Proceedings of Gesture Workshop*, York, Angleterre, p. 141-153, 1996.

BERINGER N., KARTAL U., LOUKA K., SCHIEL F., TÜRK U., « PROMISE – A Procedure for Multimodal Interactive System Evaluation », *Proceedings of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Espagne, p. 77-80, 2002.

BERNSEN N.O., DYBKJÆR H., DYBKJÆR L., *Designing Interactive Speech Systems. From First Ideas to User Testing*, Springer Verlag, Berlin, 1998.

BERNSEN N.O., DYBKJÆR L., « Evaluation of Spoken Multimodal Conversation », *Proceedings of the Sixth International Conference on Multimodal Interfaces*, Penn State University, Etats-Unis, p. 38-45, 2004.

BEUN R.J., CREMERS A.H.M., « Object Reference in a Shared Domain of Conversation », *Pragmatics and Cognition*, 6(1/2), p. 121-152, 1998.

BILANGE E., *Dialogue personne-machine : modélisation et réalisation informatique*, Hermès, Paris, 1992.

BLANCHE-BENVENISTE C., *Approches de la langue parlée en français* (seconde édition), Ophrys, Paris, 2010.

BOLT R.A., « Put-That-There: Voice and Gesture at the Graphics Interface », *Computer Graphics*, 14(3), p. 262-270, 1980.

BOBROW D.G., KAPLAN R.M., KAY M., NORMAN D.A., THOMPSON H., WINOGRAD T., « GUS, A Frame-Driven Dialog System », *Artificial Intelligence*, 8, p. 155-173, 1977.

BRANIGAN H.P., PICKERING M.J., PEARSON J., MCLEAN J.F., « Linguistic Alignment between People and Computers », *Journal of Pragmatics*, 42, p. 2355-2368, 2010.

BRENNAN S.E., CLARK H.H., « Conceptual Pacts and Lexical Choice in Conversation », *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), p. 1482-1493, 1996.

BROERSEN J., DASTANI M., VAN DER TORRE L., « Beliefs, Obligations, Intentions, and Desires as Components in an Agent Architecture », *International Journal of Intelligent Systems*, 20(9), p. 893-919, 2005.

BUNT H., « Multifunctionality in dialogue », *Computer Speech and Language*, 25, p. 222-245, 2011.

CADOZ C., « Le geste canal de communication homme-machine. La communication instrumentale », *Techniques et Sciences Informatiques*, 13(1), p. 31-61, 1994.

CAELEN J., XUEREB A., *Interaction et pragmatique. Jeux de dialogue et de langage*, Hermès-Lavoisier, Paris, 2007.

CARBERRY S., *Plan Recognition in Natural Language*, The MIT Press, Cambridge, 1990.

CHAROLLES M., *La référence et les expressions référentielles en français*, Ophrys, Paris, 2002.

CHAUDIRON S. (DIR.), *Evaluation des systèmes de traitement de l'information*, Hermès-Lavoisier, Paris, 2004.

CLARK E.V., *First Language Acquisition* (seconde édition), Cambridge University Press, Cambridge, 2009.

CLARK H.H., *Using Language*, Cambridge University Press, Cambridge, 1996.

CLARK H.H., SCHAEFER E.F., « Contributing to Discourse », *Cognitive Science*, 13, p. 259-294, 1989.

CLARK H.H., WILKES-GIBBS D., « Referring as a Collaborative Process », *Cognition*, 22, p. 1-39, 1986.

COHEN M.H., GIANGOLA J.P., BALOGH J., *Voice User Interface Design*, Addison-Wesley, Boston, 2004.

COHEN P.R., LEVESQUE H.J., « Intention is Choice with Commitment », *Artificial Intelligence*, 42, p. 213-261, 1990.

COHEN P.R., PERRAULT C.R., « Elements of a Plan-Based Theory of Speech Acts », *Cognitive Science*, 3, p. 177-212, 1979.

COLBY K.M., WEBER S., HILF F.D., « Artificial Paranoia », *Artificial Intelligence*, 2, p. 1-25, 1971.

COLE R. (DIR.), *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, Cambridge, 1998.

CORBLIN F., *Les formes de reprise dans le discours. Anaphores et chaînes de référence*, Presses Universitaires de Rennes, Rennes, 1995.

CORBLIN F., *Représentation du discours et sémantique formelle*, PUF, Paris, 2002.

DENIS A., Robustesse dans les systèmes de dialogue finalisés. Modélisation et évaluation du processus d'ancrage pour la gestion de l'incompréhension, thèse de doctorat, Université Henri Poincaré de Nancy, 2008.

DENIS A., « Generating Referring Expressions with Reference Domain Theory », *Proceedings of the 6th International Natural Language Generation Conference*, Dublin, Irlande, p. 27-35, 2011.

DE RUITER J.P., CUMMINS C., « A Model of Intentional Communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of Signals) », *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, p. 149-150, 2012.

DESSALLES J.L., *La pertinence et ses origines cognitives*, Hermès-Lavoisier, Paris, 2008.

DEVILLERS L., MAYNARD H., ROSSET S., PAROUBEK P., MCTAIT K., MOSTEFA D., CHOUKRI K., CHAMAY L., BOUSQUET C., VIGOUROUX N., BÉCHET F., ROMARY L., ANTOINE J.Y., VILLANEAU J., VERGNES M., GOULIAN J., « The French MEDIA/EVALDA Project: The Evaluation of the Understanding Capability of Spoken Language Dialog Systems », *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbonne, Portugal, p. 2131-2134, 2004.

DREYFUS H.L., DREYFUS S.E., *Mind over Machine: The Power of Human Intuition and Expertise in the Area of the Computer*, Basil Blackwell, Oxford, 1986.

DUERMAEL F., Référence aux actions dans des dialogues de commande homme-machine, thèse de doctorat, Institut National Polytechnique de Lorraine, 1994.

DYBKJÆR L., BERNSEN N.O., MINKER W., « Evaluation and Usability of Multimodal Spoken Language Dialogue Systems », *Speech Communication*, 43(1–2), p. 33-54, 2004.

EDLUND J., HELDNER M., GUSTAFSON J., « Utterance Segmentation and Turn-Taking in Spoken Dialogue Systems », dans B. Fisseni, H.C. Schmitz, B. Schröder, P. Wagner (dir.), *Computer Studies in Language and Speech*, Peter Lang, p. 576-587, 2005.

ENJALBERT P. (DIR.), *Sémantique et traitement automatique du langage naturel*, Hermès-Lavoisier, Paris, 2005.

FRASER N.M., GILBERT G.N., « Simulating Speech Systems », *Computer Speech and Language*, 5, p. 81-99, 1991.

FUCHS C., *Les ambiguïtés du français*, Ophrys, Paris, 2000.

FUNAKOSHI K., NAKANO N., TOKUNAGA T., IIDA R., « A Unified Probabilistic Approach to Referring Expressions », *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Séoul, Corée du Sud, p. 237-246, 2012.

GAONAC'H D. (DIR.), *Psychologie cognitive et bases neurophysiologiques du fonctionnement cognitif*, PUF, Paris, 2006.

GARBAY C., KAYSER D. (DIR.), *Informatique et sciences cognitives. Influences ou confluence ?*, Ophrys, Paris, 2011.

GARDENT C., PIERREL J.M. (DIR.), Dialogue : aspects linguistiques du traitement automatique du dialogue, *Traitement Automatique des Langues*, 43(2), Hermès-Lavoisier, Paris, 2002.

GIBBON D., MERTINS I., MOORE R. (DIR.), *Handbook of Multimodal and Spoken Dialogue Systems*, Kluwer Academic Publishers, Dordrecht, 2000.

GINZBURG J., *The Interactive Stance*, Oxford University Press, 2012.

GOROSTIZA J.F., SALICHS M.A., « End-User Programming of a Social Robot by Dialog », *Robotics and Autonomous Systems*, 59(12), p. 1102-1114, 2011.

GRAU B., MAGNINI B. (DIR.), Réponses à des questions, *Traitement Automatique des Langues*, 46(3), Hermès-Lavoisier, Paris, 2005.

GRICE H.P., « Logic and Conversation », dans P. Cole, J. Morgan (dir.), *Syntax and Semantics*, Vol. 3, Academic Press, p. 41-58, 1975.

GRISLIN M., KOLSKI C., « Evaluation des Interfaces Homme-Machine lors du développement des systèmes interactifs », *Technique et Science Informatiques*, 15(3), p. 265-296, 1996.

GRISVARD O., Modélisation et gestion du dialogue oral homme-machine de commande, thèse de doctorat, Université Henri Poincaré de Nancy, 2000.

GROSZ B.J., SIDNER C.L., « Attention, Intentions and the Structure of Discourse », *Computational Linguistics*, 12(3), p. 175-204, 1986.

GUIBERT G., *Le « dialogue » homme-machine. Un qui-pro-quo ?*, L'Harmattan, Paris, 2010.

GUYOMARD M., NERZIC P., SIROUX J., « Plans, métaplans et dialogue », *Actes de la quatrième école d'été sur le traitement des langues naturelles*, version mise à jour par les auteurs sur leur page *web*, 1993-2006.

HARDY H., BIERMANN A., BRYCE INOUYE R., MCKENZIE A., STRZALKOWSKI T., URSU C., WEBB N., WU M., « The AMITIÉS System: Data-Driven Techniques for Automated Dialogue », *Speech Communication*, 48, p. 354-373, 2006.

HARRIS R.A., *Voice Interaction Design: Crafting the New Conversational Speech Systems*, Morgan Kaufmann, San Francisco, 2004.

HIRSCHMAN L., « Multi-Site Data Collection for a Spoken Language Corpus: MADCOW », *Proceedings of the* DARPA *Speech and Natural Language Workshop*, New York, Etats-Unis, p. 7-14, 1992.

HORCHANI M., Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales, thèse de doctorat, Université Joseph Fourier, Grenoble, 2007.

ISSARNY V., SACCHETTI D., TARTANOGLU F., SAILHAN F., CHIBOUT R., LEVY N., TALAMONA A., « Developing Ambient Intelligence Systems: A Solution based on Web Services », *Automated Software Engineering*, 12(1), p. 101-137, 2005.

JOKINEN K., MCTEAR M.F., *Spoken Dialogue Systems*, Morgan and Claypool, Princeton, 2010.

JÖNSSON A., DÄHLBACK N., « Talking to a Computer is not like Talking to your Best Friend », *Proceedings of the Scandinavian Conference on Artificial Intelligence*, Tromsø, Norvège, 1988.

JURAFSKY D., MARTIN J.H. (DIR.), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (seconde édition), Pearson, Upper Saddle River, NJ, 2009.

KADMON N., *Formal Pragmatics*, Blackwell, Oxford, 2001.

KAMP H., REYLE U., *From Discourse to Logic*, Kluwer, Dordrecht, 1993.

KENDON A., *Gesture: Visible Action as Utterance*, Cambridge University Press, Cambridge, 2004.

KERBRAT-ORECCHIONI C., *L'implicite*, Armand Colin, Paris, 2012.

KNOTT A., VLUGTER P., « Multi-Agent Human-Machine Dialogue: Issues in Dialogue Management and Referring Expression Semantics », *Artificial Intelligence*, 172, p. 69-102, 2008.

KOLSKI C., *Ingénierie des interfaces homme-machine. Conception et évaluation*, Hermès, Paris, 1993.

KOLSKI C. (DIR.), *Interaction homme-machine dans les transports*, Hermès-Lavoisier, Paris, 2010.

KOPP S., BERGMANN K., WACHSMUTH I., « Multimodal Communication from Multimodal Thinking. Towards an Integrated Model of Speech and Gesture Production », *International Journal of Semantic Computing*, 2(1), p. 115-136, 2008.

KRAHMER E., VAN DEEMTER K., « Computational Generation of Referring Expressions: A Survey », *Computational Linguistics*, 38(1), p. 173-218, 2012.

KÜHNEL C., *Quantifying Quality Aspects of Multimodal Interactive Systems*, Springer, Berlin, 2012.

LAMEL L., ROSSET S., GAUVAIN J.L., BENNACEF S., GARNIER-RIZET M., PROUTS B., « The LIMSI ARISE System », *Speech Communication*, 31(4), p. 339-354, 2003.

LANDRAGIN F., *Dialogue homme-machine multimodal. Modélisation cognitive de la référence aux objets*, Hermès-Lavoisier, Paris, 2004.

LANDRAGIN F., « Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems », *Signal Processing*, 86(12), Elsevier, Amsterdam, p. 3578-3595, 2006.

LANGACKER R.W., *Foundations of Cognitive Grammar. Theoretical Prerequisites*, Stanford University Press, Stanford, 1987.

LARD J., LANDRAGIN F., GRISVARD O., FAURE D., « Un cadre de conception pour réunir les modèles d'interaction et l'ingénierie des interfaces », *Ingénierie des Systèmes d'Information*, 12(6), p. 67-91, 2007.

LEVINSON S.C., *Pragmatics*, Cambridge University Press, Cambridge, 1983.

LÓPEZ-CÓZAR DELGADO R., ARAKI M., *Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment*, Wiley and Sons, Chichester, 2005.

LÓPEZ-CÓZAR DELGADO R., DE LA TORRE A., SEGURA J.C., RUBIO A.J., « Assessment of Dialogue Systems by Means of a New Simulation Technique », *Speech Communication*, 40, p. 387-407, 2003.

LUPERFOY S., « The Representation Of Multimodal User Interface Dialogues Using Discourse Pegs », *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, Etats-Unis, p. 22-31, 1992.

LUZZATI D., *Le dialogue verbal homme-machine*, Masson, Paris, 1995.

MARIANI J., MASSON N., NÉEL F., CHIBOUT K. (DIR.), *Ressources et évaluations en ingénierie de la langue*, AUF et De Boeck Université, Paris, 2000.

MARTIN J.C., BUISINE S., PITEL G., BERNSEN N.O., « Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters », *Signal Processing*, 86 (12), Elsevier, Amsterdam, p. 3596-3624, 2006.

MCTEAR M.F., *Spoken Dialogue Technology: Toward the Conversational User Interface*, Springer-Verlag, Londres, 2004.

MELLISH C., SCOTT D., CAHILL L., PAIVA D., EVANS R., REAPE M., « A Reference Architecture for Natural Language Generation Systems », *Natural Language Engineering*, 12, p. 1-34, 2006.

MITKOV R., *Anaphora Resolution*, Longman, Londres, 2002.

MOESCHLER J., *Argumentation et conversation. Eléments pour une analyse pragmatique du discours*, Hatier, Paris, 1985.

MOESCHLER J. (DIR.), « Théorie des actes de langage et analyse des conversations », *Cahiers de linguistique française*, 13, Université de Genève, 1992.

MÖLLER S., SMEELE P., BOLAND H., KREBBER J., « Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study », *Computer Speech and Language*, 21, p. 26-53, 2007.

MUSKENS R., « Combining Montague Semantics and Discourse Representation », *Linguistics and Philosophy*, 19(2), p. 143-186, 1996.

OVIATT S.L., « Ten Myths of Multimodal Interaction », *Communications of the ACM*, 42(11), p. 74-81, 1999.

PAEK T., PIERACCINI R., « Automating Spoken Dialogue Management Design Using Machine Learning: An Industry Perspective », *Speech Communication*, 50, p. 716-729, 2008.

PICKERING M.J., GARROD S., « Toward a Mechanistic Psychology of Dialogue », *Behavorial and Brain Sciences*, 27, p. 169-226, 2004.

PIERREL J.M., *Dialogue oral homme-machine*, Hermès, Paris, 1987.

PINEDA L., GARZA G., « A Model for Multimodal Reference Resolution », *Computational Linguistics*, 26(2), p. 139-193, 2000.

POESIO M., TRAUM D.R., « Conversational Actions and Discourse Situations », *Computational Intelligence*, 13(3), p. 309-347, 1997.

PRÉVOT L., Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans des dialogues finalisés, thèse de doctorat, Université Paul Sabatier, Toulouse, 2004.

REBOUL A., MOESCHLER J., *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*, Armand Colin, Paris, 1998.

REICHMAN R., *Getting Computers to Talk Like You and Me*, The MIT Press, Cambridge, 1985.

REITER E., DALE R., *Building Natural Language Generation Systems*, Cambridge University Press, Cambridge, 2000.

RIESER V., LEMON O., *Reinforcement Learning for Adaptive Dialogue Systems. A Data-driven Methodology for Dialogue Management and Natural Language Generation*, Springer, Heidelberg, 2011.

ROSSET S., Systèmes de dialogue (oral) homme-machine : du domaine limité au domaine ouvert, mémoire d'Habilitation à Diriger des Recherches, Université Paris-Sud, Orsay, 2008.

ROSSET S., TRIBOUT D., LAMEL L., « Multi-level Information and Automatic dialog Act Detection in Human-Human Spoken Dialogs », *Speech Communication*, 50(1), p. 1-13, 2007.

ROSSI M., *L'intonation, le système du français*, Ophrys, Paris, 1999.

ROSSIGNOL S., PIETQUIN O., IANOTTO M., « Simulation of the Grounding Process in Spoken Dialog Systems with Bayesian Networks », *Proceedings of the 2nd International Workshop on Spoken Dialogue Systems Technology*, Gotemba, Japon, p. 110-121, 2010.

ROULET E., AUCHLIN A., MOESCHLER J., RUBATTEL C., SCHELLING M., *L'articulation du discours en français contemporain*, Lang, Berne, 1985.

SABAH G., *L'intelligence artificielle et le langage. Tome 2 : processus de compréhension*, Hermès, Paris, 1989.

SABAH G., « The "Sketchboard": A Dynamic Interpretative Memory and its Use for Spoken Language Understanding », *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, Grèce, 1997.

SABAH G., VIVIER J., VILNAT A., PIERREL J.M., ROMARY L., NICOLLE A., *Machine, langage et dialogue*, L'Harmattan, Paris, 1997.

SACKS H., SCHEGLOFF E.A., JEFFERSON G., « A Simplest Systematics for the Organization of Turn-Taking for Conversation », *Language*, 50(4), p. 696-735, 1974.

SEARLE J., *Speech Acts*, Cambridge University Press, Cambridge, 1969.

SEARLE J., VANDERVEKEN D., *Foundations of Illocutionary Logic*, Cambridge University Press, Cambridge, 1985.

SENEFF S., « TINA: A Natural Language System for Spoken Language Application », *Computational Linguistics*, 18(1), p. 62-86, 1995.

SINGH S.P., LITMAN D.J., KEARNS M., WALKER M.A., « Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System », *Journal of Artificial Intelligence Research*, 16, p. 105-133, 2002.

SOWA J., *Conceptual Structures. Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA, 1984.

SPERBER D., WILSON D., *Relevance. Communication and Cognition* (seconde édition), Blackwell, Oxford (Royaume-Uni), Cambridge (Etats-Unis), 1995.

STOCK O., ZANCANARO M. (DIR.), *Multimodal Intelligent Information Presentation*, Springer, Heidelberg, 2005.

STONE M., LASCARIDES A., « Coherence and Rationality in Grounding », *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, Poznań, Pologne, p. 51-58, 2010.

TELLIER I., STEEDMAN M. (DIR.), Apprentissage automatique pour le TAL, *Traitement Automatique des Langues*, 50(3), ATALA, 2009.

THEUNE M., « Contrast in Concept-to-Speech Generation », *Computer Speech and Language*, 16, p. 491-531, 2002.

TRAUM D.R., « 20 Questions on Dialog Act Taxonomies », *Journal of Semantics*, 17(1), p. 7-30, 2000.

TRAUM D.R., HINKELMAN E.A., « Conversation Acts in Task-Oriented Spoken Dialogue », *Computational Intelligence*, 8(3), p. 575-599, 1992.

TRAUM D.R., LARSSON S., « The Information State Approach to Dialogue Management », dans J. Van Kuppevelt, R. Smith (dir.), *Current and New Directions in Discourse and Dialogue*, Kluwer, Dordrecht, p. 325-354, 2003.

VAN DEEMTER K., KIBBLE R. (DIR.), *Information Sharing. Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, 2002.

VAN SCHOOTEN B.W., OP DEN AKKER R., ROSSET S., GALIBERT O., MAX A., ILLOUZ G., « Follow-up Question Handling in the IMIX and RITEL Systems: A Comparative Study », *Natural Language Engineering*, 1(1), p. 1-23, 2007.

VILNAT A., Dialogue et analyse de phrases, mémoire d'Habilitation à Diriger des Recherches, Université Paris-Sud, Orsay, 2005.

VUURPIJL L.G., TEN BOSCH L., ROSSIGNOL S., NEUMANN A., PFLEGER N., ENGEL R., « Evaluation of Multimodal Dialog Systems », *Proceedings of the LREC Workshop on Multimodal Corpora and Evaluation*, Lisbonne, Portugal, 2004.

WALKER M.A., « Can We Talk? Methods for Evaluation and Training of Spoken Dialogue Systems », *Journal of Language Resources and Evaluation*, 39(1), p. 65-75, 2005.

WALKER M.A., PASSONNEAU R., BOLAND J.E., « Quantitative and Qualitative Evaluation of DARPA Communicator Spoken Dialogue Systems », *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, Etats-Unis, p. 515-522, 2001.

WALKER M.A., WHITTAKER S., STENT A., MALOOR P., MOORE J., JOHNSTON M., VASIREDDY G., « Generation and Evaluation of User Tailored Responses in Multimodal Dialogue », *Cognitive Science*, 28(5), p. 811-840, 2004.

WARD N., TSUKAHARA W., « A Study in Responsiveness in Spoken Dialog », *International Journal of Human-Computer Studies*, 59(6), p. 959-981, 2003.

WARREN M., *Features of Naturalness in Conversation*, John Benjamins, Amsterdam and Philadelphia, 2006.

WEIZENBAUM J., « ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine », *Communications of the Association for Computing Machinery*, 9(1), p. 36-45, 1966.

WINOGRAD T., *Understanding Natural Language*, Academic Press, San Diego, 1972.

WRIGHT P., « Using Constraints and Reference in Task-oriented Dialogue », *Journal of Semantics*, 7, p. 65-79, 1990.

WRIGHT-HASTIE H., POESIO M., ISARD S., « Automatically Predicting Dialogue Structure using Prosodic Features », *Speech Communication*, 36, p. 63-79, 2002.

Index

Q R

S