Dialogue homme-machine multimodal : de la pragmatique linguistique à la conception de systèmes

Frédéric Landragin

MÉMOIRE

Présenté pour obtenir

L'HABILITATION À DIRIGER DES RECHERCHES DE L'UNIVERSITÉ PARIS-SUD

Spécialité: Informatique

par

Frédéric LANDRAGIN

Dialogue homme-machine multimodal : de la pragmatique linguistique à la conception de systèmes

Soutenue le 28 juin 2013 devant la commission d'examen :

Mr.	Harry BUNT	(Rapporteur)
Mme.	Sophie ROSSET	(Président du jury)
Mme.	Catherine SCHNEDECKER	(Rapporteur)
Mme.	Isabelle TELLIER	
Mme.	Mariët THEUNE	(Rapporteur)
Mme.	Anne VILNAT	

Ecole Doctorale Informatique Paris-Sud
UFR Sciences Orsay, bâtiment 650, aile nord 417
Université Paris-Sud 11
91405 Orsay Cedex

Résumé

Un des objectifs fondamentaux du dialogue homme-machine est de se rapprocher du dialogue naturel en langage naturel, c'est-à-dire de permettre une interaction entre la machine et son utilisateur humain dans la langue de celui-ci (langage naturel), avec une structure d'échanges similaire à un dialogue humain (dialogue naturel). Les recherches impliquées se nourrissent de travaux linguistiques qui analysent la langue et de travaux pragmatiques qui analysent l'usage du langage en contexte. Deux facettes importantes de la pragmatique linguistique portent ainsi sur les phénomènes de référence, par exemple les désignations des objets accessibles dans le contexte situationnel, et sur les actes de langage, ou actes de dialogue, c'est-à-dire les actions communicatives effectuées par les énoncés constituant les tours de parole. Nous présentons nos travaux de modélisation et de formalisation de ces deux facettes, avec leur application au dialogue avec support visuel et au dialogue associant parole et gestes co-verbaux (dialogue multimodal). Un autre objectif du dialogue homme-machine est de mettre en œuvre des méthodologies et des moyens, par exemple des architectures logicielles réutilisables, pour faciliter le développement de systèmes. Nous présentons nos réflexions et nos réalisations dans ce sens, à travers notamment notre participation à un ensemble de projets européens. Nous proposons enfin des perspectives de recherche qui visent à mieux intégrer au dialogue homme-machine des phénomènes linguistiques et pragmatiques telles que la saillance et l'ambiguïté.

Mots-clefs : Dialogue naturel en langage naturel, pragmatique, référence, acte de langage, architecture logicielle, évaluation de système.

Multimodal Human-Machine Dialogue: From Linguistic Pragmatics to System Design

Abstract

Human-machine dialogue aims at providing natural dialogue in natural language, i.e., allowing the user to speak using his own language (natural language), following a structure of moves and exchanges that is similar to human dialogue structure (natural dialogue). Related research works feed on linguistics, which analyses language markers, and pragmatics, which analyses language use in context. Two important sectors of linguistic pragmatics focus on referring phenomena, for instance referring to the objects that are accessible in the dialogue context, and speech acts, or dialogue acts, i.e., communicative actions carried out by utterances. We present our modelling and formalizing works dealing with these two aspects and with their application to dialogue systems, particularly when a visual scene is implied or when co-verbal gestures are processed together with speech (multimodal dialogue). Human-machine dialogue also aims at facilitating the design and development of effective systems, with methodologies and means like software architectures. We present our theoretical and practical works in this way, illustrated by our participation to several European projects. We then propose some future works that focus on the integration of linguistic and pragmatic phenomena like salience and ambiguity to human-machine dialogue.

Keywords: Natural dialogue in natural language, pragmatique, referring, speech act, software architecture, system evaluation.

Remerciements

La préparation de ce manuscrit s'est faite en lien avec la préparation d'un ouvrage de synthèse sur le dialogue homme-machine. Dans un cas comme dans l'autre, il s'agit d'une mise en perspective de dix années de recherches, c'est-à-dire depuis ma thèse de doctorat soutenue en 2003, dans le domaine du dialogue homme-machine. Plutôt que de présenter les grandes questions du domaine comme le fait l'ouvrage, le but de ce manuscrit est de reprendre les préoccupations que j'ai pu avoir pendant ces années, et de faire un point par rapport à ce que j'ai pu proposer dans mes publications.

Comment une machine peut-elle entretenir avec un utilisateur humain 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 – et prioritaires – pour de tels systèmes ? Comment implémenter et évaluer ces capacités ? Comment s'approcher du réalisme et de la fluidité du dialogue humain ? 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 Recherche et Technologie, 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 quatre ans à des étudiants de Master 2 à l'Université Paris Diderot.

Comme tout travail, celui présenté ici 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 Harry Bunt, Francis Corblin, Catherine Fuchs, Valérie Issarny, Jean-Marie Pierrel, Laurent Romary, Sophie Rosset, Jean-Paul Sansonnet, Catherine Schnedecker, Jacques Siroux, Isabelle Tellier, Mariët Theune, Bernard Victorri, et Anne Vilnat qui a (entre autres) eu le rôle de marraine pour cette habilitation. Pour l'expérience Ozone très enrichissante lors de mon post-doctorat à l'INRIA, je remercie en particulier Christophe Cérisara et surtout Alexandre Denis sur qui j'ai pu me reposer pour l'implémentation d'un démonstrateur mémorable, ainsi que Matthieu Quignard pour sa suite dans le projet Amigo. Pour l'expérience également mémorable de Thales R & T, avec notamment les projets Emode et Cocase, je remercie 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, Shirley Carter-Thomas, Michel Charolles, Thierry Poibeau, Sophie Prévost, 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.

Table des matières

Introduction

Définitions et phénomènes de dialogue

Le système OZONE (Issarny *et al.* 2005) évoqué dans les remerciements était un démonstrateur pour un service de réservation de billets de train, dans le cadre du projet européen OZONE, cf. `http://www.hitech-projects.com/euprojects/ozone/`. C'était un démonstrateur parmi d'autres dans ce projet IST (*Information Society Technologies*) de grande taille. Sa conception a pris plus d'un an, principalement en 2004, au LORIA. Il s'agissait d'un système de dialogue exécuté sur tablette tactile (*tablet PC*), avec comme tâche d'aider l'utilisateur à trouver des billets de train correspondant le mieux à son besoin, en l'occurrence repartir du centre INRIA de Rocquencourt pour aller à Paris. Le module de reconnaissance de la parole était pris en charge par Christophe Cérisara, le module de synthèse de parole était un logiciel gratuit récupéré sur Internet, et tout le reste, c'est-à-dire les capacités d'analyse linguistique, de compréhension automatique, de raisonnement et de génération automatique de messages, a été implémenté en partant de zéro. Cette implémentation a été réalisée sous ma direction par Alexandre Denis, alors en stage-ingénieur, et qui a poursuivi sur un DEA et une thèse dans le domaine du dialogue homme-machine. Ce n'est pas par hasard que je commence ce manuscrit en mentionnant ce système : d'une part il s'agissait de mon premier travail post-doctoral, dans lequel une bonne partie de ma thèse de doctorat a été implémentée ; d'autre part il a été le point de départ de plusieurs années d'interrogation et de recherche autour des aspects que je n'avais pas abordés dans ma thèse de doctorat.

La réservation de billets de train est une **application** (ou **tâche**) récurrente dans le domaine du dialogue homme-machine, et c'est dans ce cadre que je vais choisir les exemples tout au long de ce manuscrit. Le programme informatique constituant notre démonstrateur était ainsi capable de traiter une entrée audio, de transcrire en texte la parole capturée, 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. Cela se rapproche d'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 (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 simultanée 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**, mais il n'en sera pas question ici. 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**, ou tout simplement dialogue, 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 ou 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.

Approches

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-détermination** (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 manuscrit, 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 implantations informatiques. A titre d'exemples qui ont été autant de sources d'inspiration, citons, ici dans l'ordre chronologique : Reichman (1985), Pierrel (1987), Sabah (1989), Carberry (1990), Bilange (1992), Kolski (1993), Luzzati (1995), Bernsen *et al.* (1998), Reiter & Dale (2000), Asher & Lascarides (2003), Cohen *et al.* (2004), Harris (2004), McTear (2004), López-Cózar Delgado & Araki (2005), Caelen & Xuereb (2007), Jurafsky & Martin (2009), Jokinen & McTear (2010), Rieser & Lemon (2011), Ginzburg (2012), Kühnel (2012). Ce domaine du dialogue homme-machine (désormais DHM) couvre plusieurs disciplines, non seulement l'informatique et les sciences du langage déjà mentionnés, mais aussi les sciences cognitives en général, avec l'objectif de se rapprocher des capacités humaines. Il entretient des liens privilégiés avec d'autres domaines, notamment celui du traitement automatique des langues (désormais 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. Les problématiques du DHM, et aussi ses enjeux, sont liés à ceux de ces domaines.

Enjeux théoriques et techniques

Une machine peut-elle penser ? Le test de Turing et la quête du Graal de la machine capable de dialoguer 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 dans les *SmartPhones*, GPS et autres assistants personnels 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 (Weizenbaum 1966) ont pu le faire croire, rien n'est magique, et ce quelles que 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, on va discuter de 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 leur développement informatique.

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 est toujours d'actualité. En linguistique, pour la modélisation de l'interprétation, ont été mentionnées les dimensions 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 (interprétation des déictiques et

identification des actes de langage, par exemple). 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 sans être explicités par l'énoncé ; la mise à jour de l'historique du dialogue ; etc. 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. Par ailleurs, que ce soit à un niveau prosodique, syntaxique ou sémantique, on peut imaginer plusieurs analyseurs qui fonctionnent en parallèle, avec par exemple un analyseur plutôt symbolique et un autre plutôt statistique, de manière à confronter une compréhension fine avec des indicateurs de fréquence d'un phénomène. Si l'on reste dans un fonctionnement en cascade des analyses, tous ces types de mécanismes sont impossibles. Un enjeu consiste donc à explorer les implantations 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 à la Charles Morris (syntaxe, sémantique, pragmatique) ne se justifie donc plus, et l'**application des théories linguistiques au DHM** constituent toujours un enjeu 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.

Pour terminer cette liste d'enjeux théoriques, on peut souligner 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 a permis de commencer à réfléchir au concept de méta-grammaire (ou méta-modèle), avec comme but l'instanciation à partir d'une base commune d'une grammaire linguistique et d'un modèle de langage statistique.

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 il a été question 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, 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 10 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 qu'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, ainsi que les problèmes suite à des problèmes (Denis 2008). 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.

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 (cf. chapitre 2), 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 multimodales (Gorostiza & 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 celle-ci en termes d'ergonomie, de plasticité : adaptation à l'utilisateur, au terminal, à l'environnement.

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 (Rouillard 2004) est un exemple basique, mais il existe beaucoup d'autres plateformes dédiées par exemple à l'aide à la conception de systèmes de dialogue multimodaux (López-Cózar Delgado & Araki 2005). Un deuxième pas serait la mise en place d'une librairie informatique proposant un panel riche et performant d'outils de TAL et de gestionnaires de dialogue. C'est un enjeu important, amorcé 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.

Plan du manuscrit

Parmi tous ces enjeux du DHM, ceux que j'ai particulièrement abordés pendant mes dix dernières années de recherche et développement se répartissent ainsi :

- Les aspects méthodologiques : comment concevoir un système de dialogue ? comment évaluer un système ? quels types d'expérimentations peut-on envisager ? Ce sera l'objet du premier chapitre.
- Les architectures logicielles : comment matérialiser le fonctionnement d'un système de dialogue ? comment matérialiser les étapes de conception d'un système ? quel cadre l'informatique peut-elle apporter à la réalisation de systèmes ? Ce sera l'objet du deuxième chapitre.

- La modélisation des capacités de compréhension : comment faire collaborer les analyses pour traiter les signaux arrivant en entrée du système ? comment faire les liens entre ces signaux et les objets informatiques gérés par le système ? comment le système peut-il détecter les intentions sous-jacentes au message de l'utilisateur ? Ce sera l'objet des troisième et quatrième chapitres, le troisième pour ce qui concerne les liens référentiels et le quatrième pour ce qui concerne la reconnaissance des intentions. L'exemple U2 de la figure 1 servira particulièrement pour illustrer ces deux aspects.
- La modélisation des capacités de raisonnements internes au système : comment un système réfléchit-il ? comment trouve-t-il la réponse à la question posée ? comment gère-t-il un dialogue ? Ce point sera traité dans le quatrième chapitre dans la mesure où je fais un lien entre la reconnaissance des intentions et la gestion du dialogue. Le choix de la réponse S3 suite à U2, ainsi que les réponses alternatives, serviront également pour illustrer cet aspect.
- La modélisation des capacités de génération de messages : comment un système formule-t-il une réponse ? comment peut-il réagir à l'énoncé de l'utilisateur ? Ce point sera traité (notamment avec l'exemple S2) dans le premier chapitre, dans la mesure où mes propositions relèvent d'une méthodologie générale plutôt que de l'implémentation d'un système de génération automatique spécifique.

1 Questions méthodologiques

Les expériences de réalisations de systèmes auxquelles j'ai participé (notamment les projets COVEN, MIAMM, OZONE, COCASE et EMODE dont il va être question ici) m'ont amené à m'interroger sur les méthodologies de conception (§ 1.1) et d'évaluation (§ 1.2) de systèmes de dialogue. En plus de ces deux aspects fondamentaux s'ajoutent des expérimentations connexes (§ 1.3), qui peuvent être réalisées indépendamment et qui peuvent apporter un éclairage non seulement au DHM, mais aussi aux IHM, à l'ergonomie et d'une manière générale à l'étude du langage.

1.1 Etapes de conception d'un système de dialogue

1.1.1 Les étapes selon une approche idéale

Dans l'idéal et de manière un peu schématique, la conception d'un système fait intervenir : 1. la spécification de la tâche et des rôles du système ; 2. la spécification des phénomènes couverts par le système et les études de corpus ; 3. la réalisation d'expérimentations, notamment de simulation du système tel qu'il est envisagé ; 4. la spécification des processus de traitement ; 5. l'écriture et/ou la collecte des ressources et le développement ; 6. l'évaluation et le passage à l'échelle.

1. 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, cf. les dix aspects naturels du dialogue selon (Clark 1996) ou les neuf principes de (Warren 2006). Ces deux rôles sont-ils conciliables ?

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 (cf. Kolski 2010). De plus, les études de corpus (corpus SNCF, notamment) 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 & Wilson 1995) puis (Reboul & 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 statut ou **rôle d'interlocuteur de la machine**. Comme l'exprime si bien le titre de l'article de (Jönsson & Dählback 1988), 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, c'est-à-dire les simulations de systèmes pour les tester avant même de les implémenter, 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 qu'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 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. Par contre, et c'est là l'un des enjeux essentiels des concepteurs, le système doit dans tous les cas tenir le dialogue : s'il se met à montrer son incompréhension, l'utilisateur le

plus confiant va vite changer de comportement.

2. La spécification des capacités de compréhension et de dialogue du futur système (pour qu'il tienne le dialogue avec le plus grand nombre d'utilisateurs) consiste souvent en une boucle comportant 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 capacités 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. L'expérimentation de type **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 & 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, les conditions expérimentales doivent être 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 & 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. Leur 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 & 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.

Malgré toutes ces critiques, un Magicien d'Oz s'avère souvent utile, ne serait-ce que comme moyen de constitution d'un corpus de dialogue, et de manière complémentaire

avec des études de corpus de dialogues humains. Revenons sur cet aspect méthodologique important qu'est l'étude de corpus. 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 extra-linguistiques : 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 multi-niveaux 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.

4. La spécification des processus de traitement se fonde sur les résultats des trois étapes précédentes. 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 m'amener à illustrer les difficultés méthodologiques qui apparaissent alors, en prenant comme exemple un concepteur (personnage imaginaire inspiré de l'expérience Ozone), que j'imagine seul face au système qu'il est en train de concevoir. Ce 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 : 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 (Landragin 2004), 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 (cf. chapitre 3). 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.

5. Comme précédemment, c'est au cours de l'écriture des ressources et du développement informatique que les véritables problèmes se posent. En programmant l'algorithme principal d'un module, on se rend compte qu'on manque de ressources, on se rend compte 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 une caractéristique prosodique que l'on avait négligé mais qui s'avère être un paramètre important à un moment donné. On se rend compte que la sortie du module ne sera pas aussi complète ou précise que prévu. On se rend compte 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, et encore. 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 à la fin du projet MIAMM, mieux vaut commencer par le noyau du système et terminer 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 développe-

ment. En effet, une tentation est de suivre l'ordre de traitement du système : avec un découpage caricaturant ce qu'ont été plusieurs systèmes, on commence par développer le module de reconnaissance de la parole, puis l'analyseur syntaxique, puis l'analyseur sémantique, puis l'analyseur pragmatique, puis la gestion du dialogue, et enfin les modules de génération automatique et de synthèse. Or des pertes sont possibles à chaque étape, et les conséquences deviennent de plus en plus graves : au final, le système n'arrive plus à traiter correctement qu'un petit sous-ensemble des phénomènes prévus au départ.

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. La liste suivante donne une idée de la diversité des modèles impliqués dans un système de DHM :

- modèles du domaine : base de données des objets instanciés, éventuellement visibles et référables, 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 par l'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 : modèles d'émotions, modèles pour la lecture sur les lèvres, modèles pour la reconnaissance de l'écriture, etc. ;
- 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 avec la machine (orale, visuelle, multimodale, ou encore via l'IHM).

6. L'évaluation et le passage à l'échelle : dans un cycle de développement tel qu'on en voit en IHM, un **cycle en V** par exemple (Grislin & 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. Avant de revenir sur les méthodologies d'évaluation qui font l'objet du § 1.2, je me focalise ici sur le passage à l'échelle, dans la mesure où j'ai été confronté à ses caractéristiques lors de mes expériences de réalisations de systèmes à THALES, et où j'ai pu mesurer l'importance de cet aspect méthodologique.

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...

1.1.2 Les étapes dans des expériences réelles

Les pages qui précèdent montrent les préoccupations méthodologiques qui sont devenues les miennes au fur et à mesure de mes différentes expériences de conception de systèmes de dialogue qui sont les suivantes : participation au projet ACTS-AC040 COVEN (*Collaborative Virtual Environments*, `http://www.crg.cs.nott.ac.uk/research/projects/Coven/`) en 1999, pour la réalisation d'un module de résolution de la référence dans un système de dialogue multimodal sur une tâche d'aménagement d'un intérieur en 3D ; participation au projet IST-2000-29487 MIAMM (*Multidimensional Information Access using Multiple Modalities*, `http://miamm.loria.fr/`)

de 2001 à 2004, avec une participation à chacune des étapes de conception, depuis les spécifications jusqu'à l'évaluation finale, d'un système de dialogue multimodal pour l'accès à des données musicales ; participation au projet IST-2000-30026 OZONE (O_3, *Offering an Open and Optimal roadmap towards consumer oriented ambient intelligence*, `http://www.hitech-projects.com/euprojects/ozone/`), pour la réalisation avec Alexandre Denis d'un système de dialogue multimodal pour la réservation de billets de train ; participation à la première étape de spécification du projet IST-004182 AMIGO (*Ambient intelligence for the networked home environment*, `http://www.hitech-projects.com/euprojects/amigo/`) en 2004 ; coordination d'une soumission de projet, alors dénommé COCASE (*Cost-efficient, Open, Customisable, Adaptive System Environment*), à l'IST en 2005 et 2006, autour de la conception d'une plateforme de développement de systèmes de dialogue fondée sur les architectures dirigées par les modèles (cf. chapitre 2) ; participation en 2006 au projet ITEA-04046 EMODE (*Enabling Model transformation-based cost efficient adaptive multimodal user interfaces*, `http://www.itea2.org/project/index/view/?project=141`) autour des modèles impliqués dans le dialogue multimodal. Comme il s'agit de projets regroupant à chaque fois de nombreux partenaires, je n'ai pas de publications qui présentent globalement les systèmes, uniquement des publications sur des points scientifiques particuliers, que nous verrons plus loin dans ce manuscrit (cf. section suivante pour ce qui concerne l'évaluation, par exemple, ou encore le chapitre 2 pour ce qui concerne les plateformes de développement). Au-delà des publications (et des systèmes proprement dits), le résultat principal de ces expériences réside dans la prise de conscience progressive de la difficulté à mettre en œuvre un système de dialogue naturel en langage naturel. Ce résultat se traduit sous la forme, d'une part de ce manuscrit, d'autre part de l'ouvrage à paraître dont je parlais au début des remerciements.

De manière plus précise, l'expérience COVEN m'a amené à réfléchir à la résolution des références aux objets (comment faire le lien entre des objets gérés par le système et les énoncés oraux ou multimodaux de l'utilisateur) et a été le point de départ du modèle présenté dans ma thèse de doctorat. Il correspond de plus au moment où j'ai défini et commencé à caractériser la notion de saillance qui a eu ensuite une importance constante dans mon parcours. L'expérience MIAMM m'a amené à m'interroger sur les méthodologies de conception d'un système de dialogue : j'y ai découvert, par le biais des partenaires impliqués, notamment TNO Human Factors (Pays-Bas), divers protocoles expérimentaux, avant et après la réalisation du système proprement dit. J'ai pu ainsi participer à l'ensemble des étapes de conception d'un système, et commencer à identifier les manques et les problèmes potentiels. C'est là que l'importance d'un ordre privilégié pour le développement des modules est apparue. L'expérience OZONE, comme je le disais au début, est vraiment celle où j'ai pu, en coordonnant la réalisation d'un système de A à Z, vérifier à quel point le constat « c'est en implémentant qu'apparaissent les problèmes » est juste, et ce à toutes les étapes du développement. Malgré les constats réalisés suite à l'expérience MIAMM, d'autres problèmes ont été rencontrés, notamment en terme de robustesse. Alexandre Denis a ainsi poursuivi ses recherches sur la notion de robustesse (Denis 2008). Les projets OZONE et AMIGO ont par ailleurs été pour moi une source de réflexion autour de la généricité d'un système de DHM : est-il possible de concevoir des modules adaptables à plusieurs tâches différentes ? est-il possible de réutiliser un système de dialogue ? peut-on définir un gestionnaire de dialogue générique, c'est-à-dire indépendant de toute tâche ? En 2005 et 2006, les expériences COCASE et EMODE à THALES ont été pour moi une façon de donner un nouveau souffle

à ces interrogations en explorant d'autres façons de réaliser des systèmes de dialogue, et notamment en explorant les possibilités de développement semi-automatisé (§ 2.2). A l'issue de ces expériences, la méthodologie générale pour la conception de systèmes de DHM me semble bien plus claire. Malheureusement, elle s'accompagne de verrous technologiques. La proposition COCASE (classée mais non retenue pour financement) avait pour objectif de faire sauter le verrou le plus complexe : celui de la facilitation des moyens de développement. Depuis, d'autres initiatives ont pris le relai, comme nous le verrons dans le chapitre 2.

1.2 Méthodes d'évaluation pour le dialogue multimodal

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 systèmes, 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 & 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. Par contre, 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.

1.2.1 Eléments méthodologiques : expertises, questionnaires, techniques

Plus spécifiquement sur le DHM oral, un certain nombre de méthodes ont été proposées (Antoine & 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 test. Les campagnes d'évaluation (EVALDA/MEDIA : Méthodologie d'Evaluation 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 compétition (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) propose 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 & 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 (cf. 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 & 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 (chapitre de L. Devillers *et al.* dans Gardent & 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.

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 & Dybkjær 2004) présente 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. La publication suivante tente de compenser cette faiblesse, de même que le texte qui suit :

- **Landragin, F.** (2008), Vers l'évaluation de systèmes de dialogue homme-machine : de l'oral au multimodal, In : *Quinzième conférence sur le traitement automatique des langues (TALN 2008)*, Avignon, France, pp. 390–399.

Il s'agit d'un premier pas vers la proposition d'une méthodologie d'évaluation adaptée au dialogue multimodal. J'y décris les avantages et problèmes posés par l'évaluation globale, l'évaluation segmentée, l'évaluation comparative, et j'y propose quelques ébauches de questionnaires à destination de sujets, ainsi qu'une application des méthodes DQR et DCR au dialogue multimodal.

1.2.2 Enjeux pour l'évaluation des systèmes multimodaux

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. 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. 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 l'on se focalise 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.

Un deuxième 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 me 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 a pu 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 eux-mêmes le reflètent : PARADISE, PEACE, PROMISE). Heureusement, les questions posées et 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.

1.3 Expérimentations complémentaires

1.3.1 La notion de saillance

En dehors de la réalisation de systèmes, il est utile de procéder à des expérimentations afin de tester certaines notions. C'est particulièrement le cas de la notion de saillance : dans le projet COVEN, j'ai proposé une première caractérisation de la saillance visuelle qui intervient dans ce type de système où une scène en 3D est affichée à l'écran et sert de support à la communication. La tâche consistait à aménager un intérieur, les objets étaient donc des tables et des chaises dans des pièces, et la saillance d'un objet pouvait fortement influer sur la manière de produire une expression référentielle : avec une chaise en plein milieu du champ de vision et une deuxième chaise plus éloignée, dans un coin et à moitié cachée, l'expression référentielle « *la chaise* » n'est pas du tout ambiguë mais désigne de manière privilégiée la chaise saillante. Mes publications théoriques sur la saillance (cf. chapitre 5) m'ont amené à être contacté par des chercheurs du GIPSA-Lab (Grenoble), pour une collaboration qui a mené à une expérimentation sur l'un des facteurs de saillance, la couleur des objets :

- Ho-Phuoc, T., Guyader, N., **Landragin, F.** & Guérin-Dugué, A. (2012), When Viewing Natural Scenes, Do Abnormal Colors Impact on Spatial or Temporal Parameters of Eye Movements ?, *Journal of Vision* 12(2) :4, `http://www.journalofvision.org/content/12/2/4` (13 pages).

Cet article a priori complètement en marge de mes travaux sur le DHM garde un intérêt certes ponctuel mais non négligeable pour un module tel que celui de résolution de la référence dans le projet COVEN et dans ceux qui ont suivi : il donne des arguments, vérifiés expérimentalement (utilisation d'un oculomètre), permettant de hiérarchiser les facteurs de saillance dans le modèle de compréhension de la référence. De nombreuses expérimentations sont à prévoir pour otenir un modèle complet de la saillance, et j'y reviendrais dans le chapitre dédié à mes perspectives de recherche.

1.3.2 Facteurs humains et présentation d'information multimédia

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 micro-monde 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, ainsi que la transmission d'information multimédia. Pour ce dernier point, le domaine concerné est celui des systèmes de **présentation d'information multimédia** (IMMPS, *Intelligent MultiMedia Presentation Systems*, cf. Stock & 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 & 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. 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 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 & 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 (cf. chapitre 9 de Garbay & 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.

Un présentateur d'information multimédia a pour rôle de traduire le « quoi » 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 (facteurs humains). Quand l'information est amenée à être répartie sur plusieurs modalités de communication, on parle de **fission multimodale**. 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. 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.

Dans le cadre de mes activités de recherche et développement à THALES, j'ai été amené à proposer un ensemble de principes pour la conception d'un présentateur d'information multimédia. Ce travail a conduit d'une part à l'encadrement d'un mémoire de M2, en collaboration avec Claire Fraboulet-Laudy, sur le sujet (Emeline Girette, M2 en informatique à l'Université Pierre et Marie Curie, « Interface multimodale et présentation multimédia », 2006), d'autre part à la publication suivante :

- **Landragin, F.** (2008), Pragmatics and Human Factors for Intelligent Multimedia Presentation : A Synthesis and a Set of Principles, In : *Artificial Intelligence and Simulation of Behaviour (AISB) Symposium on Multimodal Output Generation (MOG 2008)*, Aberdeen, Scotland, UK, pp. 50–57.

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 1975). 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 & 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 multimédia naturels, adaptatifs et centrés sur l'utilisateur sont ainsi les suivants : 1. bien présenter en répartissant de manière pertinente les informations sur les canaux de communication ; 2. bien présenter en se préoccupant du rendu et de la valorisation de l'information sur chaque canal de communication ; 3. bien présenter en exploitant de manière pertinente le contenu sémantique du message ; 4. 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 : 5. bien présenter en exploitant de manière pertinente les moyens de présentation ; 6. 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 : 7. bien présenter avec une exploitation fine des attentes de l'utilisateur ; 8. bien présenter pour favoriser une perception adéquate du message ; 9. bien présenter pour favoriser des réactions adéquates de la part de l'utilisateur. Dans l'article cité, ces principes sont énoncés et matérialisés en tenant compte d'un ensemble de facteurs humains, tels qu'ils sont étudiés en ergonomie (cognitive) et dans les travaux en IHM qui replacent l'utilisateur au centre des préoccupations de conception.

1.3.3 Le dialogue homme-machine pour l'apprentissage d'une langue

Une dernière expérimentation un peu en marge du DHM mais pouvant apporter un éclairage complémentaire est un travail réalisé en collaboration avec Luciana Benotti (Université de Corboba, Argentine) et Alexandra Vorobyova (Université du Québec à Montréal). La publication correspondante est la suivante, sachant qu'un article de revue internationale bien plus complet est en cours de soumission :

- Vorobyova, A., Benotti, L. & **Landragin, F.** (2012), Why do we overspecify in dialogue ? An experiment on L2 lexical acquisition, In : *Sixteenth Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, Paris, France, pp. 185–186.

Les deux points de départ sont la notion de sur-spécification et celle d'alignement. Concernant la sur-spécification, il s'agit d'un phénomène visant à donner plus d'information qu'il n'est nécessaire, dans le but d'être sûr que l'interlocuteur va bien recevoir l'information essentielle. Nous avons étudié cette notion pour les phénomènes de référence aux objets, un exemple étant l'utilisation de l'expression « *la petite chaise rouge* » dans une situation où « *la petite chaise* », voire « *la chaise* », aurait largement suffi pour identifier le référent. Concernant l'alignement, les deux participants d'un dialogue s'alignent, c'est-à-dire qu'au bout de quelques tours de parole, ils emploient un vocabulaire commun, ils adaptent leur prosodie à celle de leur interlocuteur, etc. (Pickering & Garrod 2004). Ces deux notions sont étudiées dans les dialogues humains, et présentent un intérêt certain en DHM : un système qui veut se faire comprendre, par exemple après une situation de mauvaise compréhension, a tout intérêt à exploiter la sur-spécification et à s'aligner avec son interlocuteur humain. Dans le cadre de notre étude et de l'expérimentation réalisée en Argentine, c'est pour une application d'apprentissage d'une langue seconde que ces notions ont été testées. Pour cela, nous avons fait appel à la plateforme GIVE, bien connue dans la communauté du DHM, pour générer automatiquement des expressions référentielles : c'est le système qui génère des expressions, éventuellement sur-spécifiées, et qui teste la bonne ou mauvaise résolution de la référence par le sujet humain. Nous avons ainsi mis en évidence un certain nombre d'effets de la sur-spécification, et nous avons également contribué à creuser les deux notions pour leur application en DHM.

2 Architectures logicielles

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 (situation vécue dans au moins un des projets cités). 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.

2.1 Architectures *run-time*

2.1.1 Une liste de modules et de ressources

(López-Cózar Delgado & 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 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 échangés).

Dans les années 1980, les architectures suivent souvent l'ordre des traitements que j'ai 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 j'ai fait une liste page 23. 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. 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. Ce principe d'accès et de stockage à n'importe quel moment se généralise dans les systèmes actuels, d'autant plus si 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 traite 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 (notion de TRP, *transition-relevance place*). Les modules sémantiques et pragmatiques se mettent alors à traiter l'énoncé, l'enrichir, et, si le gestionnaire de dialogue l'estime pertinent, il peut décider de prendre la parole, avec comme conséquence potentielle de couper 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.

J'ai été amené à conduire de telles réflexions, notamment dans le cadre des projets OZONE et COCASE. Plus spécifiquement sur le concept d'historique du dialogue et sur la nature et l'étendue des informations à y stocker, j'ai proposé les deux publications suivantes, dont l'aspect innovant réside surtout dans la nature des informations nécessaires à la résolution des références aux objets (on stocke non seulement les énoncés de l'utilisateur avec l'interprétation qui en est faite, mais également l'état courant de la scène visuelle, de manière à pouvoir y revenir en cas de besoin, voire de manière à pouvoir relancer la résolution des références) :

- **Landragin, F.** & Romary, L. (2004), Dialogue History Modelling for Multimodal Human-Computer Interaction, In : *Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*, Barcelona, Spain, pp. 41–48.
- **Landragin, F.** (2005), Modeling Context for Referring in Multimodal Dialogue Systems, In : *Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'05, Paris, France)*, Lecture Notes in Computer Science (LNCS) Volume 3554, Springer-Verlag, Berlin, Germany, pp. 240–253.

2.1.2 Le flux des traitements

Le flux des traitements effectués par les différents modules de l'architecture peut se faire, comme on l'a 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 je viens 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 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 (Bobrow *et al.* 1977) 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 ou *cloud computing*).

(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 mon 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 multi-couches (*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'ar-

chitectures à trois couches, capables de gérer de manière séparée plusieurs aspects de la communication, par exemple avec un ACA. Des architectures multi-couches 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. C'est l'un des objets de la publication suivante, dans laquelle nous proposons une approche hybride destinée à spécifier une architecture pour concilier IHM et DHM :

- Lard, J., **Landragin, F.**, Grisvard, O. & Faure, D. (2007), Un cadre de conception pour réunir les modèles d'interaction et l'ingénierie des interfaces, *Ingénierie des Systèmes d'Information (ISI)* 12(6), Hermès–Lavoisier, Paris, France, pp. 67–91.

2.1.3 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 nous avons contribué à spécifier et que nous avons utilisé dans les projets MIAMM, OZONE, AMIGO ou encore dans la campagne d'évaluation MEDIA (Devillers *et al.* 2004), avec la propre participation d'Alexandre Denis. La publication correspondant à l'une des étapes de spécification du langage MMIL est la suivante :

- **Landragin, F.**, Denis, A., Ricci, A. & Romary, L. (2004), Multimodal Meaning Representation for Generic Dialogue Systems Architectures, In : *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 521–524.

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 multi-niveaux, 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 au niveau des modules concernés et non au niveau de l'architecture globale.

L'un des aspects de MMIL a été l'objet de 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 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.

2.2 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). C'est le pas que nous avions franchi dans le cadre du projet OZONE, avec la publication suivante :

- Gaiffe, B., **Landragin, F.** & Quignard, M. (2004), Le dialogue naturel comme un service dans un contexte multi-applicatif, In : *Actes de la journée d'étude de l'Association pour le Traitement Automatique des LAngues (ATALA) sur les relations entre systèmes multi-agents et traitement automatique des langues (AGENTAL)*, Paris, France, pp. 57–66.

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.

2.2.1 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 programma-

tion 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 plateforme 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 encore 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. Par contre, 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.

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 (j'y ai activement participé à THALES), 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. Le premier pas auquel j'ai contribué à THALES (premier pas réalisé lors de la préparation de la soumission du projet COCASE) est résumé dans l'article court suivant :

- Sedogbo, C., Grisvard, O., **Landragin, F.**, Lard, J. & Praud, S. (2006), HMI Engineering Productivity : the Poor Child of MDE/MDA Trends. A Vision for Model-Driven Human-Computer Interaction Engineering, In : *Dix-huitième Conférence Francophone sur l'Interaction Homme-Machine (Model-Driven Engineering and Human-Computer Interaction Workshop at IHM 2006)*, Montréal, Canada (2 pages).

2.2.2 *Middleware* pour l'interaction homme-machine

Une alternative, qui rejoint un peu les enjeux de type « OpenDial » décrits page 15, 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 mes 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.

Dans la publication suivante déjà citée plus haut, nous avons proposé un *middleware* implicite, basé sur une architecture multi-couches, pour faciliter le développement de systèmes d'interaction homme-machine, en incluant quelques aspects (très simplifiés) de DHM :

- Lard, J., **Landragin, F.**, Grisvard, O. & Faure, D. (2007), Un cadre de conception pour réunir les modèles d'interaction et l'ingénierie des interfaces, *Ingénierie des Systèmes d'Information (ISI)* 12(6), Hermès–Lavoisier, Paris, France, pp. 67–91.

Le principe est l'ajout d'une couche dédiée à l'interaction homme-machine dans l'architecture multi-couches 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, qui prend un peu la suite de celle réalisée avec Bertrand Gaiffe et Matthieu Quignard pour le projet OZONE (cf. plus haut), 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.

On l'a 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, on a 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 page 23, 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 que nous avons proposé, avec là aussi une prise en compte des aspects TAL dans toute leur variété et complexité. Cet enjeu peut s'ailleurs se multiplier si 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 rève 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.

3 Résolution des références en dialogue

Ce chapitre focalisé sur la référence montre 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, et la référence est ainsi parfois placée au cœur du sens.

3.1 Résolution des références à des objets

3.1.1 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. 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 vertes : 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. C'est un aspect complémentaire de celui de saillance abordé plus haut, et le modèle des domaines de référence multimodaux est ma principale proposition concrète concernant la référence.

Par rapport à la théorie de la représentation du discours et à son extension qu'est la MDRT pour le dialogue multimodal (Pineda & 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. Dans ce sens, ils apportent des capacités de compréhension à la machine, et ont de fait été utilisés par plusieurs chercheurs en DHM, voire en robotique. J'en parle 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 (chapitre 6 de Reboul & Moeschler 1998). Parmi les travaux proches se trouvent ceux de (Beun & 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 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.

Les deux points d'entrée pour ma version multimodale du modèle des domaines de référence sont les publications suivantes :

- **Landragin, F.**, Salmon-Alt, S. & Romary, L. (2002), Ancrage référentiel en situation de dialogue, *Traitement Automatique des Langues (TAL)* 43(2), Hermès–Lavoisier, Paris, France, pp. 99–129.
- **Landragin, F.** (2006), Visual Perception, Language and Gesture : A Model for their Understanding in Multimodal Dialogue Systems, *Signal Processing* 86(12), Elsevier, Amsterdam, The Netherlands, pp. 3578–3595.

Le premier de ces deux articles de revue date de ma thèse de doctorat, et a été écrit en collaboration avec Susanne Alt et Laurent Romary qui ont publié en parallèle sur les domaines de référence, notamment en faisant un lien avec l'approche des grammaires cognitives. Il pose les premières briques du modèle : modélisation linguistique, modélisation du contexte visuel (§ 3.1.2), modélisation du geste (§ 3.1.3). Le second est plus abouti dans la mesure où il intègre les trois modélisations et propose une caractérisation fine des cas de référence possibles (§ 3.1.4). Ces cas possibles ont été vérifiés un par un par N. O. Bernsen qui en a fait l'objet d'un article (Bernsen 2006), et d'une manière générale cet article a été utilisé par plusieurs chercheurs s'intéressant à la référence aux objets.

3.1.2 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. Le système de DHM connaît la nature et l'emplacement spatial de tous les objets affichés sur la scène visuelle, donc, dans mon exemple, des trajets de train qui ont été mis en valeur graphiquement. Dans ce contexte, l'utilisateur peut se focaliser sur un sous-ensemble, par exemple celui des trajets apparaissant à gauche de l'écran. 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., cf. une formalisation hiérarchique dans (Landragin 2004). 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 on l'a vu en § 1.3.1. Au niveau des domaines de référence, la saillance d'un objet ne contribue pas à construire un nouveau domaine potentiel, mais, par contre, 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.

3.1.3 Analyse des gestes de désignation

Dans le cadre d'une interaction multimodale avec écran tactile, l'utilisateur peut effectuer des gestes comme des pointages ou des entourages pour 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 (**ambiguïté sur la portée du geste**). 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 & Romary 1996), d'une analyse de la scène visuelle en termes de groupes perceptifs (§ 3.1.2), 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. Par contre, 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 pré-analyse du geste en contexte visuel, et c'est cette pré-analyse 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.

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.

En plus des publications déjà citées en § 3.1.1, la suivante porte plus spécifiquement sur les gestes de pointage et d'entourage effectués pour référer à des objets sur écran tactile. Les problèmes d'ambiguïté y sont notamment traités :

- **Landragin, F.**, De Angeli, A., Wolff, F., Lopez, P. & Romary, L. (2002), Relevance and Perceptual Constraints in Multimodal Referring Actions, In : van Deemter, K. & Kibble, R. (Eds.), *Information Sharing : Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford, CA, pp. 395–413.

3.1.4 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. 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* ». 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 mon exemple U2 impliquant l'expression référentielle « *ce chemin* » et un geste de désignation vers l'un des chemins affichés à l'écran, on est 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. 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 gestes 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 me 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 & Araki 2005). Le chapitre 4 présentera un troisième niveau, pragmatique, relatif aux actes de dialogue (§ 4.2).

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 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 (positionner un clou n'a rien à voir avec positionner 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 (cf. 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.

A propos des publications concernées par cette sous-section, je remets ici l'article paru dans la revue internationale *Signal Processing*, dans la mesure où il décrit en détails les problèmes d'appariement, de domaine sous-spécifié, et surtout des combinaisons possibles des types d'accès aux référents et des types de déterminants. Par ailleurs, d'un point de vue plus linguistique (mais toujours dans l'optique de faire des liens entre d'un côté les théories linguistiques comme celle de F. Corblin citée plus haut, et de l'autre côté les modèles computationnels), un article dans la *Revue de Sémantique et de Pragmatique* fait un écho à la discussion sur les types de référents possibles. Il présente notamment un éventail d'expressions référentielles démonstratives problématiques du point de vue de leur prise en compte dans des modèles linguistiques et dans des modèles computationnels :

- **Landragin, F.** (2004), *Dialogue homme-machine multimodal. Modélisation cognitive de la référence aux objets*, Hermès-Lavoisier, Paris, France, ISBN 2-7462-0992-6 (268 pages, version revue et publiée de ma thèse de doctorat).
- **Landragin, F.** (2005), Une caractérisation de la référence ostensive indirecte, *Revue de Sémantique et de Pragmatique (RSP)* 18, Presses Universitaires d'Orléans, France, pp. 7–22.
- **Landragin, F.** (2006), Visual Perception, Language and Gesture : A Model for their Understanding in Multimodal Dialogue Systems, *Signal Processing* 86(12), Elsevier, Amsterdam, The Netherlands, pp. 3578–3595.

3.2 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 ac-

tion 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 (§ 3.2.1), mais aussi les objets concernés et, d'une manière générale, la tâche en cours (§ 3.2.2).

3.2.1 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 en § 1.1.1 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 d'habitude avec la langue, des **ambiguïtés** peuvent survenir et compliquer la résolution de la référence. Dans un système de manipulation d'objets comme le classique et exemplaire SHRDLU (Winograd 1972), 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 extra-périphériques 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.

Je n'ai pas encore creusé l'implémentation d'un tel modèle temporel. Par contre, les publications suivantes présentent des éléments pour la résolution des références aux actions, avec la prise en compte de la valence verbale et des types d'ambiguïtés possibles, ainsi, que, de manière plus large, la représentation du sens des énoncés sur la base des résultats de la résolution des références aux objets et ceux de la résolution des références

aux actions :

- **Landragin, F.** (2004), Interface sémantique-pragmatique et domaines de référence, In : *Quatrièmes Journées d'Etudes Linguistiques de Nantes (JEL 2004)*, Nantes, France, pp. 67-72.
- **Landragin, F.** (2006), Modélisation du sens et du contexte sur la base de représentations des objets référés, Journées scientifiques de Sémantique et Modélisation, Bordeaux, France (2 pages).

3.2.2 Une 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 ré-interprétation 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-détermination 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) ; et enfin la confrontation des analyses pragmatiques, ce qui fera l'objet du chapitre 4.

Deux articles courts (une version longue beaucoup plus aboutie est en cours de soumission) viennent illustrer ces aspects, avec une mise en avant des conditions propices à l'apparition d'ambiguïtés dans un contexte de dialogue multimodal :

- **Landragin, F.** (2007), Un exemple de polysémie du geste co-verbal en situation de communication homme-machine et ses conséquences sur les analyses sémantiques et pragmatiques, In : *Typology, Gesture, and Sign, Second International Conference of the French Association for Cognitive Linguistics (AFLiCo 2)*, Lille, France, pp. 41–42.
- **Landragin, F.** (2009), Effective and Spurious Ambiguities due to some Co-verbal Gestures in Multimodal Dialogue, In : *Eighth International Gesture Workshop (GW 2009)*, Bielefeld, Germany (2 pages).

3.3 Gestion des anaphores et des coréférences

L'un des rôles de l'historique du dialogue est de retenir les objets référencés ainsi que les expressions utilisées pour ce faire, au fur et à mesure des échanges. 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 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, mes 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. De même avec « *j'ai réservé un aller pour Paris avec des préférences d'horaire et de place. Je recommence avec un aller pour Lyon* », où le verbe « *recommencer* » ne se comprend qu'à l'aide de l'antécédent explicitant une réservation. Enfin, un exemple de coréférence événementielle dans le cadre de ma 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. C'est l'originalité de ma démarche, qui fait suite à celle de l'école de Nancy (cf. par exemple le chapitre de J.-M. Pierrel et L. Romary dans Sabah *et al.* 1997), et c'est dans ce sens que vont les publications suivantes :

- **Landragin, F.** (2006), Influence de la situation lors de la résolution des anaphores dans le dialogue, In : *Treizième conférence sur le traitement automatique des langues (TALN 2006, Leuven, Belgium)*, Presses Universitaires de Louvain, Belgium, pp. 207–216.
- **Landragin, F.** (2007), Taking Situational Factors into Account when Resolving Anaphora : an Approach based on Events and Salience, In : *Sixth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007)*, Lagos, Portugal, pp. 71–76.
- **Landragin, F.** (2011), Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits, *Corpus* 10, `http://corpus.revues.org`, pp. 61–80.

Les deux premières proposent un modèle du contexte pour la résolution des anaphores en situation de dialogue, avec les éléments que nous avons décrits dans cette sous-section, mais sous la forme de paramètres hiérarchisés et d'étapes de traitement plus précises. La troisième porte sur les anaphores et les coréférences dans un tout autre contexte : il s'agit d'un projet portant sur des textes narratifs écrits (donc à l'opposé du DHM), et avec une démarche située entre la linguistique et la linguistique outillée (donc une approche différente de celle du TAL). Le contenu est complémentaire dans la mesure où il explicite un modèle linguistique conçu pour autoriser à terme des applications en linguistique computationnelle. C'est un premier pas que je suis en train d'explorer de manière plus approfondie, en collaboration avec des linguistes spécialistes de l'anaphore et de la coréférence (cf. chapitre 5), et qui me permettra à terme de mieux intégrer une prise en compte fine de ces phénomènes dans de futurs systèmes de DHM.

4 Reconnaissance des actes de dialogue multimodaux

Ce chapitre focalisé sur les actes de langage montre comment l'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. Ce 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. De fait, identification des actes de langage et gestion du dialogue vont de pair, comme on va le voir notamment dans un contexte multimodal.

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 & Wilson 1995), constituent un volet de la pragmatique 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, 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, avec les indications prosodiques, et notamment celles qui concernent le contour intonatif de l'énoncé, et avec les résultats des résolutions des références. 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 ». On va le voir, 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.

4.1 Identification et traitement des actes de dialogue

4.1.1 Actes de langage et actes de dialogue

(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. 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).

On a vu que (Sperber & 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. On a 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. 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** (Traum & Hinkelman 1992), 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, 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. 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 l'on verra plus loin sous le nom d'acte composite.

4.1.2 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 (cf. 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 (*Conditional Random Fields*) ou les HMM (*Hidden Markov Models*), la succession des actes précédents.

Comme le souligne (Jurafsky & 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.

4.1.3 Cas des actes indirects et des actes composites

Le processus d'identification des actes indirects reprend les mêmes paramètres que ceux présentés dans la sous-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 (*Beliefs, Desires, Intentions*) ou BOID (*Beliefs, Obligations, Intentions, Desires*), 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 (ou multifonctionnels) 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, car ne rien dire peut alors se comprendre comme une acceptation tacite) ;
- 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'effet contextuel (Sperber & Wilson 1995).

Dans les deux cas, les processus mis en œuvre dans les systèmes actuels reposent encore une fois sur une combinaison de règles algorithmiques classiques avec des techniques d'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 & Martin 2009). Dans les systèmes MIAMM et OZONE, ce sont des systèmes classiques de règles qui ont été implémentés : je ne suis pas encore entré dans le monde de l'apprentissage automatique, en tout cas pas au niveau des implémentations. Mon apport réside surtout dans la prise en compte des différentes facettes des actes indirects et composites, parce qu'ils sont au cœur de la gestion du dialogue, comme on le voit avec l'exemple U2 et le type de règle qu'il implique (on réagit à un acte regroupant une question et un commentaire par la réponse à la question [...], cf. ci-dessus). Dans les publications suivantes, j'ai ainsi mis en avant la complexité de ces phénomènes pragmatiques, et j'ai montré en quoi une bonne gestion notamment des actes composites pouvait entraîner un dialogue coopératif, sans avoir à revenir à de grands principes abstraits de coopérativité, comme le sont les maximes de (Grice 1975) ou des équivalents :

- **Landragin, F.** (2005), Traitement des actes de langage dans un système de dialogue homme-machine, Journées scientifiques de Sémantique et Modélisation, Paris, France (2 pages).
- **Landragin, F.** (2005), Indirect Speech Acts and Collaborativeness in Human-Machine Dialogue Systems, In : *First International Symposium on the Exploration and Modelling of Meaning (SEM-05)*, Biarritz, France, pp. 115–122.
- **Landragin, F.** (2008), Vers l'identification et le traitement des actes de dialogue composites, In : *Quinzième conférence sur le traitement automatique des langues (TALN 2008)*, Avignon, France, pp. 460–469.

4.2 Traitement des actes de dialogue multimodaux

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 (geste iconique ou quasi-linguistique). 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 (cf. page 49). 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. Par contre, en dialogue multimodal, on fait 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 » (on reste 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.

Sans encore proposer un modèle complet pour l'identification et le traitement des actes multimodaux, les publications précédentes, auxquelles s'ajoute la suivante, décrivent les éléments qui précèdent et font le point sur les processus de fusion et de fission multimodales : ces deux processus, le premier plutôt pour l'interprétation, le second pour la génération, œuvrent à trois niveaux : le niveau des signaux, ou niveau physique, avec les aspects de synchronisation temporelle ; le niveau sémantique, avec les contenus des énoncés linguistiques, gestuels voire visuels ; le niveau pragmatique, avec les actes de dialogue. Là aussi, une version plus aboutie de ce travail est en cours de soumission.

- **Landragin, F.** (2007), Physical, Semantic and Pragmatic Levels for Multimodal Fusion and Fission, In : *Seventh International Workshop on Computational Semantics (IWCS-7)*, Tilburg, The Netherlands, pp. 346–350.

5 Perspectives : la saillance en langue et en dialogue

Cette section comprend mon programme de recherche pour les années à venir. Celui-ci se fonde d'une part sur les activités décrites précédemment dans ce manuscrit, et d'autre part sur des activités plus linguistiques, en cours depuis mon arrivée dans le laboratoire LATTICE en 2006. La saillance en est le fil directeur, dans la mesure où elle représente pour moi l'écart entre une analyse sémantico-pragmatique complète et rationnelle, et la compréhension d'un sujet humain. Que ce soit en lisant un texte ou en discutant avec un interlocuteur, les effets de saillance, qui mettent en avant une partie du message transmis au point d'en occulter d'autres parties, me semblent fréquents et importants quant à leurs conséquences sur la communication. On ne percute pas toutes les informations que l'on reçoit de la même façon, et le fait que notre attention se porte sur tel point plutôt que tel autre a des répercussions sur la suite : suite du déroulement d'un dialogue, suite de la compréhension d'un texte. Les analyses linguistiques, par exemples celles de la sémantique formelle, et les analyses TAL, ne mettent pas assez en avant ces effets de saillance. Pour moi, à une représentation du sens sous forme logique ou sous la forme d'une structure de traits, il faudrait systématiquement ajouter un ou plusieurs prédicats de saillance, avec comme arguments les entités du discours mises en saillance. Ce n'est qu'ainsi (et bien entendu en exploitant ces prédicats dans les processus inférentiels qui constituent une facette de la pragmatique) que l'on peut rendre compte à leur juste mesure des effets de saillance. Je n'en suis pas encore là, je n'ai pas encore de modèle computationnel de la saillance à présenter, et c'est pourquoi j'ai choisi d'aborder ces thématiques dans le chapitre dédié à mon programme de recherche.

5.1 La saillance des entités du discours

Dans un énoncé en situation de dialogue ou dans les phrases constituant un texte écrit, on est en présence d'entités du discours, qui sont avant tout les référents. Comme on l'a vu, il peut s'agir d'individus humains, d'objets concrets, d'objets abstraits, d'actions, d'états, etc. La structure de l'énoncé ou de la phrase courante, ainsi que son rôle dans le dialogue ou le discours, tendent à affecter une certaine saillance à certaines entités plutôt qu'à d'autres. C'est ce que j'ai creusé dans les publications suivantes, avec premièrement une liste des facteurs de saillance intervenant en dialogue et en discours, deuxièmement une synthèse sur la notion de saillance en sémantique formelle, troisièmement un modèle plus complexe qu'une simple liste de facteurs mis à plat, et quatrièmement une mise au point méthodologique sur la saillance :

- **Landragin, F.** (2004), Saillance physique et saillance cognitive, *Cognition, Représentation, Langage (CORELA)* 2(2), `http://corela.edel.univ-poitiers.fr/` (24 pages).
- **Landragin, F.** (2007), Saillance, In : Godard, D., Roussarie, L. & Corblin, F. (Eds.), *Dictionnaire de sémantique*, GdR Sémantique et Modélisation, CNRS, `http://www.semantique-gdr.net/dico/` (6 pages).
- **Landragin, F.** (2010), Sur les aspects multicritères et multidimensionnels de la saillance, In : *La saillance en langue et en discours (Saillance 2)*, Strasbourg, France (4 pages).

- **Landragin, F.** (2012), La saillance : questions méthodologiques autour d'une notion multifactorielle, *Faits de Langues* 39, Peter Lang, Bern, Switzerland, pp. 15–31.

Ces publications correspondent à des interrogations et à une première tentative de modélisation de cette notion difficile à appréhender car souvent mal définie et reposant de manière floue sur d'autres notions linguistiques (Schnedecker 2011). Dans l'état actuel de mes recherches, je considère la saillance comme une notion multifactorielle, qui émerge lors de la conjonction de plusieurs facteurs (prosodiques, lexicaux, syntaxiques, sémantiques, pragmatiques) et qui joue sur l'interprétation en mettant en avant une entité du discours plutôt qu'une autre. Cette entité privilégiée va potentiellement attirer l'attention du lecteur ou de l'interlocuteur, au point d'occulter d'autres entités, voire des pans complets du message transmis. Comme la Théorie de l'Accessibilité (Ariel 1990), je considère que les formes linguistiques, et notamment les types de déterminants (pour faire écho au chapitre 3), sont liés à une saillance plus ou moins forte. Contre la Théorie du Centrage (Grosz *al.* 1995), je considère qu'il n'y a pas qu'un seul centre, mais que l'on peut considérer plusieurs centres ou saillances qui œuvrent en parallèle. C'est dans ce sens que j'ai commencé à proposer un modèle multidimensionnel de la saillance, et que je compte continuer dans cette voie :

- Saillance préalable et saillance nouvelle : j'oppose l'exploitation d'une saillance existante à la mise en saillance d'une nouvelle entité. L'exploitation d'une saillance existante intervient notamment lors d'une production linguistique fondée sur une perception visuelle et lors de l'utilisation d'un pronom anaphorique. La mise en saillance d'une nouvelle entité relève d'un mécanisme de composition du discours, avec comme but de mettre en valeur une entité particulière, en préparation d'une référence future.

- Saillance à effet immédiat et saillance à effet continu : j'oppose la saillance rapide et inconsciente (effet *pop-up*) à la saillance – plus proche de la notion de prégnance – qui se construit petit à petit, de manière incrémentale, comme lorsqu'il s'agit du personnage principal d'un roman, ou, d'une manière générale, du topique discursif.

- Saillance physique et saillance cognitive : en reprenant le titre de la première publication citée ci-dessus, il s'agit de distinguer la saillance qui a des effets physiques, c'est-à-dire ayant une trace matérielle dans le message visuel ou linguistique (saillance objective, et donc calculable dans un système de TAL) et la saillance relevant d'inférences sur l'attention, l'intention, la mémoire, les affects, et d'une manière générale tout ce qui est implicite et relève de la cognition (saillance subjective, non calculable ou alors au prix de suppositions).

- Saillance linguistique et saillance visuelle : la modalité en jeu est également à l'origine d'une dimension d'analyse de la saillance. En effet, même si les mécanismes sont similaires, il est nécessaire de retenir tout au long de l'interprétation d'un texte quelle saillance vient d'une source visuelle et quelle saillance vient du discours lui-même.

- Saillance informative et saillance rhétorique : cette dernière dimension oppose les effets interprétatifs liés à l'apport d'information (un élément nouveau est saillant de par l'information qu'apporte cette nouveauté) aux effets interprétatifs liés à la rhétorique normative ou prescriptive, ou encore aux effets des figures de style.

Les réflexions et les exemples donnés dans ces publications concernent tous des référents de type individu ou objet. Il me reste à reprendre ces travaux et à étudier la saillance des événements (actions, états, etc.). Ce n'est pas un problème simple, surtout que les événements et les référents sont liés les uns aux autres, ne serait-ce que parce qu'un événement fait intervenir des actants qui sont des référents. C'est probablement une étape nécessaire pour pouvoir aboutir à un modèle plus abouti de la saillance. Il n'est d'ailleurs pas anodin que la plus récente de mes publications sur la saillance porte sur la méthodologie : après avoir proposé différentes visions et différents modèles, je suis en train de revoir complètement la manière d'aborder cette notion. Pour ce qui concerne le TAL et l'identification automatique des entités du discours saillantes, j'y reviendrai en § 5.5. Pour ce qui concerne l'analyse linguistique, ce travail passera nécessairement par des études approfondies de corpus. Pour ce faire, une méthode consiste à annoter dans des textes l'ensemble des facteurs de saillance potentiellement impliqués, puis à confronter les résultats obtenus (via des statistiques descriptives et des statistiques inférentielles) avec la détection par des sujets des effets de saillance.

5.2 La saillance des entités du dialogue à support visuel

Dès mon expérience dans le projet européen COVEN, j'ai étudié en parallèle saillance linguistique et saillance visuelle, et ce d'autant plus que le dialogue multimodal fait intervenir ce qui est pour moi les deux matérialisations d'un même concept cognitif. Il ne s'agit plus cette fois d'entités du discours, mais d'entités visuelles, à savoir les objets que l'on identifie dans une scène. En suivant tout d'abord une approche similaire à celle de la Théorie de la Gestalt (critères d'aggrégation d'unités et de formation d'une « bonne forme »), puis en m'inspirant des travaux sur la saillance linguistique, j'ai été amené à proposer une liste de facteurs de saillance visuelle. Les nombreuses analogies possibles avec les facteurs de saillance linguistique m'ont entraîné, d'une part à explorer les points communs (analogie), d'autre part à m'interroger sur l'existence d'un mécanisme cognitif universel que serait la saillance (homologie).

Ces aspects ont été publiés dans les articles suivants, avec premièrement un ensemble simple de facteurs permettant de prendre en compte la saillance visuelle lors de la résolution des références en DHM, deuxièmement un premier modèle cognitif rapprochant saillance visuelle et saillance linguistique, modèle qui s'est affiné dans les deux publications suivantes :

- **Landragin, F.**, Bellalem, N. & Romary, L. (2001), Visual Salience and Perceptual Grouping in Multimodal Interactivity, In : *First Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 151–155.
- **Landragin, F.** (2005), Modélisation de la saillance visuelle et linguistique, In : *Sixième Colloque des Jeunes Chercheurs en Sciences Cognitives (CJCSC'05)*, Bordeaux, France, pp. 157–162.
- **Landragin, F.** (2009), De la saillance visuelle à la saillance linguistique, In : *Aspects linguistiques et communicatifs de la mise en évidence dans un texte (Saillance-1)*, Genève, Switzerland, pp. 9–13.
- **Landragin, F.** (2011), De la saillance visuelle à la saillance linguistique, In : Inkova, O. (Ed.), *Saillance. Aspects linguistiques et communicatifs de la mise en*

évidence dans un texte, Annales Littéraires de l'Université de Franche-Comté 897, pp. 67–83.

Dans cette liste, seule la première publication avait une visée informatique, c'est-à-dire ouvrait la voie à l'identification automatique des objets saillants dans une scène visuelle. De fait, cette proposition reposait en partie sur l'implémentation que j'avais réalisée dans le cadre du projet COVEN, et en partie sur l'étude d'un corpus multimodal qui a fait l'objet de mon DEA. L'avantage dans le DHM avec support visuel géré par le système, est que celui-ci dispose de la liste des objets visibles avec leurs caractéristiques. Calculer des facteurs de saillance tels qu'une singularité de couleur ou de taille est ainsi très facile à effectuer. Par contre, si l'on vise des applications en robotique, pour lesquelles le système de DHM ne connaît pas a priori les objets visibles, le travail est très différent. Il est dans ce cas nécessaire de faire appel à des techniques de vision artificielle, c'est-à-dire de détecter des contours de manière à identifier des objets dont seules des formes typiques sont connues. Aboutir à un modèle computationnel de la saillance visuelle fait partie de mes perspectives de recherche, mais ce ne peut être réalisable qu'en collaboration avec des chercheurs spécialistes de traitement du signal et de vision artificielle. Ma collaboration avec des chercheurs du GIPSA-Lab (cf. article dans la revue *Journal of Vision* en § 1.3.1) suit cette voie.

Par ailleurs, l'étude de la saillance visuelle va de pair avec des expérimentations. A l'instar de l'article dans *Journal of Vision*, il s'agit de prendre un par un les facteurs de saillance visuelle, de construire des scènes dans lesquelles ce facteur est plus ou moins mis en avant, et de tester expérimentalement les différences au niveau de la perception d'un sujet. Le protocole expérimental entraîne dans ce cas l'utilisation d'un oculomètre, et c'est pourquoi j'envisage de procéder à de telles expérimentations en collaboration avec des chercheurs dans des laboratoires spécialisés.

5.3 Saillance, coréférence et structure du discours

Les deux sous-sections précédentes concernaient surtout la saillance immédiate, qui œuvre au moment de la perception d'un énoncé oral, d'une phrase écrite ou d'une scène visuelle. Mais la saillance œuvre aussi dans la durée, dès lors qu'on considère un dialogue, un discours, ou une image animée, un film par exemple. En linguistique, on parle volontiers de topique discursif (Grobet 2002). La saillance n'est pas très loin : à partir du moment où une entité du discours telle que le personnage principal d'un roman acquiert le statut de topique, c'est aussi en raison d'une saillance maintenue. La coréférence dont nous avons parlé en § 3.3 n'est pas très loin non plus : pour maintenir la saillance d'un référent, il est nécessaire d'y référer à plusieurs reprises. Autrement dit, des liens entre saillance et chaîne de coréférence existent, même s'ils restent encore à préciser. Par ailleurs, les chaînes de coréférence entretiennent des liens avec la structure du discours, et, là aussi, les liens qui en découlent entre saillance et structure du discours restent encore à explorer.

C'était mon objectif quand j'ai coordonné entre 2009 et 2011 un groupe de travail interne au laboratoire LATTICE, puis, en 2011 et 2012, quand j'ai coordonné le projet PEPS MC4 (Modélisation Contrastive et Computationnelle des Chaînes de Coréférence), qui a fait intervenir des chercheurs des laboratoires LATTICE, LILPA et ICAR. Les résultats de ce projet sont en cours de soumission. Ils mettent en avant une façon innovante

d'appréhender les chaînes de coréférence, qui sont par ailleurs bien connues du TAL (cf. entre autres les critiques de van Deemter & Kibble 2000). Il s'agit de distinguer plusieurs degrés de contribution d'une expression linguistique à une chaîne de coréférence, avec au départ deux degrés, ce qui permet de parler de maillons forts et de maillons faibles de chaînes de coréférence. Les expressions référentielles classiques correspondent au degré le plus fort, et des formes atténuées telles que les sujets non exprimés de verbes à l'infinitif ou au participe correspondent au degré le plus faible. Derrière cette première distinction se trouve l'accessibilité, et par conséquent la saillance.

Les travaux ainsi effectués vont être prochainement poursuivis dans le cadre d'au moins deux projets faisant intervenir des chercheurs de plusieurs laboratoires français et internationaux : le projet EIIDA (Étude Interdisciplinaire et Interlinguistique du Discours Académique, responsables : Shirley Carter-Thomas et Jeanne-Marie Debaisieux) financé depuis 2012 par le Labex TransferS (`http://www.transfers.ens.fr/`) et le projet ORFEO (Outiller les Recherches sur le Français Ecrit et Oral, responsable : Jeanne-Marie Debaisieux) financé depuis 2013 par l'ANR. S'ils n'incluent ni l'un ni l'autre la saillance dans leurs thématiques, ils traitent en revanche des liens entre structure du discours et phénomènes linguistiques comme la coréférence.

5.4 Le cas des ambiguïtés et des expressions vagues

On l'a vu dans le chapitre sur la résolution des références et dans celui sur les actes de dialogue, l'ambiguïté caractérise le langage naturel. (Fuchs 2000) montre que les ambiguïtés peuvent provenir de diverses raisons (lexicales, syntaxiques, sémantiques). Avec un souci du détail et de l'exhaustivité, C. Fuchs fait la liste des structures du français propices à l'apparition d'ambiguïté. Elle distingue également l'ambiguïté (choix entre plusieurs alternatives clairement identifiables) et la sous-détermination (impossibilité d'affecter un sens à une expression). Ce sont ces deux aspects qui ont été à l'origine des publications suivantes, avec tout d'abord une synthèse sur les types d'ambiguïtés que l'on rencontre dans le dialogue homme-machine multimodal, puis deux articles courts sur l'ambiguïté et la sous-détermination lors de la résolution d'une anaphore :

- **Landragin, F.** (2003), Clues for the Identification of Implicit in Multimodal Referring Actions, In : *Tenth International Conference on Human-Computer Interaction (HCI International 2003, Heraklion, Crete, Greece)*, Lawrence Erlbaum Associates, Mahwah, NJ, Volume 2, pp. 711–715.
- **Landragin, F.** (2007), A Characterization of Underspecified Anaphora and its Consequences on the Annotation of Anaphoric Relations, In : *Theoretical and Computational Perspectives on Underspecification (Jahrestagung des SFB 732)*, Stuttgart, Germany (2 pages).
- **Landragin, F.** (2007), L'anaphore à antécédent flou : une caractérisation et ses conséquences sur l'annotation des relations anaphoriques, Journée d'étude de l'ATALA (Association pour le Traitement Automatique des LAngues) sur la résolution des anaphores, Paris, France (2 pages).

Pour ces deux derniers articles, j'envisage un approfondissement important et des perspectives de recherche qui relèvent à la fois de la linguistique descriptive, de la linguistique de corpus, de la psycholinguistique et du TAL. Par le terme « anaphore à

antécédent flou », j'entends une anaphore pour laquelle l'identification exacte de l'antécédent n'est pas nécessaire : on peut se contenter d'un antécédent sous-déterminé et continuer à dialoguer ou à lire le texte sans aucun problème de compréhension. C'est le cas dans l'exemple attesté suivant, où « *lui* » réfère à une femme : « *je lui ai dit sur un ton de plaisanterie que son idée était intéressante, qu'**elle** montrait les choses sous un angle auquel en effet on n'est pas habitué* ». Ce qui montre les choses sous un autre angle, ce peut être aussi bien l'idée que la personne, et, de fait, peu importe : le texte peut continuer avec des « *elle* » sans que déterminer s'il s'agit de l'idée, de la personne, voire d'un concept unique regroupant idée et personne, ne soit nécessaire. Cette démarche rejoint celle des interprétations *good-enough*, étudiées notamment en psycholinguistique. C'est un aspect que je compte explorer de manière à faire des liens, d'une part avec la notion de saillance, d'autre part avec l'interrogation récurrente « qu'est-ce que comprendre ? » et son homologue en TAL : « qu'est-ce que comprendre automatiquement ? ».

5.5 Linguistique-informatique et linguistique outillée

Qu'il s'agisse de saillance, d'ambiguïté ou de sous-détermination, je garde une approche linguistique à visée applicative, c'est-à-dire que je cherche à élaborer des modèles qui puissent à la fois satisfaire des linguistes non concernés par les applications potentielles des théories auxquelles ils travaillent, et permettre à terme des implémentations informatiques. C'est dans ce sens que vont les publications suivantes, avec un ensemble de propositions visant à permettre notamment des calculs de saillance à partir d'évaluation des rôles de chacun des facteurs identifiés :

- **Landragin, F.** (2003), La saillance comme point de départ pour l'interprétation et la génération, Journée d'étude de l'ATALA (Association pour le Traitement Automatique des LAngues) sur la structure communicative, Paris, France (4 pages).
- **Landragin, F.** (2004), L'utilisation de scores numériques en sémantique computationnelle, Journées scientifiques de Sémantique et Modélisation, Lyon, France (5 pages).
- **Landragin, F.** (2005), Traitement automatique de la saillance, In : *Douzième conférence sur le traitement automatique des langues (TALN 2005)*, Dourdan, France, pp. 263–272.

Par ailleurs, mes participations aux groupes de travail et aux projets du laboratoire Lattice (groupe EIOMSIT sur les éléments initiaux de phrases « Eléments Initiaux, Ordre des Mots, Structures Informationnelle et Textuelle » coordonné par Sophie Prévost et Shirley Carter-Thomas, groupe sur la coréférence, projet EIIDA, etc.) incluent souvent deux facettes complémentaires : la question scientifique traitée, comme par exemple la saillance des éléments initiaux ou les rapports entre chaînes de coréférence et structure du discours, et la méthodologie d'analyse de corpus (Habert *et al.* 1997). Sur ce dernier point, j'ai contribué à clarifier les méthodes d'annotation et d'interrogation de corpus, et à participer à l'organisation de formations – dont une école thématique du CNRS (« Annotation de données langagières », Biarritz, 2011, `http://annotationlinguistique.fr/`) – sur les outils d'annotation de corpus. Parmi ces outils, j'ai récemment pris en charge l'outil Analec (« Analyse de l'écrit », `http://www.lattice.cnrs.fr/Telecharger-Analec`) de Bernard Victorri, ce qui m'a

amené à écrire l'article suivant, premier d'une série de publications à venir sur la linguistique dite « outillée » (Habert 2005) :

- **Landragin, F.**, Poibeau, T. & Victorri, B. (2012), ANALEC : A New Tool for the Dynamic Annotation of Textual Data, In : *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquey, pp. 357–362.

C'est une voie sur laquelle je compte continuer, surtout que les travaux effectués par exemple dans le projet MC4 ont montré qu'il nous manquait des outils pour appréhender – c'est-à-dire visualiser, étudier, annoter, interroger – les données que sont les chaînes de coréférence d'un texte. Concevoir un outil dédié aux chaînes de coréférence fait ainsi partie de mes objectifs, avec comme condition de toujours garder un lien avec les questions scientifiques concernées : typologie des chaînes de coréférence, typologie des entre-croisements de chaînes, rapports entre structures discursives et chaînes de coréférence, entre autres exemples.

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 ». 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. Il n'est plus nécessaire de souligner 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. En revanche, il est encore nécessaire de souligner 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, on a donné un panorama des techniques 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 fin de l'introduction. 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. On a souligné l'importance des travaux à l'interface entre deux disciplines, avec l'exemple désormais évident de l'interface entre théories linguistiques et implantations 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 implantation, peut être critiqué aisément : seule une implantation s'accompagne de « preuves » et peut résister à la critique. Avec cette habilitation à diriger des recherches, j'espère 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. J'ai 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. J'ai 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. En conséquence, la quantité de travail nécessaire à l'élaboration d'un système de DHM dépasse le repère classique qu'est celui d'une thèse de doctorat, et des équipes s'avèrent désormais nécessaires. Les éléments constituant ces équipes 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 plateforme continuellement mise à jour. La généralisation de ce type de plateforme et des moyens de facilitation 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, cela permettrait de procéder à des évaluations plus fiables et plus comparables qu'elles ne le sont actuellement.

Références

Abbott, B. (2010), *Reference*, Oxford University Press, Oxford.

Abeillé, A. (2007), *Les grammaires d'unification*, Hermès-Lavoisier, Paris.

Allemandou, J., Charnay, L., Devillers, L., Lauvergne, M. & Mariani, J. (2007), Un paradigme pour évaluer automatiquement des systèmes de dialogue homme-machine en simulant un utilisateur de façon déterministe, *Traitement Automatique des Langues* 48(1), pp. 115–139.

Allen, J. F. & Perrault, C. R. (1980), Analysing Intention Utterances, *Artificial Intelligence* 15, pp. 143–178.

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. (1995), The TRAINS Project: A Case Study in Defining a Conversational Planning Agent, *Journal of Experimental & Theoretical Artificial Intelligence* 7(1), pp. 7–48.

Allwood, J., Traum, D. R. & Jokinen, K. (2000), Cooperation, Dialogue and Ethics, *International Journal of Human-Computer Studies* 53, pp. 871–914.

Antoine, J.-Y. (2003), Pour une ingénierie des langues plus linguistique, Mémoire d'Habilitation à Diriger des Recherches, Université de Bretagne Sud, Vannes.

Antoine, J.-Y. & Caelen, J. (1999), Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (Demande, Contrôle, Résultat), *Langues* 2(2), pp. 130–139.

Ariel, M. (1990), *Accessing Noun-Phrase Antecedents*, Routledge, London & New York.

Asher, N. & Gillies, A. (2003), Common Ground, Corrections, and Coordination, *Argumentation* 17, pp. 481–512.

Asher, N. & Lascarides, A. (2001), Indirect Speech Acts, *Synthese* 128(1–2), pp. 183–228.

Asher, N. & Lascarides, A. (2003), *Logics of Conversation*, Cambridge University Press, Cambridge.

Austin, J. (1962), *How to do things with words*, Oxford University Press, Oxford.

Baker, M. J. (2004), Recherches sur l'élaboration de connaissances dans le dialogue, Mémoire d'Habilitation à Diriger des Recherches, Université de Nancy 2.

Beaver, D. I. & Clark, B. Z. (2008), *Sense and Sensitivity: How Focus Determines Meaning*, Blackwell, Oxford.

Bellalem, N. & Romary, L. (1996), Structural Analysis of Co-Verbal Deictic Gesture in Multimodal Dialogue Systems, In: *Progress in Gestural Interaction. Proceedings of Gesture Workshop*, York, pp. 141–153.

Beringer, N., Kartal, U., Louka, K., Schiel, F. & Türk, U. (2002), PROMISE – A Procedure for Multimodal Interactive System Evaluation, In: *Proceedings of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, pp. 77–80.

Bernsen, N. O. (2006), Speech and 2D Deictic Gesture Reference to Virtual Scenes, In: André, E., Dybkjær, L., Minker, W., Neumann, H. & Weber, M. (Eds.), *Perception and Interactive Technologies*, Springer Verlag, Berlin, pp. 129–140.

Bernsen, N. O., Dybkjær, H. & Dybkjær, L. (1998), *Designing Interactive Speech Systems. From First Ideas to User Testing*, Springer Verlag, Berlin.

BERNSEN, N. O. & DYBKJÆR, L. (2004), Evaluation of Spoken Multimodal Conversation, In: *Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI)*, Penn State University, pp. 38–45.

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

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

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

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

BOBROW, D. G., KAPLAN, R. M., KAY, M., NORMAN, D. A., THOMPSON, H. & WINOGRAD, T. (1977), GUS, A Frame-Driven Dialog System, *Artificial Intelligence* 8, pp. 155–173.

BOUDREAU, S. & KITTREDGE, R. (2005), Résolution des anaphores et détermination des chaînes de coréférences. Différences entre variétés de textes, *Traitement Automatique des Langues* 46(1), pp. 41–70.

BRANIGAN, H. P., PICKERING, M. J., PEARSON, J. & MCLEAN, J. F. (2010), Linguistic Alignment between People and Computers, *Journal of Pragmatics* 42, pp. 2355–2368.

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

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

BUNT, H. (2011), Multifunctionality in dialogue, *Computer Speech and Language* 25, pp. 222–245.

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

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

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

CASSELL, J., SULLIVAN, J., PREVOST, S. & CHURCHILL, E. (Eds., 2000), *Embodied Conversational Agents*, The MIT Press, Cambridge.

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

CHASTAIN, C. (1975), Reference and Context, In: GUNDERSON, K. (Ed.), *Language Mind and Knowledge*, University of Minnesota Press, Minneapolis, pp. 194–269.

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

CLARK, E. V. (2009), *First Language Acquisition* (second edition), Cambridge University Press, Cambridge.

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

Clark, H. H. & Schaefer, E. F. (1989), Contributing to Discourse, *Cognitive Science* 13, pp. 259–294.

Clark, H. H. & Wilkes-Gibbs, D. (1986), Referring as a Collaborative Process, *Cognition* 22, pp. 1–39.

Cohen, M. H., Giangola, J. P. & Balogh, J. (2004), *Voice User Interface Design*, Addison-Wesley.

Cohen, P. R. & Levesque, H. J. (1990), Intention is Choice with Commitment, *Artificial Intelligence* 42, pp. 213–261.

Cohen, P. R. & Perrault, C. R. (1979), Elements of a Plan-Based Theory of Speech Acts, *Cognitive Science* 3, pp. 177–212.

Colby, K. M., Weber, S. & Hilf, F. D. (1971), Artificial Paranoia, *Artificial Intelligence* 2, pp. 1–25.

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

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

Corblin, F. (2002), *Représentation du discours et sémantique formelle*, PUF, Paris.

Cornish, F. (1999), *Anaphora, Discourse, and Understanding: Evidence from English and French*, Oxford University Press, Oxford.

Cosnier, J. & Vaysse, J. (1997), Sémiotique des gestes communicatifs, *Nouveaux actes sémiotiques* 52-53-54, pp. 7–28.

Dale, R. (1992), *Generating Referring Expressions*, The MIT Press, Cambridge, MA.

Denis, A. (2008), 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.

Denis, A. (2011), Generating Referring Expressions with Reference Domain Theory, In: *Proceedings of the 6th International Natural Language Generation Conference*, Dublin, pp. 27–35.

de Ruiter, J. P. & Cummins, C. (2012), A Model of Intentional Communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of Signals), In: *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, Paris, pp. 149–150.

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. (2004), The French MEDIA/EVALDA Project: The Evaluation of the Understanding Capability of Spoken Language Dialog Systems, In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisboa, pp. 2131–2134.

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

Duermael, F. (1994), Référence aux actions dans des dialogues de commande homme-machine, Thèse de doctorat, Institut National Polytechnique de Lorraine.

Dybkjær, L., Bernsen, N. O. & Minker, W. (2004), Evaluation and Usability of Multimodal Spoken Language Dialogue Systems, *Speech Communication* 43(1–2), pp. 33–54.

EDLUND, J., HELDNER, M. & GUSTAFSON, J. (2005), Utterance Segmentation and Turn-Taking in Spoken Dialogue Systems, In: FISSENI, B., SCHMITZ, H.-C., SCHRÖDER, B. & WAGNER, P. (Eds.), *Computer Studies in Language and Speech*, Peter Lang, pp. 576–587.

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

FEYEREISEN, P. (1997), La compréhension des gestes référentiels, *Nouveaux actes sémiotiques* 52-53-54, pp. 29–48.

FRASER, N. M. & GILBERT, G. N. (1991), Simulating Speech Systems, *Computer Speech and Language* 5, pp. 81–99.

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

FUNAKOSHI, K., NAKANO, N., TOKUNAGA, T. & IIDA, R. (2012), A Unified Probabilistic Approach to Referring Expressions, In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, pp. 237–246.

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

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

GIBBON, D., MERTINS, I. & MOORE, R. (Eds., 2000), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*, Kluwer Academic Publishers, Dordrecht.

GINZBURG, J. (2012), *The Interactive Stance: Meaning for Conversation*, Oxford University Press.

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

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

GRICE, H. P. (1975), Logic and Conversation, In: COLE, P. & MORGAN, J. (Eds.), *Syntax and Semantics*, Vol. 3, Academic Press, pp. 41–58.

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

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

GROBET, A. (2002), *L'identification des topiques dans les dialogues*, De Boeck-Duculot, Bruxelles.

GROSZ, B. J., JOSHI, A. & WEINSTEIN, S. (1995), Centering: a Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics* 21(2), pp. 203–225.

GROSZ, B. J. & SIDNER, C. L. (1986), Attention, Intentions and the Structure of Discourse, *Computational Linguistics* 12(3), pp. 175–204.

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

Guyomard, M., Nerzic, P. & Siroux, J. (1993–2006), 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.

Habert, B. (2005), Portrait de linguiste(s) à l'instrument, *Texto !* 104, `http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html`.

Habert, B., Nazarenko, A. & Salem, A. (1997), *Les Linguistiques de corpus*, Armand Colin, Paris.

Hardy, H., Biermann, A., Bryce Inouye, R., McKenzie, A., Strzalkowski, T., Ursu, C., Webb, N. & Wu, M. (2006), The Amitiés System: Data-Driven Techniques for Automated Dialogue, *Speech Communication* 48, pp. 354–373.

Harris, R. A. (2004), *Voice Interaction Design: Crafting the New Conversational Speech Systems*, Morgan Kaufmann Publishers.

Hirschman, L. (1992), Multi-Site Data Collection for a Spoken Language Corpus: MADCOW, In: *Proceedings of the* Darpa *Speech and Natural Language Workshop*, New York, pp. 7–14.

Hobbs, J.-R. (1979), Coherence and Coreference, *Cognitive Science* 3, pp. 67–90.

Horchani, M. (2007), Vers une communication humain-machine naturelle : stratégies de dialogue et de présentation multimodales, Thèse de doctorat, Université Joseph Fourier, Grenoble.

Issarny, V., Sacchetti, D., Tartanoglu, F., Sailhan, F., Chibout, R., Levy, N. & Talamona, A. (2005), Developing Ambient Intelligence Systems: A Solution based on Web Services, *Automated Software Engineering* 12(1), pp. 101–137.

Jokinen, K. & McTear, M. F. (2010), *Spoken Dialogue Systems*, Morgan & Claypool Publishers.

Jönsson, A. & Dählback, N. (1988), Talking to a Computer is not like Talking to your Best Friend, In: *Proceedings of the Scandinavian Conference on Artificial Intelligence*, Tromsø.

Jurafsky, D. & Martin, J. H. (Eds., 2009), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (second edition), Pearson, Upper Saddle River, NJ.

Kadmon, N. (2001), *Formal Pragmatics*, Blackwell, Oxford.

Kamp, H. & Reyle, U. (1993), *From Discourse to Logic*, Kluwer, Dordrecht.

Kendon, A. (2004), *Gesture: Visible Action as Utterance*, Cambridge University Press, Cambridge.

Kennedy, A., Wilkes, A., Elder, L. & Murray, W. (1988), Dialogue with machines, *Cognition* 30, pp. 73–105.

Kerbrat-Orecchioni, C. (2012), *L'implicite*, Armand Colin, Paris.

Kerbrat-Orecchioni, C. (1996), *La conversation*, Seuil, Paris.

Kievit, L., Piwek, P., Beun, R.-J. & Bunt, H. (2001), Multimodal Cooperative Resolution of Referential Expressions in the DenK System, In: Bunt, H. & Beun, R.-J. (Eds.), *Cooperative Multimodal Communication*, Springer, Berlin & Heidelberg, pp. 197–214.

Knott, A. & Vlugter, P. (2008), Multi-Agent Human-Machine Dialogue: Issues in Dialogue Management and Referring Expression Semantics, *Artificial Intelligence* 172, pp. 69–102.

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

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

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

KRAHMER, E. & VAN DEEMTER, K. (2012), Computational Generation of Referring Expressions: A Survey, *Computational Linguistics* 38(1), pp. 173–218.

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

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

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

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

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

LEMEUNIER, T. (2003), De la modélisation de l'activité conversationnelle des systèmes de dialogue personne-machine, *Cahiers Romans de Sciences Cognitives* 1(2), pp. 23–52.

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

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

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

LUPERFOY, S. (1992), The Representation Of Multimodal User Interface Dialogues Using Discourse Pegs, In: *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, pp. 22–31.

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

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

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

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

MELLISH, C., SCOTT, D., CAHILL, L., PAIVA, D., EVANS, R. & REAPE, M. (2006), A Reference Architecture for Natural Language Generation Systems, *Natural Language Engineering* 12, pp. 1–34.

Mitkov, R. (2002), *Anaphora Resolution*, Longman, London.

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

Moeschler, J. (Ed., 1992), Théorie des actes de langage et analyse des conversations, *Cahiers de linguistique française* 13, Université de Genève.

Möller, S., Smeele, P., Boland, H. & Krebber, J. (2007), Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study, *Computer Speech and Language* 21, pp. 26–53.

Moulton, J. & Roberts, L. D. (1994), An AI Module for Reference Based on Perception, In: *Proceedings of the AAAI Workshop on Integration of Natural Language and Vision Processing*, Seattle.

Muskens, R. (1996), Combining Montague Semantics and Discourse Representation, *Linguistics and Philosophy* 19(2), pp. 143–186.

Olson, D. R. (1970), Language and Thought: Aspects of a Cognitive Theory of Semantics, *Psychological Review* 77, pp. 257–273.

Oviatt, S. L. (1999), Ten Myths of Multimodal Interaction, *Communications of the ACM* 42(11), pp. 74–81.

Paek, T. & Pieraccini, R. (2008), Automating Spoken Dialogue Management Design Using Machine Learning: An Industry Perspective, *Speech Communication* 50, pp. 716–729.

Pickering, M. J. & Garrod, S. (2004), Toward a Mechanistic Psychology of Dialogue, *Behavorial and Brain Sciences* 27, pp. 169–226.

Pierrel, J.-M. (1987), *Dialogue oral homme-machine*, Hermès, Paris.

Pineda, L. & Garza, G. (2000), A Model for Multimodal Reference Resolution, *Computational Linguistics* 26(2), pp. 139–193.

Poesio, M. & Traum, D. R. (1997), Conversational Actions and Discourse Situations, *Computational Intelligence* 13(3), pp. 309–347.

Prévot, L. (2004), Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans des dialogues finalisés, Thèse de doctorat, Université Paul Sabatier, Toulouse.

Reboul, A. & Moeschler, J. (1998), *Pragmatique du discours. De l'interprétation de l'énoncé à l'interprétation du discours*, Armand Colin, Paris.

Reichman, R. (1985), *Getting Computers to Talk Like You and Me*, The MIT Press, Cambridge.

Reiter, E. & Dale, R. (2000), *Building Natural Language Generation Systems*, Cambridge University Press, Cambridge.

Rieser, V. & Lemon, O. (2011), *Reinforcement Learning for Adaptive Dialogue Systems. A Data-driven Methodology for Dialogue Management and Natural Language Generation*, Springer.

Rosset, S. (2008), Systèmes de dialogue (oral) homme-machine : du domaine limité au domaine ouvert, Mémoire d'Habilitation à Diriger des Recherches, Université Paris-Sud, Orsay.

Rosset, S., Tribout, D. & Lamel, L. (2007), Multi-level Information and Automatic dialog Act Detection in Human-Human Spoken Dialogs, *Speech Communication* 50(1), pp. 1–13.

ROSSI, M. (1999), *L'intonation, le système du français : description et modélisation*, Ophrys, Paris.

ROSSIGNOL, S., PIETQUIN, O. & IANOTTO, M. (2010), Simulation of the Grounding Process in Spoken Dialog Systems with Bayesian Networks, In: *Proceedings of the 2nd International Workshop on Spoken Dialogue Systems Technology*, Gotemba, Japan, pp. 110–121.

ROUILLARD, J. (2004), *VoiceXML. Le langage d'accès à Internet par téléphone*, Vuibert, Paris.

ROULET, E., AUCHLIN, A., MOESCHLER, J., RUBATTEL, C. & SCHELLING, M. (1985), *L'articulation du discours en français contemporain*, Lang, Berne.

SABAH, G. (1989), *L'intelligence artificielle et le langage. Tome 2 : processus de compréhension*, Hermès, Paris.

SABAH, G. (1997), The "Sketchboard": A Dynamic Interpretative Memory and its Use for Spoken Language Understanding, In: *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, pp. 617–620.

SABAH, G., VIVIER, J., VILNAT, A., PIERREL, J.-M., ROMARY, L. & NICOLLE, A. (1997), *Machine, langage et dialogue*, L'Harmattan, Paris.

SACKS, H., SCHEGLOFF, E. A. & JEFFERSON, G. (1974), A Simplest Systematics for the Organization of Turn-Taking for Conversation, *Language* 50(4), pp. 696–735.

SCHNEDECKER, C. (1997), *Nom propre et chaînes de référence*, Klincksieck, Paris.

SCHNEDECKER, C. (2011), La notion de saillance : problèmes définitoires et avatars, In : INKOVA, O. (Ed.), *Saillance*, Presses Universitaires de Franche-Comté, Besançon, pp. 23–43.

SEARLE, J. (1969), *Speech Acts*, Cambridge University Press, Cambridge.

SEARLE, J. & VANDERVEKEN, D. (1985), *Foundations of Illocutionary Logic*, Cambridge University Press, Cambridge.

SENEFF, S. (1995), TINA: A Natural Language System for Spoken Language Application, *Computational Linguistics* 18(1), pp. 62–86.

SINGH, S. P., LITMAN, D. J., KEARNS, M. & WALKER, M. A. (2002), Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System, *Journal of Artificial Intelligence Research* 16, pp. 105–133.

SOWA, J. (1984), *Conceptual Structures. Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA.

SPERBER, D. & WILSON, D. (1995), *Relevance. Communication and Cognition* (second edition), Blackwell, Oxford UK & Cambridge USA.

STALNAKER, R. (2002), Common Ground, *Linguistics and Philosophy* 25, pp. 701–721.

STOCK, O. & ZANCANARO, M. (Eds., 2005), *Multimodal Intelligent Information Presentation*, Springer.

STONE, M. & LASCARIDES, A. (2010), Coherence and Rationality in Grounding, In: *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, Poznań, pp. 51–58.

TELLIER, I. & STEEDMAN, M. (Eds., 2009), Apprentissage automatique pour le TAL, *Traitement Automatique des Langues* 50(3), ATALA.

Theune, M. (2002), Contrast in Concept-to-Speech Generation, *Computer Speech and Language* 16, pp. 491–531.

Traum, D. R. (2000), 20 Questions on Dialog Act Taxonomies, *Journal of Semantics* 17(1), pp. 7–30.

Traum, D. R. & Hinkelman, E. A. (1992), Conversation Acts in Task-Oriented Spoken Dialogue, *Computational Intelligence* 8(3), pp. 575–599.

Traum, D. R. & Larsson, S. (2003), The Information State Approach to Dialogue Management, In: van Kuppevelt, J. & Smith, R. (Eds.), *Current and New Directions in Discourse and Dialogue*, Kluwer, Dordrecht, pp. 325–354.

van Deemter, K. & Kibble, R. (2000), On Coreferring: Coreference Annotation in MUC and Related Schemes, *Computational Linguistics* 26(4), pp. 615–623.

van Deemter, K. & Kibble, R. (Eds., 2002), *Information Sharing. Reference and Presupposition in Language Generation and Interpretation*, CSLI Publications, Stanford.

van Schooten, B. W., op den Akker, R., Rosset, S., Galibert, O., Max, A. & Illouz, G. (2007), Follow-up Question Handling in the IMIX and RITEL Systems: A Comparative Study, *Natural Language Engineering* 1(1), pp. 1–23.

Vilnat, A. (2005), Dialogue et analyse de phrases, Mémoire d'Habilitation à Diriger des Recherches, Université Paris-Sud, Orsay.

Vivier, J. (Ed., 2001), Psycholinguistique et intelligence artificielle, *Langages* 144, Larousse, Paris.

Vuurpijl, L. G., Ten Bosch, L., Rossignol, S., Neumann, A., Pfleger, N. & Engel, R. (2004), Evaluation of Multimodal Dialogue Systems, In: *Proceedings of the LREC Workshop on Multimodal Corpora and Evaluation*, Lisboa.

Walker, M. A. (2005), Can We Talk? Methods for Evaluation and Training of Spoken Dialogue Systems, *Journal of Language Resources and Evaluation* 39(1), pp. 65–75.

Walker, M. A., Passonneau, R. & Boland, J. E. (2001), Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems, In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, pp. 515–522.

Walker, M. A., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M. & Vasireddy, G. (2004), Generation and Evaluation of User Tailored Responses in Multimodal Dialogue, *Cognitive Science* 28(5), pp. 811–840.

Ward, N. & Tsukahara, W. (2003), A Study in Responsiveness in Spoken Dialog, *International Journal of Human-Computer Studies* 59(6), pp. 959–981.

Warren, M. (2006), *Features of Naturalness in Conversation*, John Benjamins, Amsterdam & Philadelphia.

Weizenbaum, J. (1966), ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM* 9(1), pp 36–45.

Winograd, T. (1972), *Understanding Natural Language*, Academic Press, San Diego.

Wright, P. (1990), Using Constraints and Reference in Task-oriented Dialogue, *Journal of Semantics* 7, pp. 65–79.

Wright-Hastie, H., Poesio, M. & Isard, S. (2002), Automatically Predicting Dialogue Structure using Prosodic Features, *Speech Communication* 36, pp. 63–79.