Robustesse dans les systèmes de dialogue finalisé : modélisation et évaluation du processus d'ancrage pour la gestion de l'incompréhension

Alexandre A. J. Denis

Nancy-Université
Université Henri Poincaré

UFR STMIA – École doctorale IAEM Lorraine
Département de formation doctorale en informatique

Robustesse dans les systèmes de dialogue finalisé

Modélisation et évaluation du processus d'ancrage pour la gestion de l'incompréhension

THÈSE

présentée et soutenue publiquement le 24 octobre 2008

pour l'obtention du

Doctorat de l'université Henri Poincaré, Nancy 1
(spécialité Informatique Linguistique)

par

Alexandre Denis

Composition du jury

Président	Jean-Marie Pierrel	Professeur, Université Henri Poincaré, Nancy 1
Rapporteurs	Harry Bunt	Professeur, Université de Tilburg, Pays-Bas
	Jean Caelen	Directeur de recherche, CNRS, Grenoble
Examinateurs	Frédéric Béchet	Maître de conférences, Université d'Avignon
	Matthieu Quignard	Chargé de recherche, CNRS, Nancy
	Laurent Romary	Directeur de recherche, INRIA, Rocquencourt

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Remerciements

Mes profonds remerciements vont tout d'abord à Laurent Romary pour l'opportunité qui m'a été donnée d'effectuer cette thèse sous sa direction. Ensuite, je remercie chaleureusement les deux rapporteurs, Jean Caelen et Harry Bunt pour la finesse de leur analyse et ainsi que pour les précisions qui ont été apportées à cette thèse par leur concours. Je tiens également à remercier Jean-Marie Pierrel pour la présidence du jury ainsi que Frédéric Béchet pour avoir contribué à mettre en perspective certaines affirmations du document. Je souhaite exprimer en outre ma gratitude à Matthieu Quignard qui a suivi et encadré cette thèse. Il est parvenu à guider mes travaux tout en me laissant une considérable liberté d'action. Sans lui, rien n'aurait été possible. Mes parents, mes amis, mon chat et l'équipe toute entière, Langue et Dialogue puis TALARIS, doivent également être remerciés. Ils m'ont, par leur présence et leur affection, soutenu tout au long de ces quatre années de recherche. Je tiens enfin à remercier les membres du groupe sur le dialogue, Patrick Blackburn, Matthieu Quignard, Daniel Coulon, Bertrand Gaiffe, Luciana Benotti, et Guillaume Pitel pour m'avoir soutenu dans la construction du modèle présenté dans cette thèse. Cette collaboration a permis d'établir les fondations du modèle et sans la synergie de ce groupe, cette thèse n'aurait pas vu le jour en l'état. Je leur adresse à tous mes sincères remerciements.

Sommaire

La langue naturelle est sans doute le meilleur moyen que nous avons de communiquer. Son efficacité, sa facilité d'emploi et son expressivité en font un moyen privilégié pour accomplir nos buts. Mais nous n'avons en général pas idée de la complexité considérable nécessaire à son usage. Dès le berceau, nous sommes plongés dans un environnement qui produit de la langue naturelle, et sans le savoir, notre cerveau s'adapte à ces sons produits par nos pairs. Lors de l'enfance, nous intégrons les mécanismes de la communication au fur et à mesure que s'étend notre capacité à comprendre le monde. Il peut arriver que nous butions sur un mot, une construction grammaticale, ou un sens inattendu, mais nous parvenons facilement à intégrer ces nouveautés dans notre bagage linguistique. Communiquer nous est ensuite tellement aisé que la complexité de la langue disparaît totalement pour ne laisser place qu'à son emploi.

Quoi de plus naturel alors que de penser à utiliser la langue pour dialoguer avec une machine en lieu et place d'une souris et d'un clavier puisqu'il nous est si facile de communiquer ? Les systèmes de dialogue homme-machine tendent à réaliser cet objectif. Ce sont des interfaces logicielles dont le but est d'apporter un service à un utilisateur grâce à l'emploi de la langue naturelle. Cependant, la conception d'un modèle de la communication et son implémentation dans un système se heurtent fatalement à la complexité de la langue et de l'interaction verbale ; or nous ne disposons pas à l'heure actuelle de modèle complet de la communication et de la compréhension. En conséquence, les systèmes de dialogue que nous parvenons à construire aujourd'hui font bien pâle figure face à un être humain. Si on les comparait, on pourrait dire qu'un système de dialogue est en général à moitié sourd, complètement aveugle, quasiment dépourvu de facultés de raisonnement, ignorant le monde, et ne « maîtrisant » qu'une infime fraction de la langue. On pourrait bien sûr forcer la compassion de l'utilisateur, et le contraindre par exemple à n'employer que des énoncés moins complexes. Mais ce faisant, on irait à l'encontre de son utilisation quotidienne de la langue. L'objectif est au contraire de permettre une utilisation libre de la langue telle que la nature artificielle du système ne soit pas un obstacle, ni même être perceptible. Bien sûr il faut modérer cette ambition. Le rêve d'une intelligence artificielle parlante rivalisant avec l'être humain est confrontée à la complexité de la langue et de l'interaction. On peut faire néanmoins l'hypothèse que la complexité mise à l'oeuvre dans la réalisation d'un service est moindre que celle nécessaire à une conversation à bâtons rompus, et qu'alors la réalisation de systèmes de dialogue performants devient envisageable.

Cependant, même en faisant cette hypothèse, la capacité de dialoguer reste difficile à mettre en œuvre dans un système. Celle-ci nécessite en premier lieu que le système interprète l'énoncé produit par l'utilisateur et qu'il se pose des questions du type : qu'a dit l'utilisateur ? et que voulait-il dire ? Alors que les êtres humains n'ont en général aucun mal à y répondre, le système est, lui, confronté à une multitude de problèmes. Ces problèmes peuvent aboutir à des impasses dans laquelle le système ne parvient pas du tout à donner de réponse à ces questions, ou alors il ne peut obtenir que des réponses incomplètes, ou encore croire qu'il obtient les bonnes réponses alors que ce n'est pas le cas. La robustesse d'un système est la capacité à

répondre à ces questions et à faire face aux difficultés qu'elles soulèvent.

La première approche au problème de la robustesse est de s'attaquer frontalement à l'amélioration de la modélisation de la compréhension de la langue naturelle en cherchant à *éviter* les problèmes d'interprétation. Les études linguistiques par exemple cherchent à comprendre comment la langue fonctionne en modélisant des phénomènes particuliers. L'amélioration de ces modélisations, c'est-à-dire l'extension des phénomènes linguistiques pris en considération, conduit à rendre le système plus capable de donner les bonnes réponses aux questions d'interprétation, ou en tout cas de le faire plus fréquemment. Mais, même si on parvient à reproduire la compréhension telle qu'elle est mise en oeuvre chez les êtres humains, un système sera toujours susceptible de rencontrer des difficultés dans l'interprétation d'un énoncé. Il suffit de constater que l'emploi de la langue elle-même n'est pas aussi aisé que nous l'avons suggéré.

La seconde approche consiste à remarquer que nous, humains, sommes également confrontés à des difficultés d'interprétation. Nous ne pouvons pas toujours répondre de manière univoque aux questions : qu'a dit mon partenaire et que voulait-il dire ? A l'instar des systèmes de dialogue, nous sommes confrontés à des impasses, des ambiguïtés ou simplement aux réponses erronées que nous pouvons apporter à ces questions. Après tout, les individus sont différents, ils peuvent percevoir de manière différente le monde, et peuvent alors comprendre différemment les énoncés. L'idée que nous tiendrons pour centrale dans cette thèse, est que les stratégies développées par les êtres humains pour faire face aux problèmes d'interprétation peuvent servir pour améliorer la robustesse des systèmes de dialogue. Et sans nécessairement aller jusqu'à dire que les problèmes de compréhension sont une bénédiction pour le dialogue homme-machine, par exemple comme moteur de l'apprentissage, nous devrons considérer dès le départ que ces problèmes ne sont pas évitables, ou en tout cas qu'ils constituent un aspect tellement important de la communication qu'ils ne doivent pas être mis de côté.

En y regardant de plus près, nous pouvons constater que nous cherchons malgré tout à éviter ces situations problématiques. Lorsqu'un locuteur produit un énoncé, il le fait avec l'hypothèse de sa compréhensibilité. Il serait insensé qu'il ne fasse pas cette hypothèse dans la mesure où la compréhension de son énoncé est essentielle pour que son but sous-jacent puisse être réalisé. Cette hypothèse entraîne qu'un locuteur doive prendre en compte son interlocuteur dans la communication, en préférant s'appuyer sur ce qu'il croit être commun entre eux deux s'il veut minimiser les problèmes d'interprétation. A son tour, l'interlocuteur lorsqu'il cherche à interpréter un énoncé, préférera également s'appuyer sur ce qu'il croit être commun. Si un locuteur peut s'appuyer sans équivoque sur ce qui est commun entre lui et son partenaire pour produire son énoncé, et que son partenaire procède de la même façon, alors on peut supposer qu'il n'y aura pas de problème de compréhension. La problématique de la robustesse est alors très proche de la question : comment estimer le terrain commun entre deux participants du dialogue ?

Cette thèse cherche à résoudre le problème de la robustesse sous l'angle du terrain commun. Comment mettre en oeuvre un processus de gestion de terrain commun ?

Quelles en sont ses principales caractéristiques ? Comment représenter les problèmes d'interprétation ? Comment les manifester ? Comment les résoudre ? On n'explorera pas ces questions à travers tous les moyens de s'appuyer sur le terrain commun pour produire ou interpréter un énoncé. On se focalisera en revanche sur la résolution de la référence, un bon exemple d'activité collaborative nécessitant un terrain commun. D'autre part, en considérant que nous pouvons définir un processus de gestion du terrain commun, dans quelle mesure ce processus augmente-t-il la robustesse d'un système ? Est-ce qu'un système disposant de ce processus est *moins à même* de faire des erreurs qu'un système n'en disposant pas ? Nous proposons alors d'évaluer le gain d'une gestion dialogique du terrain commun en comparant les performances d'interprétation d'un système sans ce processus avec celles d'un système en disposant.

La première partie de cette thèse expose la problématique de la robustesse dans les systèmes de dialogue sous l'angle du terrain commun. Nous montrerons ce que sont les systèmes de dialogue et détaillerons les problèmes d'interprétation auxquels ils peuvent faire face. Nous proposerons deux manières de résoudre ces problèmes, d'abord dans une perspective de détection-correction puis dans la perspective du terrain commun. Dans une seconde partie nous présenterons le travail d'évaluation réalisé lors de la campagne MEDIA. Ce travail nous permettra d'estimer le terrain commun pré-existant entre un système donné et un utilisateur humain. A l'issue de cette évaluation nous aurons dégagé les principaux problèmes auxquels peut faire face le système. Enfin, dans une troisième partie, nous présenterons un processus de gestion du terrain commun dont nous procéderons à l'évaluation afin de déterminer s'il permet effectivement d'améliorer l'intercompréhension.

Gestion du terrain commun pour la robustesse

1 Communication et compréhension dans les systèmes de dialogue

1.1 Qu'est-ce qu'un système de dialogue ?

Les systèmes de dialogue sont des interfaces homme-machine dont le but est d'offrir des services réalisables par le dialogue en langue naturelle en mettant en correspondance un utilisateur humain et un service informatique. Plus généralement, les systèmes de dialogue peuvent s'appuyer sur un ensemble de modalités (parole, écrit, visuel, geste 2D, geste 3D, geste haptique, etc.) afin d'apporter le service à l'utilisateur. On considérera qu'un système de dialogue est un système d'interface complexe, qui par le biais de différents canaux de communication dont la langue naturelle, permet la réalisation du service.

Contrairement à la conversation quotidienne dans laquelle les participants n'ont pas nécessairement de but prédéfini, le dialogue dans ce contexte vise à la réalisation du service. Dès lors, les participants ont un rôle bien défini relativement à ce but : l'utilisateur demande un service et le système le lui apporte. Allen et al. (2001) formulent une hypothèse centrale dans le paradigme des systèmes de dialogue. Etant donné que le dialogue vise à réaliser le service et qu'il a donc une finalité, on peut considérer que sa mise en œuvre informatique est plus facile que celle de la conversation en général :

> *The Practical Dialogue Hypothesis* : The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence. (Allen, Byron, Dzikovska, Ferguson, and Galescu 2001)

Le contexte du dialogue finalisé est en fait une hypothèse de travail qui permet de formuler d'autres hypothèses. En particulier, si le dialogue a une finalité, alors les énoncés qui le composent tendent plus ou moins directement vers cette finalité. Comprendre les énoncés est alors plus facile dans la mesure où le contexte de leur emploi est mieux connu, et concrètement les algorithmes de compréhension et la manière de conduire un dialogue peuvent faire l'objet de simplifications plus ou moins importantes.

Quels types de services un système de dialogue est-il capable d'offrir ? Théoriquement, un système de dialogue peut apporter n'importe quel service informatique dans n'importe quel domaine d'application. Il peut s'agir de tâches purement infor-

matiques : recherche d'information, applications graphiques, jeux vidéos, etc. Mais le service peut avoir également des répercussions dans le monde physique : transactions commerciales, contrôle de machines physiques, domotique, etc. Il n'y a pas de limite *théorique* au service que peut interfacer un système de dialogue. Il y a cependant des limites *pratiques*, dûes à la complexité des tâches. Intuitivement, il semble plus simple de réaliser une tâche qui ne demande qu'un nombre très limité d'actions bien définies (par exemple rechercher une personne dans un annuaire), qu'une tâche dont le domaine d'application est vaste et dont les frontières sont floues (par exemple effectuer une réservation de n'importe quoi).

Allen et al. (2001) fournissent quelques exemples de tâches plus ou moins complexes en les reliant aux capacités nécessaires à leur gestion :

- mettre en relation des personnes, où le système peut se contenter de poser des questions et l'utilisateur de répondre à ces questions, autrement dit où le dialogue est entièrement dirigé par le système.
- connaître les heures de départ et d'arrivée d'un train, où l'utilisateur peut poser des questions, et le système requérir des clarifications, c'est-à-dire que les énoncés sont à l'initiative du système ou de l'utilisateur.
- réserver des voyages, où le thème du dialogue peut varier, d'abord réserver une chambre puis un train, ou l'inverse sans contrainte d'ordre.
- effectuer le *design* pour une cuisine, où le thème doit être défini au cours du dialogue.
- gérer un sauvetage lors d'une catastrophe, où le monde évolue, nécessitant alors l'adaptation du système à d'autres événements extérieurs que ceux générés par l'utilisateur

Face à la complexité variable des tâches, il semble difficile de construire un système de dialogue générique qui serait capable d'apporter n'importe quel type de service. Cependant, Allen formule l'hypothèse qu'il est malgré tout possible d'identifier des récurrences dans la réalisation des tâches et qu'on peut alors définir la complexité du dialogue finalisé de manière indépendante de la tâche :

> *The Domain-Independence Hypothesis* : Within the genre of practical dialogue, the bulk of the complexity in the language interpretation and dialogue management is independent of the task being performed. (Allen et al. 2001)

Cette hypothèse correspond à la direction de recherche sous-jacente au développement de systèmes de dialogue. L'objectif est de développer un système totalement indépendant de la tâche, de telle sorte qu'il suffise de lui ajouter des ressources, sans retoucher aux algorithmes pour qu'il puisse s'adapter à de nouvelles tâches. La problématique de la généricité n'est pas au cœur de cette thèse, mais correspond à la toile de fond omniprésente qui doit guider tout développement d'un système de dialogue finalisé.

Toutefois, l'hypothèse du dialogue finalisé doit être considérée avec précaution. Bien que la compétence nécessaire pour conduire un dialogue finalisé soit moindre que la compétence nécessaire pour conduire tout type de dialogue, elle n'en reste pas moins très importante. Ce n'est pas parce que le but du dialogue est bien défini

que les utilisateurs emploieront des énoncés moins complexes ou plus fréquents pour l'atteindre. Dans l'exemple de la figure 1.1 tiré du corpus MEDIA (voir chapitre 5), l'énoncé « que du bonheur » qui manifeste une acceptation n'est présent qu'une seule fois dans les 15000 énoncés que comporte le corpus.

compère : merci de patienter je vérifie ces informations
compère : l' hôtel de Lutèce possède une piscine et un jacuzzi
client : que du bonheur

FIG. 1.1 – Exemple d'énoncé peu fréquent (dialogue 331)

En conséquence, la complexité de la langue et l'interaction ainsi que son caractère imprévisible, toujours présents dans le dialogue finalisé suggèrent qu'un système de dialogue doit être développé en considérant dès le départ que sa compétence à réaliser le service sera imparfaite et qu'il pourra être confronté à des problèmes d'interprétation.

1.2 Compréhension et communication

Historiquement les premières approches sémantiques de la langue identifiaient le sens d'une phrase avec ses conditions de vérité. Par exemple la phrase « un chat dort » n'est vraie que dans les situations où il existe un chat qui dort. Le sens d'une phrase correspondrait alors à l'ensemble des situations qu'il décrit, que l'on peut représenter grâce à une formule logique à l'aide de fonctions propositionnelles (Russell 1905). La question de la vérité est alors essentielle dans ces approches (Tarski 1944). Cependant, Austin (1962) exhibe l'existence des performatifs, des énoncés dont le but n'est pas de décrire le monde mais d'agir dans et sur ce dernier. Par exemple « je vous déclare mari et femme », ne peut pas être considéré comme vrai ou faux puisque l'énoncé constitue une action qui a pour but d'établir l'état de mariage. Les énoncés sont avant tout des actions qui peuvent dès lors être satisfaites ou réussies et qui nécessitent certaines conditions sur l'état du monde, la sincérité du locuteur ou son intention d'effectuer l'action (voir Searle 1969). En dénonçant l'illusion descriptive, Austin initie le paradigme pragmatique dans lequel se situent toutes les approches modernes de modélisation du dialogue. En parlant, nous effectuons des *actes de langage* qui recouvrent plusieurs dimensions : d'abord la dimension *locutoire* qui est l'action physique de prononcer des mots, puis la dimension *illocutoire* qui correspond à l'acte que l'on fait *en* disant quelques chose (une demande, une affirmation, un ordre), et enfin la dimension *perlocutoire* qui concerne les effets que l'on produit sur l'interlocuteur ou sur le monde. Le sens d'un énoncé ne peut dès lors pas se limiter à son contenu sous forme de proposition logique (le contenu propositionnel), mais comprendre un énoncé inclut la compréhension de ces différentes dimensions, y compris celle de l'intention de son locuteur.

Bien qu'insuffisant, un des modèles de la communication le plus parlant est le modèle du code (Shannon and Weaver 1949) : lorsqu'un locuteur veut communiquer

quelque chose à son partenaire, il encode cette idée à l'aide de la langue, son partenaire décode l'énoncé et peut alors retrouver ce que voulait transmettre le locuteur. Le code utilisé correspondrait à toutes les règles de production et d'interprétation des énoncés, par exemple pour transmettre l'idée qu'il existe un chat qui dort, il faudrait prononcer les mots « un », puis « chat » puis « dort », où la combinaison de ces trois symboles désigne la situation désirée grâce au code. La nature de ce code serait alors conventionnelle, puisque les règles ne pourraient pas être établies de manière autonome mais à l'aide de conventions et arbitraire, c'est-à-dire sans relation entre les signes utilisés et les réalités qu'ils désignent.

Grice (1957) montre cependant qu'on ne peut définir ce code que vis à vis de l'intention initiale du locuteur lorsqu'on cherche à définir la signification. Dans le dialogue 1.2 ci-dessous, lorsque B dit « c'est cher », il ne veut probablement pas simplement communiquer le fait que, selon lui, la chambre à l'hôtel Ibis est chère. Pour pouvoir décoder l'acte illocutoire de refus, il faut se poser la question des relations entre le locuteur de l'énoncé, les effets désirés, et l'interlocuteur. Pour que ces effets puissent opérer, en l'occurrence pour que A comprenne que B refuse la chambre, il est nécessaire au préalable que A reconnaisse que B ait l'intention de lui communiquer quelque chose. Dans le cas contraire, A pourrait attribuer du sens à des actions de B qui n'en avaient pas pour ce dernier.

A : je vous propose l'hôtel Ibis, la chambre est à 100 euros
B : c'est cher

Fig. 1.2 – Exemple nécessitant de considérer l'intention

Grice propose alors de définir le fait qu'un locuteur A signifie quelque chose à B par x si et seulement si A a l'intention que x provoque un certain effet sur B grâce à la reconnaissance de cette intention (Grice 1957). L'intention initiale est en conséquence manifestée grâce à l'*intention communicative* du locuteur. Dans notre exemple, A signifie un refus en disant « c'est cher » car il a l'intention que cet énoncé provoque par exemple, la proposition d'un nouvel hôtel par B grâce à la reconnaissance de l'intention de communiquer un refus[1].

Pour décrire les propositions, qui sans être exprimées dans l'énoncé, font partie de l'intention du locuteur, Grice emploie le terme d'*implicature*[2]. Pour reconnaître une implicature, Grice (1975) suppose que les individus entretiennent une certaine coopération lorsqu'ils dialoguent, formulée par un *principe coopératif* :

[1]On pourrait objecter que *conventionnellement*, « c'est cher » manifeste un refus. Comment considérer alors l'exemple suivant : un individu, les bras chargés de grosses caisses fait face à une porte fermée et dit à un autre individu « la porte », signifiant alors son intention qu'il lui ouvre la porte. On ne peut pas prétendre que « la porte » exprime conventionnellement une requête d'ouverture de porte.

[2]Grice distingue les implicatures *conventionnelles*, déclenchées par l'utilisation d'éléments linguistiques à caractère conventionnel, les implicatures *conversationnelles particulières* soulevées grâce au contexte d'énonciation, et les implicatures *conversationnelles généralisées* soulevées dans tous les contextes d'énonciation mais non déclenchées par un élément linguistique particulier.

> *Cooperative principle* : Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged (Grice 1975)

Le principe coopératif impose des contraintes sur la forme et le fond de l'énoncé qu'un locuteur peut produire relativement au *but courant* de l'échange. Ces contraintes sont représentées par des maximes, issues d'une analyse empirique de la collaboration. Elles sont au nombre de quatre : maxime de *quantité* (effectuez une contribution aussi informative que nécessaire mais pas plus), maxime de *qualité* (ne dites pas ce que vous croyez faux, ni ce pour quoi vous ne disposez pas suffisamment de preuves), maxime de *relation* (soyez pertinent), et maxime de *manière* (évitez les expressions obscures, l'ambiguïté, la prolixité non nécessaire, et soyez ordonné).

Le respect de ces maximes conversationnelles par des participants rationnels et coopératifs suggère un moyen de calculer les implicatures : lorsque les participants semblent violer une maxime, alors quelque chose qui n'est pas exprimé est implicité. Calculer une implicature (conversationnelle) revient alors à calculer ce qui doit être supposé afin de préserver l'hypothèse de coopération. Un locuteur qui dit p implique (conversationnellement) q si :

1. on suppose qu'il respecte les maximes conversationnelles, ou au moins le principe coopératif
2. on suppose qu'il est conscient qu'en disant p, q est requis pour être cohérent avec 1.
3. le locuteur pense, et s'attend à ce que l'interlocuteur pense que le locuteur pense, que l'interlocuteur est capable d'inférer que 2. est requis

Dans notre exemple 1.2, lorsque A effectue sa proposition de chambre, il s'attend à ce que B accepte ou refuse cette proposition. Mais « c'est cher » ne semble pas être une réaction pertinente à cette proposition, elle n'exprime *explicitement* ni une acceptation ni un refus. Dès lors, si A suppose que B est coopératif, c'est que B doit impliciter quelque chose. Pour conserver l'hypothèse de coopérativité, A doit supposer que B accepte ou refuse la chambre. L'interprétation du refus peut être alors motivée par le fait qu'en général on cherche à minimiser le prix d'une réservation de chambre et cette interprétation conduit l'énoncé « c'est cher » à jouer l'effet désiré par B.

On peut faire deux remarques. La première est que le modèle de Grice présuppose la bonne compréhension des participants. En effet, comme le note (Johnson 2007), la mauvaise compréhension d'un énoncé peut se traduire en violation du principe coopératif[3].

Dans l'exemple 1.3, B_2 semble être non-pertinent pour A et s'il suppose B coopératif, alors B doit impliciter quelque chose pour conserver l'hypothèse de coopérativité. Or B ne semble rien impliciter. En fait B n'est pas coopératif mais de manière *non-intentionnelle*. En conséquence, une violation des maximes peut être

[3] c'est d'ailleurs un moyen employé par un certain nombre d'approches pour détecter la mauvaise compréhension, par exemple dans (Danieli 1996)

A_1 : quel jour sommes-nous ?
B_2 : jeudi
A_3 : non non, je voulais dire, quel jour du mois ?

Fig. 1.3 – Exemple de mauvaise compréhension

causée à la fois par la volonté délibérée de transmettre une implicature ou par une violation non-intentionnelle dûe à une mauvaise compréhension. Ceci suggère qu'il est nécessaire pour un participant de supposer une bonne compréhension avant de pouvoir rechercher des implicatures éventuelles. Autrement dit, le modèle de Grice présuppose la bonne compréhension de ce qui est dit avant de considérer ce qui est implicité.

La seconde remarque concerne la définition de l'ensemble des énoncés pertinents possibles. Dans notre exemple 1.3, le but courant est défini par A lorsqu'il soulève une question en A_1 : A a l'intention de connaître le jour courant de la situation d'énonciation. On peut cerner alors assez facilement l'ensemble des énoncés pertinents, ce sont ceux qui fournissent une réponse satisfaisante à A[4]. Par exemple les approches conventionnelles du dialogue décrivent des structures relationnelles sous la forme de paire adjacente (Schegloff and Sacks 1973; Sacks, Schegloff, and Jefferson 1974), où un énoncé est préférentiellement suivi d'une classe d'énoncés : un bonjour répond à un bonjour, une réponse répond à une question ; ou sous la forme de triplet (initiative, réactive, évaluative) comme le modèle genevois (Roulet et al. 1985; Moeschler 1989). Cependant, définir ce qui est une bonne réponse à la proposition de chambre dans l'exemple 1.2 est plus difficile : B n'est pas limité à l'acceptation ou au refus de la proposition. Il peut retarder son acceptation en posant un nouveau but sous la forme d'une question, revenir à un sujet antérieur, abandonner, etc. La décision de la pertinence revient aux deux participants, chaque participant adaptant ses propres buts en fonction des énoncés d'autrui, pour juger de ce qui est satisfaisant. Clark (1997) cite un exemple (figure 1.4 ci-dessous) dans lequel il abandonne son but (avoir un thé « English breakfast »), en considérant, malgré la mauvaise compréhension de la serveuse (c'est de l'« Earl Grey »), que l'interprétation de celle-ci fera tout aussi bien l'affaire.

Waitress : And what would you like to drink ?
Clark : Hot tea, please. Uh, English breakfast.
Waitress : That was Earl Grey ?
Clark : Right.

Fig. 1.4 – Exemple d'abandon de but suite à une mauvaise compréhension

La définition de l'ensemble de toutes les contraintes de pertinence possibles dans la communication humaine dépasse l'objet de cette thèse. Nous n'explorerons qu'un très petit nombre d'intentions possibles, de manières de les exprimer et de les sa-

[4]comme ceux qui choisissent un des membres de la partition opérée par la question dans (Groenendijk 1999)

tisfaire (relativement à la convergence de la compréhension des expressions référentielles). En revanche nous considérerons dès le départ que l'interprétation de l'énoncé est effectuée grâce à la participation des deux individus. En cas de divergence détectée, les participants peuvent changer le but courant pour décider de faire converger l'interprétation. Cependant, avant de déterminer à quel moment, et comment ce but de convergence est construit ou satisfait, il nous faut revenir à la problématique de la compréhension dans les systèmes de dialogue. Comment modéliser informatiquement la communication, la compréhension et ses problèmes ?

1.2.1 Compréhension et communication dans les systèmes

Toutes les approches des systèmes de dialogue s'appuient sur trois grandes parties. On retrouve classiquement une phase d'*interprétation* qui consiste à produire une représentation de l'acte communicatif de l'utilisateur, une phase de *gestion du dialogue* qui consiste à insérer cet acte dans une représentation du dialogue et à déterminer les actions communicatives ou applicatives à produire et une phase de *génération* qui consiste à verbaliser l'action communicative du système. Ce schéma peut fait l'objet de raffinement. Par exemple le système TRIPS (Allen et al. 2000; Allen et al. 2001), implémenté sous la forme d'un système multi-agents, s'appuie sur ces trois agents principaux mais utilise de plus un agent gérant le contexte du discours, et un agent gérant le contexte référentiel, communiquant tous deux avec l'agent d'interprétation et celui de génération. Le système utilise également un troisième agent qui gère la tâche sous-jacente et qui communique avec les trois agents principaux.

1.2.1.1 Interprétation

L'interprétation est le processus qui consiste à construire une représentation de l'intention du locuteur aussi détaillée que nécessaire pour les buts courants. Cette interprétation est souvent réalisée soit de manière *ascendante*, en effectuant d'abord une analyse du signal sonore, puis en construisant une représentation lexicale de l'énoncé (au niveau des mots), syntaxique (au niveau des constructions), sémantique (au niveau des significations), et pragmatique (au niveau du contexte) jusqu'à atteindre l'intention du locuteur ; soit *descendante*, en partant d'une collection d'intentions possibles et en recherchant lesquelles sont satisfaites dans l'énoncé. On peut voir les approches descendantes comme une interprétation stricte de l'hypothèse de dialogue finalisé : puisque le dialogue a une finalité bien définie, autant se focaliser sur la tâche et extraire les éléments qui n'ont de sens que dans la tâche. En conséquence, les approches descendantes s'appuient en général sur des analyses linguistiques très limitées, voires inexistantes, par exemple en recherchant des patrons de mots directement reliés à la tâche. Elles ont pour avantages d'être simples à développer, et d'offrir de bonnes performances. Cependant, elles ont pour défaut d'être peu génériques : si le module d'interprétation est dédié à une tâche, il est nécessaire d'en développer un nouveau dès que l'on souhaite porter le système à une

autre tâche. Au contraire, les approches ascendantes cherchent à rendre plus générique l'analyse linguistique et à faire le moins de suppositions possibles sur la tâche. Une représentation sémantique/pragmatique est construite dans un premier temps, puis elle est associée dans un second temps à une intention possible. Ces approches sont beaucoup plus complexes à mettre en œuvre et offrent des performances moins bonnes, mais la généricité qu'elles apportent nous conduit à les préférer pour le développement des systèmes.

Cependant, qu'elles soient ascendantes ou descendantes, toutes les approches effectuent un certain nombre de simplifications du processus interprétatif. Il y a d'abord des simplifications relatives à la synchronisation des phases d'interprétation et d'action. On considère la plupart du temps que les phases d'interprétation et d'action des deux participants sont asynchrones. Lorsque l'utilisateur produit un énoncé, le système attend la fin du tour de parole pour en effectuer l'interprétation. On considère également que ces deux phases ne peuvent être concomitantes pour un même locuteur : lorsqu'il produit un énoncé, il n'interprète pas cet énoncé. On sait cependant que l'on interprète les énoncés de manière synchrone avec leur production et que l'on effectue l'interprétation des énoncés que l'on produit. De plus les analyses syntaxiques et sémantiques ne sont pas si disjointes comme peuvent le suggérer certaines études neurolinguistiques. Friederici distingue par exemple trois étapes : une analyse syntaxique, une intégration sémantique et une analyse syntaxique tardive (Friederici and Kotz 2003; Friederici and Weissenborn 2007). D'autres phénomènes mettent en question l'antériorité d'une analyse « littérale » d'un énoncé vis à vis d'une analyse intentionnelle. Clark (1997) cite l'*illusion Moïse* (Erickson and Mattson 1981) : à la question « Combien d'animaux de chaque espèce Moïse a-t-il emmené sur son arche ? », la majorité des sujets répondent « 2 », malgré l'erreur dans la question (Moïse au lieu de Noé). Cet exemple suggère que l'analyse littérale peut parfois être court-circuitée, mais les causes et les circonstances dans lesquelles les raccourcis peuvent être pris restent à déterminer. Nous n'entrerons pas davantage dans le détail des simplifications du processus interprétatif (voir en particulier Clark 1997). Nous nous restreignons dans la suite de cette thèse, classiquement, à une interprétation linéraire ascendante, temporellement disjointe de la production (nous considérerons toutefois l'interprétation du locuteur de ses propres énoncés *a posteriori*).

1.2.1.2 Gestion du dialogue

La première question à laquelle est confronté un système de dialogue est de déterminer ce que veut le locuteur lorsqu'il effectue un acte communicatif, autrement dit, *quel est son but* ? La seconde question, en supposant que le système adopte le but de l'utilisateur, est de savoir *comment l'atteindre* ? Si le système parvient à déterminer le but de l'utilisateur, et qu'il sait comment le satisfaire, alors il peut accomplir le service demandé par l'utilisateur.

La manière la plus simple de construire un système est d'utiliser des automates à états finis. Les états représentent différentes étapes dans la réalisation du service et les informations données dans les énoncés représentent des transitions entre les états. Par exemple, dans un service de réservation de chambre, le système se situe au départ

dans un état 0, et attend de connaître la ville du séjour avant de passer dans l'état 1, où il peut proposer un hôtel. Le seul but que l'utilisateur est autorisé d'avoir est de fournir des informations au système pour lui permettre d'avancer dans la tâche. De plus, si l'utilisateur apporte des informations qui ne sont pas nécessaires pour l'état courant, ces informations sont ignorées. Ce genre d'approche est très facile à mettre en œuvre, par exemple dans VoiceXML (Rouillard 2004) mais rencontre un grand nombre de difficultés. Le dialogue est entièrement contraint par la tâche, et le système conserve l'initiative. L'utilisateur ne peut en général pas poser de questions, et ne fait que répondre aux questions du système. De plus, il est difficile, voire impossible d'adapter un tel système à une nouvelle tâche sans développer de nouveaux automates. Il n'y a enfin aucune mémoire du dialogue qui permettrait d'adapter les réponses en fonction du chemin parcouru dans l'automate.

Pour répondre aux problèmes des automates à états finis, on peut s'appuyer sur une représentation de l'état courant d'information (*information state*, appelé anciennement *blackboard*). Au lieu de modéliser une collection d'états et de transitions, on ne représente que l'état courant comme un ensemble structuré d'informations et on représente les transitions comme des règles de modification de cet état d'information (Larsson and Traum 2000). Les énoncés sont des *dialogue moves* qui sont associés à des règles de mise à jour de l'état courant. Ce type de système permet d'éviter les difficultés des automates à états finis : l'ordre des transformations est plus flexible que dans un automate à état, on peut modéliser une initiative mixte, et on peut représenter une mémoire. Le principal inconvénient de cette approche, qui est également sa plus grande force, est qu'elle est neutre vis à vis de la représentation de l'état courant. Que doit-on mettre dans l'état d'information ?

Le paradigme le plus influent qui peut apporter une réponse à cette question est celui de la planification. Dans ce paradigme, on représente l'état mental de chaque participant sous forme de croyances, désirs ou intentions. Les actes de langage sont produits dans le but d'opérer des transformations sur les états mentaux des participants, conformément aux effets perlocutoires chez Austin. Dans Cohen and Perrault (1979), Allen and Perrault (1980), les actes de langage sont représentés dans le formalisme STRIPS (Fikes and Nilsson 1971) en termes de préconditions d'intention et de capacité inspirées de Searle (1969) et d'effets sur les croyances. Par exemple si l'utilisateur demande « A quelle heure part le train 12 ? », c'est qu'il croit d'abord que le système *peut* lui donner cet horaire et qu'il a l'intention de connaître cet horaire. L'effet attendu de cette requête est que le système croit que l'utilisateur désire connaître cet horaire. Si le système adopte le but de l'utilisateur, c'est-à-dire qu'il ait l'intention de lui faire savoir l'horaire, alors il peut agir et lui donner effectivement cet horaire. Le paradigme de planification consiste à supposer que les participants au dialogue ont des buts sous-jacents, que les différents énoncés qu'ils produisent sont des étapes pour atteindre ces buts et qu'ils construisent des plans, c'est-à-dire des séquences d'actions, pour les atteindre. La reconnaissance de plan est alors l'opération qui consiste à *inférer* le plan probable du locuteur, et la construction de plan est l'opération qui consiste à *produire* un plan destiné à un certain but. La reconnaissance de plan permet en particulier de pouvoir anticiper les buts sous-

jacents et d'être plus collaboratif. Par exemple si l'utilisateur demande « A quelle heure part le train 12 ? », le système peut estimer qu'en plus de son intention de connaître l'horaire, il a probablement l'intention de prendre le train 12. Il peut alors proposer une réponse du type « le train 12 part à 15h au quai numéro 3 », où « au quai numéro 3 » est une information non-nécessaire mais que le système suppose utile à l'utilisateur.

Si le paradigme de la planification représente bien les aspects intentionnels, il nécessite en revanche des hypothèses fortes de coopérativité. En particulier pour qu'un individu puisse répondre à une question, il est nécessaire qu'il adopte l'intention du locuteur de connaître la réponse. Traum et Allen (1994) notent que cela n'explique pas pourquoi l'interlocuteur peut répondre quelque chose même s'il ne sait pas ou n'adopte pas ce but. Ils proposent alors de considérer des *obligations* issues de conventions sociales. L'idée est que les participants font face à des obligations en plus de gérer leurs buts propres, qu'ils peuvent les satisfaire ou s'y soustraire. Mais surtout, les obligations permettent de considérer un comportement réactif et local, qui ne suppose pas de planification complexe. Les auteurs insistent sur le fait que ces obligations ne remplacent pas le modèle de planification mais seulement qu'on peut considérer la réaction du système de ce point de vue : « following the initiative of the other can be seen as an *obligation driven* process, while leading the conversation will be *goal driven* » (Traum and Allen 1994).

D'autres moyens existent pour gérer le dialogue. Par exemple, (Caelen 2003) s'inspire de la théorie des jeux dans laquelle chaque participant doit maximiser son gain. Il représente le comportement rationnel des participants vis à vis de leurs propres buts et des buts de leur partenaire. Chaque participant peut avoir différentes stratégies relativement à ses buts : *réactive* lorsqu'on délègue la gestion du dialogue à autrui (en abandonnant son propre but par exemple), *directive* lorsqu'on maintient l'initiative en imposant son propre but, *constructive* lorsqu'on initie un nouveau but en laissant en suspens le but courant, *coopérative* lorsque les participants cherchent à ajuster leurs buts l'un à l'autre, de *négociation* lorsque les participants cherchent un compromis tout en maintenant leurs buts. Le calcul d'une stratégie dépend de plusieurs paramètres : la complétude de l'acte de dialogue, les conditions de réussite des buts, les buts en attente, des buts de l'usager, etc. Par exemple, au début du dialogue, le système peut avoir une stratégie directive, puis adopter une stratégie réactive en se laissant guider par l'utilisateur si le dialogue dure un certain temps. Il peut également lui proposer d'explorer des buts proches, s'il perçoit que l'utilisateur ne sait pas trop ce qu'il veut (stratégie constructive).

1.2.1.3 Génération

Une fois que le système a interprété l'énoncé et a déterminé qu'une action communicative était nécessaire pour atteindre son but, il doit être capable de verbaliser le contenu propositionnel de cette action. A l'instar de l'interprétation, on peut concevoir deux approches pour effectuer cette verbalisation. L'approche directe, consiste à associer directement l'intention communicative avec une verbalisation, par exemple au moyen de patrons à trous du type : « je vous propose l'hôtel X, la chambre est à

Y euros ». De la même manière que pour l'approche interprétative descendante, ce type d'approche dérive d'une interprétation stricte de l'hypothèse de dialogue finalisé : puisqu'on se situe dans le contexte d'une tâche déterminée, autant produire les énoncés visant à l'accomplir de manière *ad hoc*. Il est relativement facile de mettre en oeuvre une telle génération mais, comme pour l'analyse descendante, elle est peu générique, nécessitant de produire de nouveaux patrons pour chaque nouvelle tâche. Elle est de plus difficile à maintenir : si par exemple on souhaite produire un pronom au lieu d'une expression définie, il est potentiellement nécessaire de réviser de nombreux patrons.

L'autre approche consiste à détailler les différents aspects du processus de production afin d'en faciliter le contrôle, la paramétrisation ou la maintenance. Reiter and Dale (1997) distinguent six aspects impliqués dans la génération d'un énoncé ou d'un texte : la *détermination du contenu* où l'on décide quoi dire (par exemple la proposition d'un hôtel, et le détail du prix de la chambre dans cet hôtel), la *planification du discours* où l'on décide comment agencer les différentes propositions entre elles, en particulier les relations rhétoriques (par exemple d'abord proposer l'hôtel, *puis* élaborer sur le prix de la chambre), l'*aggrégation des phrases* où l'on groupe les propositions en phrases (par exemple ne faire qu'une phrase « je vous propose l'hôtel X dont la chambre est à Y euros » ou deux phrases séparées), la *lexicalisation*, spécifique au domaine, où l'on détermine les mots destinés à verbaliser les concepts (le concept HÔTEL doit être verbalisé par « hôtel »), la *génération des expressions référentielles* où il est nécessaire de décider quel type de référence (défini, indéfini, démonstratif, pronominale, etc.) on utilise pour référer à une entité donnée (par exemple « la chambre dans *cet hôtel* est à Y euros ») et enfin la *réalisation linguistique* qui applique les règles de la grammaire pour produire un énoncé syntaxiquement, et morphologiquement correct (produire « la chambre est à Y euros »). L'idéal dans cette approche est de pouvoir construire des systèmes bi-directionnels, basés sur des ressources *réversibles*, qui peuvent être appliquées à l'interprétation ou à la génération, en particulier pour l'étape de réalisation linguistique (Appelt 1987; Shieber 1988; Neumann 1994). On peut également chercher à *rendre* réversible une grammaire existante, comme dans (Kow 2007; Gardent and Kow 2007). Nous n'aborderons toutefois pas outre mesure la génération pour nous focaliser sur l'interprétation et en particulier sur les problèmes qu'elle soulève.

1.3 Problèmes d'interprétation

Les erreurs d'interprétation jalonnent les dialogues homme-homme. Nous préférons toutefois parler de *divergence* ou de *problème* pour éviter la connotation négative que le terme *erreur* peut véhiculer. En effet, si la divergence d'interprétation est un phénomène à éviter dans la communication, elle reste le point d'appui de l'apprentissage (Luzzati 1995; Lehuen 1997b) et dans cette perspective on ne peut la considérer uniquement comme un phénomène négatif. La spécificité des problèmes d'interprétation à un modèle de la langue donné rend difficile la généralisation d'une caractérisation. Nous dégageons toutefois certaines caractéristiques des problèmes

d'interprétation que l'on peut retrouver dans n'importe quel modèle.

1.3.1 Catégories de problème

Hirst et al. (1994) définisent deux catégories de problèmes d'interprétation : la *non-compréhension* et la *mauvaise compréhension*. La non-compréhension se produit lorsqu'un individu ne parvient pas à obtenir une interprétation unique et complète de l'énoncé. Cette définition est étendue par Gabsdil (2003) en ajoutant la nécessité d'obtenir une interprétation probable, c'est-à-dire dans laquelle l'individu peut avoir une certaine confiance. La mauvaise compréhension se produit lorsqu'un individu obtient une interprétation unique, complète et probable, mais que celle-ci est différente de celle que le locuteur avait l'intention qu'il obtienne. Cette distinction est également employée par Caelen and Nguyen (2006) sous des dénominations différentes : l'*incompréhension* désigne la non-compréhension de Hirst, et le *malentendu* désigne la mauvaise compréhension. Les termes sont différents mais les réalités qu'ils décrivent sont exactement les mêmes. Nous emploierons toutefois la terminologie de Hirst, plus répandue dans la littérature anglo-saxonne et appelerons *incompréhensions* les situations d'interprétation problématique.

La non-compréhension diffère de la mauvaise compréhension en ce que l'individu qui interprète l'énoncé *détecte* la présence d'un problème dès l'interprétation de l'énoncé. La mauvaise compréhension peut ne pas être détectée immédiatement et perdurer sur plusieurs énoncés, résultant en un *quiproquo*. On peut également catégoriser les problèmes selon le moment et l'auteur de la détection : *first-turn repair*, où l'auteur de la détection est le locuteur de l'énoncé qui effectue une correction immédiatement (auto-corrections), *second-turn repair* où l'interlocuteur détecte le problème juste après l'énoncé (la non-compréhension), *third-turn repair* où la détection est effectuée par le locuteur initial en recevant la réponse de son interlocuteur (la mauvaise compréhension), et *fourth-turn repair* où l'interlocuteur détecte sa propre mauvaise compréhension de l'énoncé initial (voir Schegloff 1992; Schegloff 1997; McRoy and Hirst 1995).

Il faut toutefois distinguer les *divergences* d'interprétation et les *problèmes* d'interprétation, certaines divergences peuvent en effet ne pas soulever de problème. Par exemple, si je produis « mon ami habite près de Paris », j'impose à mon interlocuteur des exigences de compréhension variables en fonction de mon but. Ici, le but d'identification de mon ami est probablement prépondérant : toute divergence sur l'identité de cet ami sera très susceptible d'être insatisfaisante et de soulever un problème. Quant à la compréhension de « près de Paris », si mon but n'est que d'indiquer sans plus de précisions la localisation géographique de mon ami, je pourrais tolérer une divergence assez importante dans l'interprétation[5]. Comme dans Clark (Clark and Schaefer 1989; Clark 1996), la bonne ou la mauvaise compréhension dépend de la satisfaction des deux participants vis à vis de l'interprétation de l'un d'entre eux. Le locuteur A estime que son interlocuteur B a bien compris son énoncé si l'interpré-

[5]Nous emploierons cependant le terme de divergence en supposant implicitement qu'il s'agit de divergence problématique.

tation de B est suffisante pour les buts de A. Dans le cas contraire, il y a mauvaise compréhension de B. L'interlocuteur B estime avoir bien compris l'énoncé de A, si son interprétation est suffisante pour ses propres buts, et dans le cas contraire il y a non-compréhension de B.

1.3.2 Niveaux de problème

Ces deux catégories ne sont cependant pas suffisantes pour caractériser un problème d'interprétation. Le niveau d'interprétation problématique est évidemment essentiel : un problème de mot inconnu ne sera pas géré de la même manière qu'un problème d'intention inconnue. Comme le note Schegloff en parlant des corrections, tout est potentiellement source de problèmes :

> By "repair" we refer to efforts to deal with trouble in speaking, hearing, or understanding talk in interaction. "Trouble" includes such occurences as misarticulations, malapropisms, use of a "wrong" word, unavailability of a word when needed, failure to hear or to be heard, trouble on the part of the recipient in understanding, incorrect understanding by recipients, and various others. Because anything in talk can be a source of trouble, everything in conversation is, in principle "repairable". (Schegloff 1987)

Il existe de nombreuses classifications des différents niveaux d'interprétation. Nous présentons ici quelques hiérarchies qui ont toutes été proposées relativement à l'expression des problèmes d'interprétation.

Pour Clark (Clark and Schaefer 1989) les individus traversent différentes phases lors de l'interprétation. Il propose alors quatre niveaux :

- Etat 0 : B n'a pas remarqué que A a prononcé un énoncé u.
- Etat 1 : B a remarqué que A a prononcé u (mais n'est pas dans l'état 2)
- Etat 2 : B a correctement entendu u (mais n'est pas dans l'état 3).
- Etat 3 : B a compris ce que A voulait dire par u.

On peut comparer ces quatre niveaux aux différentes fonctions communicatives proposées par Allwood (1995) : contact, perception, compréhension et réaction. L'état 0 correspond au contact, les états 1 et 2 correspondent à la perception, et l'état 3 à la compréhension. La réaction elle-même à l'énoncé du locuteur ne fait pas partie de l'interprétation chez Clark. Nous incluons ici d'autres hiérarchies qui considèrent le niveau de l'action. Par exemple, Brennan and Hulteen (1995) suggèrent que les systèmes devraient pouvoir manifester l'existence de problèmes lors de toutes les étapes, y compris l'action (*acting*) et la présentation du résultat (*reporting*). Ils proposent alors huit états que traversent généralement les systèmes de dialogue :

- Etat 0 : *Non-attending* (n'écoute pas)
- Etat 1 : *Attending* (écoute mais n'a pas encore identifié les mots)
- Etat 2 : *Hearing* (a identifié les mots mais n'a pas encore analysé l'énoncé)
- Etat 3 : *Parsing* (a un énoncé bien formé mais ne l'a pas encore associé à une interprétation)

- Etat 4 : *Interpreting* (a une interprétation mais n'a pas encore de commande applicative)
- Etat 5 : *Intending* (a une commande applicative mais n'a pas encore encore agi)
- Etat 6 : *Acting* (a agi mais n'a pas encore présenté les résultats)
- Etat 7 : *Reporting* (présente les résultats)

L'état 3 chez Clark and Schaefer (1989), ou 4 chez Brennan and Hulteen (1995), la compréhension, peut être davantage spécifié. Par exemple Larsson (2003) suggère de distinguer le niveau sémantique qu'il associe à un contenu indépendant du contexte de discours, et le niveau pragmatique dépendant du discours :

- Niveau sémantique : contenu indépendant du discours (mais potentiellement du domaine d'application)[6]
- Niveau pragmatique : contenu dépendant du discours et du domaine, c'est-à-dire :
  - contenu référentiel, par exemple référents des pronoms, expressions temporelles
  - contenu proprement pragmatique : pertinence de l'énoncé vis à vis du contexte courant

La hiérarchie d'interprétation peut même être davantage détaillée. Par exemple Schlangen (2004) distingue quatre niveaux principaux, et détaille particulièrement le niveau de la compréhension (figure 1.1).

Niveau	**Description**
1	mise en place d'un contact
2	reconnaissance de la parole
3a	analyse :
3aa	reconnaissance de tous les mots
3ab	détermination d'une structure syntaxique
3ac	détermination d'une unique structure syntaxique
3b	résolution de la sous-spécification :
3ba	référence
3bb	temps, portée, présuppositions, ambiguïtés lexicales, etc.
3c	pertinence contextuelle, calcul d'une connexion rhétorique
4	reconnaissance de l'intention, évaluation de la structure de discours

Tab. 1.1 – Echelle d'interprétation selon Schlangen (2004)

Bien que plus détaillée, la classification de Schlangen mélange ce qui relève du *niveau* d'interprétation et ce qui relève des *contraintes* sur cette interprétation. Par exemple « reconnaissance de *tous* les mots » ou « détermination d'une *unique* structure syntaxique ». Nous préférons toutefois distinguer ces deux aspects.

[6]C'est une simplification étant donné que le domaine peut être défini *dans* le discours, en particulier dans les systèmes multi-tâches.

Exemples de problèmes à chaque niveau Nous déroulons ici une interprétation complète de l'énoncé « est-ce que la chambre double est chère ? » en suivant les niveaux d'interprétation donnés dans Brennan and Hulteen (1995) et donnons quelques exemples de problèmes à chaque niveau :

- Etat 0 (*non-attending*) : le système demeure dans l'état 0 malgré le fait que l'énoncé a été prononcé.
- Etat 1 (*attending*) : le système croit à tort qu'un bruit est de la parole, ou retourne dans l'état 0 avant la fin de l'énonciation.
- Etat 2 (*hearing*) : le système analyse le signal mais oublie des mots, insère des mots non prononcés, ou mélange des mots. Par exemple : « ce que la chambre double Escher ».
- Etat 3 (*parsing*) : le système construit une mauvaise représentation syntaxique, ou aucune représentation syntaxique. Par exemple : « (est-ce que) (la chambre double) (est) (chère) ».
- Etat 4 (*interpreting*) : le système ne parvient pas à la bonne représentation sémantique ou référentielle. Par exemple, il peut résoudre mal « la chambre double » en associant l'expression avec un référent erroné, ou donner une interprétation de *cher* différente de celle du locuteur.
- Etat 5 (*intending*) : le système construit mal l'intention du locuteur. Par exemple, il peut considérer à tort qu'il s'agit d'une assertion et déduire un refus.
- Etat 6 (*acting*) : le système agit mal. Par exemple, il peut réserver la chambre au lieu de répondre à la question.
- Etat 7 (*reporting*) : le système produit un énoncé qui ne reflète pas son intention. Par exemple, en supposant qu'il parvient à déterminer que la chambre est chère, il produit « d'accord » au lieu de répondre « oui », ou alors s'il a l'intention de donner le prix de la chambre, il produit un énoncé présentant mal ce prix.

Une des difficultés majeures des approches ascendantes *en ligne* est qu'un niveau supérieur dépendant des niveaux inférieurs, un problème d'interprétation peut se *propager* et produire des problèmes à d'autres niveaux. Par exemple si le système analyse syntaxiquement l'énoncé comme une assertion, alors l'intention, et l'action conséquentes seront erronées. Dans la section 7.3, nous constatons que la proportion des problèmes de référence causés par les niveaux inférieurs (lexique, syntaxe et sémantique) est de 25%.

1.3.3 Types de problème de non-compréhension

Cependant, la catégorie et le niveau de problème ne sont pas suffisants pour décrire le problème. En effet comment décrire le problème lui-même ? Pour définir le type de problème nous nous appuyons sur la description de l'interprétation donnée par Allwood and Abelar (1984). L'interprétation est décrite comme une mise en relation d'une information reçue avec une information enregistrée en mémoire. D'une part cette définition très générale peut s'appliquer à différents niveaux d'in-

terprétation : on associe les mots de l'énoncé avec des mots du lexique, on associe les concepts évoqués dans l'énoncé avec des concepts présents dans l'ontologie[7], on associe les expressions référentielles avec des référents existants, etc. D'autre part, elle permet de caractériser différents types de problème de non-compréhension. Perlis (1997) par exemple décrit les problèmes en termes de contradiction entre l'attendu et l'observé. Si on définit le type de problème comme une contradiction entre une mise en relation attendue et une mise en relation observée, on peut parvenir à décrire plusieurs types de problèmes.

Par exemple Allwood distingue deux types de non-compréhension : l'absence d'informations pertinentes enregistrées (par exemple l'interlocuteur ignore ce qu'est un « Swedish May pole » étant donné qu'il n'en a jamais rencontré) ou l'absence de stratégie permettant de relier l'information reçue avec l'information existante (l'interlocuteur sait ce qu'est un lit mais ignore que le mot suédois *säng* signifie « lit »). Dans les deux cas il décrit ce que l'on pourrait appeler un *vide* d'interprétation caractérisé par l'*absence* d'une information à un certain niveau. Comme le note Allwood, le vide d'interprétation peut être *partiel* si on le caractérise relativement à l'interprétation complète de l'énoncé. Autrement dit, le vide se produit si un locuteur observe son incapacité à relier les différents aspects de l'énoncé à quelque chose qu'il connaisse. La figure 1.5 illustre la manifestation possible d'un vide au niveau de la référence.

U : je réserve cet hôtel
S : quel hôtel ?

Fig. 1.5 – Exemple de manifestation de vide référentiel

L'intérêt est qu'on peut employer cette vision de l'interprétation pour décrire d'autres types de problèmes. Une autre source d'incompréhension est l'*ambiguïté* où au contraire trop d'informations sont disponibles : soit qu'il existe une seule stratégie de connexion mais plusieurs informations pertinentes enregistrées, soit qu'il existe plusieurs stratégies de connexion. Dans l'exemple 1.6, S manifeste qu'il est incapable de relier de manière unique l'expression « cet hôtel » à un hôtel existant.

U : je réserve cet hôtel
S : l'hôtel Ibis ou le Lafayette ?

Fig. 1.6 – Exemple de manifestation d'ambiguïté référentielle

Il est également possible de relier de manière unique une information reçue avec une information enregistrée sans pour autant que cette relation soit établie avec confiance. La troisième source d'incompréhension est alors l'*incertitude*. Il s'agit d'un type d'incompréhension au même titre que le vide ou l'ambiguïté dans la mesure où une incertitude peut s'avérer bloquante pour la réalisation des buts courants (exemple 1.7).

[7]Les ontologies sont des représentations structurées des connaissances sous la forme de concepts et de relations entre ces concepts.

U : je réserve cet hôtel
S : vous voulez réserver à l'hôtel Ibis ?

FIG. 1.7 – Exemple de manifestation d'incertitude référentielle

Enfin, la seule association possible entre une information reçue et une information enregistrée peut conduire à une *incohérence*. Dans l'exemple de la figure 1.8, l'interprétation unique et certaine de « cet hôtel » entre en contradiction avec les paramètres enregistrés de la réservation.

U : je réserve cet hôtel
S : attention, l'hôtel Ibis est non-fumeur contrairement à vos souhaits

FIG. 1.8 – Exemple de manifestation d'incohérence référentielle

1.3.4 Nature de problème

Il existe enfin une différence importante quant à la nature d'un problème et on distingue les problèmes *naturels*, qui sont tous les problèmes que deux êtres humains dialoguant peuvent rencontrer, des problèmes *artificiels*, qui sont tous les problèmes qui se posent uniquement pour un système de dialogue. Un système de dialogue est susceptible de faire face à ces deux types de problèmes : les problèmes naturels parce qu'il dialogue avec un être humain à l'aide de la langue naturelle, et les problèmes artificiels parce qu'il s'appuie sur une modélisation informatique imparfaite.

Par exemple, un système qui s'appuie sur une génération par patrons à trous court le risque d'être non-réversible, c'est-à-dire de produire des énoncés qu'il ne sait pas analyser (DeVault, Oved, and Stone 2006). En conséquence, il donne l'illusion de manipuler des concepts qu'il ne connaît pas, augmentant le risque que l'utilisateur les réutilise avec confiance, entraînant alors des problèmes d'interprétation[8]. Un autre exemple représentatif est noté dans (Ginzburg 2007). L'ambiguïté de portée découlant des deux interprétations de « Tout homme aime une femme » n'est pas problématique pour un être humain car celui-ci ne perçoit même pas l'ambiguïté (pour chaque homme, il existe une femme qui est aimée par lui ou il existe une femme qui est aimée par tous les hommes).

Le dialogue 1.9 (p. 32) illustre deux types d'ambiguïtés artificielles réellement rencontrées lors de l'évaluation MEDIA (voir section 5). L'expression référentielle « la chambre » peut être résolue dans le contexte de l'hôtel Paradisius ou dans celui de l'hôtel Rio De Oro, en ce sens elle est ambiguë. Cependant cette ambiguïté est causée artificiellement par une mauvaise gestion du générique « la chambre ». La seconde ambiguïté vient du pronom « il » qui peut référer soit au premier hôtel, soit au second. Cette dernière est cependant de différente nature puisqu'un être humain pourrait y être confronté. On peut néanmoins décrire ces ambiguïtés comme

[8]Typiquement, c'est le cas dans la théorie de l'alignement (Pickering and Garrod 2004; Garrod and Pickering 2004), voir p. 48.

artificielles dans la mesure où elles n'existent pas pour U, il désire en effet simplement connaître le prix de la chambre *dans les deux hôtels*.

S : à Niort pour fin juin je vous propose deux hôtels
U : oui
S : hôtel Paradisius et hôtel Rio De Oro
U : oui au point de vue de prix prix de la chambre il est à combien

Fig. 1.9 – Exemple d'ambiguïté artificielle (dialogue 1004)

Sabah (1991) emploie le terme de point d'embarras pour décrire ces ambiguïtés artificielles :

> Nous appelons point d'embarras des situations où l'ensemble des éléments de décision ne permettent pas au programme, à un moment donné du traitement, de prendre la bonne décision. Une ambiguïté artificielle est un point d'embarras qui n'est pas dû à la langue elle-même, mais au programme. Notre argumentation selon laquelle l'usage du langage n'est pas ambigu revient à dire que, dans l'esprit du locuteur, il est possible d'utiliser des connaissances pertinentes de façon telle qu'il n'existe pas de point d'embarras. (Sabah 1991)

Cette distinction entraîne une contrainte quant à la façon de résoudre les problèmes : un utilisateur sera probablement peu enclin à gérer des problèmes purement artificiels. Comme le note Caelen, une typologie de problèmes de bas niveau (insertions de mot, omissions, substitutions) « [...] est peut-être pertinente pour le concepteur du système mais elle ne l'est pas vis-à-vis de l'utilisateur qui n'a pas à entrer dans la compréhension du système lui-même, il désire seulement poursuivre le dialogue de la manière la plus cohérente possible » (Caelen and Nguyen 2006). Tout ce qu'on peut demander à l'utilisateur est en effet de résoudre des problèmes qui peuvent faire sens pour lui.

On peut supposer que la différence entre problèmes naturels et problèmes artificiels est causée par une différence de traitement des énoncés : plus le traitement des énoncés par un système sera analogue à celui d'un être humain et plus les problèmes rencontrés dans ce traitement seront similaires. Cette remarque renforce l'idée qu'une modélisation informatique du traitement linguistique doit s'approcher le plus possible du traitement humain, et ce pour éliminer les problèmes artificiels. Par exemple, certaines approches préfèrent voir l'ambiguïté comme une non-décision (Poesio 1996; Surcin 1999; DeVault and Stone 2007). Ces approches permettent de maintenir un certain niveau d'ambiguïté dans le dialogue sans que cela ne soulève de problème. Certaines ambiguïtés peuvent en effet ne pas être caractérisées comme problème mais seulement jusqu'au moment où l'action est nécessaire : lorsqu'il faut agir, toute ambiguïté qui se répercute sur l'instanciation de l'action doit être levée, et si cela est impossible, une ambiguïté doit alors être considérée comme problème.

1.3.5 Conclusion

Nous avons présenté différents aspects de la communication et de la compréhension. En particulier que la communication d'une intention nécessite la compréhension de cette intention. Cependant la compréhension peut faire l'objet de nombreux problèmes. Afin de les caractériser nous avons distingué la *catégorie* de problème (non-compréhension ou mauvaise compréhension), le *niveau* du problème (lexical, syntaxique, sémantique, etc.), le *type* de problème (vide, ambiguïté, incertitude, incohérence) ou la *nature* de problème (naturelle ou artificielle). Nous introduisons la problématique de la robustesse, c'est-à-dire la capacité pour un système de dialogue à faire face à ces problèmes, dans le chapitre suivant.

2 Problématique de la robustesse

L'existence d'une multiplicité de problèmes d'interprétation soulève la question de la robustesse. Comment le système peut-il faire face à tous ces problèmes ? Nous introduisons ici la problématique de la robustesse sous l'angle du paradigme détection/correction dans lequel il s'agit de repérer l'existence du problème avant d'enclencher des stratégies dédiées à sa résolution. Avant de présenter les modèles existants de robustesse, nous affinons toutefois la définition de ce terme.

2.1 Définition

Il existe différentes définitions de la robustesse. Par exemple Hayes and Reddy (1983) décrivent la robustesse comme la capacité pour un système de gérer de manière appropriée n'importe quel énoncé que peut produire un utilisateur. Anderson and Perlis (2005) la décrivent plutôt comme la capacité de tolérer des perturbations dans les entrées. On peut s'appuyer sur une définition qui couvre ces deux aspects : la robustesse est la capacité d'un système de dialogue à gérer l'incompréhension. Cette définition peut être raffinée : on distinguera la capacité d'un système à résoudre l'incompréhension sans faire appel à l'interlocuteur (*robustesse interne*), et la capacité d'un système à résoudre l'incompréhension grâce à l'interlocuteur (*robustesse externe*). La problématique de la robustesse est la mise en œuvre de ces capacités dans un système de dialogue.

La capacité de robustesse interne recouvre la faculté d'un système à faire en sorte qu'un énoncé ne soulève pas d'incompréhension, ou qu'une incompréhension ne soulève pas de problème de traitement. La première approche de la robustesse interne consiste alors à étendre au préalable la couverture des phénomènes gérés par le système. Ce faisant, on réduit la probabilité d'être confronté à un énoncé dont l'interprétation est problématique. On peut étendre les ressources (rajouter des mots dans le lexique, améliorer les grammaires, augmenter les ontologies, etc.), ou améliorer les algorithmes (meilleur taux de reconnaissance, meilleure analyse syntaxique, meilleure représentation du contexte, meilleure gestion des implicatures, etc.). Ces améliorations sont réalisées en amont, lors de la modélisation ou de l'implémentation. D'une certaine manière, les disciplines telles que la linguistique, la psycholinguistique, la logique ou l'informatique en améliorant les modèles de compréhension ou de communication, améliorent la robustesse interne des systèmes.

La deuxième approche de la robustesse interne consiste à prévoir que les algo-

rithmes peuvent échouer et qu'alors il s'agit de corriger les incompréhensions lors d'une phase de post-traitements. Par exemple, en cas d'analyse syntaxique ambiguë, ne conserver que la première analyse, employer des stratégies par défaut lorsque le module d'analyse bute face à un vide d'interprétation, ou encore relâcher les contraintes d'identification du référent (Goodman 1985). Il s'agit d'un ensemble de techniques *ad hoc* plus ou moins bien fondées pour éviter les situations d'erreur (voir Allen et al. 1996).

Cependant les êtres humains sont également confrontés à des problèmes de compréhension malgré leurs capacités d'interprétation très évoluées. Même si on améliore la représentation du contexte et les capacités interprétatives, on n'est jamais garanti que l'interprétation effectuée par le système soit celle désirée par l'utilisateur. C'est au final le locuteur de l'énoncé qui possède l'interprétation correcte de ses propres énoncés, et c'est *par rapport à lui que l'on doit considérer l'incompréhension.* Améliorer la robustesse interne est certes nécessaire, mais cela n'est pas suffisant. C'est pourquoi nous suggérons que l'incompréhension doit également être étudiée sous l'angle du dialogue et des capacités nécessaires pour discuter de l'interprétation d'un énoncé, capacités que nous regroupons sous le terme de *robustesse externe.*

Loin de s'opposer, robustesse interne et robustesse externe sont complémentaires. On ne peut pas se contenter d'une très bonne robustesse interne couplée d'une faible robustesse externe en raison de l'existence de problèmes que l'on ne peut résoudre sans dialogue. Mais on ne peut se contenter non plus d'une très bonne robustesse externe couplée d'une faible robustesse interne. En effet, un système qui s'appuierait systématiquement sur l'utilisateur pour résoudre ses problèmes d'interprétation risquerait d'agacer l'utilisateur (voir Skantze 2007). Les deux types de robustesse sont nécessaires dans le dialogue homme-machine et il s'agit de trouver un équilibre entre la possibilité de résoudre un problème de manière interne, et celle de le faire de manière externe. En outre, la robustesse externe nécessite de bonnes capacités d'interprétation ou de génération pour pouvoir dialoguer à propos de l'incompréhension, et sans un certain niveau de robustesse interne, le système ne peut intégrer les réponses de l'utilisateur ou modifier son contexte pour parvenir à la convergence.

Nous choisissons toutefois de focaliser notre étude sur la robustesse externe pour deux raisons : d'abord un système sera toujours confronté à des problèmes d'interprétation qui nécessitent le dialogue. Etant donné que le dialogue est le lieu de manifestation de la divergence, il est donc pertinent d'employer le dialogue pour la recherche de convergence. Ensuite nous adhérons au point de vue selon lequel le dialogue est le moyen le plus efficace pour l'apprentissage (Luzzati 1995; Lehuen 1997b). En se focalisant sur la robustesse interne, autrement dit en « plongeant la langue dans la machine », nous risquons de fermer la porte à l'apprentissage où l'on souhaite au contraire « plonger la machine dans la langue » (Lehuen 1997a).

Pour illustrer la problématique de robustesse interne/externe, nous proposons l'exemple suivant : « Jeanne a lancé une pierre contre une fenêtre et elle s'est brisée ». L'énoncé est théoriquement ambigu, le pronom « elle » peut en effet référer à la fenêtre, à la pierre, et également à Jeanne. On peut éliminer cette ambiguïté en donnant davantage de connaissances à la machine, c'est-à-dire en améliorant sa

robustesse interne, par exemple qu'un humain ne se brise pas en lançant un objet, que lancer un objet contre un autre implique un choc, que dans un choc l'objet le plus fragile peut se casser, et qu'une fenêtre est plus fragile qu'une pierre. Hormis le fait que ces connaissances ont leurs exceptions qu'il faudrait également spécifier, ajouter plus de connaissances ne pourrait résoudre l'ambiguïté dans : « Jean a lancé un vase contre un miroir et il s'est brisé » ; les deux objets, aussi fragiles l'un que l'autre, peuvent tous les deux se briser. La seule possibilité est de pouvoir clarifier l'énoncé en posant une question, c'est-à-dire dans une démarche de robustesse externe comme dans : « le vase, le miroir ou Jean ? ». Il est préférable en outre de fournir des connaissances minimales au système afin que l'utilisateur ait moins d'effort à résoudre l'ambiguïté, par exemple que la question de clarification se résume à : « le vase ou le miroir ? » (en l'occurrence qu'un humain ne puisse pas se briser en lançant un objet). Les robustesses interne et externe sont nécessaires dans un équilibre satisfaisant l'effort demandé à l'utilisateur.

2.2 Modèles de robustesse externe

La stratégie la plus naïve pour résoudre un problème d'interprétation de manière externe est de s'appuyer sur l'utilisateur au moyen d'un énoncé du type : « Je ne comprends pas. Veuillez reformuler. ». L'emploi de cette stratégie est souvent effectué par défaut dans les systèmes, lorsqu'aucune attention particulière n'a été portée à la résolution dialogique de l'incompréhension (ce fut le cas par exemple lors du projet OZONE, Gaiffe et al. 2004). Cette stratégie présente l'avantage de la simplicité, puisqu'il suffit de faire comme si l'énoncé non-compris n'avait jamais existé et de le *remplacer* par la reformulation fournie par l'utilisateur. Cependant, elle est clairement inefficace. Sa formulation, beaucoup trop vague, ne prévient l'utilisateur que de *l'existence* d'un problème de compréhension dont il ignore le type, le niveau ou la localisation. L'utilisateur doit alors imaginer le problème de compréhension que pourrait avoir soulevé son énoncé. Moins l'utilisateur a de connaissances sur le fonctionnement du système, et moins il a de chances de se représenter correctement le problème et plus la correction du problème lui demandera d'effort. Le cas échéant, un tel énoncé n'a qu'une faible probabilité de mener à la convergence. En outre l'énoncé suggère implicitement que la faute étant dûe à une « mauvaise » formulation de l'utilisateur, il en est donc quelque part responsable, et que c'est alors à lui de corriger son énoncé en le reformulant. Enfin cette stratégie ne considère qu'une partie des phénomènes d'incompréhension, à savoir la non-compréhension du système. Elle ne résout ni la mauvaise compréhension du système, ni l'incompréhension qui proviendrait de l'utilisateur. Les défauts de cette stratégie permettent néanmoins d'éclairer les besoins d'une capacité de robustesse externe.

D'abord comme le notent Gabsdil (2003) ou Purver (2004a) un système ne devrait pas se limiter à manifester un problème d'incompréhension global à propos d'un énoncé mais devrait pouvoir manifester précisemment la cause et la couverture lexicale du problème. Ensuite en fonction du problème, le système doit pouvoir déclencher des stratégies adéquates conduisant à sa résolution ou à son abandon,

et non pas se contenter d'une unique stratégie consistant à manifester de manière passive le problème. Il est nécessaire pour le système de représenter les problèmes d'interprétation, et de gérer explicitement le processus qui vise à la convergence. Par exemple, Duff et al. (1996) proposent de distinguer quatre niveaux dans la gestion des divergences d'interprétation :

1. Détection : détection des problèmes d'interprétation
2. Diagnostic : classification du problème
3. Sélection du plan de correction : choix de la correction, interne ou externe
4. Exécution interactive du plan : poser les questions de clarification

A ces quatre niveaux, Turunen and Hakulinen (2001) ajoutent les niveaux de manifestation de l'erreur et de retour au dialogue principal :

1. Détection de l'erreur
2. Diagnostic de la cause
3. Planification de la correction
4. Exécution de la correction
5. Manifestation de l'erreur
6. Retour au dialogue principal

Les modèles qui suivent cette classification seront qualifiés de modèles détection/correction. Le but est, qu'une fois que la divergence détectée, le système enclenche des stratégies adéquates pour converger, en particulier qu'il s'appuie sur des actes communicatifs dédiés à celle-ci. Ce faisant, il entretient un dialogue, appelé métadialogue, ou dialogue de clarification visant à clarifier le dialogue lui-même. Cette fonction correspond chez Jakobson (1963) à la fonction métalinguistique du langage. Nous présentons ici trois approches qui répondent à cette problématique : d'abord la gestion des questions de clarification chez Ginzburg and Cooper (2001), Purver (2004b) et Larsson (2002), puis l'approche de métaplanification dans Litman and Allen (1984), Nerzic (1993) et Heeman and Hirst (1995) et enfin la gestion de l'incompréhension par stratégies dialogiques dans Caelen and Nguyen (2006).

2.2.1 Questions de clarification

Ginzburg and Cooper (2001) fournissent une analyse linguistique des questions de clarification, en particulier de leur caractère elliptique. Ils notent d'abord l'existence de plusieurs types de *lectures* d'une question de clarification. La lecture *clausale* d'une question vise à vérifier le contenu d'une partie de l'énoncé problématique, et la lecture *en constituant* vise à en obtenir une description alternative. Par exemple :

A : Did Bo finagle a raise ?
B : Bo ?

La question « Bo ? » peut être interprétée soit comme « Are you asking if BO finagled a raise ? » (lecture clausale) ou comme « Who is Bo ? » (lecture en constituant). Afin de modéliser ces différents types de question, les auteurs s'appuient sur HPSG (Pollard and Sag 1994) et en particulier la version de (Ginzburg 2000). Dans ce paradigme, les énoncés sont représentés comme des fonctions qui, appliquées à un contexte, renvoient un contenu complètement spécifié. L'interprétation consiste alors à résoudre les paramètres qui dépendent du contexte. Par exemple si Anne demande « Est-ce que Jo est parti ? », elle demande en fait si la propriété *être parti* peut s'appliquer à la personne à qui elle réfère en disant « Jo ». Ces paramètres ne peuvent êre résolus qu'avec le contexte, par exemple dans le contexte d'énonciation *être parti* peut signifier *avoir quitté la maison d'Anne*, et « Jo » peut référer à Jo Smith.

Le contexte est modélisé par un état d'information courant composé de FACTS (des propositions tenues pour vraies), LATEST-MOVE (l'acte de dialogue le plus récent), et QUD (*Questions Under Discussion*, les questions courantes qui attendent une réponse). La mise à jour de cet état d'information est réalisée de la façon suivante pour un énoncé d'entrée : d'abord les paramètres contextuels sont résolus grâce à l'état d'information courant, et s'ils parviennent à être instanciés de manière unique, mettre à jour FACTS et LATEST-MOVE, et produire un nouvel énoncé (une assertion ou une question applicative), sinon, soulever une question de clarification. Les questions de clarification sont vues comme des opérations de *coercion* qui consistent à produire un nouvel état d'information dans lequel la question à soulever rejoint QUD. Les deux opérations de coercion définies dans (Ginzburg and Cooper 2001) sont l'identification d'un paramètre (*parameter identification*) qui correspond à la lecture en constituant et la focalisation d'un paramètre (*parameter focussing*) qui correspond à la lecture clausale. Nous ne rentrerons toutefois pas ici dans le détail de leur modélisation en HPSG.

L'analyse de Ginzburg et Cooper ne s'applique (explicitement au moins) qu'aux noms propres et à deux types de d'interprétation des questions. Cette approche est étendue dans (Purver and Ginzburg 2004) et dans (Purver 2004b) afin de considérer différents types de questions de clarification. A partir de l'analyse du corpus BNC (Burnard 2000), Purver effectue une classification des questions de clarification que l'on peut trouver dans le dialogue homme-homme. Nous reproduisons ici les différentes formes de question possibles après l'énoncé de A : « Est-ce que Bo est parti ? » (tableau 2.1) ainsi que les différentes lectures possibles (tableau 2.2) que l'on peut trouver dans (Purver 2004b; Purver, Ratiu, and Cavedon 2006).

Le système CLARIE (Purver 2004b; Purver, Ratiu, and Cavedon 2006) implémente le mécanisme de clarification de Ginzburg en utilisant également un modèle à base d'état d'information. L'état d'information contient deux parties, une partie publique qui rassemble entre autre les questions en cours QUD, les énoncés en attente de complétion contextuelle PENDING et les énoncés saillants SAL-UTT, une partie privée qui contient les croyances propres ou l'agenda. Le traitement d'un énoncé est le suivant : d'abord l'énoncé est analysé en HPSG et la représentation de l'énoncé est placée dans PENDING, ensuite trois types de règles sont appliqués afin d'intégrer

Forme	Exemple
Conventionnelle	B : « Hein ? / Quoi ? / Pardon ? »
Non-reprise	B : « Qu'est-ce que tu as dit ? / Tu as dis "Bo" ? »
Reprise littérale	B : « Est-ce que BO est parti ? / Est-ce que Bo est PARTI ? »
Substitution QU-	B : « Est-ce que QUI est parti ? / Est-ce que Bo QUOI ? »
QU-question	B : « Qui ? / Quoi ? »
Reprise d'un fragment	B : « Bo ? / Parti ? »
Reprise « à trou »	B : « Est-ce que Bo ... ? »

Tab. 2.1 – Formes des questions de clarification

Lecture	Exemple
Clausale	« Est-ce de Bo_i dont tu me demandes si i est parti ? »
Constituant	« Que veux-tu dire par "Bo" / par "parti" ? »
Lexicale	« Est-ce que tu as dit "Bo" / "parti" ? »

Tab. 2.2 – Lectures des questions de clarification

l'énoncé dans l'état d'information.

Une première règle standard cherche à instancier les paramètres en s'appuyant sur QUD ou SAL-UTT. Une seconde règle cherche à *accommoder* une nouvelle question à laquelle l'énoncé pourrait répondre, cette question fournissant un nouveau contexte dans lequel l'énoncé devient pertinent. L'accommodation est utilisée pour gérer au moins deux phénomènes. Dans (Larsson 2002; Larsson 2003), lorsque l'utilisateur fournit des informations non requises, ces informations sont vues comme des réponses à des questions non soulevées explicitement. Par exemple, si à la question « Où désirez-vous aller ? », l'utilisateur répond « A Paris, le dix janvier », l'information « le dix janvier » répond à une question du type « Quand désirez-vous partir ? ». Cette question est alors accommodée dans QUD afin de considérer l'énoncé pertinent. L'autre type de phénomène concerne la mauvaise compréhension. Larsson distingue QUD et ISSUES, le premier contient des questions disponibles pour la résolution des ellipses (comme dans Ginzburg and Cooper 2001) et le second les questions dont la réponse reste à fournir. L'emploi de ces deux ensembles est illustré dans l'exemple suivant :

U_1 : To Lund, please.
S_2 : OK, to London.
U_3 : No !
S_4 : Not to London. So, what city do you want to go to ?

Fig. 2.1 – Dialogue tiré de Larsson (2003)

Lorsque S produit son énoncé S_2, la question de la destination est accommodée dans QUD. Mais comme elle reçoit une réponse négative de la part de U en U_3, elle rejoint ISSUES et peut alors être soulevée explicitement par S en S_4.

Si la règle standard d'intégration et les règles d'accommodation échouent, l'énoncé est considéré comme *une question de clarification de la part de l'utilisateur*, et le sys-

tème tente d'appliquer les règles de coercion. Ces règles produisent un nouvel état d'information dans lequel la question est déplacée de PENDING vers QUD et l'action de réponse placée dans l'agenda. Plusieurs heuristiques sont utilisées pour déterminer le type de lecture qu'il faut faire de la question de clarification de l'utilisateur. On teste avant tout l'interprétation conventionnelle de la question. Si ce test échoue, on essaie de l'interpréter comme une question sur un constituant (*parameter identification*), puis on essaie l'interprétation en lecture clausale (*parameter focussing*), etc.

Si toutes ces règles échouent, alors le système doit *soulever une question de clarification* en produisant un nouveau contexte dans lequel sa question de clarification rejoint QUD. Le choix du type de question dépend du résultat de l'interprétation. En l'absence d'analyse syntaxique, il est nécessaire de poser une question de constituant sur l'énoncé entier (par exemple une question de forme conventionnelle). S'il existe une analyse syntaxique mais que la valeur d'un paramètre contextuel est inconnue, soulever une question clausale si le constituant est un défini ou un pronom, une question de constituant dans le cas contraire (par exemple une question ouverte). Si la valeur du paramètre entraîne une inconsistance, poser une question clausale fermée reprenant le fragment, enfin si aucune de ces situations ne s'applique, poser une question de constituant sur l'énoncé tout entier (voir Purver 2004b; Purver et al. 2006).

De notre point de vue, outre la modélisation computationnelle des différentes questions, l'intérêt de cette approche est d'identifier deux entités qui semblent *a priori* différentes : la représentation du problème et sa manifestation dans le dialogue. En effet, la représentation du problème peut être reliée à une question que le système *se* pose à propos de l'interprétation qu'il a effectuée. Si elle ne reçoit pas de réponse satisfaisante, cette question est soulevée dans le dialogue (elle rejoint QUD). De plus, le mécanisme de traitement des questions permet de considérer le fait souvent négligé que le système n'est pas le seul à poser des questions de clarification et que l'utilisateur est également susceptible de ne pas comprendre un énoncé (voir également Cahn and Brennan 1999). Enfin, l'approche peut être étendue aux corrections (par exemple issues de la mauvaise compréhension). En particulier Ginzburg et al. (2007) appliquent des mécanismes similaires pour unifier les corrections qui proviennent de soi (auto-corrections) ou d'autrui. Cette approche élégante et complexe mériterait toutefois d'être évaluée sur corpus afin de déterminer si le traitement des questions de clarification provoque effectivement un gain de compréhension.

2.2.2 Métaplanification

La gestion de l'incompréhension a également été étudiée dans le paradigme de la planification. La métaplanification (Wilensky 1981) a été employée par Litman pour gérer des problèmes de planification (Litman and Allen 1984; Litman 1985). Il s'agit d'ajouter des plans dont l'objectif est de gérer le déroulement des plans, et pour cette raison, sont appelés des *métaplans*. Par exemple, il y aurait un métaplan pour gérer l'instanciation des préconditions d'un plan. S'il est impossible de

donner une instanciation, alors le déclenchement du métaplan permet d'orienter le dialogue par une question afin de pouvoir déclencher le plan initial. Nerzic (1993) soulève toutefois un défaut des approches classiques de planification : si l'utilisateur a une mauvaise représentation des plans possibles, ou qu'il les exécute mal, il risque d'effectuer des actions problématiques pour le système. En effet, dans les approches classiques, le système cherche à reconnaître un plan à partir de l'action effectuée par l'utilisateur, et dans le cas de plans invalides de l'utilisateur (qu'ils soient mal construits, incomplets ou faux) le système construira un plan erroné. Le modèle de Nerzic étend alors celui de Litman en considérant les plans invalides de l'utilisateur. Il s'appuie sur différents états d'un plan : plan *fini* lorsque le plan a été réalisé, plan *prêt*, lorsque tous les paramètres du plans sont instanciés et plan *échoué* lorsqu'une précondition ou une condition d'applicabilité du plan est fausse ou que le plan comporte une action manquante. Lorsque l'utilisateur effectue une action, le système recherche *au préalable* à vérifier la validité du plan. Si le plan est prêt, il peut chercher à l'exécuter. Si le plan est échoué, en revanche le système doit déclencher des actions correctives qui visent à préparer le plan. De la même façon que Litman, le modèle choisit un métaplan pour résoudre le problème.

L'idée d'utiliser des métaplans pour gérer les problèmes de compréhension a également été faite par Heeman and Hirst (1995). Leur modèle s'appuie sur la métaplanification pour modéliser la collaboration à propos de la référence. La production d'une expression référentielle est vue comme un plan dont les actions consistent à produire des parties de l'expression : la tête (le nom commun) et les modifieurs (les adjectifs, ou groupes prépositionnels). Lorsque l'interlocuteur interprète l'énoncé, il cherche à reconnaître le plan qui a conduit à la production de l'expression référentielle. Lorsqu'il a reconnu le plan, il peut procéder à son évaluation afin de déterminer s'il parvient à identifier le référent. En absence d'erreur d'identification du référent, l'interlocuteur doit accepter le plan afin d'indiquer au locuteur initial le succès de son plan. En présence d'ambiguïté ou vide, l'interlocuteur peut essayer de corriger le plan erroné en considérant un nouveau plan de manière similaire au modèle de Nerzic. Il produit alors une nouvelle action avec le but que le locuteur initial effectue cette correction. Le locuteur initial a ensuite le choix d'accepter le nouveau plan ou de le refuser.

Ces types de modèles ont en commun le fait de décrire la clarification comme des actions que l'on peut planifier. Cependant, nous estimons que la métaplanification, en rajoutant des plans au-dessus de plans, démultiplie la complexité de la gestion de l'incompréhension. Plus le système est complexe dans sa gestion de l'incompréhension, plus il court le risque d'entraîner de nouveaux problèmes en cherchant à résoudre les problèmes.

2.2.3 Stratégies dialogiques

Caelen propose d'employer différentes stratégies dialogiques pour gérer l'incompréhension dans le dialogue (Caelen and Nguyen 2006). En cas de malentendu (mauvaise compréhension), détecté par le système à l'aide d'une contradiction dans sa

base de faits, il manifeste cette contradiction en demandant à l'utilisateur de valider ou d'invalider son acte au moyen d'une stratégie directive. Si au contraire le malentendu est détecté par l'utilisateur, le système peut chercher à corriger directement sa base de faits s'il y parvient ou à clarifier le problème en le soulevant dans le dialogue au moyen d'une stratégie coopérative. Si trop de malentendus surviennent, le système notifie le problème à l'aide d'une stratégie directive.

L'incompréhension (non-compréhension) est modélisée grâce à un *marqueur dialogique d'incompréhension* qui représente l'état de l'interprétation : incompréhension totale, ou incompréhension partielle. En fonction du type d'incompréhension, le système peut déclencher différents actes : confirmation explicite, implicite, désambiguïsation, suggestion de solution, demande de répétition ou notification d'incompréhension. Le type d'acte est choisi selon la situation courante. Par exemple si le but courant est atteint, le système soulève une confirmation implicite à l'aide d'une stratégie coopérative pour une incompréhension partielle, ou directive pour une incompréhension totale. Ou encore si le système est en train de demander des informations et qu'une incompréhension survient, il suggère une solution au moyen d'une stratégie réactive. De la même façon que pour le malentendu, si le nombre d'incompréhension dépasse un certain seuil, le système applique une stratégie directive en notifiant le problème ou en demandant une répétition.

2.3 Critiques du métadialogue

Les modèles que nous avons présentés décrivent le processus dialogique qui permet de faire converger une interprétation suite à une incompréhension. L'emploi du métadialogue pour atteindre l'interprétation du locuteur ne fait toutefois pas l'objet d'un consensus. Par exemple Allen (Allen et al. 1996) déclare « ...when faced with ambiguity it is better to choose one specific interpretation and run the risk of making a mistake as opposed to generating a clarification subdialogue ». Ce principe peut être motivé pour deux raisons : d'abord, comme le note Skantze (2003, 2005), la sensation de réussite de la tâche expérimentée par des utilisateurs diminue si le système manifeste explicitement sa non-compréhension. Au contraire si le système choisit une interprétation par défaut et continue dans la tâche, au risque de devoir gérer une correction ultérieure, les utilisateurs sont davantage satisfaits de la conduite du dialogue. D'autre part, utiliser un sous-dialogue de clarification peut entraîner de nouvelles divergences d'interprétation, par exemple si la réponse à la question de clarification n'est pas comprise. Cette situation ne fait qu'empirer le problème initial et peut conduire à un abandon total du dialogue.

Cependant, gérer un problème détecté *a priori* et soulevé grâce à une question semble plus facile cognitivement que de devoir corriger *a posterori* l'interprétation d'un énoncé. En effet, si l'utilisateur détecte une mauvaise compréhension du système et le lui manifeste, le système doit *inférer* la cause de la divergence. Cette inférence a un certain coût qui n'a pas lieu d'être si le système anticipe la résolution du problème en posant une question. De plus, que faire lorsqu'il n'y a aucune analyse ? En l'absence d'interprétation, le système devrait alors revenir sur la tâche en ignorant

l'énoncé de l'utilisateur et on peut s'interroger sur la frustration de l'utilisateur face à un tel comportement. Cette remarque suggère que face à une incompréhension, le système doit évaluer ce qui est le plus important : comprendre l'énoncé ou réaliser la tâche. Cette question n'est pas anodine car le système peut toujours être confronté à des incompréhensions qu'il ne devrait pas chercher à résoudre. La figure 2.2 illustre un énoncé probablement non-compris par le système. Pour autant, il ne semble pas approprié qu'il cherche à le comprendre outre mesure étant donné qu'il ne sera probablement pas à même d'enregistrer la requête.

client : je désire également que l'on me porte une bouteille de Dom Pérignon
client : avec des biscuits r() roses de Reims à minuit tapant dans ma chambre

Fig. 2.2 – Extrait du dialogue MEDIA 1032

Le paradoxe est qu'afin de déterminer si la recherche de compréhension est nécessaire, il est indispensable de parvenir au moins à une compréhension minimale de l'énoncé. Nous supposerons alors que cette recherche de compréhension minimale devrait être systématique. Notre objectif n'est cependant pas de clarifier outre mesure la décision de la recherche de convergence (voir p. 68) mais d'identifier les caractéristiques d'un système capable de robustesse externe. Nous supposerons alors que tous les problèmes d'interprétation devront être résolus, en considérant toutefois que les problèmes doivent faire sens pour l'utilisateur si sa participation à la résolution est requise.

2.4 Conclusions

Selon nous, les modèles présentés dans ce chapitre ne traitent qu'une partie du problème de la robustesse. Ils sont basés sur un principe de détection puis de correction des problèmes. Autrement dit, la recherche de convergence n'est déclenchée *qu'en présence de divergence*. Clark and Schaefer (1989) notent que des modèles de ce type ne peuvent rendre compte que de la manifestation *négative* de la compréhension, éludant alors le problème de la manifestation *positive* de compréhension. Nous estimons, à l'instar de Clark et Schaefer, que la recherche de la convergence ne doit pas être limitée aux situations dans lesquelles un problème d'interprétation se produit mais qu'au contraire, cette recherche de convergence est sous-jacente à chaque acte communicatif. Il nous faut alors adopter une perspective plus large sur la problématique et nous pensons que le concept de terrain commun a une place centrale dans cette perspective.

3 Terrain commun et processus d'ancrage

Comme nous l'avons vu, il est possible de résoudre les divergences d'interprétation en adoptant des stratégies dédiées à la convergence, comme le fait de soulever des questions de clarification ou d'initier des plans de correction. Cependant les participants cherchent également à éviter les divergences *a priori*. Par exemple, un locuteur ne produira pas une expression tout en sachant pertinemment que son interlocuteur est incapable de le comprendre. Comment un locuteur peut-il savoir ce qui sera compris par son interlocuteur ? Nous proposons d'aborder la problématique de la robustesse dans le paradigme du terrain commun et définissons ici qu'on entend par terrain commun.

3.1 Terrain commun

Pour pouvoir communiquer, il nous est nécessaire de partager les moyens de la communication. Nous devons partager une certaine connaissance linguistique et connaissance du monde pour nous faire comprendre. Le concept de terrain commun a été introduit par les philosophes pour décrire à la fois ce qui relève des connaissances ou des règles en commun qui autorisent la communication et ce qui en constitue l'essence :

> Dans l'expérience du dialogue, il se constitue entre autrui et moi un terrain commun, ma pensée et la sienne ne font qu'un seul tissu, mes propos et ceux de l'interlocuteur sont appelés par l'état de la discussion, ils s'insèrent dans une opération commune dont aucun de nous n'est le créateur. (Merleau-Ponty 1945)

Cette notion a également été reprise par les psycholinguistes. Clark (1996) distingue par exemple différents niveaux de partage de connaissance ou de croyance et alors différents types de terrains communs. D'abord, nous sommes tous des êtres humains et conséquemment partageons des organes sensoriels. Si un son nous est audible, nous pouvons faire l'hypothèse que notre partenaire peut également l'entendre. Nous partageons alors cette perception et pouvons communiquer à ce sujet. Ensuite, les êtres humains s'organisent en communautés qui partagent des connaissances ; ce type de terrain commun est appelé terrain commun *communautaire* (*communal common ground*). Par exemple, les locuteurs d'une langue donnée partagent

la maîtrise de la langue, ses règles de phonologie, de morphologie, de syntaxe, de sémantique ou de pragmatique. Si je produis un énoncé dans une langue et que je suppose que mon interlocuteur appartient à la même communauté linguistique, je peux m'attendre à ce qu'il comprenne mon énoncé de la même manière que je l'entend. Ou encore, si je suppose que mon interlocuteur fait partie de la communauté des amateurs de musique classique, je peux supposer qu'il partage avec moi quelques connaissances sur Mozart. Enfin, deux êtres humains peuvent partager des connaissances spécifiques qui ne valent que pour eux deux. Tous les événements qui se produisent en présence des deux individus peuvent appartenir au terrain commun *personnel* (*personal common ground*). Par exemple, si deux amis se rendent à une soirée, ils pourront reparler de la soirée car cet événement appartient alors à leur terrain commun, ou encore s'ils s'accordent sur un terme pour décrire un de leurs amis communs, cette dénomination pourra appartenir à leur terrain commun.

Sur quoi peut-on fonder cette intuition ? Kingsbury (1968) a fait l'expérience de demander à des passants choisis aléatoirement la direction d'un grand magasin situé quelques rues plus loin, soit en se présentant comme un étranger, soit en se présentant comme quelqu'un de la ville. Il a observé que les descriptions de la route à suivre étaient plus détaillées lorsqu'il se présentait comme un étranger. Cette expérience illustre que les locuteurs adaptent *a priori* leurs énoncés en fonction de leurs partenaires, en prenant des raccourcis descriptifs lorsqu'ils estiment partager plus de connaissances avec leur interlocuteur.

Dans une expérience importante, Clark and Wilkes-Gibbs (1986) montrent cette adaptation au niveau de la production d'une expression référentielle : en répétant plusieurs fois une même tâche, ils observent que la longueur des expressions référentielles pour désigner les mêmes objets diminue. Leur expérience consiste à confronter deux individus séparés par un écran opaque à une même scène composées d'objets. L'un des sujets propose des expressions référentielles afin de désigner les objets que l'autre participant doit identifier. La complexité des figures implique la nécessité d'employer des expressions référentielles complexes afin de les discriminer. Cependant, lorsque la tâche est répétée plusieurs fois et que celle-ci est couronnée de succès, les participants adaptent leurs expressions référentielles en produisant des expressions plus simples pour désigner les mêmes objets. Par exemple « a person who's ice skating, except they're sticking two arms out in front » devient « the person ice skating that has two arms », puis « the person ice skating, with two arms » et enfin « the ice skater ».

L'explication avancée est que l'identification des référents repose effectivement sur le terrain commun : lorsqu'une opération de référence réussit elle établit une conceptualisation partagée du référent dans le terrain commun. Lors de la production d'énoncés ultérieurs, les locuteurs s'appuient sur cette conceptualisation partagée, et ce précédent permet de faciliter la communication en produisant des expressions plus courtes.

Cette expérience confirme l'utilisation d'un certain terrain commun, et elle éclaire en outre la manière dont les participants prennent en compte leur partenaire. Clark et Wilkes-Gibbs suggèrent qu'il ne suffit pas considérer l'effort que fait le locuteur pour

produire son énoncé mais également l'effort collaboratif que les deux participants vont dépenser pour atteindre la bonne compréhension de l'énoncé. L'établissement d'un terrain commun vise alors à diminuer cet effort. En conséquence, plus deux interlocuteurs partagent de terrain commun, moins l'effort collaboratif sera important et plus la communication sera facilitée. La notion de terrain commun semble alors s'avérer essentielle pour la problématique de la robustesse mais comment modéliser cette notion ?

3.1.1 Représentation

Le concept de terrain commun est relié par Stalnaker à celui de la présupposition : lorsqu'un locuteur produit un énoncé, il présuppose la vérité de certaines propositions (Stalnaker 1974; Stalnaker 1978). Par exemple, si A dit à B : « le roi de France est chauve », il suppose qu'il existe un roi de France. Mais il suppose également que B effectue la même supposition. L'ensemble des présuppositions correspond alors au terrain commun :

> Presuppositions are what is taken by the speaker to be the common ground of the participants in the conversation, what is treated as their common knowledge or mutual knowledge. (Stalnaker 1978)

Pour Stalnaker, effectuer une assertion provoque l'ajout de ce qui est asserté à ce qui est présupposé, autrement dit le contenu propositionnel d'un énoncé rejoint normalement le terrain commun à moins que l'assertion soit rejetée. Après avoir dit « le roi de France est chauve », si B n'objecte pas, A peut supposer que B considère la proposition correspondante vraie. De plus, B peut supposer que A la considère vraie également. Dans cette perspective, le terrain commun est un ensemble de propositions que les participants considèrent comme vraies et partagées.

Mais il ne suffit pas de croire vraie une proposition et de croire que l'interlocuteur la croit vraie également. En particulier lorsqu'un locuteur produit une expression référentielle pour désigner un objet, Clark montre que, pour que le locuteur estime que son expression puisse effectivement désigner l'objet, il est nécessaire qu'il entretienne une infinité de croyances (Clark and Marshall 1981). Par exemple, on suppose que Anne et Bob lisent ensemble un journal indiquant que le film *Monkey Business* passe au cinéma *Roxy*, et que Anne dit à Bob « est-ce que tu as vu le film qui passe au Roxy ? ». Clark s'interroge sur les conditions de réussite de l'identification du référent R *Monkey Business* grâce à l'expression t « le film qui passe au Roxy ». Il montre que, pour que Anne puisse utiliser l'expression t avec l'assurance qu'elle réfère à R, celle-ci doit entretenir un nombre infini de croyances :

1. elle doit croire tout d'abord que t réfère à R
2. elle doit croire ensuite que Bob croit que t réfère à R, car si ce n'était pas le cas Bob n'arriverait pas à identifier R grâce à t
3. elle doit croire également que Bob croit qu'elle croit que t réfère à R, car si ce n'était pas le cas, lorsque Bob interpréterait t, il s'imaginerait que Anne ne réfère pas à R

4. etc.

L'infinité des croyances résulte du fait que Anne et Bob doivent avoir une certaine croyance de premier ordre p, mais également une croyance qu'ils partagent cette croyance, et qu'ils entretiennent alors des croyances du second ordre. Mais pour savoir qu'ils partagent bien ces croyances du second ordre, il est nécessaire qu'ils entretiennent la croyance qu'ils partagent ces croyances du second ordre, et doivent avoir donc des croyances du troisième ordre, etc. Clark distingue alors les croyances *partagées*, qui sont des croyances finies, des croyances *mutuelles* qui sont des conjonctions infinies de croyances partagées :

1. $Bel_A p \wedge Bel_B p$
2. $Bel_A Bel_B p \wedge Bel_B Bel_A p$
3. $Bel_A Bel_B Bel_A p \wedge Bel_B Bel_A Bel_B p$
4. etc.

Il suggère alors que le terrain commun est constitué de l'ensemble des faits, hypothèses ou croyances mutuellement entretenus. Clark (1996) rappelle trois représentations possibles :

- la représentation *itérée* (Schiffer 1972) est composée d'un ensemble infini de croyances
- la représentation *réflexive* (Barwise 1988) est une version auto-référentielle : p appartient au terrain commun des individus C si et seulement si « (1) les membres de C ont l'information p et (1) ».
- la représentation à partir d'une *base partagée* (Lewis 1969) : p appartient au terrain commun des individus C si et seulement si :
  - tous les membres de C ont l'information qu'une certaine base b est vraie
  - b indique à chaque membre de C que chaque membre de C a l'information que b est vraie
  - b indique p aux membres de C

La représentation itérée ne semble pas être plausible cognitivement. Clark préfère s'appuyer sur la représentation à partir d'une base partagée dans la mesure où la représentation réflexive peut être inférée à partir de la base partagée et du schéma d'induction et qu'en exhiber une base permet d'expliquer comment une croyance mutuelle peut être obtenue.

3.1.2 Critiques

La notion de terrain commun ou de connaissances mutuelles n'est cependant pas consensuelle. Le premier type de critique provient de l'existence d'un paradoxe : si les deux locuteurs doivent obtenir un nombre infini de croyances pour établir une proposition dans le terrain commun, alors comment peuvent-ils construire une infinité de croyances en un temps fini ? Clark propose deux familles d'heuristiques pour résoudre ce paradoxe, les heuristiques de *troncature* et les heuristiques de *coprésence* (Clark and Marshall 1981).

Les heuristiques de *troncature* consistent à supposer que les participants n'ont pas à construire une infinité de croyances, et qu'alors ils ne cherchent pas à produire des expressions référentielles avec la certitude de leur réussite. Par exemple, un locuteur n'aurait qu'à vérifier les premières croyances pour produire une expression. Sans totalement éliminer cette possibilité, Clark doute du fait que nous vérifions ce genre de croyances, en raison de la difficulté à raisonner avec des croyances réciproques.

La seconde famille d'heuristiques, appelées heuristiques de *coprésence*, propose d'inférer la croyance mutuelle à partir de la base partagée. Les participants n'auraient alors pas à vérifier les croyances appartenant à la croyance mutuelle, mais pourraient si besoin est, les inférer à partir de la base partagée, de certaines hypothèses et du schéma d'induction. Par exemple si nous voyons une bougie, et que nous voyons que nous voyons la bougie, nous pouvons en déduire que nous avons la croyance mutuelle que nous voyons une bougie et nous pouvons alors légitimement employer l'expression référentielle « la bougie » (Schiffer 1972). Clark parle de triple coprésence : les deux individus et l'objet de la croyance mutuelle. Pour pouvoir appliquer le schéma d'induction, il est nécessaire de poser quelques hypothèses. En particulier, il est nécessaire que l'interlocuteur soit attentif, et qu'il soit également rationnel afin qu'il puisse appliquer le même principe de son côté. Plusieurs types de coprésence sont proposés qui fournissent des bases partagées différentes. Par exemple, la coprésence physique où les participants sont physiquement présents avec l'objet de la croyance mutuelle nécessite des hypothèses de simultanéité, d'attention et de rationalité. La coprésence linguistique (par exemple pour traiter l'anaphore) nécessite en outre une hypothèse de rappel de croyance mutuelle antérieure (*recallability*), ou une hypothèse de compréhensibilité (*understandability*). Cette dernière est toutefois problématique, quelles sont les conditions qui autorisent en effet d'effectuer cette hypothèse si ce n'est une croyance mutuelle existante ?

Le second type de critique porte sur la nature de ce qui est partagé et des moyens que nous avons d'y accéder. La théorie de la pertinence (Sperber and Wilson 1989) suggère que la connaissance mutuelle n'aurait pas d'utilité en vertu de l'impossibilité d'établir quoi que ce soit de certain dans la communication :

> Deux individus peuvent regarder un même objet de manière différente ; ils peuvent comprendre différemment l'information qu'ils ont reçue ensemble ; ils peuvent méconnaître certains faits. Dans tous les cas, les individus concernés auraient tort de s'attribuer un savoir mutuel. (Sperber and Wilson 1989)

On pourrait s'appuyer alors sur des suppositions (ou des croyances) qui n'ont pas à être certaines, mais plus la supposition est d'ordre élevé et moins elle est probable, et alors « la supposition de mutualité, qui coiffe toutes les autres, est la moins probable » (*ibid*). Cette critique est tout à fait justifiée, et la réponse formulée par cette théorie rejoint le concept de moindre effort collaboratif. Elle inscrit en effet les mécanismes d'interprétation et de production dans un modèle plus général de la cognition : les participants cherchent à maximiser la pertinence de leurs énoncés en s'appuyant sur des hypothèses dont les effets contextuels sont maximums et dont les efforts de traitement sont minimums. A effets constants, un locuteur préférera alors

produire des énoncés plus faciles à traiter pour son partenaire. Mais le calcul de ce coût a lui-même un coût que les locuteurs ne sont pas toujours disposés à faire.

Le troisième type de critique questionne en effet la prise en compte d'autrui dans le dialogue. Barr (2004) montre par exemple qu'on peut établir une communication conventionnelle sans prendre en compte autrui *a priori* : en communiquant les participants ajustent leur production en fonction de leur partenaire *a posteriori* et non *a priori*, procédant ainsi par essais/erreurs en établissant des conventions locales. Certains comme Koschman and LeBaron (2003) refusent même le concept de terrain commun pour expliquer la communication :

> By its name it would seem to index a place, a place where things can be stored or recorded, but this is a profoundly misleading connotation. Common ground is, after all, a place with no place.

Pourtant il est clair que la prise en compte d'autrui peut intervenir *a priori*, comme le montre l'expérience de Kingsbury, mais dans quelle mesure ? La théorie de l'alignement de Garrod et Pickering (Garrod and Pickering 2004; Pickering and Garrod 2004) propose que les participants alignent leurs représentations au fur et à mesure du dialogue. Cet alignement serait réalisé de manière automatique à différents niveaux, sans chercher à construire des croyances de la compréhension du partenaire. Ils distinguent alors un terrain commun *implicite* issu de l'alignement et un terrain commun *explicite* correspondant au terrain commun du modèle collaboratif. Cette théorie suggère alors que les participants adoptent d'abord des mécanismes *égocentriques* dans l'interprétation ou la production en s'appuyant sur le terrain commun implicite *avant* de s'appuyer sur le terrain commun explicite et de prendre davantage autrui en considération. Plusieurs expériences viennent conforter ce modèle. Horton and Keysar (1996) montrent par exemple que des locuteurs sous pression temporelle ($<$1.5s) prennent moins en compte leur interlocuteur dans la production d'un énoncé que s'ils disposent de temps pour le faire. Ou encore Keysar et al. (1998) montrent qu'un participant interprète d'abord l'énoncé de son point de vue avant de considérer le point de vue d'autrui. L'explication avancée est que la prise en compte d'autrui est plus difficile, peut-être parce qu'elle fait intervenir des croyances de plus haut niveau.

Comment réconcilier les approches égocentriques et les approches collaboratives ? Nous ne proposons pas de théorie cognitive du dialogue mais notons simplement que la prise en compte d'autrui est un effort supplémentaire que les participants n'ont pas toujours la possibilité, la nécessité ou la volonté de dépenser. La production d'un énoncé relève d'un compris entre l'effort à produire l'énoncé et l'effort à interpréter l'énoncé. Plus d'effort est dépensé dans la production de l'énoncé en considérant l'interprétation d'autrui, moins d'effort sera dépensé à l'interpréter et en retour moins d'effort collaboratif sera dépensé pour faire converger l'interprétation. Dans certains cas comme la pression temporelle, la production peut être tellement contrainte qu'autrui ne peut être pris en considération. C'est d'ailleurs une explication pour les auto-clarifications : le locuteur ne vérifie que son énoncé est compréhensible pour l'interlocuteur qu'*après* l'avoir produit, et dans le cas contraire préfère corriger lui-même son énoncé comme l'a remarqué l'analyse conversationnelle (Schegloff,

Jefferson, and Sacks 1977).

Nous ne nous attachons pas à explorer davantage la mesure avec laquelle les participants prennent en compte autrui mais remarquons que dans tous les cas une théorie valide de la communication doit considérer cet aspect. Dans cette perspective le terrain commun intervient pour faciliter la convergence de l'interprétation. Cependant, même en absence de pression temporelle, et avec toute la meilleure volonté du monde pour prendre en compte autrui, il peut exister des divergences d'interprétation. Le terrain commun n'existe en effet pas *en dehors* des individus et chaque participant entretient ce qui est selon lui le terrain commun avec son partenaire. En conséquence, les versions du terrain commun entretenues par chaque participant peuvent être différentes. Et nous pensons que c'est justement dans ces différences d'appréciation que naissent les divergences d'interprétation.

3.1.3 Divergences de terrain commun

Comme les participants s'appuient sur leur version propre du terrain commun pour interpréter ou pour produire les énoncés, et que ces versions peuvent être divergentes, l'interprétation d'un énoncé peut être divergente. Le biais cognitif appelé « faux consensus » (Tversky and Kahneman 1974; Ross et al. 1977; Gilovich 1990) entraîne peut-être cette mauvaise perception du terrain commun. Il y a faux consensus lorsqu'un participant attribue à autrui des croyances qui sont les siennes. Par extension, nous pouvons faire l'hypothèse que ce biais cognitif peut intervenir dans la production et l'interprétation des énoncés à travers une confusion entre ce qui relève du terrain commun et des croyances propres. Gilovich (1990, p. 632) souligne que la mauvaise compréhension peut intervenir lorsqu'il y a une mauvaise perception de l'ambiguïté d'un énoncé : lorsque j'emploie une expression que je sais ambiguë, je sais dans quelle mesure elle est ambiguë, autrement dit quelles en sont les interprétations possibles, et je suppose que mon interlocuteur peut parvenir à l'interprétation désirée parce que je suppose qu'il perçoit la même ambiguïté que moi. Cependant si je crois à tort que ma perception de l'ambiguïté est partagée, alors mon interlocuteur pourra mal interpréter mon énoncé. Cette remarque permet de mieux comprendre une difficulté importante de la communication homme-machine, causée par une différence considérable dans la perception de l'ambiguïté. Nous pouvons faire l'hypothèse que les problèmes de vides ou d'incertitudes peuvent relever du même ressort.

Par exemple dans la figure 3.1 (p. 52) le locuteur U veut référer à un hôtel H_1 en employant l'expression E « au clos des cyprès ». Lorsqu'il produit son expression, il suppose que son interlocuteur est susceptible d'accéder à l'hôtel en question, et qu'alors le fait « E permet d'accéder à H_1 » est mutuellement entretenu par U et S. Cependant ce n'est pas le cas, U était seul à entretenir cette croyance et l'emploi de cette expression conduit alors à la non-compréhension de S. D'une certaine manière, U est trop *optimiste* vis à vis de la compréhension de S.

A son tour, lorsque l'interlocuteur interprète l'énoncé, il s'appuie également sur le terrain commun : il suppose que certains faits qui y sont présents permettent de comprendre l'énoncé. Mais de la même façon, s'il entretient une vision erronée

U_1 : je vais au clos des cyprès
S_2 : je ne comprends pas "au clos des cyprès"

Fig. 3.1 – Exemple de non-compréhension

du terrain commun il interprétera de manière divergente l'énoncé. Dans la figure 3.2 l'interlocuteur S suppose que l'expression E « l'autre » réfère à l'hôtel Ibis H_1 et qu'alors le fait « E permet d'accéder à H_1 » appartient au terrain commun. Cependant la réaction de U montre que ce n'était pas le cas et résulte dans la mauvaise compréhension de S. En un sens, S est trop *optimiste* vis à vis de sa propre compréhension.

U_1 : bon je prends l'autre euh oui l'autre c'est ça
S_2 : ok une réservation à l'hôtel Ibis
U_3 : non non à l'hôtel Lafayette

Fig. 3.2 – Exemple de mauvaise compréhension

La théorie de Grice, en exhibant une intention communicative, entraîne que le locuteur présuppose la *compréhensibilité* de ses énoncés. Alors la croyance que l'interlocuteur est capable de comprendre l'énoncé est supposée appartenir au terrain commun. En conséquence, le locuteur présuppose tous les corollaires de la compréhension, c'est-à-dire que l'interlocuteur est capable de reconstruire les mots, le sens, la référence, les implicatures, l'intention initiale, etc. Nous faisons l'hypothèse que c'est lorsqu'une ou plusieurs de ces présuppositions sont erronées que l'interprétation peut être divergente.

3.1.4 Robustesse et terrain commun

En quoi la notion de terrain commun éclaire-t-elle la problématique de la robustesse ? La mauvaise perception du terrain commun qui existe entre deux locuteurs semble justifier la naissance des divergences d'interprétation. En admettant cette hypothèse, la problématique de la robustesse peut être décrite en termes de gestion du terrain commun :

- rechercher la cause de la divergence revient à déterminer quelles sont les présuppositions effectuées par le locuteur qui ne s'avèrent pas partagées par l'interlocuteur
- rechercher la convergence revient à corriger ou augmenter le terrain commun pour qu'il n'y ait plus de problèmes d'incompréhension

Dans cette perspective la recherche de la convergence n'a pas pour seul but de faire converger localement l'interprétation d'un énoncé donné, mais également de prévenir les problèmes ultérieurs en établissant ce qui doit être partagé pour les éviter. Le terrain commun doit alors être *géré* dans l'interaction, et les participants mettent à jour le terrain commun grâce à un processus dialogique appelé *processus d'ancrage*.

3.2 Processus d'ancrage

Clark and Schaefer (1989) critiquent les modèles dans lesquels la mise à jour du terrain commun est soit réalisée de manière automatique, soit conditionnée à l'absence de preuve négative de compréhension. Par exemple, pour Stalnaker, le contenu d'une assertion rejoint le terrain commun, à moins qu'elle soit contredite. Ce type de modèle ne parvient pas en effet à donner d'explication à la manifestation *positive* de compréhension. Le modèle des contributions de Clark et Schaefer suppose alors que la manifestation positive de compréhension peut être motivée par la volonté des participants d'atteindre une croyance mutuelle de compréhension. Ce but est décrit à l'aide d'un critère que les participants cherchent à atteindre, appelé *grounding criterion* et que nous traduirons par *critère d'ancrage* :

> *Grounding criterion* : The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes[1].

Le contenu d'un énoncé ne peut alors rejoindre le terrain commun qu'à partir du moment où le critère d'ancrage pour cet énoncé est atteint. Comme ce critère implique une croyance du locuteur à propos de la compréhension de l'interlocuteur, il permet de motiver la manifestation positive de compréhension comme *moyen* pour l'interlocuteur d'atteindre la croyance mutuelle de compréhension. Le processus dialogique qui consiste à gérer l'introduction du contenu d'un énoncé dans le terrain commun par la production de preuves de compréhension est appelé *grounding process*[2] (processus d'ancrage).

L'unité qui compose ce processus est la *contribution*, une unité collaborative qui rassemble un énoncé et tous les énoncés qui visent à en établir le critère d'ancrage. Une contribution est alors composée de deux phases : la présentation d'un énoncé et son acceptation par le partenaire.

- *Phase de présentation* : lorsque A présente un énoncé A_i, il fait l'hypothèse que s'il reçoit une preuve de compréhension de son partenaire B, il peut légitimement croire que B a compris A_i.
- *Phase d'acceptation* : lorsque B accepte A_i, il produit une preuve de compréhension en supposant qu'une fois que A reçoit cette preuve, A croit que B comprend l'énoncé.

Lorsque ces deux phases sont terminées, elles forment une base partagée pour inférer la croyance mutuelle de compréhension et alors atteindre le critère d'ancrage[3]. Comme pour les expressions référentielles, on peut estimer que cette inférence n'est

[1] Critère d'ancrage : le contributeur et les partenaires croient mutuellement que les partenaires ont compris suffisamment ce que le contributeur signifiait pour les buts courants.

[2] Le terme de *grounding process* est employé par Ginzburg et Purver dans le sens de processus qui vise à la compréhension, c'est-à-dire l'instanciation des paramètres contextuels. Au contraire Clark et Schaefer l'emploient dans un sens plus général, comme processus qui vise à la compréhension mutuelle et qui permet alors de justifier la manifestation positive de compréhension.

[3] « When these two phases are done properly, they constitute the shared basis for the mutual belief that B understands what A means by signal s » (Clark 1996)

possible que sous des conditions de simultanéité, de rationalité, d'attention, et de compréhensibilité (Clark and Marshall 1981).

La coordination de ces deux phases est modélisée par une structure de dialogue sous la forme d'un graphe associant les contributions aux énoncés qui les présentent et aux contributions ou aux énoncés qui les acceptent. Le dialogue de la figure 3.3 sera par exemple représenté par la structure de la figure 3.4.

A_1 : well wo uh what shall we do about uh this boy then
B_2 : Duveen ?
A_3 : m
B_4 : well I propose to write, uh saying. I'm very sorry

Fig. 3.3 – Dialogue tiré de Clark et Schaefer (1989)

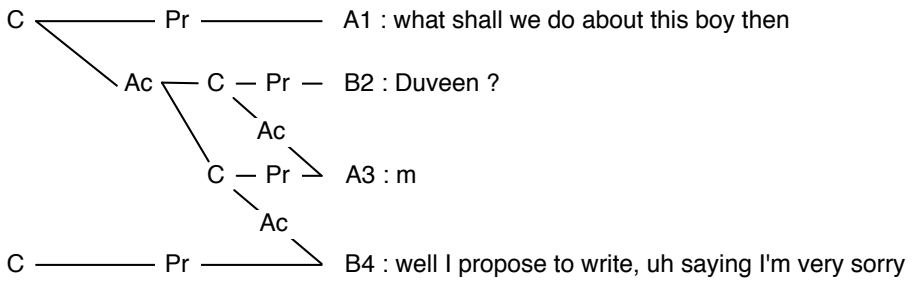

Fig. 3.4 – Structure du dialogue 3.3

Manifestation de la compréhension La manifestation de la compréhension est alors un vecteur essentiel pour transmettre la croyance de compréhension et ce vecteur peut prendre de nombreuses formes. Le modèle des contributions considère cinq catégories de manifestation, ordonnées de la plus faible à la plus forte :

1. Attention continue : B manifeste qu'il est attentif et demeure alors satisfait de la présentation de A
2. Initiation d'une contribution pertinente : B initie une contribution pertinente au même niveau que la précédente
3. Confirmation (*acknowledgement*) : B hoche la tête ou dit « ok »
4. Démonstration : B démontre qu'il a compris tout ou partie de ce que A voulait dire
5. Répétition : B répète tout ou partie de la présentation de A

Lorsqu'un locuteur produit un énoncé, il entretient certaines attentes de compréhension de la part de son partenaire, et ces attentes se traduisent directement par des attentes en termes de compréhension manifestée. Par exemple, si un locuteur transmet un numéro de téléphone par téléphone, il attendra probablement

une répétition de ce numéro par l'interlocuteur. S'il a des attentes moins fortes, par exemple s'il parle du temps qu'il fait, il se suffira peut-être d'une confirmation pour estimer que l'interlocuteur a compris. Le type de compréhension manifestée dépend de nombreux facteurs dont le but courant, le medium utilisé, la pression temporelle, etc. (Clark and Brennan 1991).

S'il est clair que cette attente de compréhension est variable, la hiérarchie des preuves du modèle des contributions est moins évidente : peut-on considérer qu'un élève qui répète les paroles d'un professeur fait preuve de *plus* de compréhension que s'il les démontrait, par exemple avec une paraphrase ? La hiérarchie semble s'appliquer plutôt bien au niveau de la reconnaissance de la parole mais peut-être moins au niveau sémantique par exemple. Bien que Clark ne le formule pas en ces termes, nous appelerons *seuil d'ancrage* la quantité ou la qualité des preuves attendues par un locuteur pour considérer que le critère d'ancrage est atteint. On appelera également *statut d'ancrage* l'état courant d'un énoncé relativement au seuil d'ancrage : si la compréhension d'un énoncé atteint ou dépasse le seuil d'ancrage, son statut sera *ancré* (*grounded*).

Différentes conceptions de l'ancrage L'ancrage n'est toutefois pas réalisé qu'au niveau de la compréhension mais on peut considérer que toute connaissance ou croyance peut faire l'objet d'un ancrage. Bunt et al. (2007) rappellent la distinction entre l'ancrage de l'énoncé (*grounding utterance*) et l'ancrage de la proposition (*grounding content*). L'ancrage d'une proposition vise à établir son statut en termes de valeur de vérité dans le terrain commun. L'ancrage de l'énoncé correspond alors à l'ancrage de la proposition : « l'interlocuteur a suffisamment compris l'énoncé pour les buts courants », et cet ancrage est une condition requise pour que la valeur de vérité puisse être établie dans le terrain commun. Par exemple Asher and Gillies (2003) considèrent l'ancrage au niveau de la vérité des propositions ou au niveau des relations rhétoriques entre ces propositions. Dans l'exemple de la figure 3.5 ci-dessous, B n'est pas d'accord avec la proposition « They gave Peter the new computer ». Bien qu'elle soit comprise, cette proposition ne peut rejoindre le terrain commun étant donné le désaccord sur sa vérité, mais la proposition « B a suffisamment compris A_1 pour les buts courants« est, elle, susceptible d'y être établie.

A_1 : They gave Peter the new computer
B_2 : No, they gave JOHN the new computer

FIG. 3.5 – Exemple de désaccord sur la vérité

Dans l'exemple de la figure 3.6 (p. 56), c'est la relation entre les deux propositions « John went to jail » et « He was caught embezzling funds from the pension plan » qui est remise en question. Bien que B soit d'accord avec la première proposition (et éventuellement avec la deuxième), il ne considère pas cette deuxième proposition comme une cause de la première.

En dehors du niveau d'ancrage, les auteurs divergent même sur ce que l'on appelle *grounding*. Dans les travaux (Gaudou et al. 2006a; Gaudou et al. 2006b), l'ancrage

A : John went to jail. He was caught embezzling funds from the pension plan.
B : Yes, John went to jail,
B : but he did so because he was convincted of tax evasion.

Fig. 3.6 – Exemple de désaccord sur la relation rhétorique

correspond au statut d'une information qui est publiquement exprimée et acceptée comme vraie par tous les participants. Cette approche vise à représenter ce statut de manière logique au niveau du groupe. On peut rapprocher ce point de vue sur l'ancrage de Saget and Guyomard (2006) qui consiste à définir la modalité du statut d'ancrage en termes d'acceptation collective et note que cette acceptation peut ne valoir que pour deux participants[4]. Ces approches visent avant tout à définir la relation entre les individus et les propositions logiques, mais ne prennent pas en compte l'élaboration dialogique des croyances de compréhension en situation de divergence d'interprétation. Au contraire, dans Ginzburg (2007) ou Purver (2004b), l'ancrage concerne les problèmes d'interprétation, mais correspond plutôt au processus d'instanciation des paramètres contextuels et ce processus peut être dialogique dans le cas où ces paramètres ne peuvent pas instanciés de manière autonome. Face aux différentes définitions de l'ancrage que l'on peut trouver dans la littérature, nous resterons fidèle à la définition de Clark : l'ancrage est le processus dialogique qui vise à établir une proposition dans le terrain commun. Nous nous focaliserons cependant sur la compréhension de l'énoncé, c'est-à-dire sur l'ancrage de la proposition « l'interlocuteur a suffisamment compris l'énoncé pour les buts courants ».

Dans cette perspective, l'intérêt du modèle des contributions est double. D'abord en exprimant un critère d'ancrage, il formule un principe motivant la manifestation *négative* mais également *positive* de la compréhension. Ensuite, en introduisant une unité dialogique, il explicite le fait que les contributions sont des actions conjointes et que le dialogue est avant tout une activité collaborative. La convergence de l'interprétation n'est cependant pas garantie lorsque le critère d'ancrage est atteint. En effet les participants peuvent croire à tort qu'un énoncé est ancré alors que ce n'est pas le cas (voir Cherubini et al. 2005), le critère d'ancrage indiquant tout au plus une *croyance* de compréhension. Cette croyance est d'autant plus nécessaire que les buts courants l'imposent, et si elle n'est pas assez certaine, les participants ne peuvent poursuivre le dialogue à cause de l'existence potentielle d'une divergence.

Le modèle des contributions souffre toutefois d'un problème majeur. Il n'est en effet pas formulé de manière computationnelle mais décrit le processus d'ancrage tel qu'il peut être observé dans des corpus. Ce défaut a entraîné de nombreuses critiques. Nous présentons ici ces critiques et les modèles qui en découlent.

[4]Par exemple deux individus peuvent s'accorder sur une dénomination sans que cette dénomination soit partagée par d'autres individus ou ne soit reliée à la réalité.

3.2.1 Modèle des *grounding acts*

Le modèle des *grounding acts* de Traum (1992, 1994, 1999) est un des premiers à critiquer les aspects non-computationnels du modèle des contributions en soulevant un problème important lié à la définition du principe d'ancrage. Celui-ci entraîne selon Traum le problème de l'*acceptation récursive* : si pour pouvoir jouer son rôle d'acceptation une contribution doit être acceptée, alors aucune contribution ne peut jamais être close. En effet pour que la phase d'acceptation d'un énoncé A_1 produit par A se termine, B doit produire une preuve de sa compréhension de A_1 dans son énoncé B_2. Mais comment savoir si cette preuve a bien été reçue par A ? B doit attendre que A manifeste sa compréhension de B_2 en A_3. Mais comment savoir si A_3 est bien compris ? B doit à nouveau effectuer un énoncé B_4, et en attendre une preuve de compréhension, etc. En fait ce problème est analogue au problème des généraux qui veulent coordonner une attaque en utilisant un moyen de communication défectueux. Halpern and Moses (1990) montrent en effet qu'il est impossible de garantir une croyance mutuelle avec un média défectueux à cause de l'incertitude liée à la réception de l'acceptation. La solution de Clark and Schaefer (1989) est de supposer la diminution de la force des preuves à donner, par exemple une répétition peut être acceptée par un *acknowledgement* et un *acknowledgement* par une attention continue. Mais cela n'explique pas pourquoi il est possible d'initier une nouvelle contribution pertinente malgré l'incertitude de la compréhension de la preuve reçue.

La solution du modèle des *grounding acts* est de partir du principe que l'ancrage est réalisé à un certain niveau communicatif à l'aide d'actes dédiés appelés *grounding acts*. Le problème de l'acceptation récursive peut être alors résolu en supposant que les *grounding acts* n'ont pas à être acceptés : leur compréhension est supposée certaine et ils terminent alors la récursion. Cette hypothèse entraîne une structure plate de dialogue : les *discourse units* (DU) sont des séquences d'énoncés qui rassemblent les énoncés adjacents qui peuvent être ancrés ensemble. Le processus d'ancrage est alors vu comme un enchaînement de *grounding acts* dont le but est de changer l'état d'une DU, jusqu'à l'amener à l'état ancré. Le modèle définit sept types de *grounding acts*, résumés dans le tableau 3.1 et chaque changement d'état est modélisé par une transition dans un automate à états finis.

Label	**Description**
initiate	crée une nouvelle DU
continue	ajoute du contenu à la DU courante
acknowledge	donne une preuve de compréhension de la DU courante
repair	corrige une mauvaise compréhension de la DU courante
ReqRepair	requiert une correction de la DU courante
ReqAck	requiert une preuve de compréhension de la DU courante
cancel	stoppe l'ancrage de la DU, en la laissant non-ancrée et non-ancrable

TAB. 3.1 – Liste des *grounding acts*

L'intérêt du modèle des *grounding acts* est de fournir une version computationnelle de l'ancrage. Cependant cette version s'éloigne significativement du modèle des

contributions sur deux aspects. Le premier tient à la nature des unités discursives employées. D'abord, une unité discursive dédiée à l'ancrage est difficile à mettre en oeuvre en raison du fait que l'ancrage ne peut être totalement indépendant du niveau intentionnel. A ce sujet Stirling et al. (2000) notent la difficulté à délimiter les frontières d'une DU en cherchant à annoter manuellement un corpus. Mais surtout, les DU sont des structures plates d'énoncés alors que les contributions sont des structures récursives. Traum (1994) justifie la structure plate comme solution au problème d'acceptation récursive. Nous estimons toutefois que ce problème est causé par une interprétation trop stricte du processus d'ancrage. Il n'est pas nécessaire en effet d'exiger qu'une preuve de compréhension soit ancrée avant qu'elle puisse ancrer un énoncé. On peut adopter un point de vue interne au processus d'ancrage pour clarifier le problème : lorsque A produit son énoncé A_1, et que B donne une preuve de sa compréhension en B_2, A peut légitimement estimer si B a compris A_1 et alors pour lui la phase d'acceptation de A_1 se termine et donc pour lui A_1 peut être ancré. Pour B en revanche, il doit attendre A_3 pour estimer si sa preuve de compréhension a été reçue, et alors s'il est satisfait de la preuve manifestée en A_3, il peut considérer la phase d'acceptation de A_1 terminée. Dans cette perspective, l'ancrage d'un énoncé ne se produit pas au même moment pour les deux participants. Pour autant, rien ne garantit que la preuve de compréhension a bien été comprise, et dans le cas contraire l'ancrage peut être alors erroné, conformément à l'objection de Cherubini (2005). Ce n'est toutefois pas le choix du modèle des *grounding acts*.

Le second aspect qui l'oppose au modèle des contributions concerne justement l'ancrage par défaut des *grounding acts*. Nous suggérons que cette hypothèse n'est pas réaliste. D'abord un participant peut ne pas comprendre ou mal comprendre un *grounding act* mais le modèle ne spécifie pas comment prendre en compte la réinterprétation nécessaire à sa réintégration. Elle ne pourrait même pas être possible dans le modèle original puisque les *grounding acts* ne font pas partie du contenu d'un énoncé et n'initient pas alors de DU. Ensuite cette hypothèse ne rend pas compte des doubles *acknowledgements*, pourquoi observe-t-on un « OK » suivi d'un « OK » si ce n'est parce que le second « OK » vient manifester la compréhension du premier « OK » ? Nous pensons alors qu'un modèle d'ancrage doit conserver l'hypothèse que toute communication peut être faillible, y compris celle des preuves de compréhension. Un modèle d'ancrage complet doit construire une unité pour chaque élément dont l'interprétation est potentiellement divergente. C'est d'ailleurs le choix qui a été fait dans l'implémentation du modèle sous forme d'état d'information, où chaque contribution y compris les *grounding acts*, initie systématiquement une nouvelle DU (Matheson, Poesio, and Traum 2000). L'idée de construire une unité pour chaque élément de la communication a été développée par le modèle des actions communicatives (Poesio and Traum 1997), formulé dans le paradigme de la DRT (Kamp and Reyle 1993), dans lequel on distingue des micro-événements communicatifs qui peuvent être ancrés indépendamment. Ce type de modèle semble pertinent mais, à notre avis, reste difficile à implémenter.

Les évolutions du modèle des *grounding acts* tendent à affiner le statut d'ancrage du modèle initial et à prendre en compte différents *degrés d'ancrage*. Par exemple

Funakoshi and Tokunaga (2006) identifient la force d'une preuve avec le degré d'ancrage qu'elle permet d'établir. Ils distinguent alors :

- niveau 0 : aucune preuve d'ancrage
- niveau 1 : la preuve fournie par A manifeste que A croit avoir compris un énoncé antérieur de B sans que B puisse l'affirmer avec certitude (typiquement un « ok »)
- niveau 2 : la preuve montre que A a compris l'information fournie par B, B peut avoir davantage confiance dans l'interprétation de A qu'au niveau 1 mais sans en être certain non plus (typiquement une répétition)
- niveau 3 : la preuve montre que A a compris l'information avec certitude (typiquement l'action qui suit l'interprétation est la bonne)

Ce degré d'ancrage est en particulier utilisé pour effectuer des préférences sur la cible d'une réinterprétation (voir p. 165).

Différents degrés d'ancrage sont également explicités dans Roque and Traum (2008). Ils sont utilisés, avec le modèle initial des *grounding acts* pour déterminer la quantité de preuves nécessaire à l'ancrage. Cette quantité est conditionnée par deux facteurs : d'abord de manière statique, la nature des informations données influe sur la quantité de preuve, ce que nous avons appelé le seuil d'ancrage (par exemple dans leur application, les cibles d'attaque militaire nécessitent un haut niveau d'ancrage), ensuite de manière dynamique, en fonction du statut d'ancrage courant, conformément au modèle initial. Neuf degrés d'ancrage sont alors définis : inconnu, non-compris, non-confirmé, accessible, signal-accepté, signal-accepté+, contenu-accepté, contenu-accepté+, supposé. Ils correspondent à un raffinement du modèle initial, par exemple la distinction signal/contenu ou le degré « supposé » (correspondant à un ancrage extérieur, dans leur application une information donnée au préalable par écrit) sont absents du modèle initial. Comme les auteurs le suggèrent, les évolutions futures du modèle devront considérer un niveau variable d'ancrage.

Le modèle des *grounding acts* capture bien l'intuition de Clark en ce qui concerne la recherche de compréhension sous-jacente à chaque énoncé : le niveau de l'ancrage est parallèle au niveau intentionnel. Il permet un ancrage bi-directionnel (prise en compte de questions de l'utilisateur ou du système) ainsi que les auto-corrections (une correction est à l'initiative du locuteur ou de l'interlocuteur indifféremment). Toutefois le niveau intentionnel peut rejoindre le niveau de l'ancrage lorsque le dialogue a justement pour but de clarifier le dialogue, et en particulier lorsqu'il vise à clarifier les preuves de compréhension. Le modèle initial des *grounding acts*, ni ses évolutions, ne considère (explicitement au moins) les problèmes de non-compréhension ou de mauvaise compréhension des preuves de compréhension. Cet aspect est pourtant indispensable dans la modélisation du processus d'ancrage.

3.2.2 Modèles structurels

Les modèles structurels de l'ancrage s'attachent à représenter explicitement la structure du dialogue afin de guider le processus d'ancrage.

3.2.2.1 Modèle des échanges

Le « modèle des échanges » (Cahn 1992; Cahn and Brennan 1999) s'appuie sur la remarque que les systèmes de dialogue gèrent habituellement leur propre non-compréhension et pas celle des utilisateurs. Dans cette optique, Cahn considère que le modèle des contributions va dans la bonne direction en prenant en compte la compréhension mutuelle. Et à l'instar de Traum, Cahn note que les aspects non-computationnels du modèle des contributions le rendent inadapté pour le dialogue homme-machine. Le premier problème vient du fait que seul le *résultat* du processus d'ancrage sous la forme d'une structure de dialogue est représenté et pas la construction de la structure. Le second problème est que cette structure de dialogue est représentée d'un point de vue *omniscient*. La solution proposée par le modèle des échanges consiste à calculer la structure de dialogue de manière incrémentale et ce avec une perspective interne. Mais plutôt que d'adopter une unité dialogique différente comme le modèle des *grounding acts*, le modèle des échanges s'appuie sur la notion de contribution et étend le modèle de Clark et Schaefer en ajoutant le niveau des échanges.

Un échange est une paire adjacente, c'est-à-dire un couple d'énoncés tels que le second membre de la paire est attendu ou préféré lorsque le premier membre a été énoncé (Schegloff and Sacks 1973). La paire adjacente est une structure fréquemment utilisée dans les systèmes de dialogue (Bilange 1992) et permet de définir une version computationnelle du modèle des contributions. Certains phénomènes ne peuvent toutefois pas être analysés en terme de paires, en particulier les énoncés évaluatifs. Le modèle genevois (Roulet et al. 1985; Moeschler 1989) intègre ces énoncés au sein de triplets, des échanges dont le premier membre est appelé l'initiative, le second la réactive, et le troisième, optionnel, l'évaluative. Dans le modèle des échanges, l'évaluation est représentée comme dans le modèle des contributions par un lien d'acceptation entre les énoncés : l'évaluation est située dans la phase d'acceptation de la réactive.

Le modèle des échanges s'appuie sur des règles dont les préconditions sont la preuve de compréhension reçue et la structure courante du dialogue. Le déclenchement de ces règles provoque la mise à jour de la structure de dialogue. Par exemple, si mon partenaire produit une preuve de sa non-compréhension, alors son énoncé initie une contribution dans un nouvel échange de la phase d'acceptation de la contribution non comprise. Le mécanisme le plus intéressant concerne le traitement de la mauvaise compréhension : si mon partenaire produit une preuve de ma mauvaise compréhension, alors ma contribution précédente est *déplacée* dans un nouvel échange dans la phase d'acceptation de la contribution mal comprise. Dans la figure 3.8 p. 61 (dialogue 3.7), ma contribution S_2 est déplacée dans la phase d'acceptation de O_1, lorsque je reçois une preuve de mauvaise compréhension donnée en O_3. Ce traitement traduit le fait que ma contribution précédente ne peut plus être considérée pertinente vis à vis de l'échange initié par mon partenaire.

Il existe toutefois deux versions du modèle des échanges qui présentent des différences importantes, Cahn (1992) auquel nous réferrerons par C92 et Cahn and Brennan (1999) auquel nous réferrerons par CB99.

O_1 : je prends cet hôtel
S_2 : OK, je réserve l'hôtel Ibis
O_3 : non non l'hôtel Lafayette

FIG. 3.7 – Exemple de mauvaise compréhension

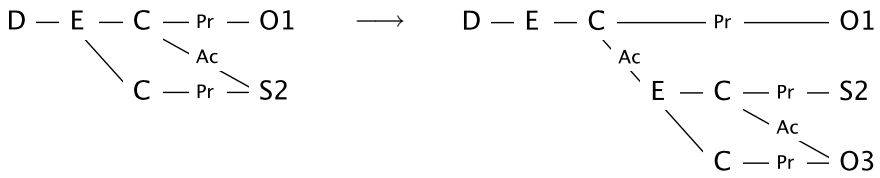

FIG. 3.8 – Modification après une preuve de mauvaise compréhension fournie en O_3

La première différence tient au point de vue selon lequel le modèle est défini. Dans C92, le modèle peut s'appliquer à l'utilisateur et au système indifféremment alors que dans CB99, le modèle est défini exclusivement du point de vue du système et est alors moins générique. En revanche, C92 ne permet d'intégrer que les énoncés du partenaire et ne considère pas les énoncés produits par le participant dont on prend le point de vue, alors que CB99 permet d'intégrer à la fois les énoncés reçus par le système et les énoncés produits par lui. Ces différences suggèrent qu'un modèle d'ancrage devrait d'une part être neutre vis à vis du point de vue et d'autre part intégrer tous les énoncés, qu'ils soient produits ou reçus.

La seconde différence est que C92 s'appuie exclusivement sur la structure de dialogue précédente et la preuve de compréhension alors que CB99 s'appuie également sur le fait que l'énoncé de l'utilisateur propose une nouvelle tâche ou non. La mise à jour dans le modèle CB99 est alors plus difficile puisqu'elle requiert de déterminer au préalable dans quelle mesure l'énoncé peut être relié à la tâche.

La troisième différence tient aux preuves de compréhension sur lesquels ils s'appuient. Le modèle C92 considère trois types de preuves :

- NOTUNDERSTOOD : le locuteur manifeste ne pas avoir compris suffisamment son partenaire, et correspond à la non-compréhension de Hirst et al. (1994)
- UNDERSTOODNOTRELEVANT : le locuteur manifeste que son partenaire ne l'a pas suffisamment compris, et correspond à la mauvaise compréhension
- UNDERSTOODRELEVANT : le locuteur manifeste être à la fois satisfait de son interprétation et de l'interprétation du partenaire de son énoncé antérieur.

Alors que CB99 n'emploie que deux catégories de preuves ACCEPTABLE et NOTACCEPTABLE, en faisant disparaître la preuve UNDERSTOODNOTRELEVANT. En fait, la preuve NOTACCEPTABLE peut être assimilée soit à un NOTUNDERSTOOD soit à un UNDERSTOODNOTRELEVANT en fonction du fait que l'utilisateur propose une nouvelle tâche ou non. Mais en conséquence, elle empêche le système de prévenir l'utilisateur de sa mauvaise compréhension.

Aussi différents qu'ils soient, les deux modèles partagent les mêmes défauts. Le

principal défaut est qu'ils ne considèrent pas que la preuve de compréhension doive être interprétée avant de pouvoir être utilisée. Ce problème a au moins deux conséquences : d'abord, en ne considérant pas l'interprétation comme étape du processus d'ancrage, ils ne représentent certaines situations que d'un seul point de vue. Par exemple, si je reçois une preuve de *ma* mauvaise compréhension, je dois réviser ma structure de dialogue (voir figure 3.8), mais rien n'est dit si moi-même je constate que mon partenaire me comprend mal et que je doive produire une preuve de *sa* mauvaise compréhension. Pour ce faire, il est nécessaire que j'interprète son énoncé et le considère comme non pertinent. En éludant le problème de l'interprétation, ni C92, ni CB99 ne rendent compte de cette situation[5].

La seconde conséquence, plus importante, est que ces deux modèles ne peuvent gérer l'incompréhension des preuves de compréhension. A l'instar du modèle des *grounding acts*, le modèle des échanges considère que les preuves de compréhension sont comprises par défaut, et ce n'est pas une hypothèse souhaitable dans un modèle d'ancrage. Pour pouvoir considérer l'incompréhension des preuves, il est nécessaire d'inclure l'interprétation comme étape du processus d'ancrage et en particulier le fait que les preuves de compréhension doivent être interprétées avant d'être utilisées. Dans ce cas seulement, on peut déterminer que faire lorsqu'une preuve de compréhension n'est pas comprise ou mal comprise. En outre, la présence de l'interprétation dans une définition du processus d'ancrage est d'autant plus justifiée que le choix de la preuve à manifester est directement relié au résultat de l'interprétation.

3.2.2.2 Modèle de Bilange

Le modèle des échanges peut être rapproché du modèle de Bilange (Bilange 1992; Bilange and Magadur 1992), inspiré du modèle genevois dans lequel on maintient une structure du dialogue. Il permet en particulier de guider le déclenchement d'un acte de dialogue en fonction de la structure. On peut concevoir dès lors des actes déclenchés en présence d'un certain motif dans la structure, par exemple s'il y a deux ou trois échanges de clarification. Le modèle s'appuie toutefois sur une séparation entre les trois fonctions, initiation, réaction et évaluation. Nous suggérons au contraire, comme dans le modèle des échanges, qu'une initiation ou réaction peut avoir valeur d'évaluation et préferons conserver l'idée que l'évaluation est avant tout une relation entre les contributions.

3.2.2.3 Modèles de Luzzati et de Lehuen

Le modèle des échanges peut également être rapproché du modèle de Luzzati (1995) dans lequel deux types d'actes de dialogue sont distingués : les actes qui ont pour but de faire progresser la tâche et qui se situent sur un axe *régissant*, et les actes qui ont pour but de clarifier l'interprétation et qui se situent sur un axe *incident*. Le modèle de Luzzati décrit une structure qui est proche de la structure du modèle des

[5]CB99 prend toutefois en compte l'interprétation dans un cas précis, lorsque l'utilisateur manifeste un Acceptable et ne propose pas de tâche. Nous estimons cependant que la prise en compte de l'interprétation doit être systématique.

échanges. Deux aspects sont particulièrement intéressants dans ce modèle. D'abord l'erreur n'est pas vue comme un phénomène à éviter à tout prix mais elle constitue au contraire le point de départ de l'apprentissage. En nos termes, une divergence d'interprétation traduit une divergence de terrain commun et en conséquence clarifier ou discuter l'interprétation d'un énoncé doit conduire à améliorer le terrain commun. Le second point concerne le fait, à l'instar du modèle de Bilange, que la perception de la structure par le système peut l'aider à éviter les situations de blocage dans lesquelles il cherche à comprendre inlassablement un énoncé. De la structure dérivent des *variables interactionnelles* qui traduisent l'état courant du dialogue, comme par exemple la profondeur de l'échange courant. Si ces variables indiquent une situation trop problématique, le système peut choisir d'abandonner l'ancrage d'un énoncé et de produire l'énoncé suivant sur l'axe régissant.

Le système COALA de Lehuen (1997b) implémente le modèle de Luzzati en s'appuyant sur un principe hypothético-déductif : le système construit son interprétation en posant des hypothèses sur sa propre compréhension. Face à une hypothèse erronée (détectée par exemple à l'issue d'une requête applicative qui a échoué) le système peut la remettre en question et déclencher des stratégies de vérification et d'apprentissage de règles d'analyse linguistique/intentionnelle. De notre point de vue, le système COALA est toutefois incomplet : il considère prioritairement la non-compréhension du système sans aborder la mauvaise compréhension (une remise en cause d'hypothèse issue d'une manifestation négative de l'utilisateur), ni les divergences d'interprétation de l'utilisateur. Il tombe alors sous la critique de Cahn and Brennan (1999) et ne peut donc pas être décrit comme une implémentation d'un processus d'ancrage. Néanmoins l'apprentissage de règles d'analyse grâce à l'interaction rejoint tout à fait notre problématique de la coordination du terrain commun et l'on peut dire que COALA est une étape importante dans cette direction.

3.2.3 Modèle des croyances faibles

En fait, ni le modèle des échanges, ni le modèle des *grounding acts* ne permettent clairement d'atteindre le critère d'ancrage du modèle des contributions car ils ne modélisent pas explicitement les croyances de compréhension nécessaires à l'ancrage et font de plus l'hypothèse irréaliste que la preuve de compréhension est automatiquement comprise. Le modèle des croyances faibles (Bunt, Morante, and Keizer 2007) apporte une solution à ces deux problèmes.

D'abord il considére que les participants effectuent des hypothèses sur la compréhension de leur partenaire et que la confirmation de ces hypothèses permet d'atteindre le critère d'ancrage. Dans des conditions normales de communication (où les participants parlent la même langue, et où il n'y a pas trop de bruit), un locuteur peut légitimement supposer, sans en être certain, que son interlocuteur a compris son énoncé. Le modèle part du principe que cette supposition, formulée par l'attitude doxastique de *croyance faible* que l'interlocuteur a compris l'énoncé, est mutuellement entretenue par les deux participants. Un mécanisme de renforcement (*strengthening principle*) permet alors d'atteindre la croyance mutuelle de compré-

hension et donc le critère d'ancrage. La première clause de ce principe, appliquée à la compréhension d'une précondition d'un acte de dialogue, est formulée par[6] :

> A dialogue participant strengthens the weak belief link in a "weak mutual mutual belief" concerning a precondition of a dialogue act that he has performed, when (1) he believes that the corresponding utterance was correctly understood ; (2) he has evidence that : (2a) the other dialogue partner also believes that ; and (2b) they both have evidence that they both have evidence that (1) and (2a) are the case. (Bunt, Morante, and Keizer 2007)

En conséquence, deux participants A et B doivent avoir tous les deux la preuve qu'ils croient que B a compris un énoncé A_1 de A, et ils doivent en outre avoir la preuve qu'ils ont cette preuve. Du point de vue de B, celui-ci doit attendre deux preuves de compréhension : il doit d'abord attendre une preuve que sa compréhension de A_1 est bonne, et celle-ci lui est donnée en A_3, mais il doit en outre faire savoir à A qu'il croit avoir bien reçu A_3 et pour savoir si c'est le cas, il doit attendre l'énoncé A_5 qui lui permet de déduire qu'ils partagent bien la croyance de compréhension de A_1. Le cas échéant, le principe de renforcement s'applique et B peut légitimement croire qu'ils ont une croyance mutuelle de compréhension. Le principe de renforcement ne pourrait par exemple pas s'appliquer dans l'exemple 3.9 car les participants n'ont pas reçu de preuve de compréhension de la preuve fournie en B_2.

A_1 : where should I insert the paper ?
B_2 : in the paper feeder
A_3 : (no) the paper to be faxed
B_4 : what did you say ?

Fig. 3.9 – Exemple d'un ancrage différé tiré de Bunt et al. (2007)

Ensuite le modèle permet d'exhiber le besoin de fournir des preuves de compréhension des preuves de compréhension contrairement au modèle des *grounding acts*, ou au modèle des échanges. La clause (2b) nécessite en effet que les participants aient la preuve qu'ils ont une preuve de compréhension.

Cependant, nous estimons que la nécessité d'avoir une preuve de la preuve présente quelques problèmes. D'abord est-ce suffisant ? Bunt (*ibid*) présente deux exemples qui vont en ce sens. Dans le dialogue 3.10, il semble impossible de revenir sur l'ancrage car B a fourni une preuve que, selon lui, le premier énoncé a été ancré, alors que cela lui est possible dans le dialogue 3.11. Il serait alors suffisant de considérer une preuve de la preuve. Mais le dialogue 3.12, où B revient sur l'ancrage *après* avoir donné une preuve de la preuve, semble pourtant plausible.

Si on suit ce raisonnement, il n'y a pas de raison de s'arrêter à la preuve de la preuve, mais on pourrait exiger une preuve de la preuve de la preuve, rejoignant alors le problème d'acceptation récursive soulevé par Traum. Le modèle des croyances faibles suppose qu'il n'est pas nécessaire d'avoir une preuve de la preuve de la preuve

[6]La seconde clause permet de l'appliquer à une information reliée à la tâche.

A_1 : The next train is at 11:02
B_2 : At 11:02
A_3 : That's correct
B_4 : Okay thanks
A_5 : You're welcome
B_6 : * I thought it would be at 11:08

FIG. 3.10 – Exemple d'impossibilité de retour sur l'ancrage tiré de Bunt et al. (2007)

A_1 : The next train is at 11:02
B_2 : At 11:02
A_3 : That's correct
B_4 : I thought it would be at 11:08

FIG. 3.11 – Exemple de possibilité de retour sur l'ancrage tiré de Bunt et al. (2007)

A_1 : The next train is at 11:02
B_2 : At 11:02
A_3 : That's correct
B_4 : Okay thanks
A_5 : You're welcome
B_6 : But wait... I thought it would be at 11:08

FIG. 3.12 – Contre-exemple de retour sur l'ancrage

pour appliquer le principe de renforcement, et alors effectue la même hypothèse que le modèle des *grounding acts* à un niveau supérieur. D'un certain point de vue, le principe de renforcement est analogue aux heuristiques de troncature de Clark. En effet, on se contente de deux niveaux : une preuve de compréhension et une preuve de la preuve, afin de pouvoir appliquer le principe de renforcement[7].

Ensuite, est-ce nécessaire ? Nous estimons que l'on peut appliquer le même raisonnement que pour résoudre le problème de l'acceptation récursive : à notre avis, il n'est pas besoin de fournir des preuves de compréhension d'une preuve de compréhension pour que cette dernière puisse jouer ses effets. Il est toutefois nécessaire de fournir une preuve pour ancrer une preuve, mais nous supposons que ses effets ne sont pas conditionnés à son ancrage. Si la preuve est mal comprise, alors les effets qu'elle provoquera relativement à l'ancrage seront erronés, et si les participants détèctent cette mauvaise compréhension, il la corrigeront, corrigeant en cela ses effets. Par exemple, dans le dialogue 3.12, nous pensons que A n'a aucune raison de douter de l'ancrage de son énoncé lorsqu'il reçoit la preuve de compréhension de B en B_2. En revanche B doit attendre l'énoncé suivant de A pour estimer s'il a bien compris

[7]Ce n'est toutefois pas exactement la même chose, car chez Clark, l'heuristique de troncature consiste à ne vérifier que quelques croyances partagées et en inférer la croyance mutuelle. Dans le modèle des croyances faibles, il existe au préalable une croyance mutuelle faible, qui est transformée en croyance mutuelle grâce au principe de renforcement. Nous notons simplement que la troncature ne s'applique pas directement sur la croyance mutuelle mais sur le principe de renforcement.

l'énoncé initial.

Le modèle des croyances faibles présente néanmoins d'énormes avantages par rapport au modèle des *grounding acts* ou au modèle des échanges. Il introduit en particulier le concept de *Feedback Chaining* qui traduit les effets de la compréhension d'une preuve de compréhension : un acte de dialogue qui permet de comprendre une preuve de compréhension provoque les effets de cette preuve de compréhension et peut alors avoir des conséquences non-locales. Par exemple dans le dialogue 3.9 (p. 64), l'énoncé A_3 manifeste la mauvaise compréhension de B de A_1 mais cet énoncé A_3 n'est pas compris par B : ses effets doivent alors différés, jusqu'à ce que la question de clarification de B en B_4 trouve une réponse. Lorsque c'est le cas, la mauvaise compréhension de B lui devient manifeste et il peut réinterpréter correctement l'énoncé A_1.

3.2.4 Stratégies de mise à jour du terrain commun

Larsson (2002) critique le critère d'ancrage du modèle des contributions et cette critique peut également s'appliquer au modèle des croyances faibles. Il affirme que la mise à jour du terrain commun n'est pas nécessairement conditionnée à la présence d'une preuve de compréhension. Il distingue alors trois types de stratégies de mise à jour qu'un système peut adopter : la stratégie *optimiste* où un énoncé est ancré dès qu'il est émis sans possibilité de révision, la stratégie *pessimiste* (celle de Clark) où un énoncé est ancré dès qu'on en reçoit une preuve de compréhension positive, et la stratégie *prudente*, utilisée dans le système Ibis (Larsson 2002), où un énoncé est ancré dès qu'il est émis mais en autorisant des révisions.

Les stratégies optimiste et prudente semblent à première vue totalement incompatibles avec le principe d'ancrage du modèle des contributions. En effet la stratégie optimiste n'est adéquate qu'en supposant une communication *certaine* puisqu'en cas de divergence les participants ne peuvent pas réviser leur interprétation, empêchant alors toute convergence. C'est le cas par exemple des *grounding acts*. Quant à la stratégie prudente, elle élimine le besoin de fournir une preuve positive de compréhension. Pourquoi le faire en effet puisque l'énoncé est ancré *par défaut* ?

Pourtant ces stratégies ne sont pas inutiles et peuvent modéliser un comportement non collaboratif : par exemple dans un monologue le locuteur produit des énoncés successivement sans en attendre de preuve de compréhension. La nécessité de fournir systématiquement des preuves de compréhension est également mise à mal par Allwood avec la prise en compte d'actes unilatéraux (Allwood 1995), ou par Koschman dans sa critique du terrain commun (Koschman and LeBaron 2003). Mais, même en contexte collaboratif, Brenner (1999) note qu'un locuteur peut produire plusieurs énoncés d'affilée et employer des expressions anaphoriques à des référents qui viennent d'être introduits, par exemple « je veux une chambre double, et je veux que *cette chambre* ait une baignoire ». Le locuteur ne semble pas attendre de preuve de compréhension de « une chambre double » pour y référer. L'hypothèse du modèle de Brenner, formulée à l'aide d'une modalité en logique épistémique, est que le locuteur *suppose* que l'interlocuteur a compris et que cette condition lui permet

de poursuivre. Pour autant, le locuteur ne peut être *certain* de la compréhension de son interlocuteur, d'où la motivation de l'interlocuteur à fournir des preuves de sa compréhension. Bien que la modélisation en soit différente, l'hypothèse du modèle de Brenner est très proche de celle du modèle des croyances faibles.

Est-ce que le modèle des contributions est capable de rendre compte de ce phénomène ? Nous estimons que oui, en supposant que le locuteur attend au moins l'*attention continue* de son interlocuteur pour poursuivre, et celle-ci correspond à la preuve minimale de compréhension dans la hiérarchie de Clark et Schaefer. En effet le locuteur ne peut pas légitimement supposer que son interlocuteur a compris si celui-ci n'est pas attentif[8]. Cependant en fonction du seuil d'ancrage, l'attention continue de l'interlocuteur peut être suffisante ou non pour considérer l'énoncé ancré. Par exemple la stratégie prudente considère que les participants ont pour seuil d'ancrage la seule attention continue[9]. Mais en conséquence, elle élimine le besoin de fournir d'autres preuves positives de compréhension puisque l'attention continue est suffisante pour ancrer un énoncé.

Stratégies dialogiques d'ancrage Les stratégies de mise à jour de Larsson ont été interprétées différemment par O'Brien (2002) qui associe les conditions d'ancrage à la stratégie dialogique du système : dans la stratégie optimiste ou prudente, le système considère l'énoncé ancré par défaut et peut poursuivre, mais dans la stratégie pessimiste le système doit demander systématiquement une confirmation de son interprétation[10]. Ce choix est souvent décrit comme une demande de confirmation *implicite* ou d'une demande *explicite* (Larsson 2003). Considérons les exemples suivants :

U : je veux aller à Paris
(1) S : Quand désirez-vous partir ?
(2) S : OK. Quand désirez-vous partir ?
(3) S : A Paris. Quand désirez-vous partir ?
(4) S : Vous voulez aller à Paris. Quand désirez-vous partir ?
(5) S : Voulez-vous aller à Paris ?

FIG. 3.13 – Différentes confirmations possibles

Les cinq exemples diffèrent quant au type de manifestation de compréhension effectuée par S. Dans les quatre premiers, S estime avoir compris suffisamment l'énoncé pour poursuivre et effectue alors une demande *implicite* de confirmation :

[8]Le locuteur ne peut déduire cela que si cette inattention est involontaire. Si au contraire, l'interlocuteur est inattentif volontairement, le locuteur peut seulement déduire que son interlocuteur ne *veut* pas comprendre, et dans ce cas on sort du modèle collaboratif.

[9]La stratégie optimiste ne peut quant à elle pas être définie dans le modèle des contributions, mais ce n'est d'ailleurs pas souhaitable dans la mesure où faire l'hypothèse de communication certaine est irréaliste (particulièrement dans les systèmes de dialogue).

[10]Le terme *stratégie d'ancrage* employé par Larsson correspond davantage à une définition du moment où l'ancrage est atteint, ici O'Brien utilise ce terme pour décrire les actions, en termes de preuves, que le système doit effectuer pour ancrer un énoncé.

(1) initiation d'une contribution pertinente, (2) *acknowledgement*, (3) répétition et (4) démonstration. Dans le cinquième, la demande de confirmation est explicite. Comment expliquer ces choix ? Le principe de moindre effort collaboratif permet de relier la décision de la preuve à manifester à la confiance du système en son interprétation. Moins le système a confiance en son interprétation, plus il estime que la probabilité de rencontrer un problème est élevée. En conséquence il suppose que la convergence nécessitera plus d'effort collaboratif, et il devra alors davantage expliciter sa compréhension pour éviter cet effort. Dans le cinquième exemple, l'incertitude est trop élevée pour pouvoir poursuivre et résulte dans une manifestation explicite de cette incertitude. Heeman et al. (1998) proposent par exemple de relier la quantité de preuves à produire directement avec le score de reconnaissance de la parole :

> The amount of evidence should depend on the speech recognition results. If the speech recognition score of the best recognition result is close to the score of the next highest competitor (or a garbage hypothesis), the system should give stronger evidence of understanding, perhaps even paraphrasing the users response, or explicitly asking for a confirmation. (Heeman, Johnston, Denney, and Kaiser 1998)

Le score de confiance peut n'être toutefois pas limité au seul score de reconnaissance de la parole mais devrait couvrir différents aspects de la compréhension. Plusieurs travaux cherchent à associer un score de confiance global à l'interprétation (Schlangen 2004; Gabsdil and Lemon 2004; Purver, Ratiu, and Cavedon 2006). Les modèles proposés fonctionnent de manière similaire : plusieurs interprétations sont générées à chaque niveau (signal, lexical, syntaxique, sémantique, référentiel, rhétorique, etc.), et chaque niveau est associé à un score de confiance. Les interprétations sont ensuite ordonnées selon un score de confiance global calculé par pondération de chaque niveau. La meilleure interprétation est ensuite sélectionnée, et si son niveau de confiance est trop bas, une question de clarification est requise afin de l'infirmer ou de la confirmer. Toutefois la confiance que le système a en son interprétation n'est qu'un des paramètres qui guident le choix de la preuve à fournir.

Modèles numériques d'ancrage Certains modèles explicitent les paramètres qui guident la production des preuves. Par exemple Traum (1999) propose de considérer le critère d'ancrage comme une mesure de l'utilité à ancrer quelque chose[11]. Cette mesure est directement associée au coût du non-ancrage en termes de dégradation de la tâche. Plus quelque chose risque de dégrader la tâche, et plus son ancrage doit s'avérer nécessaire. D'autre part, plus quelque chose est déjà ancré, et moins un nouvel ancrage est nécessaire. Enfin il faut considérer le coût de l'action destinée à réaliser l'ancrage. S'il est trop élevé à utilité constante, il réduit l'intérêt d'ancrer quelque chose. Traum donne une formule rassemblant tous ces paramètres pour

[11] Traum utilise alors « critère d'ancrage » dans un sens différent de Clark, nous préférons toutefois conserver le sens de critère d'ancrage comme *principe* motivant la recherche de la convergence, et ce principe peut être paramétré par différentes caractéristiques, dont l'utilité ou le seuil d'ancrage.

déterminer l'utilité d'une action d'ancrage. Mais la façon de les déterminer reste entière (voir toutefois Walker 1994a, Walker 1994b pour une estimation du coût cognitif de traitement des preuves).

Paek and Horvitz (2000) proposent également une théorie de l'ancrage qui va dans ce sens. Ils cherchent à modéliser l'ancrage en termes de coût de l'action destinée à l'ancrage et de probabilité d'incompréhension. L'optimum se situe lorsque le coût est minimal pour une probabilité d'incompréhension maximale. Mais ce modèle ne considère pas explicitement les différentes actions d'ancrage possibles.

Skantze (2007) propose une méthode pour évaluer l'utilité de différentes actions (accepter l'énoncé, manifester une compréhension sous forme de paraphrase, clarifier l'énoncé en posant une question, et rejeter l'énoncé) en s'appuyant sur un critère très proche de celui de Traum :

> Choose a grounding action, so that the sum of all task-related costs and grounding costs is minimised, considering the probability that the recognition hypothesis is correct. (Skantze 2007)

Skantze calcule le coût relié à une tâche de localisation spatiale en mesurant l'efficacité d'un concept par son pouvoir discriminant : dans cette tâche, un concept est d'autant plus utile qu'il discrimine mieux les lieux. Il calcule le coût de l'ancrage grâce au nombre de syllabes impliquées dans un dialogue de clarification. Grâce à ces paramètres, le système peut choisir l'action la plus adéquate à effectuer. Ces modèles sont intéressants car ils tendent de manière numérique à prévoir le moindre effort collaboratif et à déterminer la stratégie d'ancrage en fonction de celui-ci. Cependant, ils doivent impérativement être couplés à une gestion « plus symbolique » du dialogue pour parvenir à gérer les sous-dialogues de clarification et considérer les phénomènes d'incompréhension des preuves de compréhension. Il est nécessaire en effet de pouvoir intégrer les effets des preuves non comprises ou mal comprises *a posteriori* et ce n'est possible qu'en maintenant une certaine structure du dialogue.

3.3 Terrain commun, processus d'ancrage et robustesse

La définition de la robustesse sous l'angle du terrain commun et du processus d'ancrage a plusieurs avantages par rapport à une définition sous le seul angle de la résolution des problèmes. Un critère d'ancrage permet en effet de motiver la recherche de l'intercompréhension en instaurant un but de convergence du terrain commun. En conséquence, d'une part on motive la recherche de la convergence *en cas de divergence*, mais on motive en outre la recherche de la convergence *en l'absence de divergence*, autorisant alors la manifestation positive de la compréhension. D'autre part, on motive la recherche d'une convergence *locale*, c'est-à-dire à propos de l'interprétation d'un énoncé donné, mais on motive également la recherche d'une convergence *globale*, c'est-à-dire à propos de l'établissement d'un terrain commun. Enfin, cette définition permet d'adopter une hypothèse plausible sur les *causes* des

divergences d'interprétation : si l'interlocuteur n'obtient pas la même interprétation que celle désirée par le locuteur, c'est que le locuteur ou l'interlocuteur avait une vision erronée du terrain commun.

Nous estimons toutefois que les approches de l'ancrage que nous avons présentées apportent des réponses non satisfaisantes quant au critère selon lequel le contenu d'un énoncé peut rejoindre le terrain commun. D'abord, (1) on ne peut pas faire l'hypothèse d'une certaine fiabilité de la communication pour conditionner la mise à jour du terrain commun, comme c'est le cas pour le modèle des *grounding acts* ou du modèle des échanges. Toute interprétation peut être divergente, particulièrement dans le dialogue homme-machine. D'autre part, il nous est nécessaire (2) de résoudre le problème de l'acceptation récursive afin de déterminer précisemment le moment où l'on peut déclarer un énoncé suffisamment ancré. De plus, on ne peut pas adopter (3) un critère statique d'ancrage en vertu du fait que les exigences de compréhension peuvent être variables. Enfin (4) on peut mettre en question la nécessité d'une croyance mutuelle infinie comme dans le modèle des croyances faibles.

Il est nécessaire, à notre avis, (1) de considérer paradoxalement que l'ancrage ne peut être parfait. En conséquence directe du fait que la communication ne peut être fiable, l'ancrage procédant du dialogue ne peut l'être également. Nous pensons que la notion de preuve de compréhension peut permettre de modéliser un processus d'ancrage non-fiable. La preuve de compréhension du modèle des contributions est le moyen pour transmettre une croyance de compréhension et ce moyen n'est pas fiable : un participant peut ne pas comprendre ou mal comprendre cette preuve. La non-fiabilité du processus d'ancrage peut alors être décrite par la non-fiabilité de ces preuves : leur mauvaise compréhension provoque *un ancrage erroné* et leur non-compréhension *stoppe le processus d'ancrage*. En conséquence, un processus d'ancrage doit considérer ses propres conditions d'échec en gérant l'incompréhension des preuves de compréhension et leur réinterprétation. Mais, y compris dans des situations plus simples, la compréhension doit pouvoir être remise en cause, par exemple pour gérer la mauvaise compréhension, il doit être possible de croire à la convergence dans un premier temps, et prendre en compte de nouveaux indices indiquant la divergence. Nous pensons que la stratégie d'ancrage doit être conditionnée à l'évolution de ces indices de convergence. En effet le résultat d'une stratégie donnée est observable grâce aux indices de convergence qui résultent de son application : si une stratégie s'avère inefficace, on peut le savoir en notant qu'une divergence demeure ou empire, et il doit alors être possible de changer de stratégie.

Ensuite, (2) il est nécessaire de nous interroger sur le moment où un énoncé peut être déclaré suffisamment compris pour rejoindre le terrain commun en évitant l'écueil de l'acceptation récursive. Nous devons toutefois apporter à ce problème une réponse différente de celle de Traum ou de Larsson, en ne posant pas comme hypothèse un ancrage par défaut qui annulerait la motivation à fournir des preuves positives de compréhension. En particulier, les exigences de compréhension des participants peuvent être un moyen de considérer la quantité d'information nécessaire pour considérer un critère ancrage non récursif.

Il faut à notre avis (3) partir du principe que la quantité d'information nécessaire

à l'ancrage peut être variable, d'une part en fonction du type d'informations à ancrer et d'autre part en fonction du statut courant de l'information. On peut adopter un critère qui considère un énoncé ancré lorsque les indices de convergence de son interprétation sont *suffisants*. Ce niveau de « suffisant » peut être très lâche et permettre un ancrage optimiste, ou au contraire très exigeant et nécessiter des preuves particulièrement fortes. Le principe de moindre effort collaboratif permet d'éclairer la production des preuves. Plus l'interprétation risque d'être problématique et d'entraîner une divergence, plus les preuves à manifester devront être explicites afin de réduire le moindre effort collaboratif.

Enfin (4), la nécessité de croyance mutuelle infinie peut être remise en question. Si elle a une justification théorique indéniable, en particulier que les participants doivent produire leurs énoncés avec l'assurance de leur compréhension, nous estimons que sa justification pratique n'est pas assez importante. Par exemple elle ne peut traduire le fait que l'intensité avec laquelle autrui est pris en considération est variable, fonction de la nécessité, possibilité ou volonté de le faire. Parfois autrui n'est pas du tout pris en compte lors de la production d'un énoncé. L'hypothèse de croyance mutuelle infinie est selon nous beaucoup trop forte. Particulièrement dans les systèmes de dialogue, où l'intérêt de disposer potentiellement d'une infinité de croyances reste selon nous à démontrer. Cela ne remet toutefois pas en cause la nécessité d'adopter des croyances partagées, en particulier des croyances sur la croyance de compréhension d'autrui.

4 Méthodologie

Notre méthodologie consiste à explorer la problématique de l'ancrage et de la robustesse en vérifiant si le processus d'ancrage est à même d'améliorer la compréhension d'un système de dialogue. Nous proposons alors de considérer dans un premier temps une évaluation du terrain commun qui existe *au préalable* entre l'utilisateur et le système de dialogue, autrement dit de la seule capacité de robustesse interne indépendamment de ses capacités de dialogue (partie 2). Cette partie a pour objectif de déterminer avec précision les capacités d'interprétation d'un système afin de mieux appréhender en quoi un processus d'ancrage peut améliorer la robustesse. Dans un second temps, et après avoir défini un processus d'ancrage qui respecte les caractériques que nous avons évoquées, nous serons à même d'évaluer le terrain commun qui *peut* exister entre l'utilisateur et le système grâce au processus d'ancrage (partie 3).

Nous adopterons toutefois quelques restrictions dans cette démarche : nous n'évaluerons pas le système dans le cadre d'une tâche extérieure au dialogue pour concentrer nos efforts sur le processus d'ancrage lui-même. Notre intérêt est alors d'évaluer la capacité du système à atteindre l'interprétation du locuteur et non pas à réaliser le service sous-jacent.

De plus, nous n'évaluerons pas cette compétence à tous les niveaux d'interprétation mais focaliserons l'étude au phénomène de la référence. La référence est un bon exemple d'activité interprétative collaborative : le locuteur propose une expression référentielle qu'il croit désigner un référent, et l'interlocuteur interprète l'expression référentielle et l'accepte si elle semble désigner un référent pertinent. La référence s'appuie sur le terrain commun : un locuteur n'utilise une expression (dans le modèle collaboratif) que s'il croit qu'elle permet à son interlocuteur d'obtenir une représentation du référent. De plus ce choix est pratique si on considère que la relation qui existe entre la forme de surface utilisée et le référent peut être facilement remise en cause, ce qui n'est pas forcément le cas au niveau sémantique ou à d'autres niveaux.

Enfin nous restreignons également la portée du terrain commun au terrain commun *conversationnel*, c'est-à-dire l'ensemble des interprétations et des faits relatifs à un dialogue donné, et non pas cet ensemble aux frontières floues qui recouvre l'ensemble des connaissances, hypothèses ou faits mutuellement entretenus par deux participants. Nous n'effectuerons pas alors d'apprentissage de ressources linguistiques ou conceptuelles au cours d'un dialogue ou d'un dialogue à l'autre comme dans Lehuen (1997b).

4.1 Evaluation de la robustesse interne

Nous serons tout d'abord concernés par la robustesse interne et l'évaluation de la compréhension existante au préalable entre un utilisateur et un système de dialogue donné. Il existe plusieurs manières de classer les types d'évaluation que l'on peut trouver dans la littérature. La première distinction est celle de « boîte noire » opposée à « boîte transparente » (Jokinen 1996; Chaudiron and Choukri 2008). Dans le premier type d'évaluation, on considère le module à évaluer comme une boîte noire en vérifiant sa sortie pour un ensemble d'entrées données. On cherche alors à répondre à la question « est-ce que le système fait bien ce qu'il doit faire ? ». L'évaluation en « boîte transparente » donne accès au programme lui-même et cherche à comparer la structure interne du module. La question que l'on se pose est « comment est-ce le système fait ce qu'il doit faire ? ». La seconde distinction concerne le niveau de précision des résultats et on distinguera les évaluations *quantitatives*, des évaluations *qualitatives*. Pour les premières, souvent associées aux évaluations de type « boîte noire » on cherche à évaluer le module sur la plus grande quantité de données possibles en fournissant des mesures globales de performances. Pour les secondes, les jeux de tests sont plus réduits mais on peut cibler précisemment les phénomènes à évaluer.

Concernant la compréhension dans les systèmes d'interprétation, on peut par exemple vérifier la sortie au niveau de la requête applicative comme dans le programme ARPA-ATIS (Pallett et al. 1995). La proportion de bons résultats fournit une indication de la compréhension du système mais celle-ci reste très superficielle. On ne peut en particulier pas évaluer la compréhension sur des phénomènes précis, ni être certain que la compréhension est bonne (une compréhension erronée pouvant malgré tout entraîner un bon résultat « par chance »). Cette évaluation manque de *prédictivité* mais également de *généricité* comme le soulignent Antoine and Caelen (1999). Dans le paradigme DQR (Zeiliger, Caelen, and Antoine 1997) de type qualitatif « boîte noire », l'évaluation de la compréhension est réalisée en soumettant le système à un énoncé D (Déclaration) et à une question Q portant sur cet énoncé. Il s'agit alors de vérifier si la réponse R à cette question est bien celle attendue. Par exemple :

D : Ce serait pour partir demain à Vannes
Q : Aller à Vannes ?
R : oui

Fig. 4.1 – Exemple de test DQR au niveau explicite

Cette évaluation permet de cibler précisemment les phénomènes à évaluer, et les auteurs proposent sept niveaux évaluables dans cette approche, trois niveaux de compréhension (syntaxique, sémantique ou référentiel) : l'*information explicite* (illustré dans la figure 4.1) où la principale difficulté concerne la variabilité structurelle de la langue (hésitations, répétitions, corrections, etc), l'*information implicite* qui concerne les ellipses ou les anaphores, l'*inférence* (équivalence sémantique, sous-entendus et sens commun), et quatre niveaux dialogue : le *niveau illocutoire* (s'agit-il

d'une requête directe, indirecte, d'une confirmation, etc.), la *reconnaissance du but* (l'intention sous-jacente), la *pertinence de la réponse* (contraintes de production de la réponse) et la *pertinence de la stratégie* (réussite ou rapidité de la transaction). Si les premiers niveaux semblent « faciles » à évaluer puisqu'on considère les énoncés D indépendemment les uns des autres, les niveaux dialogue sont beaucoup plus difficile à évaluer puisqu'on doit fournir au système un dialogue déjà réalisé sur lequel portent les questions Q. Les deux difficultés sont l'interprétation *a posteriori* du dialogue dans lequel le système n'a joué que le rôle d'observateur et la complexité des questions de contrôle. On pourrait avoir par exemple le dialogue suivant, avec D_S les énoncés du système et D_C les énoncés du client :

D_S : Vous m'avez demandé un billet aller-retour ?
D_C : Oui pour Paris, SVP
D_S : Pour aller à Paris ?
Q : Cette question est-elle nécessaire ?
R : non

FIG. 4.2 – Exemple de test DQR au niveau pertinence de la réponse

Si le paradigme semble tout à fait crédible pour évaluer la compréhension, la complexité nécessaire à mettre en œuvre l'évaluation au niveau du dialogue est très importante. Elle requiert que le système soit capable d'avoir des raisonnements métalinguistiques à propos d'un dialogue auquel il n'a pas participé en répondant à des questions extrêmement complexes. La compréhension de la question est en effet problématique. D'abord elle ne décorrèle pas les différentes capacités interprétatives du système. Dans l'exemple 4.2, le système doit être capable de résoudre le démonstratif « cette question » avant de pouvoir répondre (niveau information implicite). Ensuite elle porte à la fois sur la *compréhension* d'un comportement, et pas uniquement sur la capacité d'avoir ce comportement. Dans l'exemple, il doit non seulement avoir une représentation de la nécessité ou non d'une question, et en plus être capable de *comprendre* le concept de « nécessaire » en tant que relation à cette représentation. Ces problèmes ont entraîné une simplification avec le paradigme DCR (Antoine and Caelen 1999; Antoine et al. 2000) qui ne propose d'évaluation que jusqu'au niveau intentionnel et dans lequel les questions sont uniquement des reformulations. La compréhension de la question se limite alors à un test d'unification entre la représentation du contrôle C et de la déclaration D. L'évaluation positive se ramène à tester si cette unification est compatible avec la réponse R. Ce paradigme nous semble tout à fait adapté mais nous estimons qu'il reste complexe à mettre en œuvre.

L'évaluation de la compréhension que nous avons conduite fut réalisée au sein du consortium MEDIA (projet TechnoLangue). Elle consiste à comparer la production sémantique et référentielle d'un système d'interprétation avec celle annotée manuellement d'un utilisateur prototypique. Il s'agit donc d'une évaluation de type « boîte noire » quantitative portant sur les mêmes niveaux que le paradigme DCR (explicite, implicite, inférence, illocutoire et intentionnel), mais contrairement à ce dernier, elle ne cible pas explicitement les différents niveaux. Elle nous permet surtout de savoir

dans quelle mesure le système et l'utilisateur partagent les mêmes moyens linguistiques et conceptuels d'interpréter un énoncé. Grâce à elle nous pourrons estimer la quantité et la qualité des problèmes de compréhension qui peuvent concrètement se poser à un système dénué de processus d'ancrage. En effet, dans cette évaluation, le système joue le rôle d'un observateur extérieur au dialogue, il ne participe pas à la résolution dialogiques des problèmes ou à l'établissement d'un terrain commun avec les participants que sont le client et le compère.

4.2 Evaluation de la gestion du terrain commun

La seconde partie de notre travail consiste à adjoindre à notre système d'interprétation un module de dialogue capable d'ancrage et à évaluer sa capacité à atteindre l'interprétation de l'utilisateur dans des conditions problématiques.

Nous serons d'abord concernés par la définition d'un processus d'ancrage qui présente les caractéristiques que nous avons évoquées. Nous clarifierons le critère d'ancrage en lui donnant deux définitions complémentaires, en termes épistémiques et en termes structurels, et présenterons ensuite le processus lui-même, inspiré du modèle des échanges ainsi que son implémentation. Nous pourrons alors estimer dans quelle mesure le système, doté d'un processus d'ancrage, *peut* atteindre l'interprétation de l'utilisateur, dans quelles circonstances et par quels moyens. Il est intéressant de nous appuyer sur l'évaluation de la robustesse interne dans la mesure où celle-ci fournit des exemples concrets de problèmes de compréhension.

La plupart des évaluations des système de dialogue en situation réelle s'appuient sur une mesure de la satisfaction de l'utilisateur en fonction de la complétion de la tâche (Walker et al. 2002; Boufaden et al. 2007). Ces évaluations traduisent bien le fait que la finalité des systèmes de dialogue est de satisfaire les utilisateurs. Mais ce type d'évaluation comporte de nombreux biais : le plus important tient à la variabilité des utilisateurs, et à leur propension à collaborer. Pour peu que l'évaluation soit réalisée avec des utilisateurs un peu moins collaboratifs, le taux de complétion de la tâche risque de varier. Dans la mesure où les résultats du système dépendent exclusivement des utilisateurs et que ces utilisateurs ne sont pas prédictibles, nous estimons que ce type d'évaluation ne peut être pertinent qu'à très grande échelle.

Nous proposons plutôt d'évaluer les capacités de convergence du système dans des conditions artificielles en simulant un dialogue de clarification entre deux systèmes de dialogue. Ce contexte d'évaluation est inspiré de travaux sur l'évaluation de l'interaction entre agents logiciels. On peut faire remonter la simulation de dialogue entre agents à Power (1979). Les deux agents cherchent à atteindre un but commun en proposant des énoncés manifestant leurs buts, ainsi que leur acceptation des buts de leur partenaire. Dans Walker (1994a, 1994b), l'efficacité et le coût de traitement de différentes stratégies communicatives sont évalués entre deux agents simulant un dialogue. Ce dialogue n'est toutefois pas réalisé en langue naturelle et ces travaux ne prennent pas en compte l'incompréhension. Le dialogue de clarification a précisemment fait l'objet d'évaluation par simulation dans Araki et al. (1997). Les deux agents cherchent à emprunter le chemin le plus court sur une carte dont ils possèdent

des versions différentes. Les agents font d'abord face à des divergences sur le moyen d'accomplir la tâche. Mais en outre, les auteurs cherchent à vérifier l'impact d'un bruit linguistique dans la simulation. Leur observation, à l'instar de Walker (1994a), est que le fait de fournir explicitement des confirmations améliore de manière importante le succès de la tâche. Plus récemment Allemandou (2007), Allemandou et al. (2007) cherchent à évaluer les systèmes grâce à la simulation déterministe d'un utilisateur. Les énoncés de l'utilisateur simulé sont produits automatiquement à partir d'un corpus annoté ou semi-automatiquement (automatiquement mais avec vérification manuelle) et peuvent inclure des ambiguïtés sur les critères de la tâche ou des hésitations. Ces évaluations de type « boîte noire » présentent un intérêt indéniable de généricité en s'appuyant sur la langue naturelle (à l'instar de DQR ou DCR). Ils présentent également un intérêt de coût, l'évaluation ne nécessitant « que » la construction d'une simulation d'utilisateur.

Nous nous inspirons de ces approches pour évaluer la capacité d'ancrage. Nous nous plaçons dans une situation asymétrique : l'un des systèmes propose des énoncés à interpréter tandis que l'autre système doit prouver qu'il a bien interprété l'énoncé. Conformément au principe d'ancrage, la décision finale de la bonne interprétation revient au premier système. Si celui-ci considère que c'est le cas, il peut proposer un nouvel énoncé. Sinon, en cas de mauvaise compréhension, il doit prévenir le second système de l'existence d'un problème d'interprétation et déclencher un sous-dialogue de clarification. Il est également possible que le problème puisse être détecté par le second système dès qu'il cherche à interpréter l'énoncé (en cas de non-compréhension), et alors celui-ci peut déclencher immédiatement une question de clarification.

Ce contexte d'évaluation permet de vérifier si le processus d'ancrage fonctionne : est-ce que les deux systèmes parviennent à s'accorder sur une interprétation donnée ? Cette question doit prendre en compte les deux systèmes. Si le premier système n'est pas à même d'évaluer correctement la compréhension de son partenaire, ou qu'il ne parvient pas à manifester correctement la mauvaise compréhension, le processus d'ancrage échouera. C'est pourquoi ce type d'évaluation n'évalue pas chaque système indépendamment mais plutôt le couple de systèmes dans leur capacité à ancrer un énoncé. En fait, cette caractéristique nous est utile puisqu'elle nous permet de considérer le système dans les deux positions : en tant que locuteur et en tant qu'interlocuteur. Au niveau de l'ancrage, ce double rôle correspond au double rôle des contributions, comme présentation et acceptation.

Toutes les caractéristiques du processus d'ancrage ne pourront pas être évaluées. En particulier on ne pourra évaluer l'impact des preuves de compréhension qui, bien que fournies dans la perspective du dialogue homme-machine, ne font pas l'objet d'interprétation par les systèmes. Nous suggérons, à l'instar de Allemandou (2007) que cette approche par simulation *ne remplace pas* une évaluation avec des utilisateurs humains. Elle permet en revanche d'isoler toutes les situations problématiques afin de guider le développeur. Nous pourrons observer si le processus d'ancrage parvient à améliorer ou non la compréhension, et surtout dans quels cas et pourquoi il n'y parvient pas. Nous examinerons également les défauts du processus d'ancrage, les problèmes de détection de la compréhension, de la manifestation de cette com-

préhension ou de la réinterprétation.

4.3 Conclusions de la première partie

Les systèmes de dialogue sont confrontés à une multitude de sources de divergence lorsqu'ils interprètent un énoncé. En cherchant à connaître l'intention de l'utilisateur, ils peuvent faire face à tous les problèmes lexicaux, syntaxiques, sémantiques, référentiels, intentionnels qui existent dans la langue elle-même. Ils sont de plus confrontés à de nombreux problèmes issus de leur modélisation imparfaite des mécanismes humains d'interprétation. Ces problèmes d'interprétation peuvent être résolus dans une perspective de détection/correction où le système déclenche des stratégies adéquates lorsqu'il détecte un problème. Cependant, nous favorisons une approche préventive dans laquelle les participants recherchent la convergence même en l'absence de divergence. Les participants coordonnent leur terrain commun afin d'éviter les problèmes ultérieurs. Dans cette perspective nous proposons d'évaluer la capacité de gestion du terrain commun conversationnel au niveau de la référence. D'abord en évaluant sur corpus la quantité et la qualité des problèmes d'interprétation (partie 2). Ensuite, après avoir proposé un critère d'ancrage non récursif et un processus d'ancrage inspiré du modèle des échanges, nous évaluerons par simulation les apports et les limites du processus modélisé (partie 3).

Robustesse interne dans un système d'interprétation

5 Paradigme d'évaluation

Avant de pouvoir considérer une modélisation et une évaluation du processus d'ancrage, il nous est nécessaire de bien appréhender le contexte dans lequel celui-ci peut être mis en oeuvre. Cette seconde partie a pour objectif de présenter une évaluation de la robustesse interne d'un système. L'idée est de confronter un système d'interprétation sémantique et référentielle à la variété des phénomènes linguistiques que l'on peut rencontrer dans le dialogue homme-machine et ce faisant, de déterminer tous les phénomènes qui provoquent des erreurs d'interprétation. L'intérêt de cette évaluation, vis à vis de l'ancrage, est qu'elle correspond à la mesure du terrain commun *communautaire* qui existe au préalable entre un système et un utilisateur. A travers l'évaluation, nous mesurons en effet dans quelle proportion le système et l'utilisateur partagent des connaissances linguistiques ou conceptuelles (au niveau lexical, syntaxique, sémantique ou référentiel). Une fois connues, les divergences d'interprétation nous permettront alors de mieux estimer dans quelle mesure et par quels moyens un processus d'ancrage sera susceptible d'améliorer l'interprétation.

Le premier chapitre de cette partie définit le paradigme d'évaluation MEDIA, le corpus et son annotation. Le second chapitre présente le système que nous avons fait concourir à l'évaluation ainsi que l'adaptation nécessaire à sa participation. Enfin, le troisième chapitre décrit la méthodologie d'évaluation et les résultats.

5.1 Evaluation MEDIA

La campagne d'évaluation MEDIA (décembre 2002 - avril 2006), au sein du projet d'évaluation des technologies de traitement de la langue TechnoLangue, avait pour but d'évaluer les capacités de compréhension des systèmes de dialogue. L'évaluation recouvre deux aspects : la compréhension hors-contexte qui concerne la capacité de construire une représentation sémantique des énoncés en dehors de tout contexte de dialogue et la compréhension en contexte constituée d'une part de la spécification du sens en fonction du contexte et d'autre part de la résolution de la référence. N'ayant pas participé à l'aspect de spécification contextuelle, nous mettons de côté cette partie de l'évaluation en renvoyant le lecteur à Bonneau-Maynard et al. (2008). De plus, étant donné que ces travaux ont déjà été publiés à de nombreuses reprises, nous décrivons ici cette campagne sous un angle critique en nous focalisant sur des aspects peu ou pas décrits dans les publications existantes, en particulier l'impact de différentes contraintes sur l'annotation, l'impact de l'annotation sur l'adaptation

des systèmes ainsi que certains aspects relatifs à la gestion de la compréhension.

Evaluer un unique système de manière indépendante n'est pas trop difficile. La vraie difficulté naît de la nécessité d'évaluer un ensemble de systèmes hétérogènes. En effet, contrairement à l'évaluation mono-système, l'évaluation multi-systèmes nécessite que le protocole d'évaluation respecte les contraintes imposées par chaque système : il doit être défini par consensus entre tous les participants. Il est nécessaire tout d'abord que l'évaluation soit *possible* pour chaque système. C'est-à-dire d'une part que chaque système dispose des outils et des ressources indispensables à l'évaluation, et d'autre part, en considérant ces outils ou ressources, que l'évaluation ne soit pas trop difficile à effectuer. Ensuite il faut que cette évaluation soit *intéressante* pour chaque système. Il serait inutilement coûteux d'évaluer un phénomène peu ou pas géré par un système. Enfin, la contrainte, peut-être la plus importante, concerne justement le *coût* de l'évaluation, coût de la production des outils ou ressources nécessaires à l'évaluation. Ces contraintes doivent être envisagées pour chacun des participants de manière à trouver le protocole optimum, et cette tâche est relativement difficile en fonction de l'hétérogénéité des systèmes.

Le protocole MEDIA était soumis à ces contraintes, et il était nécessaire de satisfaire aussi bien des systèmes de type statistique (Lamel et al. 2000; Bonneau-Maynard and Lefèvre 2005; Servan and Béchet 2006) que des systèmes de type symbolique (Villaneau et al. 2004; Denis et al. 2006b). Pour ce faire, nous nous sommes inspirés du paradigme PEACE (Devillers et al. 2002; Devillers et al. 2003) dans lequel chaque système doit projeter son interprétation dans un formalisme commun à l'aide duquel la comparaison est possible. La manière d'acquérir la représentation finale importe peu, et l'évaluation peut alors être qualifiée d'évaluation « en boîte noire ». Il s'agit d'annoter un corpus de dialogue à l'aide du formalisme commun et de tester les capacités interprétatives des systèmes sur les énoncés de ce corpus en comparant leur interprétation projetée avec l'annotation humaine. Grâce à la comparaison avec ce *gold standard*, la comparaison entre les systèmes devient possible.

5.1.1 Le corpus

Le corpus développé pour la campagne MEDIA concerne une tâche de réservation de chambre d'hôtel. Il a été enregistré en simulant le dialogue par téléphone entre un utilisateur humain, le client, et une machine, dont le rôle était joué par un compère humain selon le protocole du magicien d'Oz (Devillers, Bonneau-Maynard, and Paroubek 2002). Des directives étaient données au client concernant les modalités de la réservation : contraintes de lieux, de dates, du nombre de réservation, de prix, etc. Le comportement du compère était modulé selon sa coopérativité : présence d'erreurs, confirmations explicites ou implicites, etc. Huit catégories de complexité ont alors été définies en fonction de ces directives et des modalités de réservation.

Le corpus était soumis à deux contraintes : il devait être assez important d'une part pour exhiber une bonne diversité des phénomènes linguistiques oraux sur la tâche de réservation d'hôtel, et d'autre part pour permettre aux systèmes statistiques

d'être entraînés, leur apprentissage nécessitant une grande quantité de données. Le corpus est constitué de 1250 dialogues, enregistrés par 250 utilisateurs différents, chacun réalisant cinq scénarii, totalisant environ 15000 énoncés pour 70 heures de parole (Bonneau-Maynard et al. 2005). Il a été ensuite intégralement transcrit et annoté par ELDA selon le formalisme adopté par consensus.

5.1.2 L'annotation

Nous distinguons ici les deux aspects de l'annotation relatifs aux deux phases de l'évaluation : l'annotation hors-contexte des énoncés, et leur annotation en contexte.

5.1.2.1 L'annotation hors-contexte

L'annotation hors-contexte des énoncés correspond à une annotation sémantique littérale, décorrélée du contexte de dialogue dans lequel les énoncés ont été produits. Elle devait être soumise à toutes les contraintes que nous avons évoquées : faisabilité, intérêt et coût. Le formalisme adopté est directement inspiré de celui de PEACE qui respecte ces contraintes. Chaque énoncé est découpé en *segments* signifiants, porteurs des concepts que l'on cherche à évaluer et chaque segment est annoté comme un *triplet* du type *mode*/*attribut*:*valeur* ou +/null lorsque le segment n'apporte pas d'information que l'on désire évaluer. En raison de la contrainte de coût, chaque segment ne peut être annoté que *par un seul triplet*.

Le mode caractérise la polarité du trait sémantique : positive (+), négative (-), interrogative (?) ou optionnelle (∼). Il recouvre des fonctions différentes, illocutoires comme le mode interrogatif, ou directement relatives aux exigences du client vis à vis de la tâche comme le mode optionnel (voir tableau 5.1).

Enoncé	**Mode du trait**
je veux qu'il y ait une baignoire	+/chambre-equipement:bain
je ne veux pas qu'il y ait une baignoire	-/chambre-equipement:bain
est-ce qu'il y a une baignoire ?	?/chambre-equipement:bain
je veux qu'il y ait une baignoire si possible	∼/chambre-equipement:bain

TAB. 5.1 – Exemples d'annotation du mode

Le nom de l'attribut est composé de deux parties : un attribut de base auquel on adjoint des *spécifieurs*. Les spécifieurs raffinent le sens d'un attribut en fonction du contexte, restreint à l'énoncé lui-même lors de la phase hors-contexte. Le tableau 5.2 illustre l'emploi de spécifieurs : le trait +/nombre:1 est raffiné par le spécifieur chambre dans « je veux une chambre », ou par chambre et reservation dans « je veux réserver une chambre ».

Les différents noms d'attributs, les spécifieurs qui peuvent leur être adjoints ainsi que les valeurs possibles des traits sont définis dans une ontologie composée de deux parties. Une partie *générique* recouvre des traits indépendants de la tâche de réservation, qu'ils concernent des concepts que l'on peut retrouver dans tout type

Enoncé	Annotation
j'en veux une	+/nombre:1
je veux une chambre	+/nombre-chambre:1
je veux réserver une chambre	+/command-tache:reservation +/nombre-chambre-reservation:1

Tab. 5.2 – Exemples de raffinements par spécifieurs

de tâche (nombre, temps, nom, localisation, etc.), qu'ils concernent des fonctions dialogiques des énoncés (annulation, demande de répétition, etc.) ou référentielles (lienRef). L'autre partie est *spécifique* à la tâche et décrit les concepts de la réservation : type de chambre, services de l'hôtel, etc. L'ontologie totalise 83 attributs de base et 19 spécifieurs qui, combinés, peuvent engendrer 1121 traits différents.

Par exemple l'énoncé « euh oui l'hôtel dont le prix ne dépasse pas cent dix euros » sera annoté conformément au tableau 5.3.

Segment	Mode	Attribut	Valeur
euh	+	null	
oui	+	reponse	oui
l'	+	lienRef-coRef	singulier
hôtel	+	objetBD	hotel
dont	+	null	
le prix	+	objet	paiement-montant-chambre
ne dépasse pas	+	comparatif-paiement	inferieur
cent dix	+	paiement-montant-entier-chambre	110
euros	+	paiement-monnaie	euro

Tab. 5.3 – Exemple complet d'annotation

Règles d'annotation Les règles d'annotation comprennent une méthodologie de segmentation au niveau de l'énoncé et du syntagme, et des consignes pour l'annotation de chaque catégorie de traits (génériques ou spécifiques) qui puisse traduire la sémantique de l'énoncé. Par exemple, dans la segmentation, on cherche à faire des segments les plus longs possibles, on attache la préposition au syntagme nominal « je voudrais réserver / à Paris » ou encore on sépare les répétitions sans les annoter « je réserver / réser(ve) réserve / pour deux chambres » (« réser(ve) réserve » sera annoté par +/null par exemple). En ce qui concerne le choix des triplets, celui-ci relève d'un certain arbitraire déterminé par les contraintes d'intérêt et de coût. On ne dévoilera pas ici toutes les subtilités de l'annotation mais quelques règles représentatives de cet arbitraire.

Par exemple on considérera les chambres comme des entités dénombrables dont il est important de connaître la cardinalité. Tandis qu'au niveau des hôtels, on considérera par défaut que leur nombre n'est pas prioritaire. Ainsi « une chambre » sera annoté par un triplet indiquant le nombre (+/nombre-chambre:1) alors que « un hôtel »

sera annoté par un triplet indiquant seulement le type de l'entité (+/objetBD:hotel). Néanmoins « deux chambres » ou « deux hôtels » seront annotés de manière similaire, avec un trait d'attribut nombre spécifié par chambre ou hotel. Concernant le pluriel, celui-ci ne sera pas annoté explicitement pour les indéfinis, « des chambres » sera annoté par +/objet:chambre et « des hôtels » par +/objetBD:hotel (ne distinguant alors pas « un hôtel » de « des hôtels »).

On souhaite annoter également les expressions référentielles (voir ci-dessous 5.1.2.1) et alors les règles changent dès qu'il s'agit d'annoter autre chose qu'un indéfini. Par exemple, « cette / chambre » sera annoté par :

+/lienRef-coRef:singulier

+/objet:chambre

La présence du type d'objet est annotée dans le cas de l'indéfini par le spécifieur chambre et dans l'autre par un trait d'attribut objet. De plus la cardinalité qui était présente pour l'indéfini, disparaît et est reportée sur la valeur de l'attribut lienRef-coRef pour les autres expressions référentielles.

D'autres règles concernent la propagation des modes et des spécifieurs, par exemple faire porter le mode interrogatif sur l'ensemble du syntagme sur lequel la question s'applique, ou encore appliquer les spécifieurs selon une certaine portée (le syntagme ou l'énoncé). Les règles sont soumises à la fois à la contrainte de coût en minimisant le nombre de triplets et la contrainte d'intérêt en n'annotant que certains types de phénomènes avec le but de produire une annotation non-ambiguë. Cependant, malgré les efforts déployés pour éliminer l'ambiguïté, deux énoncés à la sémantique différente peuvent malgré tout avoir la même annotation. Par exemple : « j'en réserve deux » et « je veux deux réservations » seront tous les deux annotés par +/nombre-reservation:2, ou encore « peut-on baisser à 95 euros » et « est-ce que c'est en dessous de 95 euros » seront annotés de la même façon :

?/comparatif-paiement:inferieur

?/paiement-montant-entier:95

?/paiement-monnaie:euro.

Dans ce cas, les deux contraintes d'intérêt et de coût entrent en opposition : il était trop coûteux mais pas assez intéressant de développer des règles ou d'ajouter des attributs pour distinguer ces deux sémantiques.

Annotation hors-contexte des expressions référentielles L'annotation hors-contexte de la référence consiste à annoter la *présence* d'une expression référentielle grâce à un trait d'attribut lienRef raffiné par un spécifieur dénotant par la catégorie de l'expression. Les catégories que nous avons employées sont issues du Reference Annotation Framework, RAF (Salmon-Alt and Romary 2004). Ce format définit des catégories de données pour annoter les expressions référentielles (les *markables*), ainsi que les relations de différentes natures qu'elles entretiennent (les *referentials links*). Bien que l'évaluation ne porte pas sur les relations (voir p. 87), nous avons repris les différents types de liens référentiels définis dans RAF, fondés sur le mode d'extraction du référent : coRef, pour les extractions directes (comme « cette chambre »), elsEns, pour les extractions par discrimination (comme « le premier hôtel ») et coDom, pour

les extractions par opposition (comme « les autres hôtels »), voir tableau 5.4. Nous n'avons pas explicitement annoté les relations de partie-tout, comme les anaphores associatives, faute de consensus. A la différence des *markables* de RAF, seuls les déterminants des groupes nominaux (ou pronoms) sont associés à un trait d'attribut lienRef afin de ne pas interférer avec le reste du groupe nominal déjà annoté en sémantique. La valeur du trait lienRef correspond à la cardinalité singulier ou pluriel, en fonction du nombre de référents attendus.

Spécifieur	**Signification**	**Expressions référentielles**
coRef	coréférence : l'expression référentielle désigne son référent par référence directe	pronoms, définis, démonstratifs
elsEns	élément-ensemble : l'expression référentielle désigne son référent en vertu de propriétés sémantiques ou indexicales qui l'opposent à d'autres entités dans un ensemble	ordinaux, superlatifs, relatives, certains pronoms démonstratifs
coDom	co-domaine : l'expression référentielle désigne son référent grâce à un marqueur linguistique d'altérité	altérités

Tab. 5.4 – Types de spécifieurs de lienRef

Afin de satisfaire à la contrainte de coût, seules les expressions référentielles dont *la résolution dépasse le cadre de l'énoncé* ont été annotées. Cela exclut dès lors les référents dont l'antécédent a été introduit dans le même énoncé, les entités nommées et les indéfinis. Les figures 5.1 et 5.2 illustrent l'annotation ou la non-annotation du pronom « il ». Dans la première, l'antécédent, un indéfini est introduit dans le même énoncé et le pronom n'est pas annoté. Dans la seconde, bien qu'il y ait « cet hôtel » qui puisse jouer le rôle d'antécédent, le référent a été introduit dans un autre énoncé et le pronom sera annoté. La capacité d'annoter un énoncé hors-contexte nécessite de considérer le contexte de l'énoncé. Elle ne requiert donc pas uniquement des capacités d'ordre sémantique mais également au niveau référentiel afin de déterminer si un référent a été introduit dans le même énoncé.

je veux / un hôtel / et euh qu'il soit / près de / la mer
+/null
+/objetBD:hotel
+/null
+/localisation-distancerelative:proche
+/localisation-lieuRelatif-general-hotel:mer

Fig. 5.1 – Non-annotation des référents dont l'antécédent est introduit dans le même énoncé

Nous avons également jugé pertinent d'annoter l'indéfini avec altérité « un autre hôtel » à cause de sa fréquence. Cependant dans ce cas, ce n'est pas le référent qui sera annoté mais le référent exclu. En revanche, les définis et les démonstratifs ont été annotés systématiquement, du moins pour les entités relevant de la tâche et tous les pronoms personnels de troisième personne ont été annotés ainsi que certains pronoms possessifs (comme dans « à quel prix sont *vos* chambres ? »).

je veux / cet / hôtel / mais euh est-ce qu' / il / est près de la mer
+/null
+/lienRef-coRef:singulier
+/objetBD:hotel
+/null
+/lienRef-coRef:singulier
+/localisation-distancerelative:proche
+/localisation-lieuRelatif-general-hotel:mer

FIG. 5.2 – Annotation des référents dont l'antécédent est introduit dans un autre énoncé

5.1.2.2 Annotation en contexte de la référence

L'évaluation de la référence devait d'une part répondre aux contraintes générales que nous avons évoquées plus haut (faisabilité, intérêt et coût), et d'autre part satisfaire une contrainte additionnelle de compatibilité avec le paradigme d'évaluation hors-contexte. Ces contraintes ont fortement pesé sur la définition de l'annotation de la référence.

La plupart des approches d'évaluation des capacités référentielles évaluent les relations entre expressions référentielles (Popescu-Belis et al. 2004), et s'appuient sur des formats d'annotation qui se concentrent sur les relations, à l'instar du format des campagnes MUC-6 et MUC-7 basé sur les coréférences (Chinchor and Hirschmann 1997; van Deemter and Kibble 2000) ; ou RAF qui décrit plusieurs sortes de relation. Cependant, en raison de la contrainte de faisabilité, étant donné que tous les systèmes ne pouvaient produire ces relations mais étaient tous capables de fournir une représentation sémantique des référents, il a été décidé d'évaluer non pas les relations mais les *descriptions* des référents. L'annotation en contexte est alors vue comme une extension de l'annotation hors-contexte qui consiste à rajouter de l'information sur les traits de type lienRef.

Une référence est représentée comme un ensemble de référents, chacun décrit par un ensemble de traits sémantiques. Concrètement, on adjoint un champ *reference* à tous les traits lienRef. On notera par exemple {(t1,t2), (t3)} une expression référentielle qui fait référence à deux entités, l'une décrite par deux traits et l'autre décrite par un seul.

Exemple d'annotation en contexte :

je veux / une / chambre double / et / une / chambre simple
+/nombre-chambre:1
+/chambre-type:double
+/nombre-chambre:1
+/chambre-type:simple
est-ce que / **ces** / chambres / ont la douche ?
+/lienRef-coRef:pluriel **reference** = **{**
(+/nombre-chambre:1, +/chambre-type:double),
(+/nombre-chambre:1, +/chambre-type:simple)**}**
+/objet:chambre
?/chambre-equipement:douche

Ce formalisme montre ses limites lorsqu'on veut représenter certains phénomènes. En particulier pour correctement annoter l'ambiguïté il serait nécessaire de disposer d'un niveau supplémentaire autorisant une disjonction de groupes de référents. En l'absence d'un tel niveau, seule une approximation de l'ambiguïté a été réalisée : on annote l'union des différentes alternatives et afin de distinguer cette annotation d'un pluriel, on raffine le lienRef à l'aide du spécifieur ambigu. Cette approximation ne permet toutefois pas d'annoter correctement l'ambiguïté entre deux groupes de référents comme l'illustre l'exemple suivant[1] :

deux / chambres simples / ainsi qu'/ une / chambre simple / et / une / chambre double
    +/nombre-chambre:2
    +/chambre-type:simple
    +/nombre-chambre:1
    +/chambre-type:simple
    +/nombre-chambre:1
    +/chambre-type:double
pour ces / deux / chambres
    +/lienRef-coRef-ambigu:pluriel **reference** = **{**
        (+/nombre-chambre:2, +/chambre-type:simple),
        (+/nombre-chambre:1, +/chambre-type:double),
        (+/nombre-chambre:1, +/chambre-type:simple)**}**
    +/nombre-chambre:2
    +/objet:chambre

Enfin, on emploie également le même type de raffinement par spécifieur pour différencier les expressions d'altérités dans lesquelles le référent est inclus (comme le défini « l'autre »), et celles dans lequelles il y est exclus (comme l'indéfini « un autre »). En l'occurrence les traits lienRef-coDom sont spécifiés par inclusion dans le premier cas et par exclusion dans le second.

Règles d'annotation en contexte Nous avons collectivement défini des règles d'annotation afin de s'adapter aux spécificités de l'annotation hors-contexte tout en maintenant un coût minimal d'annotation, et un intérêt maximal (l'accord inter-annotateur est donné p. 111). Elles définissent comment annoter la description des référents au moyen du formalisme précédemment introduit et se répartissent en trois groupes :

- les règles de portée qui délimitent les énoncés fournissant les traits descripteurs
- les règles de couverture qui spécifient la manière dont sont décrits les référents
- les règles de normalisation qui précisent la forme de la description

Tout d'abord la portée de la description d'un référent est constituée uniquement des traits sémantiques présents dans le dialogue antérieur en excluant l'énoncé courant et les énoncés simultanés en cas de chevauchement. Ce choix est à mettre en parallèle avec la décision de n'annoter hors-contexte que les référents dont l'antécédent a été introduit dans un énoncé précédent.

La couverture de la description soulève le problème des référents décrits par d'autres référents. En particulier comme la chambre est l'objet de la réservation, sa

[1] Ce type d'ambiguïté ne s'est cependant pas présenté dans le corpus de test.

description recouvre l'ensemble des critères de la réservation. Plusieurs propositions ont été étudiées :

- l'annotation *maximale* constituée par la totalité des traits décrivant un référent n'a pas été retenue à cause de son coût d'annotation
- l'annotation *discriminante*, définie par la plus petite description du contexte précédent permettant d'identifier de manière non ambiguë le référent, n'a pas été non plus retenue. En effet en absence d'ambiguïté cette description se limite au type du référent (objet ou objetBD) et ne serait alors pas intéressante à évaluer.
- l'annotation *par récence* composée des traits descriptifs contenus dans l'énoncé le plus récent mentionnant le référent n'a également pas été adoptée. En effet dans la majorité des cas de coréférence, démonstrative ou pronominale, la sémantique de l'expression référentielle est vide, entraînant alors peu d'intérêt d'annoter les expressions qui y réfèrent ultérieurement.

La solution retenue est un compromis entre la contrainte de coût et la contrainte d'intérêt. Elle s'appuie sur le type des entités, qu'elles soient nommées ou non-nommées :

- les entités nommées ou assimilées (hôtel nommé, dates, prix, villes, etc.) ne sont décrites que par un ensemble très restreint de traits comme leur nom ou leur valeur ;
- les entités non nommées (hôtel non nommé et chambre) sont, elles seules, annotées avec la totalité de la description possible, y compris les traits d'autres référents.

Enfin on contraint la description des référents à être en forme normale. La forme normale d'un référent est une description non-redondante, non-contradictoire et totalement spécifiée de celui-ci. La non-redondance (ou minimalité) impose de ne pas décrire deux fois le référent avec le même trait ou d'omettre le type de l'objet lorsque celui-ci est fourni dans un spécifieur (sous peine d'aboutir à des incohérences). La non-contradiction est nécessaire lorsque certains traits de la description d'un référent sont niés à un moment du dialogue. La spécification totale oblige l'annotateur à préférer systématiquement, lorsque plusieurs instances du même trait existe, les traits les plus spécifiés pour décrire le référent.

5.2 Le corpus MEDIA et l'ancrage

On peut soulever la question de l'intérêt du corpus MEDIA vis à vis de notre problématique de robustesse : dans quelle mesure ce corpus peut-il éclairer la gestion des problèmes de compréhension dans le dialogue ? Le corpus MEDIA couvre quelques aspects de la compréhension, en termes d'annotation et de constitution de corpus, mais ces aspects sont tous deux insuffisants.

D'abord, les problèmes de compréhension de l'utilisateur n'ont été annotés que très partiellement, uniquement lorsque l'utilisateur explicite l'existence de problème à l'aide d'un énoncé du type « je ne comprends pas », « je n'ai pas entendu » ou par une question conventionnelle comme « hein ? ». Les occurrences de ces pro-

blèmes, annotées par un trait +/reponse:pasCompris, ne sont trouvées que dans 18 dialogues sur 1250. Cette très faible proportion suggère une sous-annotation des problèmes de compréhension. Deux contraintes principales sont la cause de cette limitation. D'abord l'intérêt d'annoter ces phénomènes n'était pas nécessairement partagé par l'ensemble du consortium. Ensuite la contrainte de coût qui impose l'annotation d'un segment par un unique triplet empêche toute annotation de fonction globale à l'énoncé. Par exemple, « vous avez dit / quel hôtel ? » serait annoté par deux traits +/commande-dial:repetition-demande et ?/objetBD:hotel, où le trait +/command-dial:repetition-demande couvre tous les cas de demande de répétition qui ne sont pas nécessairement causés par des problèmes d'interprétation. On ne pourrait ici pas ajouter davantage d'information en tant que triplets[2]. D'autre part, peu d'intérêt a été porté à l'annotation des marques de compréhension comme les *acknowledgements*. Par exemple tous les « oui » correspondent systématiquement à des +/reponse:oui, et tous les « non » correspondent à des +/reponse:non, qu'ils soient produits en tant que réponse à une question fermée ou pas.

Le second aspect de la compréhension concerne la constitution du corpus dans la mesure où des directives particulières ont été données au compère afin qu'il simule des erreurs de compréhension, ou manifeste de manière explicite ou implicite son interprétation. Ces directives peuvent être très intéressantes étant donné que la réaction de l'utilisateur face à ces attitudes pourrait nous renseigner sur son comportement en situation de dialogue homme-machine. Toutefois ces comportements n'ont pas été annotés localement, et il devient alors difficile d'exploiter automatiquement le corpus dans cette direction. En outre, la simulation de la mauvaise compréhension du système n'intervient en très grande majorité qu'au niveau du signal, se restreint pour la plupart du temps aux noms de ville, et est, à quelques rares exceptions près, systématiquement résolue dans l'énoncé suivant.

Un des cas pourtant que l'on suppose très fréquent semble totalement absent du corpus MEDIA : la non-compréhension ou la mauvaise compréhension des preuves de compréhension. Lorsque le compère simule une incompréhension, il parvient toujours à rétablir un dialogue compréhensible dans l'énoncé qui suit la détection par le client. Les incompréhensions ne durent pas, contrairement à ce qui peut arriver pour le dialogue homme-machine. En conséquence, les abandons dûs à la compréhension ne sont pas présents dans le corpus. Vis à vis de la compréhension, le compère du corpus MEDIA est beaucoup trop semblable à un être humain pour que cet aspect puisse être intéressant et concernant la compréhension dans le dialogue homme-homme on pourra se rapporter aux travaux de Purver (2004b).

Nous estimons donc que les phénomènes intéressants ne sont pas forcément absents du corpus MEDIA mais qu'en l'absence d'annotation correspondante, il est difficile d'évaluer leur intérêt en termes de qualité ou de quantité. Une ré-annotation du corpus pourrait exhiber les cas d'incompréhension manifeste, mais nous n'avons pas procédé à celle-ci lors de cette thèse.

[2]Pour ce faire on pourrait employer un autre moyen, les « méta-annotations » dont la portée est l'énoncé tout entier, mais cette voie n'a pas été explorée pour cet aspect (voir Bonneau-Maynard et al. 2008).

5.3 Conclusion

L'arbitraire des règles d'annotation pour circonvenir aux complexités de la langue naturelle entraîne des difficultés importantes lorsqu'on cherche à évaluer les systèmes. Ceux-ci doivent se plier aux contraintes non-naturelles en termes de faisabilité, d'intérêt et de coût exigées par ce type d'annotation. En particulier, elles ont des conséquences très importantes : les systèmes statistiques, en apprenant l'annotation à partir d'exemples, seront probablement moins sensibles à ces particularités que les systèmes symboliques, qui devront *dégrader* leur interprétation selon des règles arbitraires. Par exemple, la nécessité de n'annoter que les référents dont l'antécédent a été introduit dans un énoncé antérieur et pas dans l'énoncé courant, est un non-sens vis à vis de n'importe quelle théorie de la référence. Ces contraintes résultant du consensus nous sont pourtant obligatoires. Nous présentons le système que nous avons fait concourir à cette évaluation ainsi que sa projection dans le formalisme MEDIA dans le chapitre suivant.

6 Un système d'interprétation

Le système que nous avons développé pour cette évaluation est un système symbolique modulaire et linéaire qui s'appuie au niveau syntaxique sur un analyseur LTAG, et au niveau sémantique sur les logiques de description. La résolution de la référence est effectuée au sein du paradigme de la théorie des domaines de référence. Nous en présentons ici les principales caractéristiques ainsi que la projection de la forme d'interprétation dans le formalisme MEDIA.

6.1 Interprétation syntaxico-sémantique

6.1.1 Représentation de l'énoncé

Avant de préciser comment la forme d'interprétation est construite, nous en indiquons la nature et sa représentation. La représentation de l'énoncé doit permettre de rassembler tout ce qui peut être su d'un énoncé, à différents niveaux d'analyse. Le MultiModal Interface Language, MMIL, (Landragin et al. 2004; Landragin and Romary 2004) a été utilisé avec succès dans plusieurs projets européens MIAMM (Reithinger et al. 2005), OZONE (Gaiffe et al. 2004), AMIGO, et dans le projet MEDIA. Il permet une représentation des événements communicatifs dans un système de dialogue, 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. En particulier pour la langue naturelle, MMIL offre une représentation fine multi-niveaux du contenu propositionnel et de la force illocutoire. Il ne fournit en revanche aucune interprétation logique de l'énoncé mais seulement une *représentation* des différents aspects de l'énoncé. C'est un avantage autant qu'un inconvénient. Un inconvénient car une telle représentation ne peut être utilisée seule, mais seulement en combinaison avec une interprétation. Un avantage car elle est *neutre* vis à vis de l'interprétation ultérieure, ainsi MMIL peut être projeté tantôt comme une sémantique plate sans variables (formalisme MEDIA), avec variables (GenI[1]), comme un graphe conceptuel en logique de description (projection MEDIA), ou comme une formule Prolog (OZONE).

Un acte communicatif est représenté par un *composant*, sous forme de graphe,

[1]GenI (Kow 2006; Kow 2007; Gardent and Kow 2007) est le module de génération que nous avons utilisé pour le processus d'ancrage (voir p. 164.)

contenant les représentations des différentes entités évoquées dans l'acte (*événements* et *participants*) et de leurs relations. Chaque entité est une aggrégation de *traits* (ou attributs) morpho-syntaxiques, sémantiques et pragmatiques. A part quelques exceptions (alternatives et ensembles), les entités ne peuvent contenir d'autres entités. Les traits ou les relations qui décrivent les entités sont de deux sortes : les traits ou les relations MMIL représentent des descripteurs génériques des entités, indépendantes de la tâche (genre, nombre, type d'entité, etc.) et les traits ou les relations applicatifs décrivent des caractéristiques spécifiques à la tâche.

L'énoncé « je veux une chambre double à Paris du vingt au vingt-deux septembre » sera par exemple représenté conformément à la figure 6.1. L'événement locutoire est associé à un contenu propositionnel (*propContent*) composé d'un événement de type Vouloir. Cet événement est relié à un participant déictique correspondant au sujet « je » et à un participant correspondant à l'objet « une chambre double ». Les traits qui décrivent ce participant sont des traits MMIL (*objType*, *refType*, et *cardinality*), ainsi qu'un trait applicatif (*media_aType*).

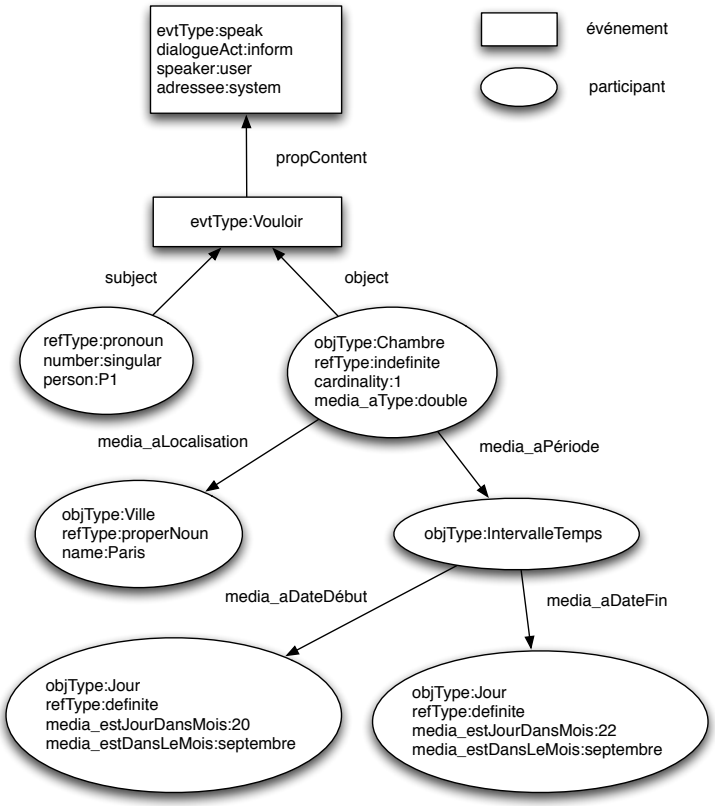

Fig. 6.1 – Composant MMIL correspondant à l'énoncé « je veux une chambre double à Paris du vingt au vingt-deux septembre »

6.1.1.1 Représentation de la dimension référentielle

Chaque expression référentielle analysée est associée à un participant décrivant le référent (qu'on appellera entité référentielle) : son type conceptuel (*objType*), sa cardinalité (*cardinality*), ses modifieurs (*modifier* ou des modifieurs de la tâche), son type référentiel (*refType*) et ses relations. La résolution de la référence consiste à associer chacune des entités référentielles d'un composant avec un identifiant unique (l'identifiant du référent). A l'issue d'une opération de référence, une entité référentielle se voit adjoindre un trait obligatoire *refStatus* indiquant le statut de la résolution (*pending*, *ambiguous*, ou *solved*), un trait optionnel désignant l'identifiant (*mmilId*) si la résolution a réussi, un trait optionnel décrivant la probabilité d'identification (*refProba*) ou une alternative désignant les différents référents possibles (*alt*).

6.1.1.2 Représentation de l'intention

Une représentation de l'énoncé est un pré-requis à la représentation de l'intention du locuteur. Plusieurs propositions existent en ce sens qu'elles soient issues du paradigme de la planification (Cohen and Perrault 1979; Allen and Perrault 1980), ou du paradigme de la logique illocutoire (Searle and Vanderveken 1985; Caelen 2003). Etant donné que nous nous restreignons à la référence une représentation poussée de l'intention n'est pas nécessaire pour notre propos, et nous pouvons nous limiter au fait que chaque participant au dialogue a l'intention de référer lorsqu'il emploie une expression référentielle, et par la suite, qu'il a l'intention de déterminer le référent d'une expression référentielle lorsqu'il en demande une clarification.

6.1.2 Analyse syntaxico-sémantique

Pour construire les composants MMIL, nous employons une analyse syntaxique profonde fondée sur le paradigme des grammaires de type LTAG (Joshi and Schabes 1997). Nous n'avons pas développé d'analyseur syntaxique mais nous nous basons sur l'analyseur LTAG de type chart développé par Lopez (1999), Crabbé et al. (2003). L'analyseur prend en entrée un énoncé analysé lexicalement et produit l'ensemble des arbres de dérivation de l'énoncé.

La construction sémantique est ensuite réalisée en plusieurs temps en s'appuyant sur plusieurs stratégies de robustesse interne : par exemple, afin de s'adapter aux contraintes de l'oral (hésitations, répétitions, corrections, etc.) nous utilisons les analyses *partielles* sorties par l'analyseur en conservant les analyses les plus couvrantes sans intersection et en éliminant toute ambiguïté. Ensuite, la forme sémantique est produite par construction compositionnelle en parcourant récursivement l'arbre de dérivation. Enfin, les formes sémantiques ainsi construites font l'objet de post-traitements, en particulier des vérifications ontologiques : les relations inconsistantes vis à vis de l'ontologie sont par exemple retirées.

Le procédé de construction conduit à représenter le contenu d'un énoncé comme une séquence de composants MMIL. Etant donné la nature du formalisme MEDIA,

nous ne nous sommes pas attachés à produire des représentations complètement relationnelles de l'énoncé. En particulier pour limiter l'ambiguïté artificielle résultant des analyses profondes, nous n'avons que très peu analysé syntaxiquement les relations prépositionnelles. Par exemple, l'énoncé « je veux une chambre double à Paris du vingt au vingt-deux septembre » sera en fait constitué de deux composants, un pour « je veux une chambre double à Paris » et l'autre pour « du vingt au vingt-deux septembre ».

6.1.3 Interprétation de la référence

L'interprétation de la référence consiste à *associer* chaque expression référentielle à une représentation mentale du référent. Nous employons le terme de représentation mentale au sens de Reboul et al. (1997), Reboul and Moeschler (1998), Gaiffe and Reboul (1999) comme une aggrégation de données multi-dimensionnelles à propos d'une entité, de nature linguistique, logique, perceptuelle, etc. La référence ne peut cependant se limiter à cette association. En effet, on ne peut restreindre la référence à l'anaphore discursive, ou à la coréférence, comme c'est le cas dans MEDIA. Nous estimons qu'une définition de la référence qui se limite à rechercher le référent dans le contexte discursif est réductrice. D'abord, elle ignore que le contexte de résolution est beaucoup plus large. Par exemple dans la référence multimodale où les expressions référentielles désignent des objets physiques, les représentations mentales n'ont pas été évoquées préalablement dans le discours mais proviennent de la scène visuelle. Ensuite, elle ignore que la référence concerne également la *restructuration* du contexte référentiel. Il est impossible par exemple de considérer les expressions d'altérité telle que « l'autre chambre » si on se limite à la recherche du référent sans prendre en compte les effets structurants de l'interprétation des expressions référentielles. Une des caractéristiques les plus importantes de la référence concerne son pouvoir discriminant (Olson 1970) : la référence ne se limite pas à l'association de l'expression référentielle et du référent, mais à la discrimination du référent parmi un ensemble de candidats alternatifs. Cette approche permet par exemple de prendre en compte les altérités, dans la mesure où ces dernières désignent justement les candidats alternatifs d'une précédente résolution.

Notre position est de dire que la référence concerne à la fois la relation entre l'expression référentielle et la représentation du référent, en tant que *résultat* de l'opération mais également la gestion du contexte référentiel dans la diversité de restructuration de représentations mentales : construction, modification, fusion, groupement, etc. (Reboul and Moeschler 1998; Salmon-Alt 2001). Ainsi nous considérerons par exemple qu'un indéfini qui provoque la création d'une nouvelle représentation mentale est une opération de référence. En termes d'intercompréhension, l'emploi d'une expression référentielle suppose que l'interlocuteur soit capable non seulement d'identifier le référent évoqué mais également de procéder à la restructuration nécessaire de son espace référentiel.

Nous nous plaçons alors dans la théorie des domaines de référence (Reboul and Moeschler 1998; Corblin 1987; Salmon-Alt 2001; Kumar, Salmon-Alt, and Romary

2003; Pitel 2003; Landragin, Salmon-Alt, and Romary 2002; Denis, Pitel, and Quignard 2006b) qui vise à décrire les opérations de référence dans ce paradigme. Dans cette théorie, l'opération de référence principale est l'*extraction* d'un référent à partir d'un contexte local appelé domaine de référence. Cette extraction provoque la restructuration du domaine, autrement dit, elle « laisse des traces » qui vont au-delà de la simple mise en relation. L'extraction du référent dans des domaines de référence est justifiée par de nombreux exemples issus de corpus multi-modaux (Salmon-Alt 2001; Landragin, Salmon-Alt, and Romary 2002; Landragin 2003). La théorie des domaines de référence peut être rapprochée alors des espaces focaux de Grosz and Sidner (1986) en s'éloignant des théories purement sémantiques comme la DRT (Kamp and Reyle 1993)[2]. Nous ne rentrons pas ici outre mesure dans les justifications théoriques des domaines de référence mais présentons plutôt leur implémentation dans notre système d'interprétation.

Etant donné qu'une représentation mentale peut servir de domaine de référence (par exemple pour l'anaphore associative) ou qu'un domaine de référence peut servir de représentation mentale (par exemple pour les pluriels), nous définirons un domaine de référence comme une représentation mentale dénotant un ensemble d'autres représentations mentales, à laquelle on adjoint une structure de différenciation. Un domaine de référence sera alors constitué d'un *support*, un ensemble de représentations mentales défini extensionnellement ou intensionnellement (comme les domaines issus des indéfinis pluriels), et d'un ensemble de critères de différenciation qui en discriminent les éléments en *partitionnant* le support. Les partitions sont autant de points de vue différents possibles sur un domaine.

Par exemple, on pourra considérer le domaine de référence des films Star Wars dans la figure 6.2 (p. 98). Ce domaine (@7) illustre l'ambiguïté de différenciation d'une expression telle que « le premier Star Wars », il peut s'agir d'une différenciation sur la base de la date (et référer à @1) ou du numéro de l'épisode (et référer à @4). Ce type d'ambiguïté peut être interprété comme deux façons de différencier les éléments du domaine. Etant donné qu'il existe un très grand nombre de façon de partitionner les éléments d'un domaine, il paraît peu plausible de considérer que ces partitions existent au préalable mais plutôt qu'elles sont *projetées* sur un domaine lors de la mention d'une expression référentielle, en fonction de la différenciation possible des éléments du domaine. Par exemple la mention de « le premier Star Wars » ne suppose pas que la partition soit déjà construite, mais seulement qu'elle le soit à l'issue de l'interprétation, et en l'occurrence la construction des deux partitions entraîne l'ambiguïté d'extraction.

A l'issue de l'interprétation, les éléments du domaine sont d'une part discriminés selon le point de vue apporté par l'expression référentielle mais d'autre part cette discrimination s'accompagne d'une *focalisation* du référent ou des référents à l'intérieur du domaine. Cette focalisation, dans le contexte local du domaine permet par exemple d'effectuer des altérités à l'intérieur de ce même domaine. Dans la théo-

[2]Bien que dans la DRT, il y ait également des espaces locaux, les DRS, qui contraignent les interprétations ultérieures, la construction de ces espaces peut avoir des origines non-discursives telles que la perception, la mémoire, etc.

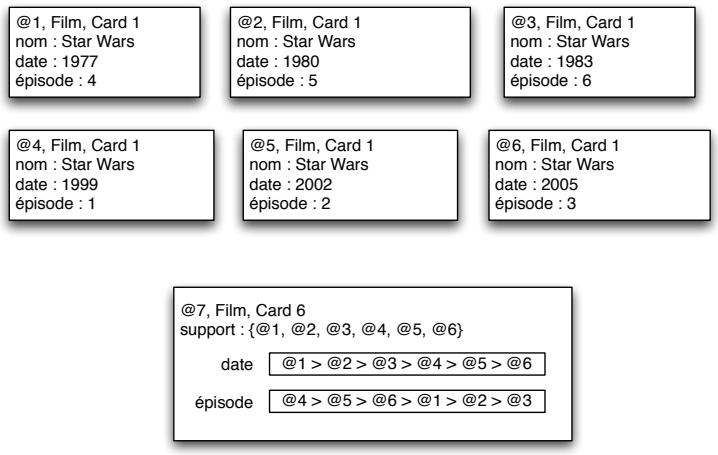

Fig. 6.2 – Exemple d'ambiguïté de différenciation : « le premier Star Wars » peut référer au premier épisode ou au quatrième

rie originale, il n'y a qu'une seule opération de discrimination de type partie-tout. Cette unique opération permet de discriminer à la fois un sous-ensemble du support (par exemple « deux hôtels » suivi de « le premier »), et de discriminer une sous-partie du domaine conceptuellement reliée (par exemple « l'hôtel Ibis » suivi de « la chambre »). Nous avons préféré remettre en cause la notion de partition qui ne s'applique pas au second cas. Nous préférons donc parler de *point de vue* sur le domaine, un point de vue pouvant soit donner accès à un sous-domaine par une relation d'inclusion, soit à d'autres domaines par une relation conceptuelle.

Le contexte global, appelé espace référentiel, sera représenté comme un ensemble de domaines de référence muni d'une relation d'ordre large en fonction de la saillance. Toutefois, par mesure de simplification, nous ne considérerons dans l'implémentation qu'un ordre strict. Cette limitation a des conséquences importantes puisqu'elle nous empêche de représenter correctement l'ambiguïté domaniale et, bien que ce cas puisse se produire, nous considérerons que deux domaines ne peuvent avoir la même saillance[3].

6.1.3.1 Mécanisme de résolution

Le mécanisme de résolution, conformément à la théorie originale des domaines de référence est composé de deux étapes : d'abord la *recherche* d'un domaine satisfaisant les contraintes de l'expression référentielle, c'est-à-dire un domaine dans lequel elle soit susceptible de discriminer un référent, puis, une étape de *restructuration* lors de laquelle on procède à la création de nouveaux domaines, et à la focalisation du ou des référents dans un point de vue.

[3]C'est un défaut notable de l'implémentation qui peut être corrigé en considérant la saillance des domaines indépendamment de leur ordre, voir par exemple une modélisation à l'aide de dendogrammes (Landragin 2003).

Les contraintes associées à l'expression référentielle sont représentées sous la forme d'un domaine de référence *sous-spécifié*. Par exemple les définis « le N » chercheront de préférence à être interprétés dans des domaines fournissant une opposition N versus non-N. Chaque type de référence impose des contraintes sur la sémantique ou la structure de focalisation du domaine dans lequel on cherche à l'extraire. Par exemple l'altérité « l'autre N » recherche un domaine composé de N dans lequel il y ait un élément déjà focalisé. Dans la modélisation originale des domaines de référence (Salmon-Alt 2001), chaque type d'expression n'est associé qu'à un seul type de domaine sous-spécifié. Nous avons cependant rencontré le besoin d'associer *plusieurs* domaines sous-spécifiés à chaque type d'expression. Par exemple le défini peut également être interprété en anaphore associative non-coréférente : dans « je vous propose l'hôtel Ibis, *la chambre* est à vingt euros », « la chambre » doit être interprétée relativement à l'hôtel. Les différents types de sous-spécifications peuvent être ordonnés en fonction de la fréquence d'apparition. Dès lors, on pourra associer une probabilité de résolution en fonction de chaque type de sous-spécification : les domaines sous-spécifiés les moins contraints entraîneront une confiance moins importante dans la résolution. Ce n'est toutefois pas le seul facteur, puisque l'ambiguïté d'extraction joue un rôle important dans la probabilité de résolution. En particulier l'impact d'une incertitude au niveau sémantique peut entraîner une incertitude au niveau de la résolution. Nous ne représenterons pas cet aspect, par exemple comme dans (DeVault and Stone 2007), mais laisserons ouvert la possibilité d'attribuer une probabilité de résolution fondée sur d'autres facteurs grâce à la présence d'un attribut *refProba* associé à chaque entité référentielle après résolution.

Pour une expression référentielle donnée, nous testons, pour chaque domaine présent dans l'espace référentiel, l'ensemble des domaines sous-spécifiés correspondants. Si aucun ne vérifie les contraintes de sous-spécification, nous testons le domaine suivant par ordre de saillance jusqu'à épuiser l'ensemble des domaines existants. Si aucun domaine n'est trouvé, une restructuration particulière intervient pour construire un domaine d'interprétation satisfaisant si l'expression le permet. Si un domaine de référence est trouvé et que l'extraction d'un référent y est possible, alors la phase de restructuration consiste à focaliser le référent dans un point de vue qui soit capable de le discriminer par rapport aux autres éléments du domaine. Les entités référentielles du composant sémantique sont parcourues selon l'ordre d'énonciation, avec l'exception des modifieurs relationnels tels que « une chambre dans cet hôtel » où « cet hôtel » doit être résolu *avant* « une chambre ». L'algorithme est alors le suivant :

1. pour chaque entité référentielle, construire le solveur référentiel associé, et
2. pour chaque domaine existant du contexte D_C par ordre de saillance,
3. pour chaque domaine sous-spécifié du solveur D_S, tester si D_C vérifie les contraintes de sous-spécification D_S, si c'est le cas restructurer D_C en extrayant le référent, sinon tester le domaine suivant
4. si aucun domaine n'est trouvé, tenter de construire un domaine et un référent valide (accommodation)

5. enfin, effectuer des restructurations globales comme les groupements (Denis et al. 2006b; Denis et al. 2006a)

6.1.3.2 Implémentation

Nous décrivons ici l'implémentation du modèle des domaines de référence en détaillant les différents types de solveurs référentiels, la définition des domaines et des contraintes de sous-spécification.

Solveurs référentiels Pour implémenter ces principes nous nous sommes largement appuyés sur l'héritage du langage objet JAVA en décrivant une classe principale dont héritent les différents types de solveurs référentiels. Les trois méthodes génériques sont :

- `UnderspecifiedDomain getUnderspecifiedDomain(int i)`
  renvoie le i-ème domaine de référence sous-spécifié
- `restructurate(RefSpace space, IdentifiedDomain dom)`
  restructure l'espace référentiel en fonction du domaine trouvé
- `IdentifiedDomain newDomain(RefSpace space)`
  renvoie un nouveau domaine

Ainsi, chaque type d'expression référentielle est associé à ces trois méthodes. Les solveurs implémentés sont : Definite, DefiniteAlterity, DefiniteOrdinal, Indefinite, IndefiniteAlterity, Demonstrative, Pronoun, ProperNoun, Interrogative (voir p. 170), et MetaReference (voir p. 157). Chacun d'entre eux regroupe différents domaines sous-spécifiés, différentes méthodes de restructuration et différentes façons de créer un nouveau domaine. Ils permettent de résoudre des anaphores directes, associatives, des altérités, des ordinaux ou des ellipses. Comme le montrent les résultats de l'évaluation (cf section 7.3), les cas partiellement traités recouvrent l'anaphore associative multiple (mention de plusieurs hôtels, suivi de « la chambre »), les restrictions de sélection pronominale « ils acceptent les animaux » ou les modifieurs relationnels « la chambre dans cet hôtel ».

Implémentation des domaines Nous représentons les domaines de référence comme des concepts en logique de description auxquels on adjoint des points de vue potentiellement focalisés. Nous ne détaillons pas ici les logiques de description, en renvoyant le lecteur à Franconi (1994). Les logiques de description sont particulièrement adaptées à l'implémentation des domaines de référence en tant que concepts. D'abord, au niveau de la représentation, la sémantique d'un concept est un ensemble d'individus défini en intension ou en extension à l'instar du support d'un domaine de référence. Ensuite, au niveau de la recherche d'un domaine, nous pouvons directement traduire la sous-spécification du type en tant que subsomption de concepts et facilement représenter l'anaphore hypéronymique comme dans « J'ai vu un chien. L'animal était féroce. ». On peut de plus considérer les rôles afin de gérer l'anaphore associative : dans « Je vous propose un hôtel, la chambre est à vingt euros », l'expression « la chambre » recherche un domaine susceptible d'être relié au concept de

Chambre par un rôle donné[4]. On peut également gérer les restrictions de sélection fournies par le cotexte, par exemple pour les pronoms comme dans « ils acceptent les animaux » où l'on doit rechercher quelque chose susceptible d'accepter les animaux.

Le principal problème posé par cette modélisation vient de la gestion des pluriels. Classiquement en logique de description les individus sont représentés comme des instances de la Abox (voir par exemple l'application des logiques de description à la génération d'expressions définies dans Gardent and Striegnitz 2007). Cependant, si l'on procède ainsi, les ensembles issus d'expressions plurielles doivent également être représentés comme des instances puisqu'on peut y référer, par exemple en les reliant par un rôle *aElement* avec les instances qu'ils contiennent. Il est nécessaire toutefois de gérer certaines propriétés des ensembles, comme l'inclusion et cela s'avère difficile. On peut s'appuyer sur une modélisation adéquate des pluriels en logique de description, comme $\mathcal{ALCS}$ dans Franconi (1993), mais l'absence d'implémentation nous a conduit à trouver une solution alternative, en représentant les pluriels comme des concepts. Il est toutefois impossible de décrire la cardinalité des concepts dans le langage et cette caractéristique nous est pourtant nécessaire pour représenter « deux hôtels ». Nous avons alors utilisé le fait qu'il est possible de représenter la cardinalité d'un rôle afin d'introduire des concepts décrivant la pluralité :

- $\mathsf{Cardn} \doteq \exists n\, \mathsf{aElement}$
- $\mathsf{Plural} \doteq \exists > 1\, \mathsf{aElement}$

Les concepts Cardn sont introduits dès qu'une expression cardinale est mentionnée. Par exemple la mention de « deux hôtels » provoque la création d'un concept Card2 dans la Tbox et sa représentation est alors $\mathsf{Hotel} \sqcap \mathsf{Card2}$. Ce faisant on peut bénéficier des opérations d'inférence définies sur les concepts comme la subsomption (par exemple $\mathsf{Card2} \sqsubset \mathsf{Plural}$).

La représentation des points de vue peut également s'appuyer sur les concepts et les rôles. Il suffit de considérer que, pour les points de vue opérant une partition, deux sous-domaines sont des sous-concepts disjoints d'un concept support donné, et que pour les points de vue opérant une relation conceptuelle, les sous-domaines correspondent à des concepts qui peuvent être reliés par un rôle au concept support[5]. La focalisation néanmoins nécessite de devoir conserver la trace des opérations de référence et il nous est nécessaire de représenter cette focalisation à l'extérieur de la Tbox.

Il est cependant indispensable de pouvoir considérer la focalisation de groupes d'individus, mais cet aspect n'est pas modélisé dans la théorie originale des domaines de référence. Pour modéliser la focalisation de domaines pluriels nous avons introduit dans (Denis, Pitel, and Quignard 2006b; Denis, Pitel, and Quignard 2006a) des domaines contenant l'ensemble des domaines pluriels d'un type donné (appelés

[4]Il est toutefois nécessaire de restreindre l'ensemble des rôles permettant de faire des anaphores associatives, cela est possible par exemple en utilisant des hiérarchies de rôles et en décrivant les rôles susceptibles d'anaphores associatives comme des sous-rôles d'un rôle anaphoreAssociative.

[5]Pour les pluriels, nous avons supposé une lecture distributive de la relation conceptuelle (voir Corblin 1987). Cette lecture a d'ailleurs soulevé d'importants problèmes pour les anaphores associatives multiples avec génériques, typiquement « l'hôtel Ibis et l'hôtel Lafayette » suivi de « la chambre » (voir section 7.3 p. 115).

super-domaines). La construction des domaines pluriels est effectuée *a posteriori*, mais elle est limitée aux référents de même type, introduits dans le même énoncé et uniquement aux référents dont on suppose une existence extra-linguistique[6]. Les super-domaines permettent de conserver des traces de l'activation des groupes de référents, mais ils nécessitent en contre-partie de ne pas faire l'hypothèse de disjonction puisque les sous-domaines correspondant aux groupes peuvent ne pas être disjoints.

La création de nouveaux domaines doit s'accompagner de la création de nouveaux concepts. Cependant, un problème de performance nous a conduit à simplifier ce traitement. Le moteur d'inférence Racer (Haarslev and Möller 2003) que nous avons utilisé était trop lent vis à vis de la reclassification de l'ontologie nécessaire à cause de l'insertion de nouveaux concepts issus de nouveaux domaines. C'est pourquoi nous n'avons associé que des concepts préalablement existants aux domaines sans ajouter de nouveaux concepts. Cette simplification a des conséquences importantes puisque l'on ne peut pas exprimer la négation correctement : un hôtel différent de l'hôtel Ibis sera représenté comme un hôtel qui ne s'appelle pas « Ibis » alors qu'il devrait être représenté comme un hôtel disjoint de l'hôtel Ibis. Nous n'avons effectué l'assertion de nouveaux concepts que dans le cas des pluriels énumérés, où il s'avère nécessaire de représenter conceptuellement le groupe d'objets afin d'autoriser les reprises plurielles comme « les hôtels ».

Au final les domaines de référence seront des objets JAVA auxquels on associe un type (l'entrée logique de la représentation mentale) formé du concept et de sa cardinalité, une source (indiquant la modalité de création du domaine, linguistique, conceptuelle, gestuelle ou visuelle), un ensemble de points de vues décrivant chacun une alternative de sous-domaines (soit par une relation de subsomption, soit par un rôle conceptuel) ainsi que la focalisation d'un de ces sous-domaines et une entrée linguistique correspondant à la dernière entité référentielle qui a activé le domaine.

Implémentation des contraintes Les contraintes sur lesquelles nous avons basé les domaines sous-spécifiés sont indiquées dans la figure 6.3. Nous distinguons les contraintes *booléennes* qui vérifient si un domaine possède une caractéristique donnée, des contraintes d'*identification* qui vérifient une caractéristique et permettent d'identifier un sous-domaine (par exemple pour l'altérité, il ne suffit pas de savoir qu'il existe un sous-domaine focalisé mais également de savoir lequel). Les contraintes booléennes incluent des contraintes de source (l'origine du domaine), des contraintes conceptuelles vérifiées à l'aide de Racer (TypeSubsumer, TypeEqual ou TypeRelated), des contraintes sur la cardinalité (Cardinal, Plural ou UnspecifiedPlural — qui vérifie que le domaine est pluriel sans que ces sous domaines soient spécifiés), des contraintes de genre (contraintes sur l'entrée linguistique) ou une contrainte de description (qui vérifie le type et la cardinalité en même temps). Les contraintes

[6]Cette contrainte naît de la cohabitation de représentations mentales dénotant des objets du monde ou non. On peut comparer l'énoncé « je veux un hôtel près de la mer » et « je vous propose l'hôtel Ibis ». On ne souhaite en l'occurrence pas pouvoir grouper des représentations de différentes natures.

d'identification ciblent l'existence d'une point de vue capable de discriminer un sous-domaine sur la base de sa description, d'une focalisation antérieure ou de son index (pour les ordinaux).

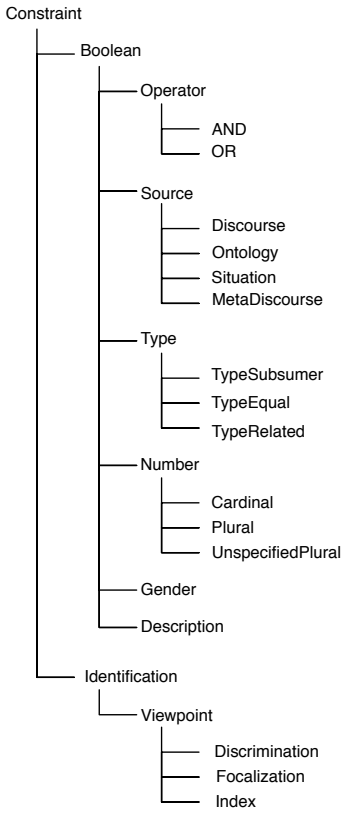

FIG. 6.3 – Contraintes utilisées pour la gestion de la référence

6.1.3.3 Exemple de résolution

L'exemple de la figure 6.4 illustre une résolution. L'expression « l'hôtel Ibis et le Lafayette » entraîne la création de quatre domaines : @0, le super-domaine correspondant aux hôtels (créé si celui-ci n'existait pas), @1 le domaine correspondant au groupe formé des deux hôtels, discriminés par leur index mais sans focalisation, @2 et @3 les domaines correspondant à l'hôtel Ibis et à l'hôtel Lafayette. L'expression « le premier hôtel » recherche un domaine contenant plusieurs hôtels discriminés par leur index : @1 est trouvé, @2 est extrait et focalisé. Ensuite la mention de « l'hôtel Ibis » recherche un domaine d'hôtel contenant une discrimination par le nom (non représentée ici) : @1 est trouvé, @2 est extrait et la focalisation ne change pas. Enfin l'expression « l'autre » recherche un domaine quelconque contenant un point de vue focalisé : @1 est trouvé et le complément @3 est extrait et focalisé.

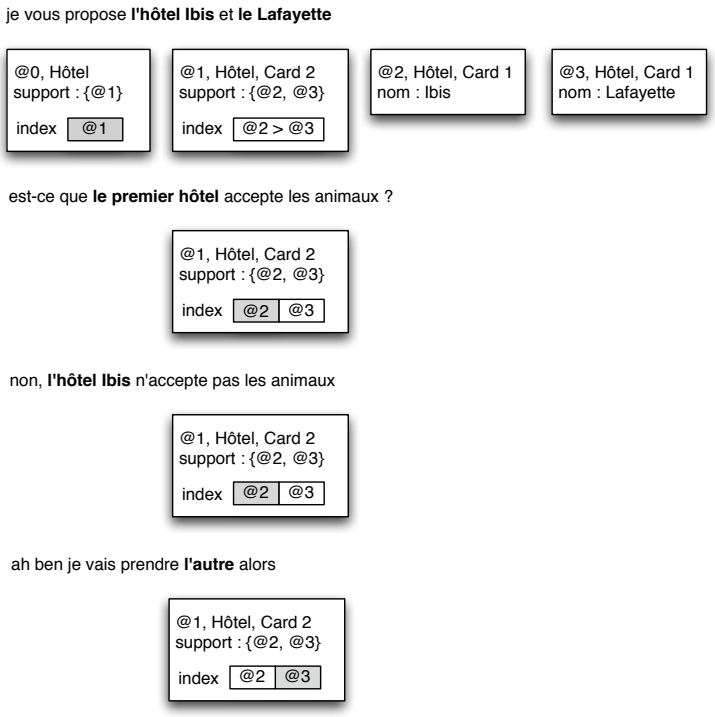

Fig. 6.4 – Exemple de gestion de la référence

6.2 Projection dans le formalisme MEDIA

6.2.1 Projection hors-contexte

La projection hors-contexte est particulièrement difficile étant donné les différences entre le formalisme MEDIA sous forme de sémantique plate sans variable et notre représentation interne sous forme de graphe conceptuel.

Ordre des triplets La première caractéristique du formalisme MEDIA à prendre en compte concerne l'*ordre* des triplets à produire. Afin d'être capable de conserver l'ordre des informations, nous avons simplement ajouté un attribut correspondant à l'index sur chaque trait ou participant MMIL. Cet index est récupéré à partir de l'analyse syntaxique, nous le conservons et l'associons à la forme d'interprétation. La facilité avec laquelle cette opération a été réalisée en MMIL peut être comparée à la difficulté rencontrée par les autres approches symboliques pour conserver l'ordre, dans la mesure où ces dernières produisaient directement une forme logique, cf. Bonneau-Maynard et al. (2008, page 218). Etant donné que MMIL fournit une représentation neutre, et n'est interprété sous forme logique que dans un second temps, l'ajout des indices des traits était relativement aisé.

Projection du composant sémantique La projection hors-contexte consiste ensuite à déterminer quels sont les triplets mentionnés dans la forme d'interprétation. Pour ce faire nous avons fondé la projection sur deux ontologies : une ontologie du domaine, appelée ontologie *interne* définie en OWL, qui décrit les concepts tels qu'ils peuvent apparaître dans les énoncés et dans les composants sémantiques. Par exemple, les instances du concept Reservation seront associées par le rôle aObjetReservation à des instances de Chambre, elles-mêmes associées à des instances d'Hotel par le rôle estDansHotel. Et d'autre part une ontologie *externe* spécifique à MEDIA, qui décrit des concepts plus proches du formalisme dans les termes de l'ontologie interne. Cette ontologie traduit les règles d'annotation sous forme d'assertions de subsomption. Par mesure de clarté nous appelerons l'ontologie interne décrivant les concepts et les rôles présents dans les composants MMIL, l'ontologie MMIL, et l'ontologie externe décrivant les concepts proches de traits à produire, l'ontologie MEDIA.

Le tableau 6.1 montre trois types de règles en décrivant le concept MEDIA à produire, le concept MMIL et un moyen de calculer la valeur associée au triplet final. La première règle correspond au trait du type +/objet:chambre (comme dans « cette chambre »), la seconde au trait du type +/chambre-type:simple (comme dans « une chambre simple ») et la troisième au trait du type +/chambre-type:double (comme dans « une chambre pour deux personnes »).

Concept MMIL	**Concept MEDIA**	**Valeur**
Chambre $\sqcap \exists$refType$\neg$Indefinite	$\sqsubset$ chambre	OBJECT
TypeDeChambre	$\sqsubset$ chambre-type-1	CLASS
Chambre $\sqcap \exists R^{\top}$(Personne $\sqcap \exists$cardinality2)	$\sqsubset$ chambre-type-2	double

Tab. 6.1 – Exemples de règles de projection

La production des triplets est alors réalisée en trois étapes : d'abord le composant MMIL est intégralement traduit sous forme de Abox. Chaque entité, et chaque valeur d'attribut sont associées à des instances de concepts MMIL et chaque relation ou chaque attribut est associé à un rôle. Ensuite, pour chaque instance de la Abox résultant, on recherche l'ensemble des concepts MEDIA *les plus spécifiques* dont elle

relève à l'aide du moteur d'inférences Racer (Haarslev and Möller 2003). Enfin, on construit les triplets sur la base du nom du concept MEDIA, et de la valeur associée dans la règle correspondante. En l'occurrence, on produit quatre types de triplets en fonction du champ Valeur de la règle :

- OBJECT, où le triplet doit être de la forme +/objet:X où X est le nom du concept trouvé,
- CLASS, où la valeur du triplet correspond au nom du concept MMIL le plus spécifique dont relève l'instance,
- un nom de rôle, où la valeur correspond à la valeur associée au rôle, utile pour produire les triplets du type +/nom:lafayette
- ou une chaîne de caractère, lorsque la valeur est constante

Règles additionnelles D'autres règles doivent néanmoins être ajoutées. En effet, il est nécessaire de s'adapter aux spécificités de l'annotation évoquées p. 84. Par exemple « la chambre » doit être annotée par :

+/lienRef-coRef:singulier

+/objet:chambre

alors que « la chambre simple » doit être annotée par :

+/lienRef-coRef:singulier

+/chambre-type:simple

Afin d'éliminer le trait +/objet:chambre déclenché par la présence du concept MMIL Chambre dans le cas de « la chambre simple », il suffit de nous appuyer sur le fait qu'on ne recherche que les concepts MEDIA les plus spécifiques. On définit alors des règles additionnelles de subsomption dans l'ontologie MEDIA pour éliminer les triplets indésirables. En l'occurrence, il suffit des règles fournies dans le tableau 6.2.

Concept MEDIA	**Concept MEDIA**
chambre-type-1	$\sqsubset$ chambre
chambre-type-2	$\sqsubset$ chambre

Tab. 6.2 – Exemples de règles additionnelles de subsomption

Les différentes polarités du mode, et les spécifieurs sont produits selon le même principe, en distinguant toutefois les spécifieurs globaux dont la portée est l'énoncé entier et les spécifieurs locaux qui ne portent que sur un participant MMIL. Une dernière étape de post-traitements intervient enfin pour ordonner les triplets selon les indices, normaliser les attributs et les valeurs, et réécrire certaines séquences de triplets (essentiellement les dates). L'algorithme complet de projection est résumé dans la figure 6.5.

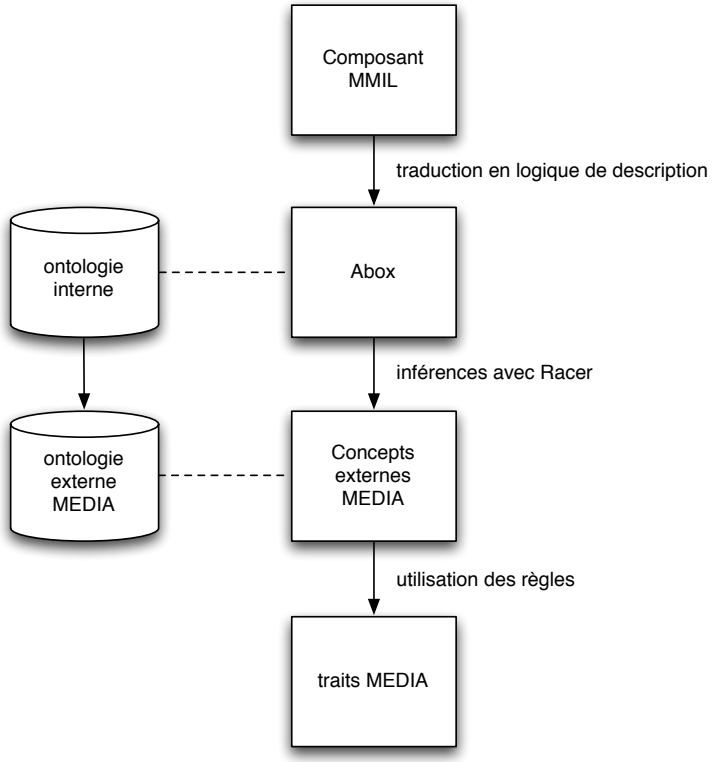

FIG. 6.5 – Algorithme de projection hors-contexte

Exemple de projection Le composant sémantique correspondant à l'énoncé « la chambre pour deux personnes » serait traduit conformément à la figure 6.6.

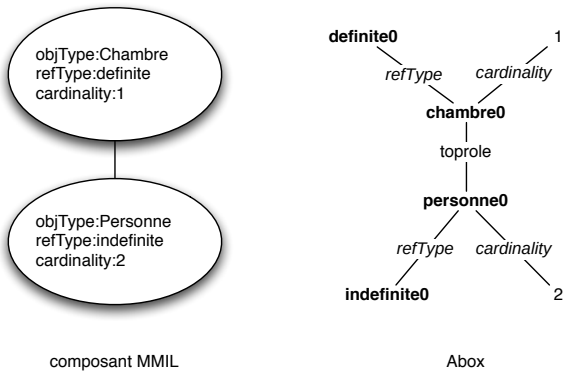

FIG. 6.6 – Transformation MMIL vers Abox

Les concepts MEDIA dont relève l'instance **chambre0** sont alors {chambre, chambre-type-2} et comme chambre-type-2 est plus spécifique que chambre, seul

chambre-type-2 est pris en compte. La phase de post-traitements enlève ensuite le postfixe -2 et produit un trait du type chambre-type:double.

6.2.2 Projection en contexte

La projection en contexte consiste à produire la description d'un référent dans le formalisme MEDIA. Etant donné que les domaines de référence ne conservent que le point de vue courant sur les référents et non pas les expressions référentielles et le contexte de leur emploi, il était nécessaire de combiner le modèle avec l'historique de résolution. Parallèlement à la structuration domaniale de l'espace référentiel, nous avons conservé la structure sémantique MMIL des énoncés à laquelle nous avons ajouté des relations référentielles : coréférence, anaphore associative, sous-domaniale et altérité. L'exemple de la figure 6.7 illustre deux de ces relations. Ces liens référentiels nous permettent de parcourir les chaînes de coréférence afin d'annoter les descriptions des référents.

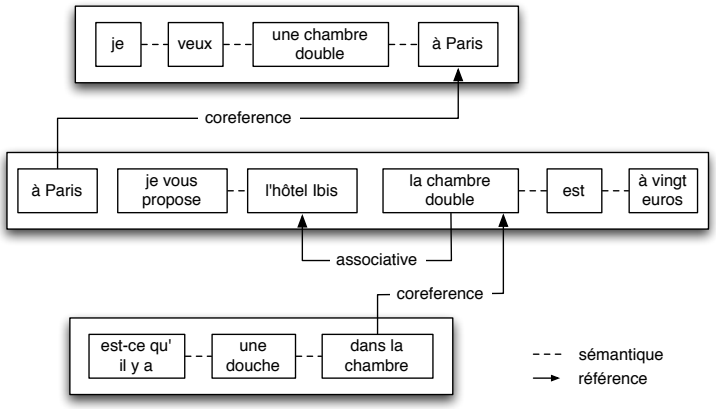

Fig. 6.7 – Exemple d'historique MMIL avec relations référentielles

Les chaînes de coréférence sont parcourues en agglomérant les projections hors-contexte des entités participant à la description des référents concernés. Par exemple un hôtel est décrit par la ville, une chambre par l'hôtel, etc. Plusieurs types de contraintes d'aggrégation ont été considérées en fonction du types des entités : présence d'une relation sémantique, co-occurrence dans le composant sémantique, co-occurence dans l'énoncé et présence dans le dialogue. Dans la condition du test *avecHC* (cf section 7.1.2), au lieu de s'appuyer sur la projection hors-contexte, nous utilisons les traits sémantiques fournis. Nous ne tirons partie des informations données par ces traits à aucun moment pour guider le processus de résolution lui-même.

6.2.3 Ressources

Le système s'appuie uniquement sur des ressources hétérogènes écrites manuellement ou générées semi-automatiquement : *un lexique morphologique* extrait du

lexique Multext (5400 mots et 3000 lemmes) auquel nous avons ajouté les noms propres (lieux, hôtels ou événements), une petite *grammaire LTAG* de 80 arbres, *un lexique sémantique* utilisé pour produire le graphe conceptuel (150 schémas de construction), une *ontologie interne* pour vérifier le graphe conceptuel et représenter le composant MMIL en logique de description (220 concepts), et une *ontologie externe* utilisée pour la projection MEDIA hors-contexte (130 concepts).

6.3 Conclusion

Le système d'interprétation que nous avons présenté produit une forme sémantique résolue en référence à partir d'un énoncé en langue naturelle. Afin de satisfaire aux contraintes de généricité que nous nous sommes imposées, nous avons clairement décorrélé la capacité de construire cette forme d'interprétation et la capacité de la projeter dans le formalisme MEDIA. Bien que le module de projection soit spécifique au formalisme MEDIA, les principes présentés, liés à la projection d'ontologie, peuvent être théoriquement réutilisés dans un système de dialogue. En effet, le module permet d'effectuer des inférences conceptuelles à partir d'une forme sémantique *proche de la langue* pour la traduire sous une forme *proche de l'application*, utilisable par le gestionnaire de dialogue. On pourrait dès lors concevoir une ontologie générique de la langue, susceptible de couvrir plusieurs applications, dont les concepts pourraient être ré-écrits pour satisfaire les besoins d'une application donnée.

Toutefois, on peut s'interroger sur la robustesse interne de notre système. D'abord l'utilisation d'un analyseur profond plutôt que d'un analyseur de surface (Abney 1991) risque d'entraîner des difficultés importantes liées à la spécificité de l'oral. Ensuite la construction manuelle des ressources risque de limiter la couverture des phénomènes pris en compte. En effet, pour construire les ressources (lexicales, syntaxiques ou conceptuelles), nous ne nous sommes appuyés que sur le manuel d'annotation ainsi que sur des exemples représentatifs issus du corpus d'apprentissage. Pour ces deux raisons, il est probable que la capacité à produire une forme sémantique correcte soit beaucoup plus faible que celle des systèmes statistiques, capables d'entraîner leurs modèles sur la totalité du corpus. Il est en revanche plus difficile d'estimer quels vont être les résultats de l'évaluation en contexte. Il semble que notre système symbolique puisse être avantagé grâce à la gestion explicite d'un contexte, contrairement aux systèmes statistiques qui voient le contexte comme un ensemble non structuré de traits. Le chapitre suivant qui aborde la méthodologie d'évaluation ainsi que les résultats apporte des réponses à ces questions de robustesse.

7 Méthodologie et résultats

Nous présentons ici la méthodologie d'évaluation ainsi que les résultats obtenus. Nous n'effectuerons cependant pas d'analyse comparative poussée entre les systèmes et on pourra se reporter au compte-rendu complet fourni dans Bonneau-Maynard et al. (2008). De plus, nous n'aborderons que superficiellement la méthodologie et les résultats de la phase hors-contexte pour nous concentrer sur la phase en contexte.

7.1 Méthodologies d'évaluation

7.1.1 Evaluation hors-contexte

L'évaluation hors-contexte repose sur la distance de Levenshtein grâce à une mesure des insertions, omissions ou substitutions de triplets. On mesure le score de *rappel* qui correspond au pourcentage de triplets correctement trouvés sur l'ensemble des triplets à trouver et le score de *précision* qui correspond au pourcentage de triplets correctement trouvés sur l'ensemble des triplets trouvés. La *f-mesure* est une moyenne harmonique de la précision et du rappel.

Plusieurs conditions de test ont été effectuées : une évaluation *complète* lors de laquelle on testait les triplets complets en termes de mode, attribut spécifié et valeur, une évaluation *relâchée* lors de laquelle les spécifieurs n'ont pas été pris en compte et une évaluation du mode *simplifié* lors de laquelle on considère les modes interrogatifs ou optionnels comme des modes positifs. Dans toutes les conditions les triplets +/null ne seront jamais évalués et seront concrètement retirés de l'annotation de référence et des annotations des systèmes.

7.1.1.1 Corpus et accord inter-annotateurs

Le corpus d'apprentissage était constitué de 11 000 énoncés et le corpus de test de 3 000 énoncés. Plusieurs tests d'accord inter-annotateurs ont été effectués sur des paquets de dix dialogues annotés par deux annotateurs différents. Le dernier test d'accord donnait 87.8% pour l'évaluation avec mode complet et 89.1% pour l'évaluation avec mode simplifié.

7.1.2 Evaluation en contexte

Deux conditions de test ont été organisées, afin de décorréler les capacités hors et en contexte : la première (appelée *sansHC*) consiste à n'utiliser que les dialogues transcrits, les erreurs de l'annotation hors-contexte se cumulant à celles de l'annotation en contexte. La deuxième condition (appelée *avecHC*) consiste à fournir aux participants les dialogues accompagnés de l'annotation hors-contexte effectuée par les annotateurs.

La condition *avecHC* implique que les systèmes soient capables de prendre des traits en entrée. On peut s'appuyer sur les traits de deux façons au moins : d'abord pour connaître les descriptions des entités, c'est-à-dire pour ne former la description des référents que sur des traits corrects, ou pour guider la résolution elle-même. Etant donné que le contexte antérieur de dialogue est annoté hors-contexte, toutes les expressions référentielles correspondent effectivement aux lienRef, et il n'existe que celles-ci. Dès lors on peut imaginer que le processus référentiel puisse être aisément guidé par les traits, ou tout au moins contrôlé par ceux-ci. Mais les traits étant insuffisants pour notre résolution (en particulier les types de référence[1]), il nous était nécessaire d'interpréter l'énoncé syntaxiquement et sémantiquement. Il s'agissait alors d'associer *a posteriori* les traits et les entités référentielles correspondantes. D'autre part, un système de dialogue ne disposant pas de ces annotations *en conditions réelles*, notre système n'a dès lors considéré que l'apport descriptif et pas référentiel de la phase *avecHC*.

7.1.2.1 Méthode d'évaluation de la référence

Etant donné l'annotation en contexte, fondée sur les descriptions sémantiques des référents, nous pouvons reprendre les mesures hors-contexte pour l'évaluation en contexte, en effectuant la comparaison des traits sémantiques pour chaque référent. Nous observons que pour décrire un référent, il faut préalablement l'avoir identifié. De même, pour l'identifier, il faut préalablement avoir repéré l'expression référentielle. Comme ces tâches impliquent potentiellement des capacités différentes, nous avons jugé intéressant d'évaluer la résolution de la référence selon quatre niveaux, chacun donnant lieu à des scores de rappel, précision et f-mesure :

IER Capacité à repérer (ou identifier) les expressions référentielles. Il s'agit d'une capacité hors-contexte, qu'on évalue tout de même car l'évaluation de la référence en dépend.

DER Capacité à décrire les expressions référentielles identifiées, c'est-à-dire, à fournir les bons spécifieurs (coRef, coDom, elsEns, mais aussi inclusion, exclusion et ambigu). Cette capacité est évaluée sur la base des expressions correctement repérées en IER.

IREF Capacité à identifier les référents, c'est-à-dire à fournir pour chaque référent suffisamment de traits corrects pour qu'il soit couplable avec un référent

[1]Par exemple comment intégrer les indéfinis si ceux-ci ne sont pas annotés en référence ?

à trouver. Comme la précédente, cette capacité n'est évaluée que sur les expressions référentielles correctement repérées en IER.

DREF Capacité à décrire *in extenso* les référents. Cette évaluation n'est calculée que sur les référents corrects en IREF.

En procédant ainsi par niveau, nous pouvons mieux apprécier les différentes capacités qu'implique la résolution de la référence. Nous notons qu'il est possible d'avoir un score global de rappel (resp. de précision) en DREF, c'est-à-dire le nombre de traits corrects fournis par un système rapporté au nombre de traits fournis par l'annotation manuelle (resp. fournis par le système), en multipliant les scores de rappel (resp. de précision) obtenus en IER, IREF et DREF.

Pour chaque niveau, les algorithmes suivants ont été développés :

IER Le but est d'aligner les traits lienRef sans s'appuyer ni sur leurs spécifieurs (évalués en DER), ni sur leur contenu référentiel (évalué en IREF et DREF). Or, si l'on retire ces annotations, le trait lienRef comporte trop peu d'information pour pouvoir effectuer l'alignement sans risque dans le cas d'une omission ou d'une addition de lienRef. Nous effectuons donc un alignement des lienRef sur la base des autres traits sémantiques qui sont dans l'intervalle. Nous avons adapté l'algorithme de Levenshtein pour que le gain d'un appariement de lienRef soit proportionnel à la valeur d'appariement des traits (non référentiels) qui se trouvent entre un lienRef et le suivant[2].

DER Pour chaque couple de lienRef appariés selon le procédé ci-dessus, on calcule le nombre d'erreurs qui apparaissent lorsqu'on rajoute les modes et surtout les spécifieurs.

IREF Pour chaque couple de lienRef appariés en IER, on effectue un couplage maximal de poids maximal entre référents hypothèse et référents de l'annotation manuelle. La matrice de couplage donne un poids proportionnel aux nombres de traits partagés[3].

DREF Pour chaque couple de référents formé en IREF, on effectue un couplage maximal entre traits. En effet, il n'y a pas d'ordre prescrit pour décrire les référents.

7.1.2.2 Corpus et accord inter-annotateurs

Pour la campagne d'évaluation en contexte, le corpus d'apprentissage disponible contient 814 dialogues, 11 800 énoncés utilisateurs et 38 800 segments sémantiques dont 2 294 liens référentiels. Le corpus de test contient 174 dialogues pour 2 650 énoncés utilisateurs et 7 780 segments dont 455 liens référentiels. A l'instar de l'annotation hors-contexte, l'annotation en contexte a fait l'objet de tests d'accord

[2]Ce procédé évite aux systèmes de produire une double annotation (HC et EC), l'IER devant être *a priori* calculée sur les lienRef corrects en HC et non sur l'annotation EC.

[3]Il suffit donc qu'un référent ait un trait correct pour être candidat. Pour être plus précis, il faudrait ne conserver que les traits permettant de discriminer un référent parmi les autres référents proposés.

inter-annotateurs. Les tests ont également été conduits par une double annotation de paquets de dix dialogues. L'accord en IER est de 100%, en DER de 95%, en IREF de 95% et en DREF 82%. Les scores IER, DER et IREF sont globalement très bons. Le score en DREF, plus faible que les autres, traduit la difficulté de fournir unanimement une description complète des référents.

7.2 Résultats

Les résultats contredisent nos prévisions. Nous estimions les systèmes symboliques bien moins robustes que les systèmes statistiques sur l'annotation sémantique hors-contexte, en particulier à cause de la construction manuelle des ressources. Or les résultats hors-contexte sont comparables. Au contraire, nous prédisions que notre système symbolique, en gérant explicitement le contexte serait plus capable de comprendre le niveau référentiel que les systèmes statistiques. L'évaluation en contexte montre que ce n'est pas le cas. Nous présentons ci-dessous les tableaux comparatifs que l'on peut trouver dans Bonneau-Maynard et al. (2008) et dans Denis et al. (2007) sans aborder outre mesure les résultats des autres systèmes.

7.2.1 Résultats hors-contexte

Les systèmes qui ont participé à l'évaluation hors-contexte sont nommés selon le laboratoire correspondant : Lia (statistique), Limsi-1 (statistique), Limsi-2 (une passe symbolique, puis statistique), notre système Loria (symbolique) et Valoria (symbolique). Les résultats en termes de taux d'erreur sont fournis pour les conditions *complète*, et *relâchée* considérant les modes simplifiés ou non (2 modes ou 4 modes) dans le tableau 7.2.

	Complet		Relâché	
	4 modes	2 modes	4 modes	2 modes
Lia	41,3	36,4	29,8	24,1
Limsi-1	**29,0**	23,8	**27,0**	21,6
Limsi-2	30,3	**23,2**	27,2	**19,6**
Loria	36,3	28,9	32,3	24,6
Valoria	37,8	30,6	35,1	27,6

Tab. 7.1 – Résultats de l'évaluation hors-contexte

Les taux d'erreur de notre système sont environ de 5 points supérieurs aux taux du meilleur système. En conséquence, les systèmes statistiques sont plus robustes que les systèmes symboliques mais de manière beaucoup moins significative que nous ne l'avions envisagé.

7.2.2 Résultats en contexte

Seuls deux systèmes ont participé à l'évaluation en contexte, le système du Lia et notre système. Le système statistique du Loria ne s'appuie pas sur une repré-

sentation structurée de l'espace référentiel. Il recherche dans l'ensemble des triplets antécédents les triplets correspondants à la description des référents (voir Denis et al. 2007). Nous donnons ici les résultats pour les conditions *sansHC* et *avecHC*.

sansHC	DER		IREF		DREF	
	prec	*rappel*	*prec*	*rappel*	*prec*	*rappel*
LIA	71.4	71.4	74.1	61.9	67.3	55.2
LORIA	50.9	50.9	65.2	44.3	68.9	48.3
avecHC	DER		IREF		DREF	
	prec	*rappel*	*prec*	*rappel*	*prec*	*rappel*
LIA	86.5	86.5	77.1	73.8	74.1	64.0
LORIA	86.5	86.5	62.4	40.8	76.5	43.3

TAB. 7.2 – Résultats de l'évaluation en contexte

Notre système symbolique obtient des scores globalement inférieurs au système du LIA. En particulier les scores de rappel sont très bas (40.8% en IREF *avecHC*). La condition *avecHC* dégrade même globalement notre score (à l'exception de la précision en DREF), en raison de la nécessité d'intégrer les triplets en entrée. D'autre part, notre système ne s'appuie sur les traits que pour décrire les référents une fois que leur résolution a été effectuée et pas pour guider ou contrôler la référence. Au contraire, le système statistique parvient bien à intégrer les annotations hors-contexte et augmente ses scores de rappel d'une dizaine de points sur les trois catégories.

7.3 Les erreurs du système

Afin de déterminer les erreurs de notre système, nous nous sommes appuyés exclusivement sur le score le plus bas, c'est-à-dire le score de rappel en IREF de la condition *avecHC*. Un référent est erroné en IREF lorsqu'aucun trait correct n'est fourni pour sa description. Nous avons procédé à une classification *manuelle* de toutes les erreurs de rappel IREF en nous appuyant sur une ré-évaluation systématique de chaque dialogue et un examen des logs du système. Normalement, seul le niveau de la référence aurait du faire l'objet d'erreurs mais étant donné que nous n'avons pas considéré les triplets en entrée, tous les niveaux sont susceptibles de provoquer des erreurs référentielles. La classification s'appuie en conséquence sur le niveau où l'erreur apparaît pour la première fois. Nous distinguons alors les erreurs lexicales, syntaxiques, sémantiques, référentielles, ainsi que les erreurs de projection en contexte ou les erreurs dans l'annotation de référence (media)[4]. Au total, nous avons dénombré 118 erreurs dont le tableau 7.3 détaille la proportion en fonction du niveau.

[4]Contrairement à la phase hors-contexte qui a fait l'objet d'une adjudication détaillée, ce ne fut pas le cas pour la phase en contexte. Subsistent alors quelques erreurs marginales au niveau de l'annotation en contexte du corpus.

Niveau	Erreurs IREF
référence	57%
projection	13%
sémantique	11%
syntaxe	11%
media	5%
lexical	3%

Tab. 7.3 – Erreurs de rappel IREF par niveau

Le premier résultat significatif est que 43% des erreurs de référence sont provoquées par des causes extérieures au module de référence. Si on ôte les problèmes de projection ou d'annotation du corpus, 25% des erreurs sont provoquées en amont du traitement de la référence. Cela n'a rien d'étonnant si on considère l'interprétation *stricte* de la référence que nous avons adoptée. Par exemple si l'analyse syntaxique oublie un hôtel parmi trois hôtels, toute référence telle que « ces trois hôtels » échouera et ne retournera rien. Le système statistique du Lia est beaucoup plus lâche dans sa résolution en recherchant tous les référents possibles. Nous pensons toutefois qu'il est nécessaire de pouvoir soulever une erreur et qu'alors une interprétation stricte de la référence reste intéressante dans le contexte du dialogue homme-machine.

Nous nous sommes intéressés dans un second temps aux différentes erreurs de référence. Nous avons classé ces dernières en deux groupes, rassemblant onze catégories[5].

Le premier groupe est constitué de tous les phénomènes non pris en compte (36%). Le cas le plus fréquent est celui que nous avons appelé *anaphore associative multiple*, c'est-à-dire des cas d'anaphores associatives avec antécédent pluriel énuméré. Par exemple le générique « la chambre » interprété dans le contexte de plusieurs hôtels soulève des problèmes d'inconsistance logique : une chambre ne peut appartenir logiquement qu'à un unique hôtel. Le second cas le plus fréquent est celui appelé *double altérité* et soulève une question importante liée au traitement de l'altérité. Il se produit lorsque le compère produit un énoncé du type « souhaitez-vous un autre hôtel ? » et que le client répond « oui un autre hôtel ». La première altérité exclut un hôtel donné en construisant un nouveau référent, et la seconde exclut ce nouveau référent, perdant alors le lien avec l'hôtel exclut initialement[6].

Le second groupe est constitué des réelles erreurs et bugs (64%) et correspond à un fonctionnement anormal du module de référence. Les cas les plus fréquents sont les modifieurs relationnels mal gérés et les restrictions de sélection pronominales. Dans

[5](Denis et al. 2007) mentionne vingt catégories, mais nous avons ici regroupés les catégories ne contenant qu'une seule occurrence sous l'étiquette *divers*.

[6]On peut adopter diverses stratégies de robustesse interne pour résoudre ce problème. Dans (Denis, Quignard, and Pitel 2006), nous proposons de considérer des contraintes additionnelles en associant des marqueurs aux focalisations en fonction du type d'extraction, par exemple *ordinalité* ou *altérité*. Ici, la seconde altérité ne doit pas provoquer de nouvelle création de référent mais être considérée comme une simple répétition de confirmation, et entraîner alors une coréférence.

Type	Statut	Occurrences
anaphore associative multiple	non traité	12
modifieurs relationnels	bug	8
restriction de sélection	bug	8
divers bugs	bug	6
problèmes de groupements	bug	5
altérités	bug	5
doubles altérités	non traité	5
focalisations	bug	5
divers non traités	non traité	4
conséquence erreur antérieure	bug	3
problème de genre	bug	3
groupements extra-énoncés	non traité	3

TAB. 7.4 – Causes d'erreur de référence

le premier cas, les relations qui contraignent la résolution sont mal délimitées, par exemple dans l'énoncé « je voudrais plus de détails sur les hôtels », la relation entre le participant MMIL correspondant à « plus de détails » et le participant correspondant à « les hôtels » contraint à tort la résolution et entraîne la recherche d'« hôtels avec plus de détails ». Dans le cas des restrictions de sélection, la distinction entre présupposition et assertion, bien que représentée, est mal traduite dans la résolution. Typiquement dans « ils acceptent les animaux », ce qui est asserté (« acceptent les animaux ») est inclut à tort dans ce qui est présupposé et conduit à rechercher quelque chose dont on sait qu'ils acceptent les animaux, et pas quelque chose qui potentiellement accepte les animaux. Les autres erreurs recouvrent par exemple des problèmes de groupements, ou des focalisations erronées.

7.4 Critique de l'évaluation de la robustesse interne

7.4.1 Critique de l'annotation en contexte

Les résultats de l'évaluation en contexte doivent cependant être compris en considérant la mesure qui a été adoptée. En effet cette dernière définit l'identité de deux référents comme une identité de *description*, traduisant alors mal le fait que deux référents peuvent être différents tout en se ressemblant. Par exemple, le référent de « une chambre double à l'hôtel Ibis paris » et celui de « une chambre double à l'hôtel Lafayette paris » sont similaires à 75% en termes de DREF bien qu'ils représentent deux référents distincts. Le système du LIA est insensible au fait qu'il s'agisse de deux référents puisqu'il s'appuie directement sur les traits sémantiques de bon type dans le contexte antérieur, alors que notre système qui construit une représentation structurée des référents y est au contraire très sensible. Cette mesure ne permet alors d'évaluer que les capacités descriptives des référents, nécessaires dans un système de dialogue mais pas suffisantes. En effet la mesure ne permet pas de comparer les

capacités référentielles pour lesquelles il est indispensable d'identifier avec précision le référent (en l'occurrence ne pas réserver une chambre dans le mauvais hôtel). Afin de considérer ces capacités, on peut envisager dans une future évaluation d'améliorer le couplage IREF pour qu'il ne couple que des référents décrits par les mêmes traits s'il s'agit d'entités nommées ou que des référents décrits par les mêmes traits discriminants s'il s'agit d'entités non-nommées.

7.4.2 Critique de l'évaluation MEDIA

Ce type d'évaluation peut être qualifié de quantitatif : on évalue les systèmes sur une large variété de phénomènes et on donne une appréciation globale de la précision et du rappel. L'intérêt est que les mesures sont en général très simples (distance de Levenshtein). Une fois que les systèmes ont produit leurs sorties, il est très facile de fournir une mesure de leur robustesse interne. Cependant ce type d'évaluation présente plusieurs défauts. Le premier, et le plus important, est qu'elle est biaisée. On n'évalue pas *seulement* la capacité d'interprétation des systèmes mais également leur capacité de projection dans le formalisme de comparaison. Ainsi, les systèmes dont la représentation interne est plus proche du formalisme de sortie seront naturellement favorisés par rapport aux autres systèmes dans lesquels la projection peut être difficile. Le second défaut est que, malgré toutes les précautions qui ont été prises, elle reste coûteuse puisqu'elle requiert l'enregistrement, la transcription, et l'annotation — avec un bon accord inter-annotateurs — d'un corpus d'une taille conséquente. Le troisième défaut est qu'elle court le risque d'être peu précise : à moins de disposer d'une double annotation, les phénomènes à évaluer et l'interprétation qui en est faite, l'évaluation ne fournit qu'une mesure globale décorrélée des phénomènes particuliers. Par exemple, on ignore comment se comportent les systèmes sur chaque type d'expression référentielle (type de référence, présence de modifieurs relationnels ou non, subordonnées relatives, etc.) car ces aspects n'ont été annotés que partiellement, via le spécifieur de lien référentiel. Enfin, on peut s'interroger dans ce type d'évaluation sur la pertinence du consensus lors de l'élaboration du manuel d'annotation. En effet, étant donné que les contraintes de faisabilité, d'intérêt et de coût reposent sur les exigences partagées du consortium, une grande hétérogénéité dans le consortium entraîne nécessairement des contraintes plus importantes sur la constitution d'un manuel d'annotation. Les difficultés de mise en oeuvre expliquent alors certains des choix pratiques adoptés, l'annotation des référents extra-énoncés ou le choix d'une annotation descriptive des référents par exemple.

Malgré tous les défauts de l'évaluation de la robustesse interne, l'annotation qu'elle définit permet de fournir un *gold standard* qui, lorsqu'il n'est pas atteint permet d'exhiber les problèmes d'interprétation (ou de projection) d'un système donné. Nous utiliserons cette caractéristique dans le chapitre 10 pour vérifier si le processus d'ancrage est capable de résoudre un problème d'interprétation donné.

7.5 Amélioration attendue par l'ancrage

Estimer le gain de compréhension grâce à un processus d'ancrage est relativement difficile. On peut toutefois supposer que les aspects problématiques de la résolution de la référence *causés par le contexte* pourront être résolus dans la mesure où l'ancrage, en focalisant le dialogue sur l'incompréhension des référents, s'appuie sur un contexte plus simple. Au niveau syntaxique, lorsque par exemple l'analyse oublie un hôtel parmi trois hôtels *à cause du cotexte*, on peut espérer que l'ancrage, en s'appuyant sur un cotexte plus simple entraîne la disparition du problème. Le même raisonnement peut être employé pour toutes les catégories d'erreurs purement référentielles ou non relevant du contexte. Par exemple, pour toutes les erreurs dans lesquelles on contraint trop la résolution, comme les restrictions de sélection ou les modifieurs relationnels, bien que le contexte sémantique puisse contraindre à tort la résolution, on peut espérer qu'en cherchant de manière alternative le référent, il soit possible de le trouver, comme illustré dans le dialogue 7.1.

U : je voudrais plus de détails sur l'hôtel
S : quel hôtel ?
U : l'hôtel Ibis
S : à l'hôtel Ibis, la chambre est à vingt euros

FIG. 7.1 – Exemple d'ancrage espéré

A contrario, pour les problèmes qui ne relèvent pas du contexte, on peut être assurés que l'ancrage ne parviendra pas à augmenter la compréhension. Les erreurs lexicales, que nous mettrons de côté faute d'apprentissage, ou les erreurs d'annotation du corpus ne pourront certainement pas être résolues par l'ancrage. On ne pourra également pas s'attendre à ce que le problème d'anaphore associative multiple puisse être résolu étant donné que dans tous les contextes possibles, le problème d'inconsistance logique se posera[7].

Mais de manière générale, il serait malavisé de pronostiquer un gain de compréhension de façon systématique. En effet, comme nous l'avons évoqué, le processus d'ancrage en soi peut échouer. Les problèmes de compréhension peuvent ne pas être détectés ou mal détectés, ou encore entraîner de nouveaux problèmes. Nous proposons, en troisième partie de cette thèse, une évaluation méthodique du gain de compréhension à l'issue du processus d'ancrage.

7.6 Conclusions de la seconde partie

Nous avons présenté la méthodologie d'évaluation MEDIA des capacités sémantiques et référentielles des systèmes d'interprétation. Celle-ci propose d'évaluer la

[7]Nous avons toutefois remis en cause la lecture distributive de l'anaphore associative multiple dans le cas de « la chambre » à cause du fait qu'elle était la première source d'erreur. Cette stratégie correspond toutefois à une pure stratégie de robustesse interne.

compréhension des systèmes par comparaison de leur interprétation avec l'interprétation d'un humain prototypique. Des énoncés d'un large corpus de dialogue oral ont été transcrits et annotés selon un formalisme commun de représentation et afin d'autoriser la comparaison, chaque système devait *projeter* sa forme d'interprétation dans ce formalisme.

Le système d'interprétation symbolique que nous avons présenté obtient des scores légèrement inférieurs aux systèmes statistiques sur la représentation sémantique. Sa robustesse dans la construction de la forme sémantique est alors analogue à celle des systèmes statistiques. Toutefois, une différence importante avec le système statistique a été relevée dans la capacité à résoudre les référents. Cette différence a plusieurs causes. La première concerne la forte dépendance de notre module de résolution de la référence vis à vis des niveaux inférieurs, une analyse syntaxique échouée entraîne par exemple l'échec de la description des référents. En tout, 25% des erreurs sont purement dûes aux niveaux inférieurs de traitement. La seconde cause est la nature de l'évaluation. Nous n'avons évalué que la capacité de *description* des référents. Celle-ci peut, comme le système statistique l'a montré, ne s'appuyer que sur une représentation non structurée du contexte. En contrepartie cette évaluation ne mesure pas la capacité d'identifier précisemment les référents. Enfin, la troisième cause concerne la capacité de projeter la forme d'interprétation dans le formalisme commun. Il s'agit en effet de la première cause d'erreur hors-référence, et on peut supposer que le système statistique y était moins sensible que notre système symbolique.

Le défaut principal de ce paradigme est d'évaluer à la fois ce qui relève de la projection et ce qui relève de la compréhension sans pouvoir distinguer les deux sources d'erreur automatiquement. Malgré ce défaut, l'examen manuel des cas problématiques de rappel, nous a permis d'identifier les principales sources d'erreur et les phénomènes non traités. Ces phénomènes nous autorisent à estimer qu'un modèle d'ancrage *pourrait* améliorer la compréhension. Nous proposons dans la partie suivante la définition d'un critère et d'un processus d'ancrage dont nous procéderons à l'évaluation. Le but de cette évaluation est de vérifier automatiquement si notre système d'interprétation à laquel on adjoint un module de dialogue dédié à l'ancrage peut parvenir à améliorer sa compréhension.

Modélisation et évaluation d'un processus d'ancrage

8 Modélisation du processus d'ancrage

8.1 Introduction

Le modèle que nous proposons se situe au niveau de la partie gestion du dialogue des systèmes de dialogue. Il spécifie l'action communicative nécessaire pour ancrer l'intention du locuteur dans le terrain commun : il prend en entrée une interprétation déjà réalisée et retourne une spécification de l'action à effectuer en termes de preuves de compréhension que doit manifester le système. Pour déterminer ces preuves, le modèle s'appuie deux types de compréhension, la compréhension de l'énoncé par celui qui l'écoute et la compréhension que l'énoncé manifeste. Ces deux aspects de la compréhension correspondent au double rôle des énoncés dans le modèle des contributions, un rôle d'acceptation et un rôle de présentation.

Notre méthodologie consiste à nous appuyer sur le système d'interprétation existant (section 6) qui possède ses propres limitations (décrites section 7.3), et à lui adjoindre un module de dialogue avec ancrage et un module de génération. La tâche sur laquelle nous nous appuyons est une tâche métadiscursive de clarification de la référence. Le module de dialogue est alors réduit à sa plus simple expression : la construction du contenu sémantique à produire pour ancrer les énoncés de l'utilisateur au niveau de la référence. Néanmoins les principes présentés dans cette section peuvent s'appliquer à d'autres niveaux d'analyse.

8.1.1 Rappel des problèmes

Nous avons présenté les modèles existants d'ancrage et leurs caractéristiques. En fait, tous ces modèles peuvent être unifiés sous la définition originale de l'ancrage proposée par Clark et Schaefer : si on met de côté la nécessité d'atteindre une croyance mutuelle de compréhension, les participants peuvent ancrer un énoncé dès qu'ils ont des preuves *suffisantes* qu'il n'y a pas de problème de compréhension de l'énoncé. Par exemple pour le modèle des contributions (Clark) ou le modèle des échanges (Cahn), il est suffisant de recevoir une preuve de compréhension ; pour le modèle des *grounding acts* (Traum), il est suffisant de produire ou interpréter un *grounding act* pour qu'il soit ancré et il est suffisant de recevoir un *acknowledgement* pour clore une phase d'acceptation ; pour la stratégie prudente (Larsson), il est suffisant de comprendre un énoncé pour l'ancrer ; pour le modèle des croyances

faibles (Bunt), il est suffisant de considérer qu'on a une preuve de la compréhension de la preuve de compréhension. Nous avons déjà effectué les critiques de ces modèles et désirons ici proposer un critère d'ancrage à même de les unifier.

Rappelons les points essentiels : un modèle d'ancrage doit pouvoir, à l'instar du modèle original des contributions, considérer les preuves de compréhension positives mais aussi négatives. Il doit pouvoir expliquer pourquoi les participants produisent des preuves de compréhension et maintenir un principe qui en motive l'emploi. Ensuite un modèle d'ancrage doit éviter de tomber dans le problème de l'acceptation récursive en explicitant clairement le moment où la phase d'acceptation d'un énoncé peut être déclarée terminée. Enfin, un modèle d'ancrage doit considérer que les preuves de compréhension ne sont pas automatiquement comprises mais que, comme toute communication, l'ancrage ne peut être fiable : une non-compréhension des preuves entraîne le blocage du processus d'ancrage, et une mauvaise compréhension des preuves entraîne un ancrage erroné.

Puisqu'on dispose d'une définition qui repose sur la notion de compréhension *suffisante*, le problème peut se ramener à déterminer la nature des *exigences* de compréhension, la façon dont elles peuvent être satisfaites et ce qui doit être fait lorsqu'elles ne sont pas ou ne peuvent pas être satisfaites.

8.2 Modèle d'ancrage théorique

8.2.1 Deux types d'exigence de compréhension

La difficulté principale pour définir les exigences de compréhension entretenues par les participants est qu'elles ne sont pas établies de manière autonome. L'exigence de compréhension que le locuteur impose sur l'interlocuteur dépend des croyances du locuteur sur les capacités d'interprétation de son interlocuteur, par exemple si je sais que quelque chose ne sera pas compris par mon partenaire, je ne peux pas en exiger la compréhension. A l'inverse, les exigences de compréhension que l'interlocuteur se pose à lui-même dépendent du locuteur initial. Par exemple, si je sais que le locuteur attend de moi une compréhension bien précise, je ne peux me satisfaire d'une compréhension moindre. Les deux participants influencent mutuellement leurs exigences de compréhension de manière complexe, celles-ci pouvant dépendre des rapports sociaux entre les participants, de leurs buts propres/partagés, etc. Aborder dans sa globalité le problème de l'influence d'autrui dans les exigences de compréhension est sans doute difficile. Nous effectuons alors la simplification suivante : l'influence d'un participant sur les exigences de compréhension d'autrui se limite à la représentation réciproque des exigences de compréhension. C'est-à-dire que chaque participant entretient d'une part des exigences de compréhension propres sur un énoncé et d'autre part une représentation plus ou moins explicite des exigences de son partenaire, mais qu'on ne considère pas ici l'influence d'autrui dans la constitution des exigences propres. Cette simplification peut être motivée par le coût de prise en compte d'autrui dans la constitution des exigences propres (voir section 3.1.2 p. 48), dans l'incapacité d'effectuer cet effort, les participants peuvent

s'appuyer de manière autonome sur leurs propres exigences.

La distinction des deux types d'exigence permet de motiver le besoin de manifester de manière positive la compréhension : puisque mon interlocuteur a des exigences sur ma compréhension qui doivent être satisfaites pour atteindre le critère d'ancrage, et que je souhaite atteindre ce critère, alors je dois chercher à satisfaire ses exigences en montrant explicitement ou implicitement que je crois avoir bien compris l'énoncé. En soumettant l'ancrage à l'exigence de compréhension d'autrui, on motive l'expression des preuves positives de compréhension comme moyen de satisfaire à cette exigence, comme dans le modèle de Clark. Toutefois l'ancrage demeure problématique : comment connaître les exigences du partenaire ? Si je dois en effet attendre sa réaction et comprendre suffisamment cette réaction pour connaître ses exigences de compréhension, le problème d'acceptation récursive se pose : ai-je compris suffisamment l'énoncé ? Pour le savoir je dois comprendre suffisamment l'énoncé suivant, etc. Nous adopterons un point de vue interne à l'ancrage, en notant S (*Self*) le participant dont on prend le point de vue et O (*Other*) son partenaire. Le critère d'ancrage récursif d'un énoncé O_i du point de vue de S correspond alors à la définition suivante :

Définition récursive du critère d'ancrage du point de vue de S. *L'énoncé O_i est suffisamment compris du point de vue de S pour être ancré si et seulement si les exigences de compréhension propres de S sont satisfaites et que les exigences de O selon S sont également satisfaites. S croit connaître les exigences de O, si l'énoncé O_k ($k > i$) qui les manifeste est suffisamment compris par S pour être ancré.*

Etant donné que nous avons limité l'influence d'autrui dans la constitution des exigences propres de compréhension, la récursivité du critère d'ancrage ne tient qu'à l'estimation des exigences d'autrui. Dans le modèle des *grounding acts* de Traum, ces exigences sont automatiquement acquises, S les connaît dès qu'il reçoit O_k, ce qui coupe la récursion. Mais cette approche ne convient pas pour l'incompréhension des exigences. Dans la stratégie prudente de Larsson, l'ancrage n'est pas conditionné *a priori* par les exigences d'autrui, la satisfaction négative de ces exigences n'est seulement prise en compte qu'*a posteriori* lorsque l'énoncé O_k est reçu. Mais cette approche ôte la nécessité de produire des preuves positives de compréhension. Dans le modèle des croyances faibles de Bunt, les exigences d'autrui sur son énoncé O_i sont conditionnées à ses exigences de compréhension de l'énoncé O_k qui les manifeste. Cette approche convient pour l'incompréhension des exigences et requiert de fournir des preuves positives, mais nous estimons qu'un critère d'ancrage plus simple est possible.

8.2.2 Solution au problème d'acceptation récursive

La solution que nous proposons au problème d'acceptation récursive est alors de supposer que répondre à la question « ai-je compris suffisamment l'énoncé O_i ? » nécessite à la fois satisfaire des exigences propres et des exigences d'autrui, mais que comprendre les exigences d'autrui peut ne nécessiter qu'une exigence propre.

Autrement dit, on peut s'appuyer avec une certaine confiance sur sa propre estimation des exigences d'autrui. Cette hypothèse peut être motivée par le fait qu'il ne soit pas nécessaire d'ancrer un énoncé pour qu'il puisse avoir des effets, et qu'alors une preuve de compréhension peut jouer des effets dès que l'interlocuteur croit la comprendre.

Définition non-récursive du critère d'ancrage du point de vue de S. *$L'énoncé$ O_i est suffisamment compris du point de vue de S pour être ancré si et seulement si les exigences de compréhension propres de S sont satisfaites et que les exigences de O selon S sont également satisfaites. S croit connaître les exigences de O, si l'énoncé O_k qui les manifeste satisfait les exigences propres de S.*

En premier lieu, l'exigence de compréhension de l'interlocuteur lorsqu'il reçoit un énoncé est relative aux principes génériques de la communication : est-ce que, selon moi, j'ai suffisamment interprété les mots, le sens, et la référence de l'énoncé ainsi que l'intention de mon interlocuteur pour mes buts courants ? Je peux estimer de manière autonome ma bonne compréhension de l'énoncé si je parviens à interpréter de manière unique l'énoncé et que cette interprétation m'est satisfaisante. Dans le cas contraire, il s'agit d'une non-compréhension et je peux d'ores et déjà considérer que l'énoncé n'est pas ancré. Si je suis satisfait de ma propre interprétation et que mes exigences individuelles sont remplies, l'énoncé peut jouer les effets que je lui prête. En particulier, les effets relatifs à l'ancrage d'autres énoncés peuvent être joués. Grâce à lui je peux estimer si oui ou non les exigences de compréhension de mon partenaire concernant d'autres énoncés sont satisfaites, et donc si je peux considérer certains énoncés antérieurs ancrés (ce contexte correspond par exemple à la contribution, à la *discourse unit* ou aux QUD).

Du point de vue du locuteur de l'énoncé, la situation est plus simple. Celui-ci peut considérer son propre énoncé ancré lorsqu'il croit que son énoncé a été suffisamment compris (exigence propre) et qu'il croit que son interlocuteur le croit également (exigence attribuée). De la même façon, on peut estimer qu'il suffit pour le locuteur initial de croire comprendre la réaction de son partenaire pour pouvoir décider si le critère d'ancrage est atteint. Le critère peut alors s'écrire de son point de vue :

Définition non-récursive du critère d'ancrage du point de vue de O. *$L'énoncé$ O_i est suffisamment compris du point de vue de O pour être ancré si et seulement si les exigences de compréhension propres de O sont satisfaites et que les exigences de S selon O sont également satisfaites. O croit connaître les exigences de S, si l'énoncé S_j qui les manifeste satisfait les exigences propres de O.*

Une preuve de compréhension doit-elle être acceptée pour pouvoir jouer son rôle d'acceptation ? Tout dépend de quel point de vue on se place : si on se place du point de vue de celui qui *produit* la preuve, alors oui, il doit attendre de recevoir l'énoncé suivant pour estimer si selon lui sa preuve a joué ses effets. Si on se place du point de vue de celui qui *reçoit* la preuve, alors non, la preuve, dès qu'elle est crûe comprise, peut jouer des effets d'acceptation. Cette solution permet de terminer la récursion, mais en contrepartie l'ancrage ne peut pas être fiable puisque les preuves de compréhension peuvent être mal comprises.

8.2.3 Critère d'ancrage épistémique

Pour modéliser le critère d'ancrage, nous ne représenterons pas directement les exigences de compréhension mais plutôt le résultat de la satisfaction de ces exigences en terme de croyances sur la compréhension. Nous noterons $Und_S\, O_i$ la proposition correspondant à l'identité[1] de l'interprétation de O_i par S et O : $Und_S\, O_i \equiv (Int_S\, O_i = Int_O\, O_i)$. Cette proposition pourrait être évaluée d'un point de vue omniscient comme dans le modèle des contributions, mais nous plébiscitons un point de vue interne à l'ancrage, à l'instar du modèle des échanges, plus proche d'une modélisation computationnelle. Nous n'aurons donc accès qu'aux croyances des participants. Nous noterons alors la *croyance* que S a suffisamment compris l'énoncé O_i par $Bel_S\, Und_S\, O_i$. Cette croyance doit être mise en parallèle avec la croyance que O entretient sur la compréhension de O_i par S que nous noterons $Bel_O\, Und_S\, O_i$. Cette croyance très importante traduit la véritable compréhension de S : c'est O, le locuteur de O_i qui détermine au final si oui ou non la compréhension de S est suffisante.

Cependant les participants doivent également entretenir une croyance à propos de la croyance d'autrui, correspondant à la satisfaction des exigences qu'ils attribuent à autrui. Nous noterons alors la croyance de S à propos de la croyance de O en la bonne compréhension de l'énoncé par : $Bel_S\, Bel_O\, Und_S\, O_i$, c'est-à-dire que « S croit que O croit que S a suffisamment compris O_i ». A son tour, O peut considérer la croyance de compréhension de S : $Bel_O\, Bel_S\, Und_S\, O_i$. Cette croyance est utile pour distinguer les cas où S manifeste explicitement sa non-compréhension (c'est-à-dire qu'il croit ne pas avoir compris), des cas où il comprend mal sans le savoir.

Le critère d'ancrage peut alors être formulé de chaque point de vue comme la conjonction des deux types de croyances pour chaque participant :

$$Bel_S\, Grounded\, O_i = \underbrace{Bel_S\, Und_S\, O_i}_{(1A)} \wedge \underbrace{Bel_S\, Bel_O\, Und_S\, O_i}_{(2A)}$$

$$Bel_O\, Grounded\, O_i = \underbrace{Bel_O\, Und_S\, O_i}_{(1B)} \wedge \underbrace{Bel_O\, Bel_S\, Und_S\, O_i}_{(2B)}$$

Cette définition du critère d'ancrage a plusieurs intérêts. Le premier est qu'elle fait participer les deux types d'exigence de compréhension : sa propre exigence de compréhension et l'exigence d'autrui. Et elle le fait en explicitant l'auteur du jugement et l'auteur de l'interprétation. Conséquemment elle permet une prise en compte unifiée de la non-compréhension et de la mauvaise compréhension : la non-compréhension se produit lorsqu'un participant juge négativement sa propre interprétation ; la mauvaise compréhension survient lorsqu'un participant juge négativement l'interprétation d'autrui. Ce critère d'ancrage permet alors de représenter les cas suivants d'incompréhension (ici $1B$ et $2B$ sont formulées du point de vue de S sur son énoncé S_j) :

[1] Pour être plus précis, il serait nécessaire de considérer une équivalence relative aux buts courants des participants, en effet les interprétations n'ont pas à être identiques mais seulement suffisamment proches pour ne pas soulever d'obstacles à la réalisation des buts courants.

- ma non-compréhension : $Bel_S \neg Und_S O_i$ (1A-)
- ma mauvaise compréhension : $Bel_S\, Bel_O \neg Und_S O_i$ (2A-)
- la mauvaise compréhension de mon partenaire : $Bel_S \neg Und_O S_j$ (1B'-)
- la non-compréhension de mon partenaire : $Bel_S\, Bel_O \neg Und_O S_j$ (2B'-)

Ces jugements ne sont pas totalement disjoints : par exemple on pourrait considérer que (2A) entraîne (1A), c'est-à-dire que si autrui manifeste ma bonne compréhension (2A) alors je peux considérer que j'ai bien compris (1A). On pourrait, dans cette direction, spécifier tous les axiomes nécessaires à un critère d'ancrage purement logique. Cependant, selon nous, un système purement logique court le risque d'être coupé d'une base procédurale, en particulier (1A) et (2A) représentent des types de jugements bien distincts, acquis par des biais différents. Ils ne représentent pas tant des croyances, que des exigences différentes pour déterminer le statut d'ancrage, d'où notre préférence au terme de « jugement » : une croyance est une attitude épistémique interne, un jugement représente à la fois cette attitude (une exigence propre) et sa manifestation dans le dialogue (une exigence attribuée). C'est pourquoi ce terme nous semble plus approprié pour décrire la dimension interne ou externe de la compréhension.

Tous ces jugements peuvent intervenir dans la gestion de la compréhension. Notre hypothèse est que le participant qui reçoit un énoncé se pose plusieurs questions de compréhension relatives à l'établissement des jugements de compréhension et à l'ancrage des énoncés antérieurs. Formulées du point de vue de S qui reçoit l'énoncé O_i, ces questions sont :

- Est-ce que je juge avoir suffisamment compris l'énoncé de mon partenaire ? ($Bel_S\, Und_S\, O_i$)
- Est-ce que je juge que mon partenaire a suffisamment compris mon énoncé antérieur ? ($Bel_S\, Und_O\, S_j$)
- Est-ce que (je crois que) mon partenaire juge avoir suffisamment compris mon énoncé antérieur ? ($Bel_S\, Bel_O\, Und_O\, S_j$)
- Est-ce que (je crois que) mon partenaire juge que j'ai suffisamment compris son énoncé antérieur ? ($Bel_S\, Bel_O\, Und_S\, O_k$)

Ces quatre types de questions découlent de la combinatoire *auteur du jugement* (J) / *auteur de l'interprétation* (I) et sont résumés dans le tableau 8.1. On appelera par exemple SO le jugement de S sur l'interprétation de O ou OO le jugement de O sur sa propre interprétation.

J\I	S	O
S	SS	SO
O	OS	OO

Tab. 8.1 – Combinatoire des jugements

Pour être parfaitement rigoureux, on devrait noter le jugement OS par SOS, c'est-à-dire le jugement OS du point de vue de S. Cependant, étant donné que les jugements ne seront évalués que d'un point de vue interne, celui de S, nous omettrons cette distinction en supposant toujours que OS réfère à SOS et OO réfère à SOO.

Polarités des jugements En adjoignant une polarité positive ou négative aux jugements, nous pouvons retrouver les catégories de preuves du modèle des échanges. Nous notons toutefois que le modèle des échanges ne considère que les jugements manifestés OO et OS, et pas les jugements propres SS et SO. Mais on peut utiliser les catégories du modèle des échanges pour décrire ces deux types de jugements. Les jugements positifs ou négatifs d'un participant sur sa propre interprétation (SS et OO) correspondent à la preuve UNDERSTOOD (U) ou NOTUNDERSTOOD ($\overline{U}$). Les jugements positifs ou négatifs d'un participant sur l'interprétation de son partenaire (SO et OS) correspondent à la preuve RELEVANT (R) ou NOTRELEVANT ($\overline{R}$). Le tableau 8.2 résume les différentes polarités des jugements. La combinaison de ces jugements détermine à la fois le jugement propre d'un énoncé et le jugement attribué, manifesté dans l'énoncé. Nous dirons alors qu'un énoncé *est* UR lorsque le jugement par celui qui l'interprète est U et R, et qu'il *manifeste* UR lorsqu'il exprime un jugement U et R de la part d'autrui, et nous noterons dans ce cas UR/UR pour représenter à la fois la compréhension propre de l'énoncé (le premier terme) et la compréhension manifestée dans l'énoncé (le second terme).

Jugement	**Positif**	**Négatif**
SS, OO	U	$\overline{U}$
SO, OS	R	$\overline{R}$

TAB. 8.2 – Polarités des jugements

Par exemple, on pourrait annoter le dialogue suivant en termes de jugements de compréhension[2] :

Enoncé	**Jugements de S**	**Jugements de O**
O_1 : je veux une chambre à Paris	O_1 :UR/UR	
S_2 : à Nancy, je vous propose l'hôtel Ibis		S_2 :$U\overline{R}/UR$
O_3 : non, à Paris	O_3 :$\overline{U}/U\overline{R}$	
S_4 : vous voulez aller à Paris ?		S_4 :$UR/\overline{U}$
O_5 : oui c'est ça	O_5 :UR/UR O_3 :$UR/U\overline{R}$	
S_6 : à Paris, je vous propose l'hôtel Lafayette		S_6 :UR/UR

TAB. 8.3 – Exemple de dialogue annoté

Le premier énoncé O_1 est jugé compris et pertinent par S (par défaut le premier énoncé sera toujours considéré pertinent et manifestant la pertinence) et manifestant une bonne compréhension (UR/UR). O juge ensuite l'énoncé S_2 non pertinent vis à vis de la compréhension de « Paris » ($U\overline{R}/UR$). A son tour S juge O_3 comme non compris mais manifestant sa propre mauvaise compréhension ($\overline{U}/U\overline{R}$). S n'est

[2]Ce dialogue est annoté en ne faisant pas l'hypothèse de prépondérance des jugements propres, on ne construira en effet pas de jugement du type $U\overline{R}/UR$ ou $\overline{U}/U\overline{R}$ mais ceux-ci sont donnés ici à titre indicatif (voir section 8.3).

cependant pas certain de ce qu'il faut comprendre et pose une question de clarification que O juge alors $UR/\overline{U}$ (la question est jugée comprise et pertinente par O mais manifestant un problème de non-compréhension). Lorsque S reçoit la réponse, il juge O_5 bien compris et répondant à sa question S_4, ne manifestant aucun problème (UR/UR). De plus, il peut désormais considérer que sa propre compréhension de O_3 est satisfaisante (O_3 :$UR/U\overline{R}$). Enfin sa réponse S_6 est jugée UR/UR par O.

8.2.3.1 Asynchronicité du processus d'ancrage

Le processus d'ancrage d'un énoncé se termine avec succès dès que les participants obtiennent des jugements de compréhension positifs. En fait, le moment précis où les participants peuvent considérer un énoncé ancré est différent selon qu'on se place du point de vue du locuteur ou de l'interlocuteur. De ce fait, l'ancrage peut être qualifié d'*asynchrone* (Denis, Pitel, Quignard, and Blackburn 2007). L'exemple suivant permet d'illustrer le moment à partir duquel on peut considérer l'énoncé ancré de chaque point de vue. Le tableau correspondant (figure 8.4) montre les croyances et le statut d'ancrage après réception des énoncés par chaque participant.

O_1 : je veux une chambre double à Paris
S_2 : à Paris, je vous propose l'hôtel Lafayette
O_3 : à combien est la chambre ?

Fig. 8.1 – Exemple de dialogue sans incompréhension

Enoncé	**Croyances et ancrage de S**	**Croyances et ancrage de O**
O_1	$Bel_S\, Und_S\, O_1$	
S_2		$Bel_O\, Und_S\, O_1 \wedge Bel_O\, Bel_S\, Und_S\, O_1$ $\Rightarrow Bel_O\, Grounded\, O_1$
O_3	$Bel_S\, Und_S\, O_1 \wedge Bel_S\, Bel_O\, Und_S\, O_1$ $\Rightarrow Bel_S\, Grounded\, O_1$	

Tab. 8.4 – Croyances et ancrage du dialogue 8.1

Du point de vue de S, l'énoncé O_1 satisfait les exigences propres de compréhension et son jugement est alors U ($Bel_S\, Und_S\, O_1$). Il manifeste ensuite sa compréhension en initiant une contribution pertinente S_2. Ce n'est que lorsqu'il reçoit O_3 qu'il peut estimer que les exigences de compréhension de O sont satisfaites ($Bel_S\, Bel_O\, Und_S\, O_1$) et qu'alors l'énoncé O_1 lui paraît ancré. Du point de vue de O, dès qu'il reçoit l'énoncé S_2, il peut estimer d'une part s'il croit que S a compris O_1 ($Bel_O\, Und_S\, O_1$) et d'autre part si S le croit également ($Bel_O\, Bel_S\, Und_S\, O_1$). Comme c'est le cas, O peut considérer que son énoncé O_1 est ancré dès S_2.

8.2.3.2 Pourquoi s'arrêter au rang 2 ?

Notre critère d'ancrage estime que deux croyances sont suffisantes pour ancrer un énoncé d'autrui : une croyance de rang 1 sur sa propre compréhension et une

croyance à propos de la croyance d'autrui en cette compréhension (rang 2). Lorsque ces deux croyances sont satisfaites de manière positive, un énoncé est déclaré ancré pour un participant. Les croyances d'ordre supérieur interviennent dans des croyances des participants sur le statut d'ancrage d'un énoncé vu par leur partenaire : $Bel_A\, Bel_B\, Grounded\, A_i$ et $Bel_B\, Bel_A\, Grounded\, A_i$. Ces croyances semblent indispensables lorsqu'un participant croit que les deux participants ne sont pas d'accord sur le statut d'ancrage d'un énoncé, par exemple il serait contradictoire qu'un participant considère un énoncé ancré alors qu'il croit également que son partenaire pense le contraire (exemple 8.2).

O_1 : je veux cet hôtel
S_2 : ok l'hôtel Lafayette
O_3 : non l'hôtel Ibis
S_4 : à l'hôtel Lafayette je vous propose une chambre à 100 euros
O_5 : ah non non je parle de l'hôtel Ibis

FIG. 8.2 – Détection immédiate par O de la mauvaise compréhension de sa preuve

Ici S comprend mal la preuve de mauvaise compréhension donnée en O_3 et croit à tort à l'ancrage de O_1, mais pour représenter le fait que O détecte que S a fait cet ancrage à tort, O devrait manipuler explicitement des croyances d'ordre 3 (ici que $Bel_O\, Bel_S\, Bel_O\, Und_S\, O_1$).

En fait ces cas correspondent à la mauvaise compréhension d'une preuve de compréhension : si l'ancrage n'est pas le même pour les individus c'est qu'ils se sont attribués à tort des exigences de compréhension et qu'ils ont donc mal compris une preuve de compréhension à un certain niveau. Le schéma 8.3 (p. 132) illustre le point de vue de chaque participant sur les croyances de niveau inférieur de leur partenaire. Une divergence à un certain niveau peut être traduite par l'attribution erronée d'une croyance d'un niveau inférieur.

Les participants peuvent ne jamais percevoir la mauvaise compréhension d'une preuve, entraînant alors un ancrage erroné et potentiellement des quiproquos. Au contraire, si l'un des deux participants détecte qu'une preuve a mal été comprise, grâce au principe collaboratif il doit chercher à atteindre le critère d'ancrage de cette preuve. Donc, même s'ils ne détectent pas la divergence de statut de l'énoncé initial, les participants peuvent détecter la mauvaise compréhension des preuves qui ont conduit à cet ancrage erroné. Dans notre exemple, bien que O ne se représente pas explicitement le fait que S croit à tort O_1 ancré, il conserve le jugement que O_1 et O_3 n'ont pas été compris, il peut alors chercher à ancrer O_3 qui aura pour effet d'ancrer O_1.

Nous estimons que les croyances d'ordre supérieur ne sont pas nécessaires dans la mesure où on arrive à relier les preuves de compréhension aux énoncés auxquels elles se rapportent. Si une preuve de compréhension est mal comprise, mais qu'on parvient *a posteriori* à la corriger, et qu'on sait de plus à quel énoncé elle se rapporte, alors le critère d'ancrage de cet énoncé pourra être révisé. Cette remarque suggère que le maintien d'une structure de dialogue est nécessaire pour pouvoir considèrer

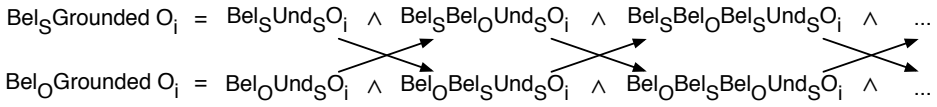

Fig. 8.3 – Réciprocité des croyances partagées

la réinterprétation des preuves de compréhension. Toutefois le problème de la réinterprétation des preuves est difficile, particulièrement dans le cas de la mauvaise compréhension de la preuve de compréhension.

Gestion de la mauvaise compréhension de la preuve La gestion de la mauvaise compréhension de la preuve de compréhension est en effet compliquée. Comprendre une preuve de compréhension peut entraîner des effets complexes en termes de réinterprétation : par exemple une mauvaise compréhension nécessite de réinterpréter l'énoncé initialement mal compris. Si cette preuve est mal comprise, les effets qu'elle jouera seront erronés, par exemple l'énoncé pourrait être mal réinterprété ou pas réinterprété du tout. Le problème est que cette mauvaise compréhension de la preuve peut n'être détectée que tardivement, nécessitant alors de pouvoir se replacer dans le contexte initial pour effectuer la correction. L'exemple de la figure 8.4 montre une mauvaise compréhension d'une mauvaise compréhension qui n'est pas détectée immédiatement.

O_1 : je veux cet hôtel
S_2 : ok l'hôtel Lafayette
O_3 : non l'hôtel Ibis
S_4 : dans cet hôtel, je vous propose une chambre à 100 euros
O_5 : ok, bon et dans l'autre hôtel ?
S_6 : à l'hôtel Ibis, je vous propose une chambre à 50 euros
O_7 : ah mais euh je voulais dire le Lafayette

Fig. 8.4 – Détection différée par O de la mauvaise compréhension de sa preuve

S comprend mal la preuve de mauvaise compréhension en O_3 (un $U\overline{R}$ compris en UR) et initie une nouvelle contribution qu'il croit pertinente (S réfère à l'hôtel Lafayette en S_4). Lorsqu'il reçoit S_4, O n'a aucun moyen de savoir si oui ou non sa correction a été prise en compte puisque S ne mentionne pas explicitement sa compréhension. Notre modèle déclare alors l'énoncé O_1 ancré pour les deux participants dès S_4. En fait O ne peut détecter la mauvaise compréhension de la preuve donnée en O_3 qu'après avoir reçu S_6. Il peut juger par exemple que O_5 a mal été compris, qu'alors O_3 l'a été également et que donc O_1 n'est pas ancré.

Pour autant, il semble difficile d'admettre que la réinterprétation et le ré-examen du statut d'ancrage soit systématique, surtout comme dans l'exemple 8.4 où la divergence de statut ne peut être détectée que tardivement. Le coût de la réinterprétation est certainement très élevé. En l'occurrence, pour savoir que O_1 n'est pas ancré, O

devrait effectuer les opérations suivantes :

1. S a interprété mon expression en O_5 « l'autre hôtel » comme référence à l'hôtel Ibis,
2. donc pour lui lorsqu'il a dit en S_4 « cet hôtel », il faisait référence à l'hôtel Lafayette,
3. donc il a mal compris ma correction en O_3,
4. donc il a mal interprété O_1,
5. et donc O_1 n'est pas ancré.

Reconnaître que O_1 n'a pas été ancré semble très coûteux. Il est probable qu'un participant puisse détecter qu'il y a un problème vis à vis de ses attentes (« l'autre hôtel » est censé référer à l'hôtel Lafayette pour O), mais à notre avis tracer la cause du problème sur plusieurs énoncés est un effort trop important pour pouvoir être effectué systématiquement, particulièrement lorsqu'il est nécessaire, comme dans l'exemple, de se projeter dans l'autre et de tenter d'analyser les expressions de son point de vue. Il est très probable que dans ces circonstances O tente de revenir sur l'axe régissant, par exemple en considérant qu'étant donné qu'il connaît les prix, il puisse choisir son hôtel (comme dans le modèle de Luzzati ou de Lehuen). De deux choses l'une, si O s'est rendu compte que O_1 n'était pas ancré mais a poursuivi dans la tâche, on peut considérer qu'il *abandonne* l'ancrage de l'énoncé. Si O ne s'est pas rendu compte du statut d'ancrage erroné, ce statut demeurera.

La difficulté d'établir le statut d'ancrage dans les cas de mauvaise compréhension de la preuve nous incite à écarter ces situations de notre modèle. Les capacités de réinterprétation nécessaires sont complexes et nécessitent probablement de devoir conserver le contexte mais également les différentes évolutions de ce contexte. Par exemple, si dans le contexte C_1 on reçoit une preuve de mauvaise compréhension, ce contexte est transformé en C_2. Si cette preuve s'avère mal comprise, il faut pouvoir la réinterpréter dans le contexte C_1 et non C_2. Or le modèle des échanges, sur lequel nous nous sommes appuyés ne considère que des modifications du contexte courant. Mais comme nous l'avons noté, le coût de maintien et de réinterprétation peut s'avérer coûteux et de plus amples études cognitives seraient nécessaires afin de déterminer dans quelle mesure nous effectuons la réinterprétation. Nous serons toutefois à même de gérer la non-compréhension des preuves, étant donné que celle-ci *interrompt* l'ancrage d'un énoncé donné pour consacrer le dialogue à la compréhension de la preuve. L'ancrage n'est alors que différé puisqu'une fois que la preuve est crûe comprise, elle peut jouer ses effets vis à vis de l'énoncé initial.

8.2.4 Critère d'ancrage dans le modèle des échanges

Etant donné la nature dialogique des preuves, le modèle doit maintenir une structure courante du dialogue qui lui permet de conserver les relations d'acceptation entre les énoncés. Le modèle des contributions n'est pas tout à fait en mesure de représenter correctement cette structure (à cause de son point de vue omniscient entre autre), et nous préférons nous tourner vers le modèle des échanges qui, en

ligne directe avec le modèle des contributions, fournit des moyens computationnels de construire incrémentalement cette structure.

8.2.4.1 Un critère d'ancrage structurel

Avant de procéder à la définition du processus d'ancrage, il nous est nécessaire de considérer la relation entre la structure et le critère d'ancrage. En effet, ni le modèle des échanges, ni le modèle des contributions ne relie clairement la structure avec le statut d'ancrage. Autrement dit, quelles sont les conditions en termes de structure qui nous indiquent qu'un énoncé est ancré ? Pour pouvoir répondre, nous rappelons la structure du modèle des échanges dans la figure 8.5.

$$
\begin{array}{lcl}
\text{D} & \rightarrow & \text{E}^* \\
\text{E} & \rightarrow & \text{C} \mid \text{C C} \\
\text{C} & \rightarrow & \text{Pr} \mid \text{Ac} \\
\text{Pr} & \rightarrow & \text{U} \\
\text{Ac} & \rightarrow & \text{U} \mid \text{E}+
\end{array}
$$

Fig. 8.5 – Structure du modèle des échanges

Le dialogue (D) est composé d'échanges (E), eux-mêmes composés d'une contribution (C) lorsque l'échange n'est pas terminé ou de deux contributions lorsqu'il l'est. Conformément au modèle des contributions, chaque contribution a une phase de présentation (Pr) et une phase d'acceptation (Ac). La phase de présentation se limite à un seul énoncé (U) et la phase d'acceptation peut soit contenir un unique énoncé, soit une séquence non vide d'échanges. Nous disons que la phase d'acceptation d'une contribution est *close* si elle ne contient qu'un énoncé, ou si la phase d'acceptation de la seconde contribution du dernier échange d'acceptation est close.

Cette définition de la phase d'acceptation en termes de structure est toutefois différente de la phase d'acceptation en termes d'ancrage : dans le modèle des échanges, la clôture de la phase d'acceptation survient lorsqu'on reçoit un énoncé UR/UR. Cependant, la réception d'un énoncé UR/UR ne peut être suffisante pour ancrer un énoncé. Si on se place du point de vue du locuteur O de l'énoncé O_i, un énoncé S_j qui est UR/UR peut certes suffire à déterminer le statut d'ancrage : S_j est R donc les exigences propres de O sur O_i sont satisfaites, et il manifeste U donc les exigences de S sur O_i sont également satisfaites. Toutefois si on se place du point de vue de l'interlocuteur S, manifester un UR/UR dans son énoncé S_j ne saurait suffire à déterminer le statut d'ancrage de l'énoncé O_i : S doit attendre de (croire) comprendre les exigences de O, dans un énoncé ultérieur O_k. Autrement dit, il doit attendre la clôture de la phase d'acceptation de son énoncé en termes de structure.

Nous pouvons alors donner une version structurelle de l'ancrage :

1. pour ancrer le contenu d'une contribution initiée par soi, il faut et il suffit que :
   (a) sa phase d'acceptation soit close par un énoncé d'autrui,
   (b) et que cet énoncé soit jugé UR/UR

2. pour ancrer le contenu d'une contribution initiée par autrui, il faut et il suffit que :
   (a) sa phase d'acceptation soit close par un de ses propres énoncés,
   (b) et que la phase d'acceptation de la contribution présentée par cet énoncé soit close par un énoncé d'autrui,
   (c) et que ce dernier énoncé soit jugé UR/UR

8.2.4.2 Prise en compte de la tâche

D'autre part, le modèle des échanges mélange ce qui relève de la tâche et ce qui relève de la compréhension. Ce n'est pas absurde car établir la compréhension peut être une tâche en soi. Toutefois ce principe entraîne des problèmes si on ne définit l'ancrage qu'en termes d'ancrage structurel. Par exemple, comment considérer les sous-dialogues qui visent à établir une précondition (voir dialogue 8.6) ?

O_1 : je voudrais une chambre d'hôtel
S_2 : où désirez-vous aller ?
O_3 : à Paris
S_4 : à Paris je vous propose l'hôtel Ibis

FIG. 8.6 – Initiation d'un échange de satisfaction de précondition

On peut considérer que l'énoncé S_2 doit initier un échange incident à O_1 dans la mesure où déterminer la ville du séjour est indispensable pour pouvoir répondre à la requête. Il est difficile de soutenir que cette incidence vise à clarifier l'interprétation de l'énoncé. Elle semble plutôt avoir pour but de clarifier l'intention du locuteur vis à vis de la tâche. Initier un échange incident pour traduire la relation entre un but et un sous-but a des conséquences importantes en termes d'ancrage : un énoncé peut être parfaitement compris sans être ancré, puisque sa phase d'acceptation n'est pas encore close. Si on choisit cette direction, l'ancrage doit alors être étendu à la réalisation des buts.

En quoi est-ce gênant ? Nous remarquons que l'ancrage structurel du modèle des échanges revêt un caractère *maximal* : certains aspects de l'énoncé peuvent très bien être ancrés sans pour autant que le critère d'ancrage structurel soit atteint. Ce critère n'est atteint qu'à la condition que l'énoncé satisfasse *toutes* les exigences de compréhension de soi et d'autrui. Le critère d'ancrage structurel ne descend pas au niveau des différents aspects de l'énoncé mais considère l'énoncé dans sa globalité. Si on veut pouvoir effectuer l'ancrage de manière plus fine, comme dans (Poesio and Traum 1997), il nous est nécessaire de prendre en compte également le critère d'ancrage épistémique, par exemple en supposant que la proposition $Und_S\, O_i$ puisse s'appliquer à différents aspects de l'énoncé.

Dès lors il n'y a rien de contradictoire à considérer que les exigences puissent couvrir également la satisfaction des buts : il ne s'agit que d'exigences supplémentaires à satisfaire pour un ancrage *maximal*. Dans notre exemple, rien n'empêche de considérer d'une part que la phase d'acceptation de O_1 n'est pas close mais qu'en même

temps le critère d'ancrage épistémique puisse être atteint au niveau de la compréhension de l'intention communicative. En l'occurrence, le niveau de la gestion des buts correspond au niveau de la conversation chez Clark (1996), où le locuteur propose une activité et l'interlocuteur considère cette activité.

Il nous a paru alors nécessaire d'introduire un nouveau type de jugement, absent du modèle des échanges, à propos de l'exécution d'une requête : les participants peuvent très bien comprendre le but présenté par un énoncé mais peuvent être dans l'incapacité de l'exécuter, soit qu'ils en sont incapables, soit qu'il leur manque des informations. On dira qu'un énoncé est jugé UnderstoodRelevantAccepted (URA) lorsqu'il satisfait les exigences de compréhension et les exigences d'action, et UnderstoodRelevantNotAccepted ($UR\overline{A}$) lorsqu'il ne satisfait que les exigences de compréhension mais pas d'action. Ce jugement correspond à la question « est-ce que je parviens à exécuter la requête présentée en O_i ? ». Nous décorrélons volontairement ce jugement du critère d'ancrage épistémique dans la mesure où celui-ci peut être atteint même en présence d'un jugement $UR\overline{A}$. En outre, nous n'avons besoin que d'une modification mineure du critère d'ancrage structurel tel qu'il a été défini plus haut : au lieu de ne considérer qu'un jugement UR, on peut ancrer un énoncé si l'on en a un jugement URA ou UR.

En conséquence, on ne pourra plus parler à proprement dit d'un axe « régissant » (niveau dialogue) et d'un axe « incident » (niveau acceptation) relativement à la position d'un échange dans la structure mais seulement d'échanges régissants ou incidents, les premiers traduisant des problèmes d'action et les seconds des problèmes de compréhension. Nous continuerons toutefois d'employer le terme d'« axe », qui bien que non reflété dans la structure, exprime bien la direction du dialogue.

Un autre problème important concerne le rôle des contributions vis à vis de l'échange : dans le modèle des échanges une contribution n'a que deux rôles possibles, soit elle *présente* une tâche et initie alors un échange, soit elle *exécute* une tâche et clôt un échange. On ne peut pas considérer les contributions qui jouent un double rôle. Par exemple la réponse à la question « que désirez-vous ? » exécute la tâche qui consiste à connaître le désir de l'utilisateur, mais en présentant ce désir, initie également une tâche. Nous serons contraint de ne représenter qu'un des deux rôles, en l'occurrence l'exécution de la tâche, au risque d'adopter une structure qui ne représente que partiellement la relation à la tâche.

8.2.4.3 Abandon du processus d'ancrage

En cas de problème trop important pour satisfaire les exigences propres ou les exigences d'autrui, ou que l'intérêt de poursuivre l'ancrage est trop faible, il n'est d'autre choix que d'abandonner le processus d'ancrage. Comme nous l'avons montré pour la gestion de la mauvaise compréhension, cet abandon peut être implicite. Il consiste à continuer sur l'axe régissant bien que le participant ait conscience de l'existence d'un problème d'interprétation et du non-ancrage d'un énoncé. Cet abandon peut être également explicite (par exemple « laisse tomber », voir le *Never mind* de Cahn and Brennan 1999) et manifester au partenaire que l'ancrage ne désire pas être atteint. L'abandon traduit alors un relâchement du principe collaboratif

qui entraîne que les exigences de compréhension doivent être suspendues pour un énoncé. Pour représenter ces situations nous introduisons un nouveau type de jugement ABANDONED (*Ab*) qui peut conceptuellement être rapproché de l'$U\overline{R}$. A la différence toutefois des autres jugements qui manifestent des exigences, l'abandon manifeste que les exigences ne peuvent pas être satisfaites ou n'ont pas besoin de l'être. Cette catégorie peut être assimilée au statut *ungroundable* dans le modèle des *grounding acts*, mais modéliser ses effets dans le modèle des échanges est beaucoup plus difficile que le faire dans le modèle des *grounding acts* (voir p. 148).

8.2.5 Conclusion

Nous avons défini le moment où l'on peut déclarer un énoncé ancré : un énoncé est ancré pour un participant donné lorsque ses exigences propres et les exigences qu'il attribue à autrui sont satisfaites. Ce critère d'ancrage a d'abord été modélisé à partir des exigences de compréhension en reliant logiquement les jugements de compréhension au statut d'ancrage. Cependant il manque à ce critère les relations qu'entretiennent les énoncés entre eux et il nous a paru intéressant de modéliser l'ancrage en terme de structure, à l'instar du modèle des échanges. Nous avons défini ensuite un critère d'ancrage structurel, mais celui-ci revêt un caractère maximal tel que le critère épistémique s'avère nécessaire si l'on souhaite prendre en compte l'ancrage au niveau des différents aspects de l'énoncé. Afin de compléter le modèle d'ancrage, il nous reste à spécifier comment la structure est mise à jour, et comment les actions à effectuer pour ancrer un énoncé sont déterminées.

8.3 Processus d'ancrage

La mise à jour de la structure dans le modèle des échanges souffre de nombreux défauts exposés p. 60. Le principal problème est que le résultat de l'interprétation n'est pas pris en compte. Les conséquences sont, d'une part que les preuves de compréhension sont toujours considérées comprises et d'autre part que le modèle s'appuie uniquement sur le jugement d'autrui et pas sur le jugement propre de celui qui interprète l'énoncé. Ces problèmes sont résolus grâce à notre définition des jugements de compréhension. En effet, les jugements propres traduisent directement le résultat de l'interprétation. Le problème de la non-compréhension des preuves peut alors être résolu en considérant le *niveau* du jugement : on peut effectuer un jugement à propos du contenu ou un jugement au niveau de la preuve. Le problème de la prise en compte des jugements propres est résolu en appuyant le processus d'ancrage d'abord sur ces jugements, puis sur les jugements manifestés. La prise en compte de ces deux types de jugements fournit une bonne base à un modèle d'ancrage indépendant du point de vue et nous permet de facilement symétriser le modèle d'échange initial.

8.3.1 Hypothèses sur le processus d'ancrage

Nous effectuons cependant quelques hypothèses sur le processus. On se place avant tout dans le modèle collaboratif où les participants manifestent leur compréhension, et tous les énoncés apporteront au moins une preuve de compréhension potentiellement implicite (à l'exception justement de la non-compréhension de la preuve). Ensuite, toutes les preuves négatives *provoqueront nécessairement* la création d'un contexte dédié à la convergence. Nous ne modéliserons pas le choix de résoudre un problème d'interprétation comme dans (Traum 1999; Paek and Horvitz 2000; Skantze 2007) mais nous le poserons comme nécessité[3].

Nous supposerons également que les jugements SS et SO sont *prépondérants* dans le processus d'ancrage : en cas de jugement négatif de SS ($\overline{U}$) ou de SO ($\overline{R}$), nous considérerons qu'il n'est pas nécessaire de prendre en compte la preuve de compréhension apportée dans l'énoncé. Ce choix peut être motivé par le fait que les exigences propres priment sur les exigences d'autrui : ce n'est que lorsque les exigences propres sont satisfaites que les exigences d'autrui peuvent être considérées.

Enfin nous supposerons que la structure courante et le jugement de compréhension manifesté suffisent à déterminer le rôle de l'énoncé qui vient d'être reçu. Par exemple, si un énoncé ouvre un nouvel échange et que l'énoncé suivant est considéré UR/UR, alors cet énoncé clôt l'échange. Cela exclut toutefois tout raffinement vis à vis des buts courants comme la modification d'un but, issue d'une assertion par exemple. Cette hypothèse nous est imposée par le modèle initial des échanges (C92) dans lequel on ne calcule pas le but d'un énoncé, mais ce but est donné par la structure courante et le jugement manifesté. Le second modèle des échanges (CB99) inclut le but présenté par un énoncé, mais le conflit qui peut exister avec la structure courante n'est pas considéré : par exemple, si un énoncé a initié un échange et qu'on s'attend à une réponse à cet échange, que faire si le nouvel énoncé introduit un nouveau but applicatif parallèle au premier, c'est-à-dire ne satisfaisant ni la relation de dominance ni la relation de précédence selon la terminologie de Grosz and Sidner (1986) ? Nous adopterons toutefois la simplification effectuée par le premier modèle des échanges en ayant conscience de ses limites.

8.3.2 Déroulement du processus

Le processus d'ancrage est composé des étapes suivantes :

1. interprétation de l'énoncé O_n,
2. évaluation des jugements SS, SO, OS et OO,
3. intégration de la contribution présentée par O_n,
4. éventuellement ré-ancrage d'une contribution antérieure,
5. production de la contribution suivante S_{n+1} sur la base des preuves de compréhension à manifester

[3]Toutefois, on peut insérer facilement ce choix dans la constitution des jugements, en améliorant le module d'évaluation à l'aide de paramètres modélisant l'intérêt de l'ancrage. Il suffit concrètement qu'au lieu de considérer une divergence d'interprétation $\overline{U}$, on la considère U.

La première étape pour déterminer les actions à accomplir pour l'ancrage, hormis l'interprétation proprement dite, est d'établir les jugements de compréhension de l'énoncé et manifestés dans l'énoncé. Nous ne donnons pas ici de manière générique d'élaborer ces jugements mais explicitons dans la partie 8.4.2.2 (p. 159) comment on peut les construire à propos de la résolution de la référence. Nous supposerons ici que ces jugements ont déjà été acquis lorsque le système cherche à déterminer l'action adéquate[4].

8.3.2.1 Choix des règles

Le processus de décision a été choisi exclusif : les jugements de compréhension sont logiquement disjoints, et on peut s'appuyer alors sur un simple arbre de décision qui, en fonction des jugements de compréhension et de la structure courante, procède à la mise à jour de la structure, à l'exécution des requêtes et aux réinterprétations nécessaires.

Nous distinguons trois types de règles[5] : les règles de divergence détectée par soi-même (Ab, $\overline{U}$ ou $U\overline{R}$), les règles de divergence détectée par autrui (UR/Ab, $UR/\overline{U}$ ou $UR/U\overline{R}$) et les règles de convergence (UR/UR). Les règles de divergence partagent le fait que l'action suivante du système est entièrement dédiée à la résolution de la divergence. Les différentes situations problématiques sont résumées dans le tableau 8.7. Au contraire, les règles de convergence déterminent l'action du système correspondant au jugement UR/UR, c'est-à-dire en l'absence de problème d'interprétation. Ces dernières ont la particularité d'entraîner l'intégration de nouvelles informations (crûes comprises) et en particulier de provoquer l'ancrage à d'autres niveaux. Elles s'appuient alors exclusivement sur la structure courante (tableau 8.8).

Chacune des règles de mise à jour s'appuie sur deux énoncés : l'énoncé de l'utilisateur O_n qui vient d'être reçu et face auquel le système doit réagir, et l'énoncé antérieur S_k du système par rapport auquel l'utilisateur a réagi. Etant donné la nature récursive de la structure de dialogue, O_n ne correspond pas toujours à l'énoncé qui vient d'être reçu, et S_k ne correspond pas toujours à l'énoncé directement antérieur.

Test	**Méthode**
O_n est Ab	selfAbandon(S_k, O_n)
O_n est $\overline{U}$	selfClarificationRequest(S_k, O_n)
O_n est $U\overline{R}$	selfReject(S_k, O_n)
O_n est UR/Ab	otherAbandon(S_k, O_n)
O_n est $UR/\overline{U}$	otherClarificationRequest(S_k, O_n)
O_n est $UR/U\overline{R}$	otherReject(S_k, O_n)

FIG. 8.7 – Cas de divergence

[4] A l'exception du jugement URA ou $UR\overline{A}$, tous les jugements peuvent être acquis lors de la phase d'interprétation.

[5] On appelera les règles aussi *méthodes*, étant donné leur implémentation en JAVA.

Test	Méthode
O_n est une requête niveau dialogue	integrateDialogueRequest(S_k, O_n)
O_n est une réponse niveau dialogue	integrateDialogueAnswer(S_k, O_n)
O_n est une réponse niveau acceptance	integrateClarificationAnswer(S_k, O_n)

Fig. 8.8 – Cas de convergence, O_n est UR/UR

8.3.2.2 Méthodes du gestionnaire d'application

Le gestionnaire d'application (ou directement l'application) est représenté de manière minimale grâce à trois méthodes que nous n'avons instanciées que pour l'évaluation :

- une méthode *executeRequest* qui exécute une requête applicative, en retourne une réponse ainsi qu'un jugement URA ou $UR\overline{A}$
- une méthode *integrateAnswer* qui intègre une réponse à une requête et retourne un jugement URA ou $UR\overline{A}$ correspondant au résultat de l'intégration
- une méthode *nextAction* qui retourne l'action suivante conformément à l'état courant

Nous ne spécifions pas le formalisme de représentation de l'état courant, on peut s'appuyer tout aussi bien sur une approche de type état d'information, sur une représentation du plan courant ou sur un automate à états finis. La seule contrainte sur le gestionnaire d'application est qu'il doit produire des jugements positifs (URA) ou négatifs ($UR\overline{A}$) à propos des tentatives d'exécution des requêtes ou d'intégration de réponse.

8.3.2.3 Méthodes du module d'interprétation

Nous supposons que le module d'application et le module d'interprétation sont similaires d'un certain point de vue. Leurs méthodes renvoient en effet des jugements positifs ou négatifs d'action (sauf la méthode *interpret* qui calcule également les jugements de compréhension) :

- une méthode *interpret* qui interprète un énoncé et effectue l'évaluation des jugements grâce à un appel de sous-module (voir la section de l'architecture 8.4.1 p. 156)
- une méthode *executeClarificationRequest* qui exécute une requête de clarification, en retourne une réponse et un jugement URA ou $UR\overline{A}$
- une méthode *integrateClarificationAnswer* qui intègre une réponse à une requête de clarification et retourne un jugement URA ou $UR\overline{A}$ correspondant au résultat de l'intégration
- une méthode *integrateReject* qui intègre une correction issue d'un $U\overline{R}$ et retourne un également jugement URA ou $UR\overline{A}$

8.3.3 Détail des règles d'ancrage

8.3.3.1 integrateDialogueRequest

Méthode : intégration d'une requête niveau dialogue
Précondition : O_n est UR/UR et est une requête niveau dialogue
Action : insère O_n dans un nouvel échange niveau dialogue

Cette première règle correspond à l'initiation d'une nouvelle requête au niveau dialogue (le niveau régissant chez Luzzati). Elle introduit alors un nouveau but que le système doit satisfaire. L'action à accomplir pour satisfaire les exigences d'autrui peut se résumer à exécuter l'action correspondant à la réalisation du but et à en manifester le résultat. Etant donné que l'énoncé O_n est compris et qu'il apporte une preuve de compréhension UR de l'énoncé antérieur S_k, il peut être intégré à la structure : il initie la première contribution d'un nouvel échange au niveau dialogue et clôt la phase d'acceptation de S_k (et incidemment la phase d'acceptation d'autres énoncés). La méthode de l'application, *executeRequest*, est ensuite déclenchée et son résultat permet à la méthode *testAcceptance* d'insérer l'énoncé suivant S_{n+1}.

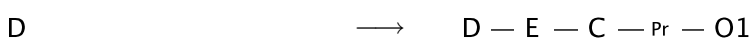

FIG. 8.9 – integrateDialogueRequest après O_1

8.3.3.2 integrateDialogueAnswer

Méthode : intégration d'une réponse à un échange niveau dialogue
Précondition : O_n est UR/UR et est une réponse niveau dialogue
Action : insère O_n comme seconde contribution d'un échange niveau dialogue

Cette règle correspond à la réception d'une réponse à une requête initiée par le système au niveau dialogue. Comme la méthode *integrateDialogueRequest*, étant donné que l'énoncé O_n a été jugé UR, il doit pouvoir être intégré à la structure. En l'occurrence, il présente la seconde contribution de l'échange initié par S_k et peut en clore la phase d'acceptation. Cependant, le résultat de l'intégration de la réponse peut être négatif. La méthode de l'application *integrateAnswer* est alors déclenchée et son résultat détermine où insérer S_{n+1} grâce à la méthode *testAcceptance*.

O_1 : je voudrais une chambre d'hôtel à Paris
S_2 : je vous propose l'hôtel Ibis

FIG. 8.10 – Echange régissant

Fig. 8.11 – integrateDialogueAnswer après O_2

8.3.3.3 selfClarificationRequest

Méthode : initiation d'une requête de clarification portant sur O_n
Précondition : O_n est $\overline{U}$
Action : insère O_n comme contribution flottante et S_{n+1} dans un échange d'acceptation de O_n

Cette situation correspond à la non-compréhension du système. Deux situations peuvent se produire en fonction du niveau de la non-compréhension : si la non-compréhension concerne une preuve de compréhension, il est impossible que l'énoncé O_n puisse jouer un quelconque effet relativement à l'ancrage, sinon O_n peut jouer normalement ses effets. On peut comparer l'exemple 8.12 et l'exemple 8.13. Dans le premier exemple, la preuve de non-compréhension donnée en O_2 n'est pas du tout comprise par le système, les effets de O_2 ne peuvent être pris en compte et O_2 ne peut pas être intégré dans la structure. En revanche dans le second exemple, la non-compréhension porte sur un autre niveau et O_2 est suffisamment compris pour pouvoir être intégré. Cependant, gérer cette intégration est relativement difficile. Nous avons préféré considérer que les deux cas sont similaires, autrement dit que comprendre la preuve de compréhension est conditionnée à la compréhension entière de l'énoncé. Cette hypothèse n'est pas problématique dans la mesure où, lorsque le problème reçoit une réponse satisfaisante, l'énoncé initial O_n doit faire l'objet d'un ré-examen complet et lors de ce ré-examen, l'évaluation de la compréhension peut être effectuée et l'énoncé être réinséré (voir la règle *integrateClarificationAnswer*).

S_1 : je vous propose l'hôtel Ibis
O_2 : l'hôtel quoi ? (+ bruit)
S_3 : quoi ?

Fig. 8.12 – Non-compréhension de la non-compréhension

S_1 : je vous propose l'hôtel Ibis
O_2 : non non pas l'hôtel Ibis, l'autre hôtel plutôt
S_3 : quel hôtel ?

Fig. 8.13 – Non-compréhension à un autre niveau

Pour pouvoir gérer ces cas, nous sommes toutefois contraints d'ajouter un nouvel état d'une contribution (Denis, Pitel, Quignard, and Blackburn 2007). Nous disons

alors qu'une contribution est *flottante* lorsque l'énoncé qu'elle présente n'est pas suffisamment compris pour en extraire une preuve de compréhension. En fait, nous supposons par simplification que c'est systématique en cas de non-compréhension de S. Les contributions flottantes sont des contributions sans parent dans la structure de dialogue, en attente d'une clarification nécessaire pour qu'elles puissent jouer leurs effets relativement à l'ancrage. Lors de la réinterprétation, l'évaluation de la compréhension est effectuée à nouveau et les contributions flottantes peuvent être réintégrées.

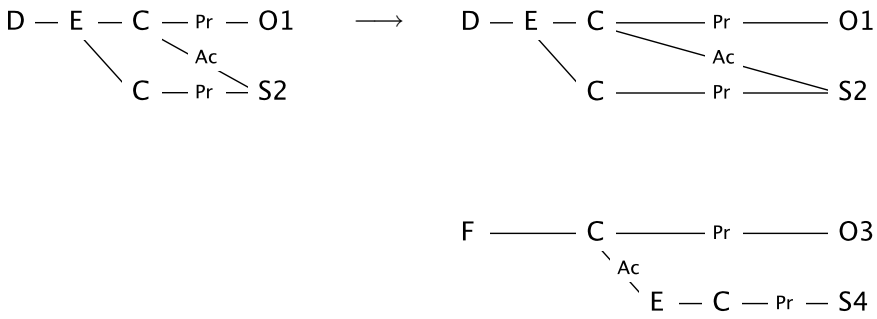

FIG. 8.14 – selfClarificationRequest après O_3 et S_4

8.3.3.4 integrateClarificationAnswer

Méthode : intégration d'une réponse à une question de clarification soulevée par soi
Précondition : O_n est UR/UR et est une réponse niveau acceptation
Action : insère O_n comme deuxième contribution d'un échange niveau acceptation et ré-ancre l'énoncé initial

Cette règle est une des plus intéressantes et correspond à la réponse à un problème d'interprétation de l'énoncé O_i soulevé par le système lors d'un précédent *selfClarificationRequest*. Lorsque le système reçoit une réponse UR/UR à un échange incident qu'il a lui-même initié, il doit chercher à résoudre son problème d'interprétation en mettant en correspondance le type de problème et les nouvelles informations apportées dans l'énoncé O_n. Cette méthode appelle la méthode *integrateClarificationAnswer* du module d'interprétation qui tente de résoudre la question. L'intégration de la réponse doit permettre de réviser l'interprétation d'un énoncé antérieur O_i (nous n'avons cependant implémenté qu'une révision d'ordre référentiel, voir section 8.4.2.4). Si cette révision est un échec, on insère l'énoncé suivant dans l'acceptation de O_n (voir méthode *testAcceptance*).

Si le problème est effectivement corrigé, le processus d'ancrage entier doit à nouveau être appliqué à l'énoncé initialement problématique O_i. En effet, le problème étant résolu, l'énoncé O_i peut potentiellement jouer ses effets, en particulier ceux

relatifs à l'ancrage. C'est-à-dire premièrement on doit le réinterpréter, deuxièmement on doit ré-évaluer la compréhension et troisièmement on doit le réinsérer dans la structure[6].

De nombreux cas peuvent se produire : l'énoncé initialement problématique demeure problématique (son jugement reste $\overline{U}$) par exemple si d'autres problèmes se présentent ; l'énoncé initialement problématique devient non-pertinent (son jugement devient $U\overline{R}$) ; l'énoncé manifeste désormais une non-compréhension (son jugement devient $UR/\overline{U}$) ou une mauvaise compréhension (son jugement devient $UR/U\overline{R}$), etc. L'énoncé suivant doit alors pouvoir manifester au moins deux preuves : une preuve de la compréhension de O_n et une preuve de la nouvelle compréhension de O_i. La sortie du module d'ancrage doit non pas être un simple jugement mais un ensemble de jugements relatifs à plusieurs énoncés antérieurs.

Dans tous les cas, l'énoncé O_n peut être inséré dans la structure puisqu'il est UR, il présente alors la seconde contribution de l'échange de clarification initié par l'énoncé S_k. L'énoncé suivant est conditionné par le rappel récursif au processus d'ancrage appliqué à l'énoncé O_i (mais on doit maintenir les jugements à manifester au cours de cet appel récursif).

O_1 : je prends cet hôtel
S_2 : quel hôtel ?
O_3 : l'hôtel Ibis

Fig. 8.15 – Echange incident issu d'une non-compréhension

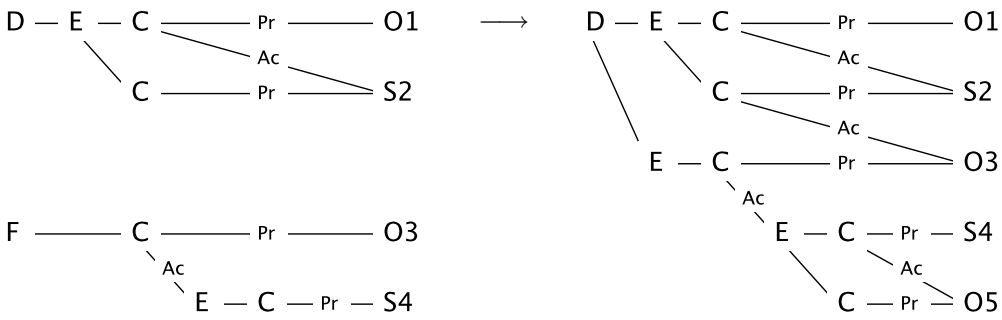

Fig. 8.16 – integrateClarificationAnswer après O_5

[6]Pour faciliter le traitement, on appuie toutes les méthodes sur une méthode *obtainContribution* qui construit une contribution présentée par un énoncé, ou récupère une contribution existante présentée par l'énoncé. Ainsi, l'intégration d'une contribution flottante (pré-existante) ou d'une nouvelle contribution peut être modélisée de la même manière.

8.3.3.5 otherReject

Méthode : intégration d'une preuve de sa propre mauvaise compréhension
Précondition : O_n est $UR/U\overline{R}$
Action : déplace S_k dans un nouvel échange d'acceptation de l'énoncé problématique et insère O_n comme seconde contribution de cet échange

Une autre règle intéressante est celle du *otherReject* où le système reçoit une preuve de sa mauvaise compréhension. Comme dans le modèle des échanges, lorsque le système reçoit une telle preuve, il doit modifier la structure du dialogue afin qu'elle reflète le fait qu'un de ses énoncés ne doit plus être considéré pertinent vis à vis d'un énoncé de l'utilisateur. Typiquement le système a mal compris une requête de l'utilisateur et n'a pas correctement répondu.

Le *otherReject* partage de nombreux points communs avec le *integrateClarificationAnswer*. Dans les deux cas le système est face à un problème d'interprétation, une non-compréhension pour le *integrateClarificationAnswer* et une mauvaise compréhension pour le *otherReject*. Le fonctionnement de ces deux règles est alors similaire en ce qu'ils conduisent tous deux à une altération de l'interprétation et le rappel du processus d'ancrage sur un énoncé révisé O_i. Cependant les deux situations sont différentes sur un point : dans le cas de la non-compréhension, le problème d'interprétation est détecté *a priori* et par le système, dans le cas de la mauvaise compréhension, le problème est détecté *a posteriori* et par l'utilisateur. Cette remarque entraîne que le système doit être capable d'inférer la cause du problème en croisant l'interprétation qu'il a effectuée de O_i avec les nouvelles informations reçues dans l'énoncé O_n.

Cette méthode appelle la méthode *integrateReject* du module d'interprétation. Si la révision de l'interprétation est un échec ($UR\overline{A}$), alors l'énoncé suivant est inséré dans l'acceptation de O_n (voir *testAcceptance*). Si l'altération de l'interprétation est un succès (URA) alors, comme le *integrateClarificationAnswer*, l'énoncé O_i doit faire l'objet d'un nouvel ancrage : réinterprétation, ré-évaluation et réinsertion dans la structure.

La modification de la structure est la suivante : lorsque S comprend que son énoncé S_k n'est pas pertinent vis à vis d'un énoncé O_i, il déplace la contribution initiée par S_k dans un nouvel échange incident à O_i. La seconde contribution de l'échange est la contribution présentée par O_n qui vient d'être reçue. En fonction du succès ou de l'échec de la réinterprétation, S effectue alors sa contribution suivante.

O_1 : je voudrais une chambre d'hôtel à Paris
S_2 : à Nancy, je vous propose l'hôtel Lafayette
O_3 : non à Paris
S_4 : à Paris, je vous propose l'hôtel Ibis

FIG. 8.17 – Echange incident issu d'une mauvaise compréhension

Il existe toutefois un problème d'ambiguïté de portée de l'$U\overline{R}$ illustré dans le dialogue 8.19 et la structure correspondante (figure 8.20). Ici l'énoncé S_k par rapport

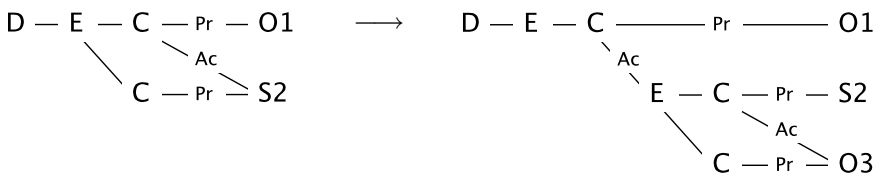

Fig. 8.18 – otherReject après O_3

auquel l'$U\overline{R}$ manifesté en O_5 doit être considéré n'est pas l'énoncé précédent S_4 mais l'énoncé S_2. Pour trouver la bonne portée de l'$U\overline{R}$, nous avons dédoublé la méthode *integrateReject* du module d'interprétation : d'abord une méthode qui *teste* si l'altération de l'interprétation est possible, et ensuite une méthode qui effectue l'altération proprement dite. En l'occurrence, le premier candidat S_4 vis à vis de O_3 invalide le test (impossible de réinterpréter « pardon vous avez dit quoi ? » avec « non non j'ai dit à Paris ». Le second candidat en remontant dans la structure est S_2 vis à vis de O_1 et le test est positif. La méthode *otherReject* s'appuie alors sur cet échange pour effectuer la restructuration.

O_1 : je voudrais une chambre double à Paris (+ bruit)
S_2 : à Nancy, je vous propose l'hôtel Ibis (+ bruit)
O_3 : pardon, vous avez dit quoi ?
S_4 : à Nancy, je vous propose l'hôtel Ibis
O_5 : non non, j'ai dit à Paris

Fig. 8.19 – Ambiguïté de portée de la mauvaise compréhension

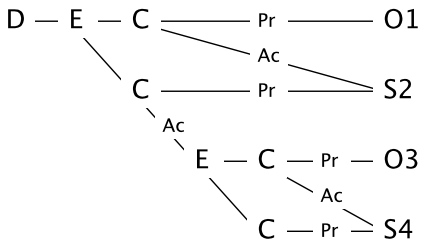

Fig. 8.20 – Structure correspondant au dialogue 8.19 avant O_5

8.3.3.6 selfReject

Méthode : manifestation de la mauvaise compréhension d'autrui
Précondition : O_n est $U\overline{R}$
Action : insère O_n dans un nouvel échange d'acceptation de S_k et S_{n+1} comme seconde contribution de cet échange

Le *selfReject* concerne la production d'une preuve de mauvaise compréhension de la part de l'interlocuteur. C'est un cas basique mais pourtant absent du modèle des échanges. Pour modéliser cette situation nous nous sommes appuyés sur l'autre cas ($UR/U\overline{R}$) qui, lui, est décrit dans le modèle des échanges. Un énoncé qui est jugé $U\overline{R}$ initie alors un échange incident à S_k et traduit directement le fait que O_n n'est pas pertinent vis à vis de S_k. L'énoncé suivant de S doit pouvoir systématiquement jouer le rôle de seconde contribution de cet échange en manifestant la mauvaise compréhension de O.

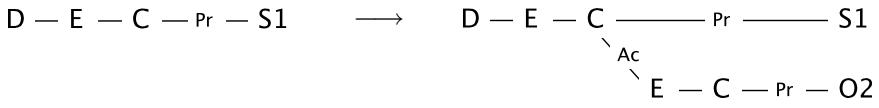

FIG. 8.21 – selfReject après O_2

8.3.3.7 otherClarificationRequest

Méthode : intégration d'une question de clarification d'autrui
Précondition : O_n est $UR/\overline{U}$
Action : insère O_n dans un nouvel échange d'acceptation de S_k

Cette règle recouvre les cas de non-compréhension de l'utilisateur. Comparée aux autres règles, celle-ci est relativement simple. Etant donné que l'énoncé O_n manifeste un $\overline{U}$, il initie systématiquement un échange incident dans la phase d'acceptation de l'énoncé S_k auquel il se rapporte. Ensuite, la méthode *testAcceptance* permet de déterminer que faire si le système ne peut résoudre le problème ou s'il y parvient.

La méthode *otherClarificationRequest* est toutefois sujette au même type d'ambiguïté de portée que la méthode *otherReject*. Nous pouvons appliquer le même raisonnement : en cas de résultat négatif du test de portée, le modèle peut tenter de résoudre la question de clarification en s'appuyant sur un autre énoncé que l'énoncé précédent S_k. Toutefois, étant donné la manière dont les questions sont résolues, il nous a été inutile de modéliser l'ambiguïté de portée des questions de O (voir 8.4.3 p. 170).

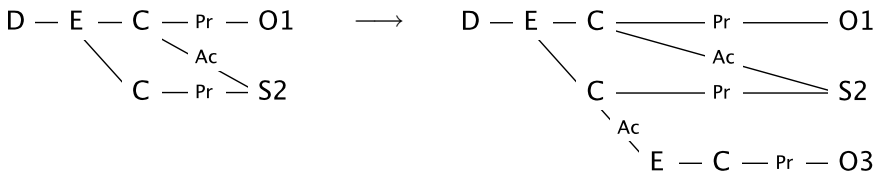

Fig. 8.22 – otherClarificationRequest après O_3

8.3.3.8 otherAbandon

Méthode : intégration d'un abandon manifesté par autrui
Précondition : O_n est UR/Ab
Action : insère O_n dans un nouvel échange d'acceptation de S_k, clôt l'acceptation de toutes les contributions et insère S_{n+1} dans un nouvel échange niveau dialogue

Cette situation correspond à un abandon de l'ancrage effectué par l'utilisateur. Il est relativement difficile d'estimer à quel niveau l'abandon joue ses effets : s'agit-il de l'échange courant, ou de la phase d'acceptation toute entière d'une contribution au niveau dialogue ? Nous considérons alors, par simplification, une interprétation *maximale* de l'abandon, où l'ancrage de toutes les contributions incidentes doit être abandonné. Nous supposons qu'en abandonnant le but d'ancrage, l'utilisateur abandonne également le but régissant sous-jacent. En conséquence, toutes les phases d'acceptation de toutes les contributions de tous les échanges encore ouverts dans l'incidence d'une contribution sont clôts artificiellement par une contribution nulle. L'introduction de cet artefact est problématique puisque le critère d'ancrage structurel s'appuie la clôture de la phase d'acceptation, et qu'en conséquence les énoncés intermédiaires sont considérés structurellement ancrés sans que le critère épistémique soit atteint. Pour pallier à ce problème, nous avons précisé pour chaque échange si celui-ci devait être considéré comme abandonné ou non. Ainsi, on peut déterminer que le critère d'ancrage a été abandonné si l'échange auquel participe un énoncé a été abandonné. Ce marqueur d'abandon nous permet également de modéliser des abandons locaux (voir dans l'exemple complet de la section 9.6 p. 180).

8.3.3.9 selfAbandon

Méthode : manifestation d'un abandon soulevé par soi
Précondition : O_n est Ab
Action : insère O_n dans un nouvel échange d'acceptation de S_k, et S_{n+1} comme seconde contribution de cet échange et clôt l'acceptation de toutes les contributions

Ce cas correspond au propre abandon de l'ancrage de O_n. De la même façon que pour le *otherAbandon* nous avons considéré que l'abandon avait un caractère maximal et définitif. Autrement dit, O_n est ajouté dans un nouvel échange incident à

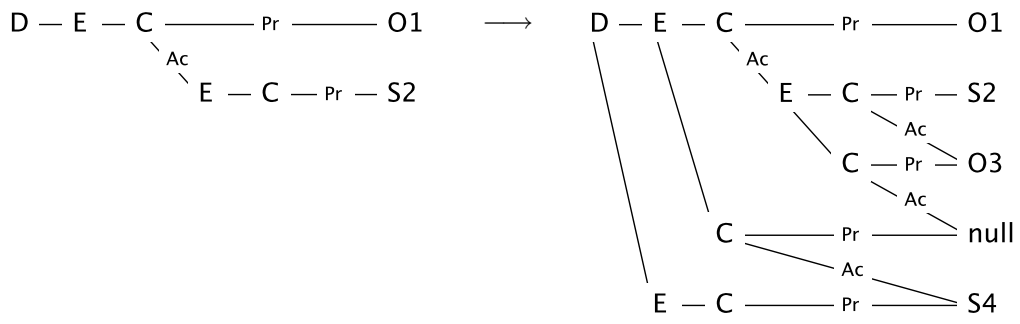

FIG. 8.23 – otherAbandon après O_3 et S_4

S_k et tous les échanges ouverts sont clôts artificiellement par une contribution nulle. Etant donné que le fonctionnement est similaire au *otherAbandon*, les critiques que l'on peut faire sont exactement les mêmes. En outre, ce type de modification de la structure entraîne une complexité importante dans la gestion d'un $U\overline{R}$ ou d'un $\overline{U}$ suivant. Comme dans l'exemple 8.25, il est impossible actuellement de revenir sur un abandon qui, dans notre modèle a un caractère définitif : il devrait être possible de ré-ouvrir les échanges artificiellement clôts. Nous ne pouvons prétendre alors que la solution de l'abandon proposée est satisfaisante.

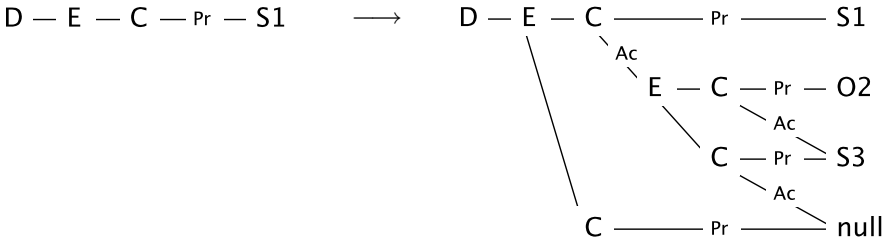

FIG. 8.24 – selfAbandon après O_2 et S_3

S_1 : je vous propose l'hôtel Ibis
O_2 : quel hôtel ?
S_3 : laisse tomber (+ bruit)
O_4 : quoi ?

FIG. 8.25 – Abandon suivi d'un $\overline{U}$

8.3.3.10 testAcceptance

Méthode : insertion de l'énoncé suivant après exécution d'une requête
Précondition : O_n est UR

La dernière règle que nous présentons est particulière puisqu'elle ne concerne pas l'insertion de l'énoncé O_n mais exclusivement de l'énoncé ultérieur du système S_{n+1}. Elle ne s'appuie pas sur la compréhension mais recouvre tous les cas où la compréhension paraît suffisante, mais que l'action correspondante ne peut être exécutée ou reçoit une erreur d'exécution. Il peut s'agir par exemple d'un problème de préconditions manquantes, de problèmes dans la génération d'une réponse à une question de clarification, ou encore de problèmes dans la réinterprétation. Contrairement aux autres règles qui s'appuient sur les jugements de compréhension et qui peuvent être déclenchées immédiatement après l'évaluation de ceux-ci, cette règle ne peut être déclenchée qu'après avoir tenté l'action correspondante. C'est pourquoi elle est appelée après l'exécution de chacune des règles de type UR à l'exception de l'abandon ($UR/\overline{U}$, $UR/U\overline{R}$ et UR/UR) afin de déterminer, une fois que les effets de l'énoncé ont été pris en compte, où la contribution suivante doit être insérée.

Nous avons imposé à chaque action (exécution d'une requête, ou réinterprétation) de retourner un jugement déterminant le succès (URA) ou l'échec de l'action ($UR\overline{A}$). Si l'action est un succès, la contribution présentée par S_{n+1} doit être insérée *après* O_n. C'est-à-dire que si O_n ouvre un échange, S_{n+1} doit clore cet échange. Si O_n clôt un échange, alors S_{n+1} doit être insérée dans l'échange ouvert le plus proche et s'il n'y en a aucun, S_{n+1} doit initier un nouvel échange au niveau dialogue (comme dans la figure 8.27).

Au contraire, si l'action est un échec, la contribution présentée par S_{n+1} doit être insérée dans la phase d'acceptation de O_n en initiant un nouvel échange manifestant implicitement ou explicitement le problème d'exécution (comme dans la figure 8.28).

O_1 : je voudrais une chambre d'hôtel
S_2 : où désirez-vous aller ?
O_3 : à Paris
S_4 : à Paris, je vous propose l'hôtel Ibis

Fig. 8.26 – Initiation d'un échange incident résultant d'un $UR\overline{A}$

Le dialogue 8.26 permet d'illustrer les deux situations. D'abord, O_1 est jugé UR/UR. A l'issue de la méthode *integrateDialogueRequest*, la méthode *testAcceptance* est exécutée pour vérifier que le système peut obtenir une chambre d'hôtel. Celle-ci échoue en renvoyant un $UR\overline{A}$. Dès lors l'énoncé suivant du système doit être inséré comme nouvel échange dans la phase d'acceptation de O_1. Ensuite, O_3 est également jugé UR/UR mais cette fois c'est la méthode *integrateClarificationAnswer* qui est exécutée puisque S_2 a initié un échange incident. Elle provoque la réinterprétation de O_1, qui se déroule sans problème. Le système ré-exécute alors la méthode *integrateDialogueRequest* ainsi que la méthode *testAcceptance*. Cette fois, l'action correspondante réussit et S doit réagir par rapport à O_1 (l'échange ouvert le

plus proche). Il produit alors sa réponse au même niveau que O_1 et S_4 clôt l'échange initié par O_1.

FIG. 8.27 – testAcceptance positive après S_2

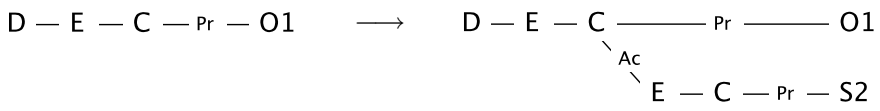

FIG. 8.28 – testAcceptance négative après S_2

8.3.3.11 Gestion des énoncés purement évaluatifs

Certains énoncés purement évaluatifs (par exemple « OK ») ne peuvent initier d'échange car ils ne proposent pas de nouveau but ni n'exécutent de but soulevé lors d'un échange. Dans le modèle des *grounding acts*, ces contributions sont faciles à modéliser puisqu'elles changent simplement le statut d'une *discourse unit*. En revanche, le modèle des échanges les gère avec difficulté puisque toute contribution doit soit initier un but, soit exécuter un but. Les « OK » initient alors de nouveaux échanges (Cahn 1992) mais cette solution n'est pas satisfaisante : si un « OK » présente un nouveau but, par quoi est-il exécuté ? Nous proposons plutôt de considérer les énoncés purement évaluatifs qui n'introduisent pas de but à l'instar du modèle genevois, comme troisième membre d'un échange. Un échange sera alors constitué de l'introduction d'un but, de l'exécution de ce but et de son évaluation lorsque celle-ci est explicite. Le dialogue 8.29 sera alors modélisé avec la structure de la figure 8.30.

O_1 : je voudrais une chambre à Paris
S_2 : je vous propose l'hôtel Ibis
O_3 : ah très bien ça

FIG. 8.29 – Exemple d'énoncé purement évaluatif

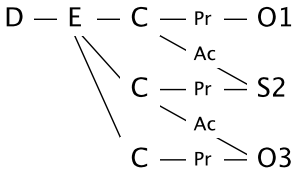

Fig. 8.30 – Structure ternaire d'un échange

8.3.4 Résumé du processus

Le schéma 8.31 résume le processus d'ancrage pour un énoncé O_n en précisant le moment où l'énoncé O_n et l'énoncé S_{n+1} sont insérés dans la structure. Le schéma ne détaille pas les différentes méthodes mais les regroupe en fonction de leur similarité vis à vis de l'insertion de O_n et de S_{n+1}. D'abord l'énoncé est interprété et évalué. Ensuite les méthodes associées aux jugements Ab, $\overline{U}$, $U\overline{R}$ et UR/Ab sont testées et leur exécution provoque l'insertion à la fois de O_n et de S_{n+1}. Sinon, la méthode associée au jugement $UR/\overline{U}$ (*otherClarificationRequest*) est testée et O_n est inséré, puis la méthode *testAcceptance* est appelée et permet d'insérer S_{n+1} dans tous les cas (URA ou $UR\overline{A}$). Sinon, les méthodes associées aux jugements $UR/U\overline{R}$ et UR/UR sont testées et O_n peut être inséré. Ensuite, si le résultat de l'action est négatif ($UR\overline{A}$), S_{n+1} peut directement être inséré. S'il est positif (URA), il est nécessaire de procéder à la réinterprétation de l'énoncé O_i (s'il y en a un) et de n'insérer S_{n+1} qu'à l'issue de cette réinterprétation.

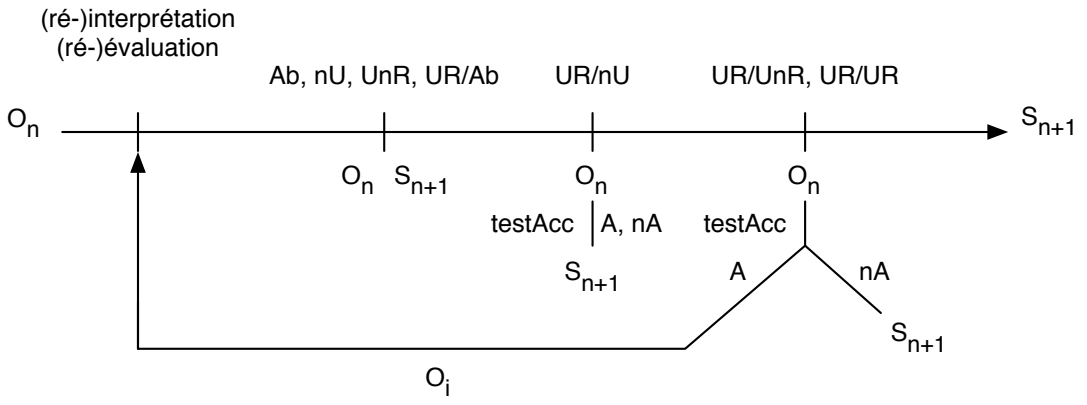

Fig. 8.31 – Résumé du processus d'ancrage

8.3.5 Bilan

Nous avons modélisé des règles pour mettre à jour la structure de dialogue en fonction de son état courant et des jugements de compréhension ou d'action évalués par le système. Nous nous sommes inspirés du modèle des échanges en améliorant les catégories de compréhension. En tant que jugements, ces catégories recouvrent désormais la compréhension propre autant que la compréhension manifestée et permettent d'intégrer facilement le résultat de l'interprétation. Les jugements propres nous ont permis entre autre de symétriser le modèle et d'ajouter les méthodes *selfClarificationRequest* et *selfReject* qui étaient absentes du modèle initial. Enfin, nous avons rajouté deux nouvelles catégories de jugement décrivant d'une part les échecs d'exécution ou de réinterprétation et d'autre par les échecs du processus d'ancrage à travers l'abandon.

8.3.6 Production de la réponse

Le second aspect important du processus d'ancrage concerne la production de la réponse. Nous avons insisté sur la prise en compte *a posteriori* des exigences de compréhension d'autrui grâce à la prise en compte des preuves de compréhension. Toutefois, lors de la production, ces exigences doivent être également prises en compte *a priori* afin de déterminer quelle qualité et quelle quantité de preuves est nécessaire pour satisfaire ces exigences. Nous avons appelé en première partie (section 3.2) cette quantité de preuve, le seuil d'ancrage. Ici, ce seuil d'ancrage doit être évalué par l'interlocuteur lorsqu'il produit son énoncé, et on peut parler de *seuil d'ancrage attribué*, correspondant à la quantité ou qualité des preuves que l'interlocuteur estime nécessaire au locuteur. D'autres aspects entrent en jeu lors de la production, en particulier la nécessité de manifester des jugements à propos de *plusieurs énoncés* à cause de la réinterprétation ou la possibilité de manifester l'évolution des jugements. Enfin, nous devons considérer également le choix de la preuve à produire en fonction de la confiance que le système a en son interprétation, pour distinguer les confirmations explicites des confirmations implicites. Ces différents aspects ne sont pas au coeur de la thèse, mais nous effectuons ici quelques propositions afin de les prendre en compte.

8.3.6.1 Preuves multiples et évolution des jugements

Tout d'abord, la manifestation de la compréhension peut concerner *plusieurs* énoncés. Le cas typique est celui de la clôture d'un échange incident qui entraîne la réinterprétation d'un énoncé antérieur (méthode *integrateClarificationAnswer*). Le système doit alors être capable de manifester une preuve de compréhension de l'énoncé qui clôt l'échange mais également une preuve de l'énoncé réinterprété (et ce récursivement). Nous proposons de manifester la liste des différentes preuves au moyen de connecteurs discursifs (voir Byron and Heeman 1997). Les connecteurs correspondent à une information supplémentaire qui permet de traduire le résultat de la réinterprétation. Par exemple, si la réinterprétation réussit et que l'énoncé

antérieur, préalablement $\overline{U}$ devient U, alors le système doit coordonner les deux preuves par un « donc ». Si en revanche elle échoue, ou qu'un problème à un autre endroit se pose, alors il doit coordonner les deux preuves par un « mais ». Le dialogue 8.32 illustre la réinterprétation réussie de O_1 : O_1 qui était $\overline{U}$ est UR après la réception de O_3. L'énoncé S_4 doit apporter deux preuves : une preuve que O_3 a été jugé UR et une preuve que désormais O_1 est jugé UR. La première est traduite par un « OK » et la seconde par une paraphrase et les deux sont coordonnées par un « donc ».

O_1 : je prends cet hôtel
S_2 : quel hôtel ?
O_3 : l'hôtel Ibis
S_4 : OK donc l'hôtel Ibis

Fig. 8.32 – Exemple de manifestation de réinterprétation réussie

Indépendemment de la nécessité de produire *plusieurs preuves*, il est intéressant de manifester comment la compréhension évolue ou n'évolue pas. Si l'on souhaite pouvoir manifester cette évolution, il est nécessaire de conserver un historique des différents jugements portés sur un énoncé (voir méthode *integrateClarificationAnswer*). Par exemple : deux $\overline{U}$ d'affilée qui portent sur le même problème pourront être traduits par « je ne comprends *toujours* pas », ou un U qui devient $\overline{U}$ pourra être traduit par un « je ne comprends *plus* ».

La production de tous ces marqueurs dépasse la simple manifestation de la compréhension et traduit l'*évolution* de la compréhension. Manifester l'évolution de la compréhension revient à expliciter l'ancrage en tant que processus. Il ne s'agit pas seulement d'atteindre la convergence, mais de manifester comment cette convergence est atteinte, autrement dit de manifester l'état courant *vis à vis* de l'état antérieur : le « donc » permet à l'utilisateur de déduire qu'il y a eu une progression effective, le « je ne comprends *toujours* pas » traduit l'absence de cette progression. Nous n'avons pas détaillé outre mesure cette dimension, mais considérons que cette direction de recherche serait très prolifique.

8.3.6.2 Seuil d'ancrage attribué

Il est nécessaire également de considérer le seuil d'ancrage attribué à l'utilisateur. Produire *trop* de preuves de compréhension irait à l'encontre de la maxime de quantité de Grice ou du principe de moindre effort collaboratif. Produire des preuves *insuffisantes* entraînerait l'injonction probable de l'utilisateur au système de démontrer sa compréhension. Estimer ce seuil est relativement difficile. Nous proposons de faire l'hypothèse que la quantité de preuves à produire est directement corrélée à l'existence de divergences : le système suppose alors qu'en présence de divergence, l'utilisateur exigera davantage de preuves. Dès lors plus il existe de divergences, plus le système doit manifester précisemment sa compréhension. Pour modéliser ce fonctionnement nous proposons un algorithme naïf : on associe à chaque type de

preuve un niveau de verbosité ; en présence de divergence ($\overline{U}$ ou $U\overline{R}$) augmenter un certain compteur γ, et en présence de convergence diminuer ce même compteur. On ne produit alors que les preuves dont le niveau de verbosité est inférieur à γ.

O_1 : je prends cet hôtel
S_2 : quel hôtel ?
O_3 : le lac
S_4 : je ne comprends pas. Quel hôtel ?
O_5 : l'hôtel du lac
S_6 : je ne comprends pas « l'hôtel du lac », ça ne réfère à rien. Quel hôtel ?

FIG. 8.33 – Quantité de preuve variable

Le dialogue 8.33 illustre la variation du seuil d'ancrage attribué. Le premier $\overline{U}$ (S_2) est manifesté comme une unique requête de clarification. Le second $\overline{U}(S_4)$ explicite la non-compréhension de S. Le troisième $\overline{U}$ (S_6) affine la manifestation de la non-compréhension en exhibant la localisation du problème et le type de problème.

8.3.6.3 Score de confiance

Nous avons montré (partie 3.2.4) que la confiance d'un participant en son interprétation influait sur la manière de manifester une même preuve. Par exemple on souhaite distinguer :

U : je veux aller à Paris
(1) S : Quand désirez-vous partir ?
(2) S : OK. Quand désirez-vous partir ?
(3) S : A Paris. Quand désirez-vous partir ?
(4) S : Voulez-vous aller à Paris ?

FIG. 8.34 – Différentes confirmations possibles

Les trois premières réponses peuvent être distinguées de la quatrième dans le type de jugement manifesté (UR opposé à $\overline{U}$). Et les trois premières peuvent être distinguées entre elles en fonction de la confiance qu'a le système en son interprétation. Pour chaque type de preuve on peut associer une *manière* de la manifester en fonction du score de confiance. Pour les UR, on pourrait à titre d'exemple s'appuyer sur un tableau du type suivant :

Score	**Type de preuve**
1.0	implicite
$>$0.8	*acknowledgement*
$>$0.6	paraphrase

TAB. 8.5 – Preuve UR à manifester en fonction du score de confiance

8.4 Implémentation dans un système de dialogue

Nous présentons ici comment le modèle d'ancrage peut être instancié sur la référence en spécifiant trois aspects de l'ancrage : l'interprétation, l'évaluation de l'interprétation et la ré-interprétation. Nous supposons que ces trois aspects peuvent être considérés à tous les niveaux mais la généralisation restant difficile, nous nous focalisons sur la résolution de la référence[7]. Il nous faut tout d'abord nous doter d'une représentation flexible du résultat du traitement de la référence. Ensuite il nous est nécessaire de déterminer comment les jugements de compréhension de la référence peuvent être construits et manifestés. Enfin, nous devons considérer comment le système est capable de réinterpréter les expressions référentielles. Avant d'expliciter l'implémentation de chacun de ces aspects, nous présentons l'architecture du système dans sa globalité.

8.4.1 Système de dialogue avec ancrage

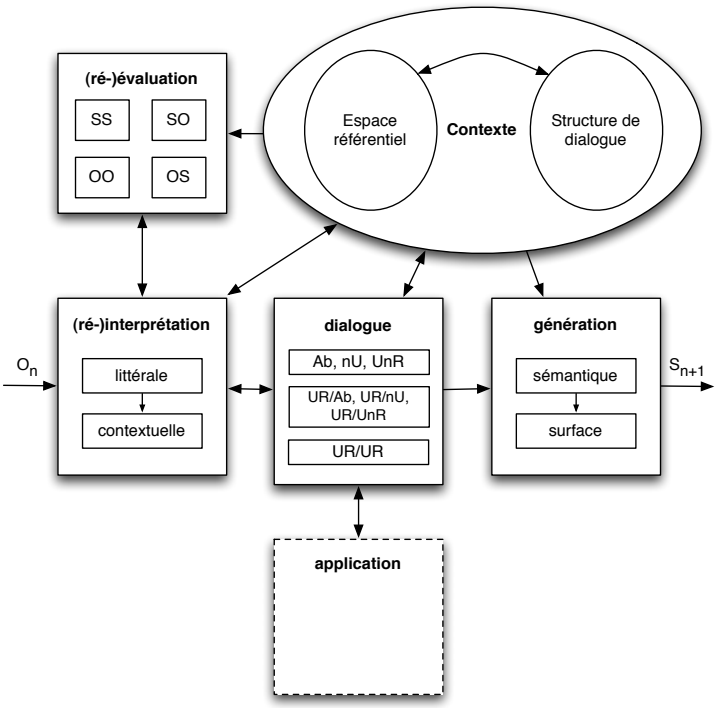

Fig. 8.35 – Architecture

Le système est classiquement composé de trois modules principaux : le module d'interprétation, de dialogue et de génération. Le module d'interprétation a déjà été présenté dans la section 6. On lui adjoint un module d'évaluation destiné à acquérir

[7]Voir par exemple (Poesio and Traum 1997; Larsson 2002; Ginzburg, Fernandez, and Schlangen 2007) qui sont autant d'approches tendant à la gestion de l'ancrage à plusieurs niveaux.

les jugements d'interprétation SS, SO, OO et OS. Ce module, relativement indépendant, s'appuie sur le contexte afin d'estimer les jugements. Il reçoit des requêtes d'évaluation du module d'interprétation et en retourne les résultats. Le module d'interprétation communique dans les deux sens avec le module de dialogue : d'une part le résultat de l'interprétation évalué fait l'objet du processus d'ancrage et d'autre part le module de dialogue peut requérir la réinterprétation ou la ré-évaluation des énoncés. Le module de dialogue communique avec l'application dans les deux sens : pour exécuter des requêtes applicatives, ou pour recevoir les résultats de ces requêtes. Enfin le module de génération produit la réponse en fonction des preuves de compréhension à produire.

La modularité de l'architecture facilitera l'évaluation de l'ancrage. Il nous suffira en particulier d'associer des modules d'évaluation et d'application spécifiques. Elle permet en outre de considérer des développements séparés de chaque module. De plus, étant donné que le module de dialogue ne perçoit que l'*existence* d'un problème en tant que jugement de haut niveau, celui-ci est relativement indépendant du contenu du problème. Seule lui importe la catégorie ($\overline{U}$, $U\overline{R}$, $UR\overline{A}$, ou URA) pour déterminer où insérer les énoncés. La résolution même des problèmes est effectuée directement par le module d'interprétation (pour l'$U\overline{R}$), ou par les instances qui représentent les questions de clarification (pour le $\overline{U}$).

8.4.2 Instanciation sur le problème de la référence

8.4.2.1 Représentation de la référence

Comment représenter l'ancrage de la référence ? Nous avons défini l'opération de référence d'abord comme une mise en relation d'une expression référentielle (ER) avec un référent (ensuite comme un ensemble d'opérations de restructuration). Notre approche consiste à considérer les deux niveaux, celui des ER, et celui des référents, comme dans Heeman and Hirst (1995). Plusieurs phénomènes suggèrent que les deux niveaux sont indispensables à l'ancrage, ou même à la gestion de certains cas « normaux » de référence.

Il est d'abord possible de référer à des expressions référentielles, grâce à une opération que nous avons appelée métaréférence. Le dialogue 8.36 (p. 158) illustre la référence à une ER par le système, la métaréférence « "cet hôtel" » réfère, grâce au contenu lexical, à l'expression « cet hôtel » utilisée en U_1. On peut également référer à une expression sans passer par son contenu, par exemple « je ne comprends pas cette expression »[8]. La métaréférence est un cas particulier d'un phénomène plus vaste appelé *métalangage*. Son utilisation dans la gestion du dialogue est massive. Par exemple Anderson and Lee (2005) citent le nombre de 54.93% d'énoncés dédiés à la gestion du dialogue qui incluent des phénomènes d'ordre métalinguistique.

Il est souvent fait mention du besoin de localiser la source d'une erreur. La métaréférence est en effet nécessaire lorsque plusieurs ER de tête identique sont

[8]L'utilité d'avoir un niveau représentant le matériau linguistique est flagrante dans des exemples du type « je ne connais pas ce mot ».

U_1 : je prends cet hôtel
S_2 : je ne comprends pas
U_3 : qu'est-ce que tu ne comprends pas ?
S_4 : « cet hôtel »

Fig. 8.36 – Exemple de métaréférence

mentionnées dans un énoncé et que l'une d'entre elle requiert une clarification en excluant les autres. Le dialogue 8.37 illustre la préférence à la mention explicite de l'expression problématique plutôt qu'à la seule question « quel hôtel ? » qui serait ambiguë. Selon nous les énoncés « je ne comprends pas "l'autre hôtel". Quel hôtel ? » prennent la valeur de « quel est le référent de type hôtel de "l'autre hôtel" ? ».

U_1 : à combien cet hôtel et euh l'autre hôtel aussi ?
S_2 : je ne comprends pas « l'autre hôtel ». Quel hôtel ?

Fig. 8.37 – Exemple de multiples ER de tête identique

D'autre part, il semble que certains phénomènes référentiels nécessitent de considérer la forme de l'expression référentielle indépendemment de son contenu, en particulier les répétitions, comme le suggère l'exemple 8.38. Le traitement de l'altérité suppose l'identification d'un objet-repère (Salmon-Alt 2001), mais dans l'exemple 8.38, cet objet-repère est le même pour le compère et le client malgré la focalisation opérée par la première expression d'altérité (phénomène que nous avons appelé double altérité). Selon nous, l'expression employée par le client peut désigner le même référent que le compère en vertu de l'identité de forme. On peut supposer que le client emploie cette expression en se plaçant du point de vue du compère et qu'alors la coréférence est possible. Ce phénomène très intéressant mériterait toutefois une analyse plus approfondie[9].

compère : souhaitez-vous réserver dans l'autre hôtel ?
client : oui à ce moment-là dans l'autre hôtel

Fig. 8.38 – Exemple de double altérité (dialogue 1359)

Ces exemples suggèrent que les deux niveaux sont indispensables. Pour les représenter dans les domaines de référence, nous avons considéré que l'interprétation des ER provoquait la création de deux représentations mentales, une pour l'ER et une pour le référent. Les ER appartiennent à un domaine de référence de type ontologique ReferringExpression, dont les éléments ont deux attributs : hasReferent et hasContent. La valeur de l'attribut hasReferent est l'identifiant d'une représentation mentale existante (dénotant potentiellement un domaine), et celle de hasContent, une chaîne de caractères. Chaque interprétation d'une ER conduit à la création d'une nouvelle entité dans le domaine des ReferringExpression discriminée par son

[9]Il pourrait par exemple relever d'un alignement à la fois lexical et référentiel (au sens de Pickering and Garrod 2004).

index. Nous introduisons également un solveur référentiel associé (`MetaReference`), très contraint pour éviter toute résolution non désirée, comme les références pronominales par exemple. La résolution d'une expression du type « le référent de "l'autre hôtel" » est effectuée de manière similaire à « le prix de la chambre ». Nous ne tirons toutefois pas partie du potentiel d'expressivité de la métaréférence, en particulier avec des expressions du type « cette expression », mais nous nous appuyons essentiellement sur ces deux niveaux pour considérer la réinterprétation de la référence en tant que révision de la relation entre une ER et un référent.

8.4.2.2 Constitution des jugements de compréhension

Nous présentons ici comment représenter et obtenir les différents jugements de compréhension.

Niveaux considérés Le niveau d'interprétation principal de notre étude est le niveau de la référence. Nous serons concernés prioritairement par le niveau du référent et ne considérerons pas tous les problèmes de *construction* de l'expression référentielle au niveau lexical, syntaxique ou sémantique. Nous pourrons néanmoins considérer des expressions référentielles mal construites étant donné que ces divergences influencent l'identification des référents mais nous ne traiterons pas des problèmes à ces niveaux.

Lors du processus d'ancrage d'autres niveaux doivent être pris en considération. D'abord, nous avons noté qu'un des principaux défauts du modèle des échanges était d'estimer que les preuves de compréhension étaient toujours bien comprises. Etant donné que nous préférons considérer que les preuves peuvent être non-comprises, il nous est nécessaire de prendre en compte le niveau de la compréhension de la preuve. Ensuite, nous avons montré l'importance du résultat de la réinterprétation en tant que tel et qu'alors les participants peuvent effectuer des jugements de compréhension au niveau de la réinterprétation de la référence. Enfin, nous devons également prendre en compte la compréhension des requêtes de clarification et considérer le résultat d'exécution de ces requêtes. Les niveaux considérés seront alors : référence, réinterprétation de la référence, preuve de compréhension et requête.

Jugements de S La compréhension issue de son propre jugement est effectuée en deux temps : SS, son propre jugement de sa propre interprétation, puis SO, son propre jugement de l'interprétation d'autrui. Dans les deux cas on peut modéliser cette compréhension comme une attente : attente d'une interprétation unique, complète et certaine pour SS et attente d'une interprétation pertinente pour SO.

Construire le jugement SS Le système construit le jugement SS sur la base d'attentes génériques liées à la bonne formation de l'énoncé ou à sa relation à la tâche. La première difficulté lorsqu'on cherche à construire ce jugement est la présence d'une collection de problèmes : d'abord il peut y avoir des problèmes à différents niveaux, par exemple s'il existe un problème lexical, il y aura probablement un problème

syntaxique, qui entraînera un problème sémantique puis référentiel. Comme l'ont noté Caelen and Nguyen (2006), le système devrait construire une représentation des problèmes qui peut faire sens pour l'utilisateur. Nous estimons alors que le système devrait se limiter aux aspects sémantiques, référentiels ou intentionnels et en l'occurrence nous nous restreignons au niveau référentiel.

Ensuite, il peut exister plusieurs problèmes à différents endroits de l'énoncé. En toute logique, le jugement SS devrait alors prendre en compte ces différents problèmes, et le système devrait être capable de chercher à les résoudre itérativement. Nous préferons alors considérer que le jugement SS porte sur *un unique problème*. Etant donné que les problèmes peuvent être reliés, résoudre un problème à un endroit peut conduire à résoudre un problème à un autre endroit. Par exemple dans « je voudrais une chambre double et qu'elle ait la douche », s'il existe un problème sur « une chambre double » et sur « elle », il est probable que corriger le problème sur « une chambre double » conduira à la disparition du problème sur « elle ». En fait, lorsqu'un problème de compréhension est résolu, le système doit ré-évaluer la compréhension lors de la phase de réinterprétation. Dès lors on peut se contenter d'associer au jugement SS un unique problème, le problème courant, de préférence le plus important si l'on parvient à hiérarchiser les problèmes. Nous nous sommes contentés cependant de ne considérer que le premier problème rencontré.

L'attente de compréhension de la référence vis à vis de sa propre interprétation d'un énoncé est d'associer à chaque expression référentielle de l'énoncé un unique et probable référent. Une interprétation de la référence qui est soit *ambiguë* (plusieurs référents potentiels), soit *vide* (aucun référent), soit *incertaine* (un référent avec une probabilité faible) est associée à un jugement $\overline{U}$, dans tous les autres cas le jugement associé est U[10]. On s'appuie sur le nombre d'attributs hasReferent associé à une ER. Nous avons décrit ici cette attente en termes de résultat de l'opération de référence et pas en termes d'attentes sur le processus de résolution lui-même. Par exemple, on pourrait considérer l'identification du domaine de référence et la restructuration comme deux étapes distinctes pouvant présenter des problèmes (voir section 6.1.3.1). Mais étant donné qu'*in fine*, les problèmes qui se présentent à ces niveaux se répercutent sur le résultat de la résolution, nous nous sommes limités à décrire l'attente en termes de résultat.

Nous devons ajouter à cette classification les problèmes de compréhension de la preuve, les problèmes d'exécution d'une requête et les problèmes de réinterprétation. Nous avons toutefois beaucoup limité la représentation de ces problèmes. En ce qui concerne la compréhension de la preuve, ne pas comprendre une preuve de compréhension revient à dire qu'on ne connaît pas les effets de l'énoncé qui la manifeste relativement à l'ancrage. Nous avons effectué à ce sujet l'hypothèse de prépondérance des jugements propres, qui implique que lorsqu'un énoncé est non-compris, la preuve de compréhension qu'il porte ne peut être calculée. Représenter la non-compréhension de la preuve dans ce cas n'est pas utile puisqu'elle est conditionnée à la non-compréhension du contenu. Il peut toutefois arriver que tous les niveaux

[10]Nous n'avons pas implémenté l'*incohérence*. Elle nécessite en particulier de considérer des niveaux de description du contexte que nous n'avons pas représentés.

soient effectivement compris mais que seul le niveau de la preuve pose problème. Dans notre implémentation ce cas n'arrive pas étant donné que nous construisons les preuves sur la base de la force illocutoire, toujours présente (voir plus bas la construction des jugements d'autrui). Au final, dans l'implémentation, la preuve est toujours crûe comprise lorsqu'aucun autre problème d'interprétation ne se présente. Cela ne signifie toutefois pas qu'elle soit effectivement comprise, et nous avons déjà abordé la difficulté de la mauvaise compréhension de la preuve. Afin de représenter les problèmes au niveau de la preuve, il serait nécessaire de mieux calculer cette preuve au préalable en prenant plus en compte le contenu propositionnel. On pourrait alors considérer l'incertitude de la compréhension de la preuve en manifestant explicitement un énoncé du type « ai-je bien compris l'énoncé ? » par exemple. Nous n'avons pas implémenté ce type de requête mais le modèle est ouvert à cette possibilité. Nous avons également limité les problèmes d'exécution d'une requête à l'existence ou l'absence d'un problème (URA ou $UR\overline{A}$). Le seul problème dont nous avons spécifié le contenu est une requête inconnue UnknownRequest. Pour aller plus loin on pourrait s'inspirer du modèle de Nerzic (1993) dans la description de ces problèmes : précondition manquante, condition d'applicabilité fausse, ou action manquante. Les problèmes de réinterprétation (Unchanged, Missed, ou Unprovided) seront quant à eux explicités dans la section 8.4.2.4 (p. 165).

Nous représentons alors les problèmes de SS ($\overline{U}$ ou $UR\overline{A}$) par :

- un type de problème
- un niveau de problème (ici référence, réinterprétation, preuve ou requête)
- une localisation de problème (un énoncé, un composant ou une entité MMIL)

Les différents types de problèmes et leur niveau sont illustrés dans le tableau 8.6.

Type	**Niveau**	**Description**
Empty	référence	référence vide
Ambiguous	référence	référence ambiguë
Uncertain	référence	probabilité de résolution insuffisante
Uchanged	réint. référence	référence inchangée
Unprovided	réint. référence	réinterprétation impossible
Missed	réint. référence	expression référentielle manquante
UnknownEvidence	preuve	preuve de compréhension manquante
UnknownRequest	requête	requête inconnue

Tab. 8.6 – Type de problèmes par niveau

Construire le jugement SO De la même manière, on peut représenter la compréhension SO comme une attente sur l'énoncé O_n, cette attente est cependant de différente nature puisqu'elle est reliée à la pertinence de l'énoncé vis à vis de ses propres énoncés antérieurs. Dans un modèle de paires adjacentes, le premier membre de la paire impose des contraintes sur le second membre (par exemple une question impose une obligation de réponse). Si ces contraintes ne sont pas satisfaites alors on construit une compréhension de jugement $\overline{R}$, si elles le sont, alors le jugement est R.

Au niveau de la référence, nous avons implémenté plusieurs types de contraintes. D'abord la contrainte de type imposée par les questions de clarification : si une question qui attend un certain type d'entité, par exemple « Quel hôtel ? » ne reçoit pas une réponse du bon type (*modulo* la subsomption), alors la réponse est jugée $\overline{R}$, sinon elle est jugée R. La réponse est également jugée R si elle invalide l'ER, par exemple « aucun » en réponse à « Quel hôtel ? ». On pourrait néanmoins voir la réponse « aucun » à une question d'identification comme une manifestation de la non-pertinence de la question, c'est-à-dire la manifestation d'un $\overline{R}$: le système a construit à tort une expression référentielle qu'il croit référer à un hôtel mais ce n'est pas le cas. Nous n'avons toutefois pas approfondi la gestion de la non-pertinence des questions de clarification, ni du point de vue du locuteur, ni de celui de l'interlocuteur, car elle nécessite probablement d'identifier les présuppositions fausses effectuées par l'un ou par l'autre, comme dans les travaux de Hirst et al. (1994), McRoy and Hirst (1993).

Ensuite, nous avons implémenté deux contraintes de type paraphrase utilisées uniquement lors de l'évaluation. Dans le contexte de l'évaluation (chapitre 10), les participants doivent manifester *explicitement* leur compréhension. Les contraintes utilisées sont donc beaucoup plus fortes que celles nécessaires dans le dialogue. La première est une contrainte faible qui impose, pour chaque référent mentionné dans un de ses propres énoncés S_k, l'existence d'une expression référentielle qui y réfère dans l'énoncé O_n. Si c'est le cas l'énoncé est jugé R, sinon $\overline{R}$. A cet égard, l'énoncé O_2 de la figure 8.39 sera considéré pertinent. Cette contrainte n'était toutefois pas suffisante pour déterminer si la compréhension était bonne puisqu'il suffit de produire des pronoms du type « ça » pour considérer l'énoncé pertinent. La seconde contrainte, plus forte, inclut la contrainte faible mais impose en outre que l'expression référentielle corresponde à une description du référent. L'énoncé O_2 figure 8.39 ne pourrait être considéré pertinent car ni « la chambre », ni « cet hôtel » ne satisfont la contrainte de description. En revanche l'énoncé O_2 figure 8.40 satisfait ces contraintes grâce à la présence de « double » et du nom de l'hôtel.

S_1 : je vous propose une chambre double à l'hôtel Ibis
O_2 : à combien est la chambre dans cet hôtel ?

Fig. 8.39 – Exemple d'énoncé pertinent avec contrainte faible

S_1 : je vous propose une chambre double à l'hôtel Ibis
O_2 : à combien est la chambre double à l'hôtel Ibis ?

Fig. 8.40 – Exemple d'énoncé pertinent avec contrainte forte

L'abandon A l'instar du modèle de Luzzati (1995), nous considérons des variables interactionnelles afin de décider s'il doit y avoir abandon du processus d'ancrage. Plusieurs caractéristiques entrent en jeu : la profondeur de l'échange courant, le nombre d'échanges dans la phase d'acceptation d'une contribution donnée et le

nombre de contributions flottantes. L'abandon ne peut être déclenché que si l'énoncé O_n qui vient d'être reçu est $\overline{U}$ ou qu'il manifeste un jugement $\overline{U}$. Dans ce cas, si le nombre d'échanges directement dans l'acceptation de O_n est supérieur à un seuil θ, ou que la profondeur de l'acceptation de O_n est supérieure à θ, ou que le nombre de contributions encore flottantes est supérieur à θ, le jugement propre de O_n est Ab. Nous n'avons implémenté qu'un seul seuil θ pour la largeur ou la profondeur de l'acceptation, ainsi que pour le nombre de contributions flottantes mais nous pouvons supposer que celui-ci peut être différent dans chacun des cas, par exemple en donnant un poids plus important à la profondeur. Etant donné que le critère d'abandon inclut le nombre de contributions flottantes, nous avons observé des situations où le système ne cherchait plus à comprendre lorsqu'un nombre supérieur de contributions flottantes à θ était atteint. Autrement dit, le système quitte définitivement le modèle collaboratif. Ce comportement n'est sans doute pas souhaitable et il serait nécessaire de pouvoir moduler l'influence des contributions flottantes dans le choix de l'abandon.

Jugements de O Pour construire les jugements manifestés dans un énoncé d'autrui (OO et OS), nous suggérons une association entre la force illocutoire de l'acte de dialogue et la compréhension manifestée (voir tableau 8.7). Les actes de dialogues utilisés ici sont ceux de MMIL, eux-même étant basés sur DAMSL (Allen and Core 1997).

Force illocutoire	**Preuve manifestée**
Request	$\overline{U}$
Reject	$U\overline{R}$
Inform, Accept	UR
Opening, Closing	UR

TAB. 8.7 – Association de la force illocutoire et de la preuve manifestée

Cette association représente ce qui est minimalement nécessaire pour conduire un processus d'ancrage. Il s'agit d'une simplification importante. Tout d'abord la preuve manifestée ne se résume pas à la force illocutoire mais peut être également décrite par le contenu propositionnel de l'acte de dialogue. Par exemple tous les $\overline{U}$ ne sont pas manifestés au moyen d'un Request (par exemple « je ne comprends pas ») et tous les $U\overline{R}$ ne sont pas manifestés par des Reject comme dans le dialogue 8.41. On peut étendre également la critique à toutes les manifestations extra-linguistiques de l'incompréhension, par exemple les mimiques faciales peuvent indiquer la non-compréhension, ou un mouvement latéral de la tête peut indiquer un $U\overline{R}$ (Nakano et al. 2003).

Nous avons également considéré une hiérarchisation des preuves afin de simplifier le traitement des énoncés multi-contributifs. Certains énoncés peuvent en effet apporter *plusieurs* preuves de compréhension : « OK mais de quel hôtel vous parlez ? » manifeste un UR et un $\overline{U}$. Prendre en compte tous les effets est relativement

O_1 : je veux un billet pour Paris
S_2 : à Paris, quel type de chambre voulez-vous réserver ?
O_3 : je n'ai pas parlé de chambre !

Fig. 8.41 – Manifestation d'un $U\overline{R}$ à l'aide d'un Inform

difficile puisqu'il faut réussir à identifier précisemment les énoncés ou les aspects des énoncés sur lesquels portent les preuves. Nous avons préféré simplifier le traitement de ces énoncés en considérant qu'ils n'apportaient qu'une seule preuve de compréhension, la plus « forte » si on considère une hiérarchie de preuves : $Ab > U\overline{R} > \overline{U} > UR$. L'énoncé « OK mais de quel hôtel vous parlez ? » sera alors considéré comme manifestant un $\overline{U}$, en ignorant l'UR à cause de la force plus importante du $\overline{U}$. Ou encore « non non c'est pas ça, bon laisse tomber » sera considéré comme Ab malgré la présence de l'$U\overline{R}$.

8.4.2.3 Manifestation des jugements

La manifestation des jugements propres doit être potentiellement la plus explicite possible (sans dépasser le seuil d'ancrage attribué). Nous proposons de distinguer les cas de non-compréhension ($\overline{U}$ et $UR\overline{A}$), des cas de mauvaises compréhension ($U\overline{R}$) et ceux de bonne compréhension (UR et URA), chacun d'entre eux nécessitant des manifestations différentes.

Pour le $\overline{U}$ ou l'$UR\overline{A}$, il s'agit de manifester un problème de compréhension ou d'action. Etant donné que nous conservons dans la représentation du jugement SS les informations relatives au problème, il suffit de les verbaliser. L'énoncé suivant illustre tous les aspects verbalisables : « je ne comprends pas "cet hôtel". Ca ne réfère à rien. Quel hôtel ? ». On distingue :

- l'existence du problème par « je ne comprends pas »,
- la localisation du problème grâce à la métaréférence à l'expression problématique « "cet hôtel" »,
- le type de problème « Ca ne réfère à rien » (ici le vide),
- et la question de clarification « Quel hôtel ? ».

Nous n'avons implémenté que deux types de question de clarification : les questions ouvertes du type « Quel N ? » (ou « Quoi ? » lorsque le type est inconnu) et les questions fermées du type « Est-ce que le référent de "X" réfère à Y ? ».

Pour l'$U\overline{R}$ nous avons considéré le Reject « non » suivi d'une paraphrase (acte qu'on appelle Reject+Inform). Enfin pour l'UR ou l'URA, nous n'avons considéré que la paraphrase.

Pour générer les paraphrases nous avons utilisé le générateur de surface GenI (Kow 2006; Kow 2007; Gardent and Kow 2007). Il s'appuie sur trois ressources : une grammaire TAG, un lexique syntaxico-sémantique qui décrit l'association entre les termes sémantiques et les arbres de la grammaire et un lexique morphologique. Etant donné que notre grammaire initiale avait été conçue pour l'interprétation, celle-ci surgénérait. Nous nous sommes appuyés alors sur une grammaire « racine » à partir de laquelle nous dérivons deux grammaires, une pour l'interprétation et

une pour la génération. Nous avons ensuite traduit les ressources qui existaient déjà pour l'interprétation pour les adapter au format requis par GenI. Au final, GenI produit à partir de la forme sémantique MMIL, traduite en sémantique plate[11], une séquence de mots morphologisée correspondant à la verbalisation de la forme sémantique d'entrée.

Afin de produire la forme sémantique MMIL correspondant à un référent nous nous sommes appuyés sur la représentation interne en logique de description. Cette représentation permet de générer des descriptions relationnelles étant donné que nous nous appuyons uniquement sur des relations non réversibles[12]. Nous nous sommes restreints toutefois aux paraphrases non relationnelles lors de l'évaluation (voir p. 191).

8.4.2.4 Réinterprétation

La réinterprétation consiste à modifier une interprétation antérieure en fonction de nouvelles informations issues d'un énoncé de l'interlocuteur qui clôt un échange d'acceptation : soit que cet énoncé est une réponse UR/UR à un $\overline{U}$ initié par soi (*integrateClarificationAnswer*), soit que cet énoncé manifeste un $U\overline{R}$ (*otherReject*). Les deux cas sont très similaires : l'énoncé, appelé *source*, apporte une information sur l'interprétation qui doit être faite d'un énoncé antérieur de son partenaire, appelé *cible* (figure 8.42). La différence principale tient au fait que dans le cas de la question de clarification, l'interprétation a été jugée problématique *a priori*, alors que pour l'$U\overline{R}$, l'interprétation de l'énoncé cible est jugée problématique *en recevant l'énoncé source*. Dans les deux cas, la similarité se traduit par une réinterprétation qui consiste à altérer la forme d'interprétation. Cependant celle-ci ne se limite pas à modifier la forme d'interprétation. En effet puisqu'un énoncé peut avoir des effets sur le contexte (c'est le cas en particulier de toutes les sémantiques dynamiques dans lesquelles l'interprétation des énoncés est une fonction qui modifie le contexte), il est nécessaire de pouvoir altérer également l'espace référentiel.

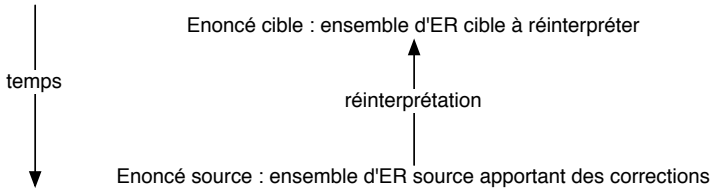

FIG. 8.42 – Schéma de réinterprétation

[11]La traduction est relativement aisée puisqu'il suffit de produire pour chaque entité MMIL un prédicat correspondant *participant*(?p0) ainsi qu'un ensemble de prédicats associés à chaque trait sémantique. Par exemple « une chambre double » sera représentée par l'ensemble {*participant*(?p0), *refType*(?p0 indefinite), *cardinality*(?p0 1), *media_ aType*(?p0 double)}.

[12]On évite ainsi des problèmes de récursion tels que « le livre sur la table sur laquelle il y a le livre... », voir (Areces et al. 2008).

Nous suggérons donc que l'altération du contexte est réalisée en deux temps : d'abord les effets d'une interprétation erronée sont annulés ensuite la nouvelle interprétation produit de nouveaux effets sur le contexte. Enfin étant donné que l'interprétation peut changer, il est nécessaire de ré-évaluer les jugements de compréhension. C'est pourquoi nous distinguons deux phases de la réinterprétation :

- altération de l'interprétation, et annulation des effets divergents
- réinterprétation, et ré-évaluation des jugements

Altération de l'interprétation Nous indiquons ici un moyen d'altérer la résultat de la référence mais ne développons pas de mécanisme générique d'altération. L'altération de la forme d'interprétation résolue en référence consiste à forcer l'association entre une expression référentielle et un référent. On distingue deux mécanismes différents correspondant aux deux situations de réinterprétation : le $\overline{U}$ et l'$U\overline{R}$. Dans les deux cas par hypothèse l'énoncé source apportant de nouvelles informations est crû compris et pertinent (UR, voir les préconditions des méthodes *integrateClarificationAnswer* et *otherReject*).

Altération issue d'un $\overline{U}$ On distingue le cas des questions ouvertes ou des questions fermées. Les questions ouvertes recherchent à *identifier* le référent d'une ER tandis que les questions fermées cherchent à *vérifier* le référent d'une ER[13].

Les questions ouvertes attendent deux types de réponse pertinente, soit une paraphrase du référent, soit « aucun ». Dans le premier cas, les ER source et cible sont identifiées. Il suffit d'altérer la forme d'interprétation et l'espace référentiel en *forçant* l'association entre l'ER source et le référent de l'ER cible. L'altération du contexte revient à modifier la représentation mentale de l'ER en modifiant la relation hasReferent. Si le système reçoit « aucun », l'ER qui a été construite ne réfère à rien, autrement dit, elle n'existe pas pour l'utilisateur. Il serait donc nécessaire de la supprimer à la fois de la forme d'interprétation et de l'espace référentiel. Nous avons cependant simplement considéré que l'ER ne référait à rien en supprimant la relation éventuelle avec son référent et en assertant un statut de résolution *solved*.

La réponse positive à une question fermée ne fait qu'établir la résolution de la référence en changeant sa probabilité de résolution à 1. La réponse négative provoque la suppression de la relation entre l'ER et le référent, à la fois dans la forme d'interprétation et dans l'espace référentiel. Cependant, cette réponse a un effet supplémentaire qui est d'*exclure* le référent pour toute résolution ultérieure de cette ER.

L'exemple de la figure 8.43 illustre le besoin d'interdire une association, le système ne peut réinterpréter « l'autre hôtel » sur l'hôtel Ibis en raison de la réponse négative de O. A cause de notre simplification dans la représentation des domaines de référence (voir p. 100), nous ne pouvons pas directement représenter cette exclusion. Nous avons ajouté à la place un nouvel attribut MMIL, *refExclude* dont

[13]Nous mettons de côté les questions de description. Les questions fermées de description cherchent à vérifier si une description s'applique bien à un référent et les questions ouvertes de description cherchent à établir une description du référent.

la valeur correspond à l'identifiant du référent exclus et dont la prise en compte intervient dans le module de résolution de la référence.

O_1 : je vais prendre l'autre hôtel
S_2 : est-ce que « l'autre hôtel » réfère à l'hôtel Ibis ?
O_3 : non pas du tout

FIG. 8.43 – Exemple d'exclusion

Altération issue d'un $U\overline{R}$ Plus complexe, l'$U\overline{R}$ peut être manifesté de deux manières différentes : par un simple Reject (typiquement « non non ») ou par un Reject+Inform (par exemple « non, l'hôtel Lafayette »). Dans les deux cas la difficulté est que l'expression cible n'est pas identifiée *a priori* et doit être inférée.

Le simple Reject ne fournit pas d'expression source pour corriger l'interprétation, et pas davantage de précision concernant l'expression cible (cas *Unprovided*). Etant donné que la cible de l'$U\overline{R}$ n'est pas indiquée, toutes les ER de l'énoncé cible peuvent être potentiellement mal comprises. Nous suggérons alors de traduire l'incertitude de la correction par la diminution de la probabilité de réussite de chaque référence. La ré-évaluation de la compréhension permet ensuite de pouvoir soulever un $\overline{U}$:Uncertain (SS) puis des questions à propos des expressions référentielles. Le dialogue 8.44 illustre la diminution de probabilité de réussite de l'expression « cet hôtel » ainsi que la question liée à l'incertitude.

O_1 : je prends cet hôtel du vingt au vingt-deux septembre
S_2 : ok, l'hôtel Ibis, du vingt au vingt-deux septembre
O_3 : non non
S_4 : est-ce que « cet hôtel » réfère à l'hôtel Ibis ?

FIG. 8.44 – Mauvaise compréhension sans expression source

Au contraire, le Reject+Inform a pour avantage de fournir une ou plusieurs expressions sources permettant de réviser l'interprétation de l'énoncé cible. A l'oral ces expressions sont en général marquées d'une intonation différentes (voir le dialogue 8.45). Le parallélisme syntaxique, l'intonation et le type sont des moyens d'identifier l'expression cible problématique.

O_1 : je veux une chambre à Paris
S_2 : à Nancy, je vous propose l'hôtel Ibis
O_3 : non, à PARIS

FIG. 8.45 – Intonation particulière lors de la correction

Dans (Funakoshi and Tokunaga 2006), la cible de la réinterprétation est choisie par son niveau d'ancrage. Les cibles de bon type sont ordonnées selon leur niveau d'ancrage et celle qui a le niveau le plus faible est préférée. Nous considérerons

toutefois un algorithme plus simple de recherche de l'expression cible : l'expression cible est l'ER de l'énoncé cible dont le référent possède un type compatible avec le référent de l'expression de l'énoncé source[14]. Le couplage doit également prendre en compte les pluriels : deux ER source peuvent correspondre à deux référents désignés par une unique ER cible plurielle comme dans la figure 8.46, et à l'inverse une ER source plurielle peut référer aux référents introduits par deux ER cible (figure 8.47).

O_1 : à combien étaient les premiers hôtels ?
S_2 : à l'hôtel Ibis, la chambre est à 50 euros,
: à l'hôtel Lafayette, la chambre est à 100 euros
O_3 : non, non je parle de l'hôtel Mercure et de l'hôtel Paradisius

Fig. 8.46 – Exemple d'une ER cible plurielle

O_1 : à combien étaient le premier et l'autre hôtel ?
S_2 : à l'hôtel Ibis, la chambre est à 50 euros,
: à l'hôtel Lafayette, la chambre est à 100 euros
O_3 : non, non je parle des deux autres

Fig. 8.47 – Exemple d'une ER source plurielle

Retrouver l'ER cible peut soulever des problèmes complexes. Nous distinguons trois cas problématiques :

- une ER source n'a pas d'ER cible correspondante, autrement dit, il doit exister un référent dans l'énoncé cible dont l'ER n'a pas été préalablement construite (*Missed*). Nous pensons que ce problème est en général résolu en effectuant une nouvelle analyse de l'énoncé qui conduise à la création de l'ER manquante. Etant donné la difficulté de considérer les contraintes qui peuvent guider la réinterprétation (la ré-analyse est guidée par l'ER source, mais dans quelle mesure ?), nous avons considéré que ce problème était résolu par une stratégie de robustesse interne en *ajoutant* l'ER source à l'interprétation de l'énoncé cible. Ce nouveau référent devrait être inséré en considérant une reconnexion des relations sémantiques, mais nous n'avons pas pris celles-ci en compte. Toutefois, la cause de ce problème peut également être une analyse défectueuse et non détectée de l'énoncé source qui conduit à générer une fausse ER source (jugement propre U au lieu de $\overline{U}$). Nous avons préféré éluder ce cas étant donné que s'il existe une divergence dans une ER de l'énoncé source, et que celle-ci est détectée et manifestée par l'utilisateur, la réinterprétation de l'énoncé source entraînera sa correction et par *feedback chaining* la réinterprétation de l'énoncé cible initial.

[14]Pour les pronoms ou les ellipses de tête nominale, le type compatible doit être défini par les restrictions de sélection mais nous avons simplifié ce traitement en supposant que les pronoms pouvaient être réinterprétés avec n'importe quel type de référent.

- aucun couplage ER source/ER cible n'entraîne d'altération, autrement dit la réinterprétation n'a aucun effet (*Unchanged*). Nous avons estimé que ce problème était causé par des divergences similaires dans l'interprétation de l'énoncé source et de l'énoncé cible. En conséquence on peut traduire ce problème de réinterprétation comme un problème d'interprétation de l'énoncé source. Nous soulevons un problème de type $UR\overline{A}$ indiquant l'entité problématique. Sa manifestation entraîne une *explicitation* de l'interprétation avec un énoncé du type « Pour moi, "X" réfère à Y » que l'on assimile à une question fermée indirecte. La réponse positive à cette question entraîne l'abandon de la réinterprétation : l'utilisateur accepte que "X" réfère à Y et que la réinterprétation n'ait pas lieu d'être (voir l'exemple section 9.6). La réponse négative a les mêmes effets qu'une réponse négative à la question fermée correspondante.
- une ER source correspond à *plusieurs* ER cible, autrement dit, la réinterprétation est ambiguë. De façon similaire que pour le *Unchanged* on peut traduire cette ambiguïté au niveau d'un problème d'interprétation de l'énoncé manifestant l'$U\overline{R}$. Nous n'avons pas modélisé explicitement ce cas mais on pourrait de la même façon soulever une question $UR\overline{A}$ manifestant cette ambiguïté, entraînant une réinterprétation de l'$U\overline{R}$ provoquant la réinterprétation de l'énoncé initial à l'instar du dialogue 8.48 inspiré de (Funakoshi and Tokunaga 2006).

U : Put the red ball on the green table
S puts the red ball on the green table
U : Sorry, I meant blue
S : the ball or the table ?

FIG. 8.48 – Exemple d'ambiguïté de réinterprétation

Altération du contexte Dans les cas où l'altération de l'interprétation réussit on doit également altérer le contexte. C'est un problème difficile et nous n'avons pas trouvé de travaux qui détaille l'altération contextuelle sur plusieurs énoncés. Nous proposons un algorithme naïf pour altérer le contexte : tous les référents en relation avec le référent de l'expression problématique qui ont été introduits *après* dans les énoncés de l'utilisateur sont effacés. Le dialogue 8.49 (p. 170) illustre ce besoin. Le système associe « cet hôtel » à l'hôtel Ibis, et alors construit « une chambre double » en relation avec l'hôtel Ibis. Cependant, O n'a jamais référé à une chambre à l'hôtel Ibis : cette entité n'existe que pour S. Pour éviter de pouvoir y référer, il est nécessaire alors d'effacer ce référent et de construire à la place une nouvelle « chambre double » dans l'hôtel Lafayette.

Cependant l'altération du contexte ne consiste pas toujours à effacer le référent comme le montre le dialogue 8.50. En effet, l'$U\overline{R}$ n'intervient qu'en O_5 et manifeste que l'interprétation de « cet hôtel » en O_1 est divergente. Dans ce cas il ne faut pas annuler la création de l'instance de chambre puisqu'elle a été mentionnée en S_2 et qu'elle est présente dans le terrain commun conversationnel. Au contraire il faut simplement annuler l'*association* entre l'expression « la chambre » en O_1 et la

O_1 : je voudrais une chambre double dans cet hôtel
S_2 : je suis désolé, il n'y a plus de chambres disponibles à l'hôtel Ibis
O_3 : non, je voulais dire l'hôtel Lafayette
S_4 : à l'hôtel Lafayette, il y a des chambres disponibles

Fig. 8.49 – Nécessité d'altérer le contexte

mauvaise instance. La réinterprétation est encore plus difficile si on se place du point de vue de O : celui-ci va parcourir à nouveau les énoncés de S_2 à S_4 en considérant que S parlait de l'hôtel Ibis. Il va annuler à son tour l'association entre « cet hôtel » et l'hôtel Lafayette en S_2. La réinterprétation conduira O à acquérir de nouvelles informations : étant donné qu'il peut savoir qu'en S_2 « cet hôtel » réfère à l'hôtel Ibis au lieu de l'hôtel Lafayette, il peut savoir qu'à l'hôtel Ibis la chambre coûte 20 euros. Il peut ensuite référer à cette chambre en O_7. Nous ne sommes toutefois pas descendus à ce niveau de complexité dans l'altération contextuelle.

O_1 : à combien est la chambre dans cet hôtel ?
S_2 : dans cet hôtel, la chambre est à 20 euros
O_3 : ok, et il y a la douche ?
S_4 : à l'hôtel Ibis, il y a la douche
O_5 : non, je voulais dire l'hôtel Lafayette
S_6 : à l'hôtel Lafayette, la chambre est à 100 euros et n'a pas de douche
O_7 : ah ben non, je prends l'autre chambre alors

Fig. 8.50 – Complexité de l'altération contextuelle

8.4.3 Résolution des questions de l'utilisateur

Les questions de clarification ne sont pas résolues de manière complexe étant donné qu'on se limite à l'identification des référents au moyen de questions du type « Quel N ? » ou « Quoi ? ». Nous ajoutons simplement un solveur référentiel avec deux types de domaines sous-spécifiés dont les contraintes sont TypeSubsumer et ViewPointFocal pour l'un et TypeSubsumer et Discriminating pour l'autre (voir section 6.1.3.2). Etant donné que nous basons la résolution de la question sur le module de traitement de la référence, on peut appuyer le jugement URA ou $UR\overline{A}$ de la question sur le jugement de la référence. Ainsi, la question peut conduire à une unicité de réponse, une ambiguïté, un vide ou une incertitude. L'unicité et le vide ne sont pas considérés comme des problèmes, le premier étant manifesté par la description du référent et le second par « aucun(s) ». En revanche, l'incertitude ou l'ambiguïté conduisent à un jugement $UR\overline{A}$, la requête ne pouvant pas aboutir, et sont manifestés de la même manière que pour la résolution de la référence en explicitant le problème.

8.5 Conclusions

Nous avons présenté un modèle d'ancrage qui motive le besoin de fournir des preuves positives de compréhension comme moyen de satisfaire les exigences de compréhension d'autrui, à l'instar du modèle original de Clark. Notre modèle ne pose toutefois pas le problème de l'acceptation récursive en précisant que les exigences de compréhension d'autrui peuvent être estimées dès la réception de ses énoncés ultérieurs. La récursion peut être terminée de cette façon mais en contrepartie le critère d'ancrage risque d'être atteint de manière erronée. Cette situation se produit lorsque les preuves de compréhension sont mal comprises. Mais plutôt que de renforcer le critère d'ancrage, par exemple à l'aide de croyances mutuelles, nous avons suggéré que les participants cherchaient à résoudre les divergences d'interprétation des preuves de compréhension et en cela rétablir un critère d'ancrage partagé. Nous n'avons toutefois que décrit la non-compréhension de ces preuves et pas leur mauvaise compréhension.

Notre modèle généralise la notion de preuve en tant que jugement de compréhension, que l'on peut former soi-même ou que l'on peut manifester dans le dialogue. Ces jugements de compréhension permettent d'une part de résoudre les différents problèmes posés par le modèle des échanges, en particulier l'absence de prise en compte de l'interprétation. D'autre part, ils permettent une représentation unifiée de la non-compréhension ou de la mauvaise compréhension en tant que jugement négatif de sa propre interprétation ou jugement négatif de l'interprétation d'autrui. Nous n'avons toutefois considéré qu'un petit nombre d'incompréhensions liées à la résolution de la référence. Afin d'évaluer la pertinence du modèle nous proposons une évaluation du processus d'ancrage en termes de gain de compréhension dans le chapitre 10.

9 Exemples de traitements

Nous illustrons ici le processus d'ancrage sur un dialogue artificiel qui démontre le comportement du système face à différents problèmes de compréhension (figure 9.1). Nous proposons un traitement incrémental du dialogue, du point de vue de S, en isolant chaque situation problématique. Pour produire ce dialogue, nous ne supposons que l'existence préalable d'un hôtel, l'hôtel Ibis, dans l'espace référentiel. Tous les problèmes issus de l'utilisateur, une non-compréhension et une fausse mauvaise compréhension, seront simulés (on peut supposer par exemple qu'ils sont causés par des bruits).

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : hein ? quel hôtel ?
S_4 : l'hôtel Ibis
U_5 : ah non non l'autre
S_6 : "l'autre" ne réfère à rien. Quel hôtel ?
U_7 : le Lafayette
S_8 : OK OK donc le Lafayette (+ bruit)
U_9 : mais non le Lafayette !
S_{10} : je ne comprends pas. Pour moi "le Lafayette" réfère au Lafayette
U_{11} : ah oui pardon j'avais mal entendu
S_{12} : OK donc le Lafayette

FIG. 9.1 – Exemple de dialogue complet

9.1 Bonne compréhension

U_1 : **je prends cet hôtel**
S_2 : **OK l'hôtel Ibis (+ bruit)**

FIG. 9.2 – Exemple de bonne compréhension

Dans l'extrait de la figure 9.2, S résout « cet hôtel » de manière unique et probable comme référence à l'hôtel Ibis. Etant donné que U_1 ne manifeste aucun problème de compréhension, l'énoncé U_1 est alors UR/UR. Comme U_1 est au niveau

dialogue, S déclenche alors la méthode *integrateDialogueRequest*. Cette méthode conduit à l'initiation d'un échange niveau dialogue, conformément à la figure 9.3. La méthode *testAcceptance* est ensuite exécutée et son résultat URA entraîne la clôture de l'échange par S_2 (figure 9.4). S_2 manifeste la bonne compréhension de U_1 par la production d'un « OK » par défaut et la génération d'une paraphrase.

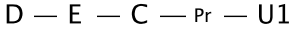

Fig. 9.3 – Dialogue après intégration de U_1

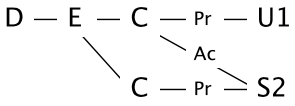

Fig. 9.4 – Dialogue après intégration de S_2

9.2 Non-compréhension de la part de U

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : **hein ? quel hôtel ?**
S_4 : **l'hôtel Ibis**

Fig. 9.5 – Exemple de non-compréhension de U

Dans l'extrait de la figure 9.5, U manifeste sa non-compréhension au moyen d'une question conventionnelle « hein ? » et d'une question ouverte « quel hôtel ? » mais nous ne considérons que la seconde question. Etant donné que S croit comprendre la question et la juge pertinente[1], le jugement de U_3 est $UR/\overline{U}$. Le système déclenche alors la méthode *otherClarificationRequest* qui entraîne l'insertion de U_3 dans la phase d'acceptation de l'énoncé précédent S_2 (figure 9.6).

Nous n'avons pas représenté l'ambiguïté de portée pour les questions étant donné que nous utilisons un solveur référentiel particulier pour la résolution qui ne s'appuie donc pas sur la structure du dialogue mais sur l'espace référentiel. La question ReferenceOpenRequest est résolue grâce à ce solveur, et conduit à un résultat URA

[1]Nous n'avons pas considéré les questions de clarification non-pertinentes dans l'implémentation. Ce traitement est plutôt difficile puisqu'il nécessite de considérer les présuppositions erronées effectuées par le locuteur lorsqu'il pose sa question de clarification.

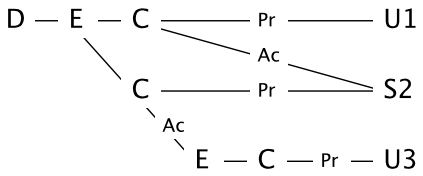

FIG. 9.6 – Dialogue après intégration de U_3

de la méthode *testAcceptance* : la requête U_3 est non seulement jugée comprise et pertinente mais également résolue avec succès. L'énoncé suivant est alors inséré comme seconde contribution de l'échange de clarification (figure 9.7).

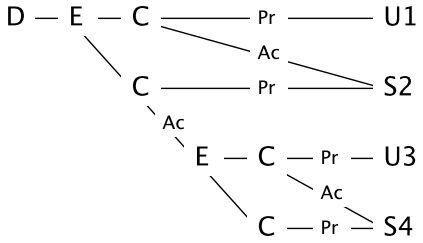

FIG. 9.7 – Dialogue après intégration de S_4

9.3 Non-compréhension de S de sa mauvaise compréhension

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : hein ? quel hôtel ?
S_4 : l'hôtel Ibis
U_5 : **ah non non l'autre**
S_6 : **"l'autre" ne réfère à rien. Quel hôtel ?**

FIG. 9.8 – Exemple de mauvaise compréhension de S

Dans l'extrait de la figure 9.8, U manifeste la mauvaise compréhension de l'énoncé U_1 en U_5 : étant donné que U n'avait pas compris la preuve donnée en S_2, il ne pouvait pas détecter cette mauvaise compréhension avant la réponse à sa question U_3. Toutefois, cette preuve de compréhension est problématique pour S puisqu'il

ne parvient pas à résoudre « l'autre ». Son jugement propre de U_5 est $\overline{U}$:Empty. En conséquence, étant donné que nous avons fait l'hypothèse de prépondérance des jugements propres, il lui est nécessaire d'abord de résoudre la non-compréhension de U_5 avant de prendre en compte ses effets. Son jugement déclenche la méthode *selfRequestClarification* qui insère l'énoncé U_5 en tant que contribution flottante et S_6 comme initiant un échange incident à cette contribution (figure 9.9). L'énoncé S_6 manifeste deux preuves de $\overline{U}$: une explicitation du problème avec couverture lexicale « "l'autre" ne réfère à rien » et une question ouverte de type ReferenceOpenRequest « Quel hôtel ? ».

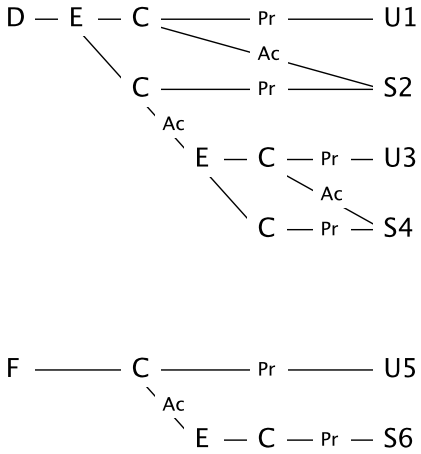

Fig. 9.9 – Dialogue après intégration de U_5 et de S_6

9.4 Mauvaise compréhension de S et succès de réinterprétation

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : hein ? quel hôtel ?
S_4 : l'hôtel Ibis
U_5 : ah non non l'autre
S_6 : "l'autre" ne réfère à rien. Quel hôtel ?
U_7 : **le Lafayette**
S_8 : **OK OK donc le Lafayette (+ bruit)**

Fig. 9.10 – Exemple d'intégration de la réponse avec *feedback chaining*

Dans l'extrait de la figure 9.10, U fournit une réponse à la question soulevée en S_6. Cette réponse est jugée UR/UR par S, et son intégration entraîne de nom-

breux effets. Tout d'abord, S déclenche la méthode *integrateClarificationAnswer* étant donné qu'elle vient clore un échange incident. Celle-ci provoque une tentative de réinterprétation de U_5 : l'association entre « l'autre » et l'hôtel Lafayette est forcée. Son succès entraîne un jugement URA de U_7 et l'intégration de U_7 comme seconde contribution de l'échange initié par S_6 (figure 9.11). Grâce à ce succès, S peut désormais prendre en compte les effets de U_5.

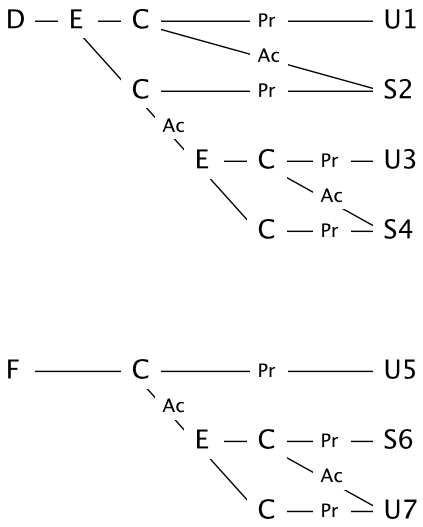

FIG. 9.11 – Dialogue après intégration de U_7

La compréhension de U_5 est donc ré-évaluée et son jugement est alors $UR/U\overline{R}$. En conséquence S déclenche la méthode *otherReject*. Cependant, S est confronté à l'ambiguïté de portée de l'$U\overline{R}$: il peut s'agir d'une mauvaise compréhension de la question en U_3 ou de l'énoncé U_1. S teste tout d'abord la réinterprétation de U_3, mais celle-ci échoue étant donné que U_3 ne fournit pas d'expression référentielle source susceptible d'être le support de la réinterprétation. L'énoncé U_1 au contraire autorise cette réinterprétation et l'association entre « cet hôtel » et l'hôtel Ibis peut être remise en question. Dès lors, S effectue trois opérations :

- d'abord il déplace la contribution présentée par S_2 dans un nouvel échange d'acceptation de U_1 puisque S_2 ne doit plus être considérée pertinente (figure 9.12),
- puis il insère U_5 comme seconde contribution de cet échange (figure 9.13),
- enfin il effectue la réinterprétation proprement dite de U_1. Celle-ci fonctionne et force l'association entre « cet hôtel » et le référent de « l'autre », c'est-à-dire l'hôtel Lafayette.

Enfin, il peut produire son énoncé suivant. Celui-ci est composé de trois preuves : une preuve que U_7 a été jugé UR sous la forme d'un « OK », une preuve que U_5 a été jugé UR (également un « OK »), et une preuve de la compréhension UR de U_1 sous la forme d'une paraphrase, le tout coordonné par un « donc » traduisant la réussite

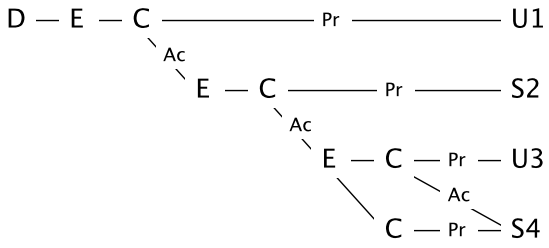

Fig. 9.12 – Dialogue après restructuration causée par U_5

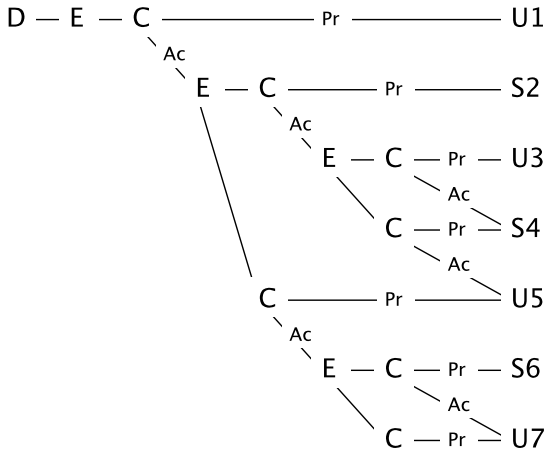

Fig. 9.13 – Dialogue après réintégration de U_5

de la réinterprétation. L'énoncé S_8 est alors inséré comme seconde contribution de l'échange initial (figure 9.14).

9.5 Mauvaise compréhension de U et échec de réinterprétation

Cependant un nouveau bruit empêche U de bien comprendre S_8 (figure 9.15) et il croit à tort à une autre mauvaise compréhension de S. C'est le seul cas de mauvaise compréhension de la preuve que nous prenons en compte. Celle-ci entraîne un échec de réinterprétation. En effet, S juge que l'énoncé U_9 est $UR/U\overline{R}$, il déclenche alors la méthode *otherReject*. La réinterprétation de U_1 semble possible et alors S déplace S_8 dans un nouvel échange incident à U_1 (figure 9.16 p. 180), insère U_9 comme seconde contribution de cet échange (figure 9.17 p. 180), et procède à la réinterprétation de U_1.

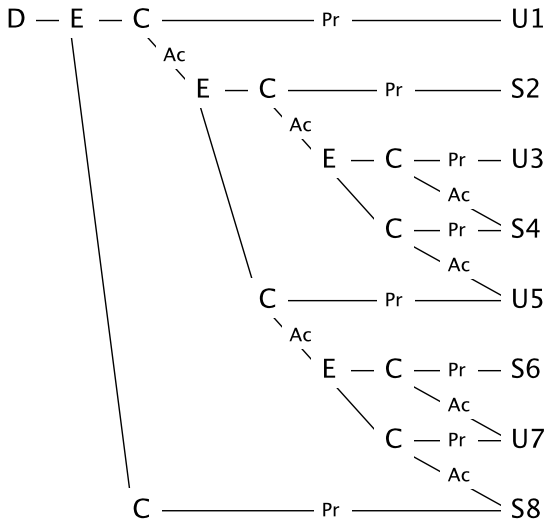

FIG. 9.14 – Dialogue après réintégration de S_8

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : hein ? quel hôtel ?
S_4 : l'hôtel Ibis
U_5 : ah non non l'autre
S_6 : "l'autre" ne réfère à rien. Quel hôtel ?
U_7 : le Lafayette
S_8 : OK OK donc le Lafayette (+ bruit)
U_9 : **mais non le Lafayette !**
S_{10} : **je ne comprends pas. Pour moi "le Lafayette" réfère au Lafayette**

FIG. 9.15 – Echec de réinterprétation

Mais celle-ci échoue : impossible de réinterpréter « cet hôtel » puisque le référent est identique pour l'expression cible et source. Cet échec entraîne un jugement $UR\overline{A}$ de U_9 avec pour type de problème *Unchanged*. S produit donc une requête de clarification de U_9 du type UnchangedReferentClosedRequest en s'attendant soit à une confirmation soit une infirmation (figure 9.18). L'énoncé S_{10} est alors inséré dans un échange d'acceptation de U_9 sous la forme d'une non-compréhension explicite « je ne comprends pas » et d'une explicitation du problème « Pour moi "le Lafayette" réfère au Lafayette », qui ressemble à une tautologie sans en être une.

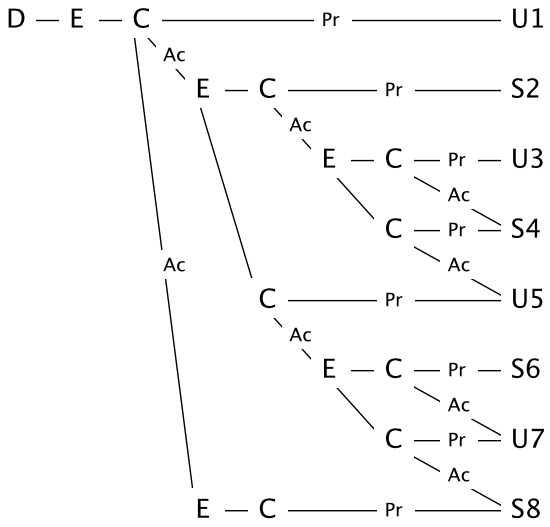

Fig. 9.16 – Dialogue après restructuration causée par U_9

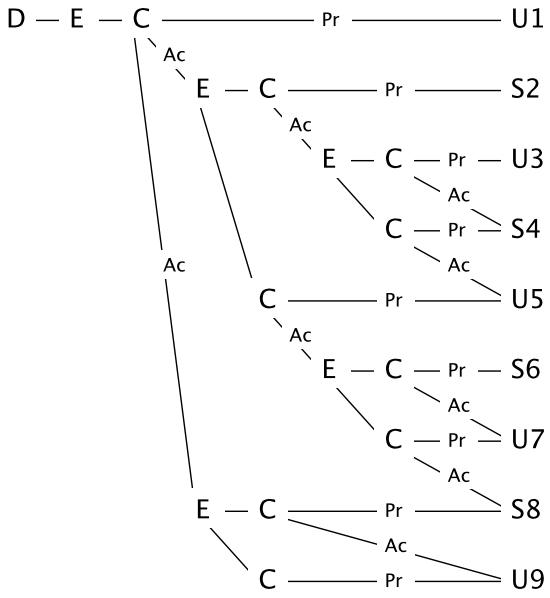

Fig. 9.17 – Dialogue après intégration de U_9

9.6 Abandon d'une mauvaise compréhension

L'énoncé S_{10} permet à U de reconnaître son erreur. Il produit alors U_{11} qui est jugé UR/UR par S (figure 9.19). Nous n'avons pas pris en compte la manifestation du problème « j'avais mal entendu » ou encore la forme de politesse « pardon », mais

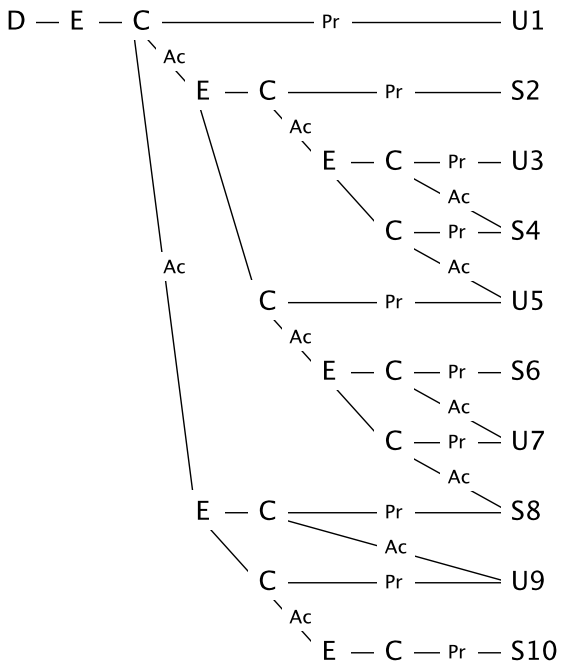

FIG. 9.18 – Dialogue après intégration de S_{10}

U_1 : je prends cet hôtel
S_2 : OK l'hôtel Ibis (+ bruit)
U_3 : hein ? quel hôtel ?
S_4 : l'hôtel Ibis
U_5 : ah non non l'autre
S_6 : "l'autre" ne réfère à rien. Quel hôtel ?
U_7 : le Lafayette
S_8 : OK OK donc le Lafayette (+ bruit)
U_9 : mais non le Lafayette !
S_{10} : je ne comprends pas. Pour moi "le Lafayette" réfère au Lafayette
U_{11} : **ah oui pardon j'avais mal entendu**
S_{12} : **OK donc le Lafayette**

FIG. 9.19 – Exemple de dialogue complet

seulement l'acceptation « oui ». S déclenche alors la méthode *integrateClarificationAnswer*, ce qui d'une part insère U_{11} comme réponse à S_{10} (figure 9.20) et d'autre part conduit à la résolution de la question UnchangedReferentClosedRequest. Cette résolution est particulière puisqu'elle doit *annuler* la preuve de mauvaise compréhension donnée en U_9.

Nous avons alors considéré que la réponse positive à un UnchangedReferentClosedRequest avait pour effet d'abandonner le but présenté par l'énoncé (ici U_9). Cet

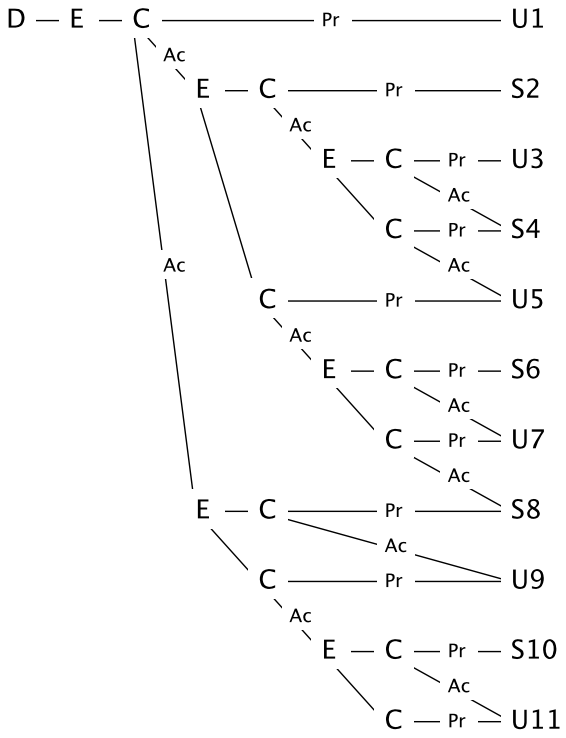

Fig. 9.20 – Dialogue après intégration de U_{11}

abandon est toutefois délicat à gérer puisque les jugements de compréhension que nous construisons traduisent soit des jugements propres, soit des jugements manifestés. L'abandon, par exemple, doit être explicitement manifesté. Ici, le jugement n'est pas porté par un énoncé en particulier mais par l'échange incident $\langle S_{10}, U_{11}\rangle$. Les effets de cet échange sont qu'au final, U reconnaît qu'il avait mal compris S_8 et qu'également S avait bien compris U_1.

D'abord, il est nécessaire de considérer que U_9 n'apporte plus de mauvaise compréhension et nous avons traduit cet abandon par la révision de la preuve de compréhension apportée par U_9, initialement $U\overline{R}$ en UR. Ce n'est pourtant pas suffisant. En effet, étant donné que U_9 est désormais UR/UR et qu'il clôt un échange d'acceptation, il déclenche par erreur la méthode *integrateClarificationAnswer*. Afin de modéliser complètement l'abandon, il est nécessaire également de marquer l'échange $\langle S_8, U_9\rangle$ comme abandonné. L'abandon nous permet d'éviter de reconsidérer cet échange et par défaut, la méthode *testAcceptance* renverra un jugement URA.

Au final, l'énoncé suivant, S_{12} peut être inséré comme seconde contribution de l'échange initié par U_1 (figure 9.21). En l'occurrence, il manifeste un « OK » comme preuve que U_{11} a été considéré UR, et une paraphrase du référent de « cet hôtel », coordonnées par un « donc ».

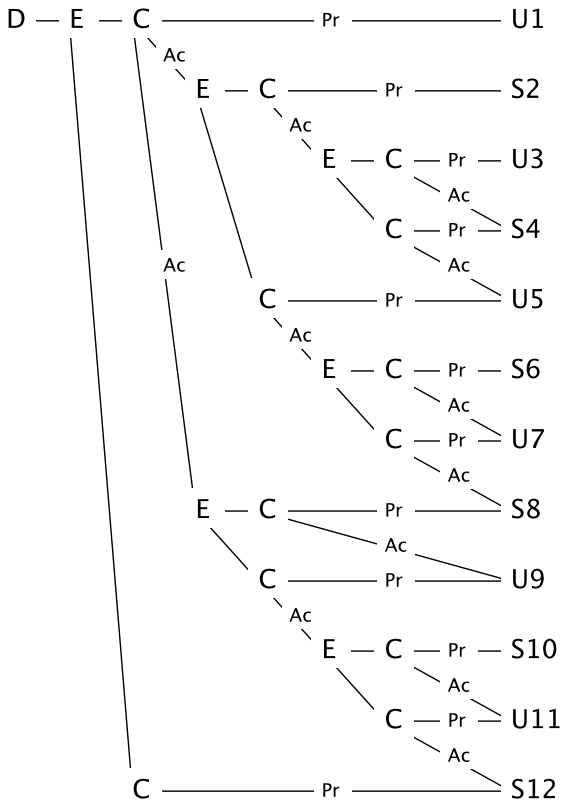

FIG. 9.21 – Dialogue après intégration de S_{12}

10 Evaluation du processus d'ancrage sur corpus

Nous avons défini la robustesse interne comme la capacité d'un système à résoudre ses problèmes par lui-même et la robustesse externe comme sa capacité à les résoudre dans le dialogue. Cette définition est toutefois problématique. Par exemple, la faculté de détecter la mauvaise compréhension fait-elle partie de la robustesse interne (en tant que capacité propre d'interprétation) ou de la robustesse externe (en tant que détection d'un problème soulevé dans le dialogue) ? Evaluer la seule robustesse externe en faisant abstraction de la robustesse interne est relativement difficile. Contrairement à la robustesse interne, il est nécessaire en effet de pouvoir évaluer la robustesse externe *dans l'interaction* et de fournir un partenaire au système que l'on cherche à évaluer. Cependant, cette hypothèse biaise l'évaluation puisque la convergence dépend de l'attitude des deux partenaires. Si le partenaire n'est pas suffisamment robuste, son influence sur l'échec du processus d'ancrage risque d'introduire un biais important dans la mesure de la robustesse externe, et on ne peut pas pénaliser le système à évaluer pour les erreurs de son partenaire. Nous proposons, plutôt que de chercher à définir un partenaire qui soit parfaitement robuste, de chercher à découpler les capacités nécessaires à l'ancrage dans deux systèmes et de n'évaluer non pas la seule capacité de robustesse externe (dans la faculté de manifester les problèmes) mais d'évaluer le processus d'ancrage lui-même, c'est-à-dire à la fois la faculté d'établir les jugements de compréhension et de les manifester afin d'atteindre l'ancrage d'un énoncé.

Ce paradigme peut être rapproché du paradigme SIMDIAL d'Allemandou (2007) qui permet d'évaluer les systèmes de dialogue par la simulation d'un utilisateur. L'idée est de simuler un utilisateur déterministe et d'évaluer la capacité du système à fournir le service demandé malgré certaines perturbations (ambiguïtés, hésitations, etc.). Dans notre contexte, la tâche n'est cependant pas extérieure au dialogue mais vise au contraire à ancrer un énoncé précédent. De plus, contrairement à SIMDIAL, l'analogie entre le dialogue machine-machine et le dialogue homme-machine sera limitée. Chaque système possède en effet différentes capacités disjointes liées à l'ancrage mais dans le dialogue homme-machine, toutes ces capacités doivent être réunies dans un unique système.

Deux aspects sont à considérer. D'une part, il est nécessaire tout de même d'intégrer la dimension de la langue naturelle telle qu'elle est produite par des locuteurs humains. Dans le cas contraire, il n'y aurait qu'une portée limitée à faire converser

deux machines dans un langage qu'ils maîtrisent parfaitement. D'autre part il est nécessaire d'introduire des divergences d'interprétation dans les énoncés afin d'observer si le processus d'ancrage permet d'établir la convergence. C'est pourquoi nous proposons de nous appuyer sur le corpus MEDIA qui fournit en effet des énoncés en langue naturelle dont on connaît les résultats d'interprétation grâce à la campagne d'évaluation MEDIA. Grâce à l'annotation correcte de ces énoncés, on peut fournir à l'un des systèmes la bonne interprétation et évaluer si le couple permet d'atteindre cette interprétation par le dialogue.

On peut faire l'analogie suivante : deux amis appelés *Client* et *Compère* discutent ensemble autour d'un verre. Sur la table voisine un individu assis à proximité qu'on appelera Contrôle écoute cette conversation et la rapporte à une quatrième personne assise plus loin qu'on appelera Test. L'évaluation MEDIA correspond à une situation où Test est seul assis à sa table, et tente de comprendre seul la conversation malgré la distance (robustesse interne). Nous proposons d'évaluer le processus d'ancrage en vérifiant si la compréhension du dialogue entre *Client* et *Compère* par Test est améliorée grâce au dialogue qu'il peut entretenir avec Contrôle.

Nous appelons ce protocole, le protocole miroir, en raison du fait que Contrôle et Test sont deux instances issues du même système. Cependant plusieurs caractéristiques importantes les différencient. D'abord, ils ne possèdent pas la même interprétation du dialogue entre *Client* et *Compère*. Le Contrôle a en effet un accès direct à la bonne interprétation issue de l'annotation, tandis que le Test doit s'appuyer sur le module d'interprétation symbolique présenté au chapitre 6. Ensuite nous ne doterons pas le Contrôle et le Test des mêmes capacités d'ancrage. Par exemple, seul le Contrôle sera à même d'établir des jugements de mauvaise compréhension du Test et seul le Test sera habilité à manifester sa non-compréhension. La décision de l'abandon sera uniquement effectuée par le Contrôle. Conformément au principe d'ancrage celui-ci détermine si oui ou non la bonne compréhension est atteinte ou peut être atteinte. Enfin, si les deux systèmes ont pour but de faire converger l'interprétation du Test, leur rôle sera différent : le Contrôle propose un énoncé à interpréter et vérifie l'interprétation du Test tandis que le Test doit prouver qu'il a effectivement compris l'énoncé initial.

Cette évaluation permet d'une part de vérifier la présence d'un gain de compréhension. Est-ce que, grâce au dialogue avec Contrôle, le Test peut parvenir à améliorer sa compréhension ? Ensuite elle permet de dégager des situations suffisamment riches pour pouvoir évaluer la présence de problèmes d'ancrage : quelles sont les difficultés que peuvent encourir le Contrôle et le Test à ancrer un énoncé ? On peut supposer en effet que les problèmes d'ancrage qu'ils vont se poser, pourront se retrouver dans le dialogue homme-machine. Nous présentons le protocole d'évaluation ainsi que les résultats et les problèmes d'ancrage dans ce chapitre et discutons de l'évaluation dans le chapitre suivant.

10.1 Protocole d'évaluation

Le protocole d'évaluation pour un dialogue donné est le suivant :

1. le CONTRÔLE sélectionne l'énoncé suivant d'un dialogue annoté du corpus MEDIA (e_1)
2. il construit son interprétation *sur la base de l'annotation* en mettant à jour l'espace référentiel
3. il produit l'énoncé à l'adresse du TEST sous la forme « Interprète "e_1" »
4. le TEST interprète l'énoncé du CONTRÔLE et produit une manifestation de sa compréhension au moyen de la stratégie T (e_2)
5. le CONTRÔLE interprète la réponse et s'il juge l'énoncé ancré, il se rend à l'étape 1
6. sinon, s'il estime que l'ancrage ne peut pas être atteint, il manifeste un abandon et retourne à l'étape 1
7. sinon, la procédure d'ancrage continue, il produit un énoncé destiné à ancrer l'énoncé problématique au moyen de la stratégie C, et se rend à l'étape 4

Ce paradigme d'évaluation autorise plusieurs conditions de test données par les stratégies d'ancrage T et C. Nous n'avons effectué toutefois qu'une seule évaluation, où la stratégie T consiste à manifester sa compréhension de manière explicite au moyen d'une paraphrase ou de questions d'identification résolvant l'ambiguïté et la stratégie C à effectuer des $U\overline{R}$. Nous illustrons le protocole miroir dans la figure 10.1 (p. 188). Nous rappelons également le protocole d'évaluation en contexte dans la figure 10.2 (p. 189).

10.1.1 Mesures

Nous évaluons avant tout la capacité à obtenir la bonne interprétation, indépendemment de la réussite ou de l'échec du processus d'ancrage en termes d'abandon. Les mesures d'évaluation sont en conséquence exactement les mêmes que celles utilisées lors de l'évaluation en contexte (section 7.1.2 p. 112) et nous mesurons ici le *gain* en IREF et DREF des compétences interprétatives avec et sans ancrage. Le score *sans* ancrage correspond au score de l'évaluation MEDIA, que nous recalculons ici sur les dialogues du corpus de test afin de prendre en compte les modifications éventuelles par rapport au système de la campagne. Le score *avec* ancrage correspond à la même mesure mais évaluée sur un dialogue dont les énoncés ont fait l'objet de réinterprétations issues de l'ancrage.

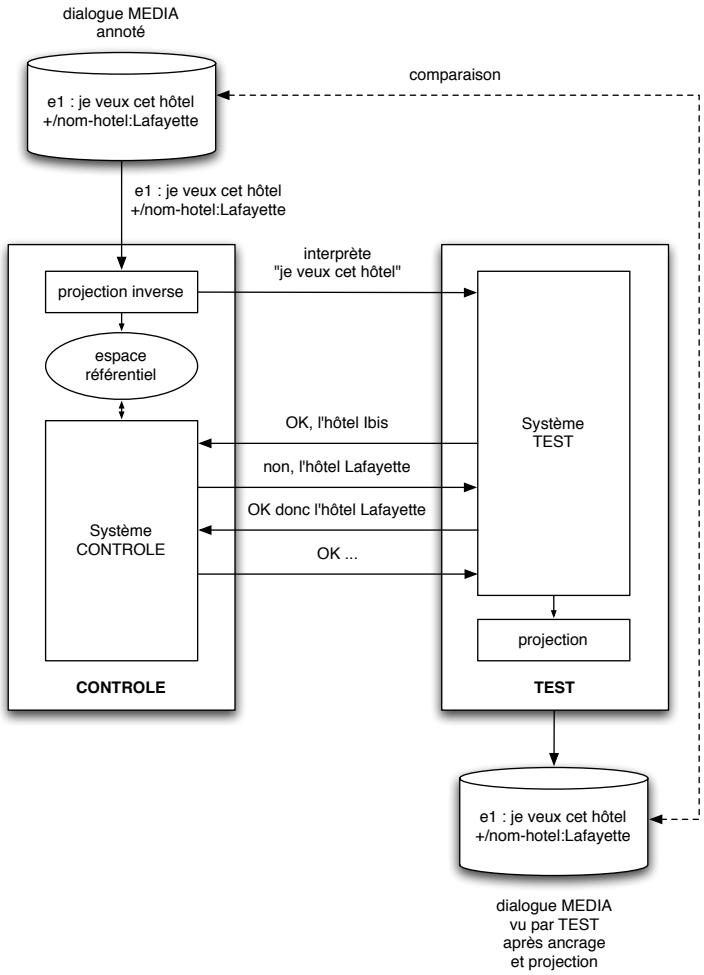

Fig. 10.1 – Illustration du protocole miroir

10.2 Mise en œuvre de l'évaluation

Plusieurs modifications ont été effectuées sur le système afin de rendre possible l'évaluation.

10.2.1 Construction d'une forme sémantique interne de référence

La plus importante de ces modifications concerne la réutilisation du corpus pour constituer la bonne interprétation du Contrôle (étape 2 du protocole). Il s'agissait de construire un composant MMIL correct à partir de l'annotation, autrement dit d'effectuer l'inverse de la projection présentée à la section 6.2. Nous avons constitué des règles de projection inverse, qui, à partir d'une séquence de triplets, construisent des composants. Les patrons reconnus les plus longs conduisent à la formation d'un composant, et de ses participants. Deux des principales difficultés ont été les relations

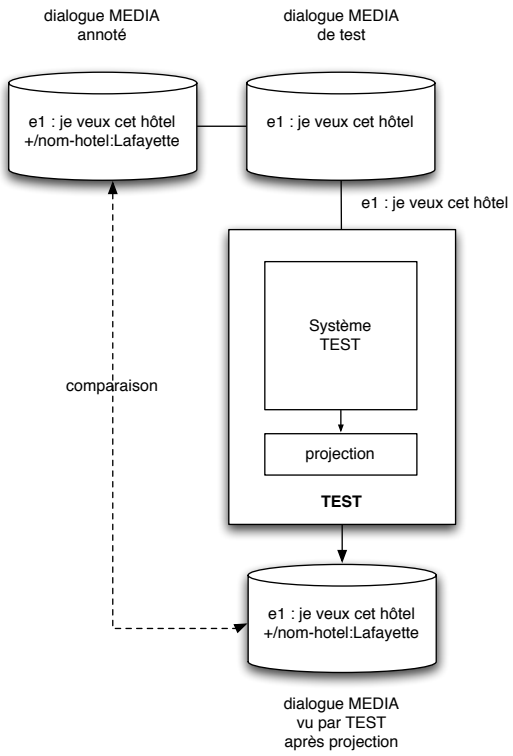

FIG. 10.2 – Rappel du protocole d'évaluation en contexte

et les types de référence. Nous n'avons que très partiellement considéré les relations, potentiellement reconstructibles à l'aide des spécifieurs, en nous focalisant sur les relations locales à une séquence de triplets. Par exemple la séquence +/objetBD:hotel, +/localisation-ville-hotel:paris correspond très probablement à « un hôtel à Paris », et une relation *aLocalisation* peut être construite entre les deux participants. Cet aspect n'a été que superficiellement pris en compte.

Les types de référence (défini, indéfini, démonstratif, etc.) étaient plus problématiques car ils n'ont pas été annotés explicitement mais nous sont pourtant nécessaires. Les noms propres ne soulevaient pas de problème puisque l'annotation permet de les identifier (par exemple +/nom-hotel:lafayette). Lorsqu'aucune indication du type d'objet n'était précisée (par exemple en cas de lienRef non suivi du trait objet), nous avons adopté le type pronominal par défaut. Lorsqu'il y avait des indications de nombre, nous avons supposé qu'il s'agissait d'indéfinis, sauf dans certains cas comme « ces deux hôtels » où le marqueur de lienRef nous permet de savoir qu'il ne s'agit pas d'un indéfini. La figure 10.3 illustre une de ces règles, sélectionnée en présence d'un trait +/lienRef-coRef:pluriel suivi d'un trait d'attribut nombre-hotel. Les deux participants construits à l'aide de cette règle sont fusionnés afin de ne former qu'un seul participant dont le type de référence est démonstratif. Le choix du démonstratif est arbitraire, il pourrait s'agir d'un défini par exemple.

En tout 164 règles de projection inverse ont été définies. Nous ignorons cepen-

```
<pattern-group id="ces n hôtels" reference="true">
    <pattern match="lienRef-coRef" with-value="pluriel">
        <participant id="p1">
            <feature ns="mmil" name="refType" value="demonstrative"/>
            <feature ns="mmil" name="number" value="plural"/>
        </participant>
    </pattern>
    <pattern match="nombre" with-specif="hotel">
        <participant>
            <feature ns="mmil" name="objType" value="Hôtel"/>
            <feature ns="mmil" name="cardinality" value="@value"/>
        </participant>
    </pattern>
</pattern-group>
```

Fig. 10.3 – Exemple de règle de projection inverse

dant dans quelle mesure ces règles parviennent à la construction d'un composant correct. Nous avions tout d'abord estimé possible la vérification automatique des composants, par exemple en testant la projection inverse suivie d'une projection hors-contexte et en vérifiant qu'on retrouve bien la séquence de triplets, nous assurant alors de la bonne construction du composant. Cette vérification n'est pas envisageable pour deux raisons : d'abord la projection hors-contexte n'étant pas parfaite, elle ne permet pas de déterminer avec certitude les triplets correspondants à un composant MMIL et nous pourrions cumuler les erreurs de projection inverse et les erreurs de projection hors-contexte. Ensuite et surtout, même dans le cas où cette vérification restitue les bons triplets, nous n'aurions aucune garantie que le composant produit soit suffisant relativement aux exigences du module de référence. L'annotation MEDIA étant sous-spécifiée vis à vis de l'entrée attendue de notre module de référence, garantir une projection inverse correcte à l'aide de l'annotation MEDIA serait tout à fait insuffisant. Seule la vérification *manuelle* des composants pourrait alors être effectuée. Nous n'avons toutefois pas procédé à cette dernière. En conséquence, le Contrôle sera également susceptible de commettre des erreurs d'interprétation du corpus de référence.

Des travaux de reconstruction de forme sémantique ont également été conduits par Meurs et al. (2008). L'objectif est de construire une annotation sous la forme de *frames* sémantiques du corpus MEDIA inspirées du projet FrameNet (Fillmore, Johnson, and Petruck 2003). L'approche est comparable puisque la construction des *frames* s'appuie sur des patrons, reconnus à partir de séquences de triplets. En revanche, peu d'intérêt a été porté dans ce projet à la construction d'entités référentielles, indispensables pour fournir une interprétation de référence au Contrôle.

10.2.2 Soumission des énoncés à tester

Dans notre évaluation, les énoncés produits par le CONTRÔLE sont de deux types : les énoncés issus du corpus (étape 1) et les énoncés destinés à la convergence (étape 7). L'interprétation de ces énoncés ne doit toutefois pas être la même pour le TEST. En effet, étant donné que les énoncés du corpus peuvent porter du contenu assimilable à des marques d'ancrage, des Rejects ou des questions par exemple, il était nécessaire d'*opacifier* ces énoncés afin qu'ils ne perturbent pas le processus d'ancrage.

Nous avons alors choisi de traduire cette opacification par des requêtes d'interprétation de la forme « Interprète "..." ». Ainsi, le contenu à interpréter est protégé d'une interprétation relative à l'ancrage tout en traduisant la nécessité de manifester le résultat d'interprétation comme résultat d'une requête. En fait, cette approche est particulièrement intéressante puisqu'on pourrait considérer que le système est confronté à deux requêtes lorsqu'il reçoit un énoncé : une requête d'interprétation et une requête d'action. Dans une certaine mesure nous avons rendu compte de ces deux niveaux étant donné que les jugements de compréhension sont construits avant les jugements d'action mais cette direction mériterait d'être explorée davantage.

10.2.3 Instanciations TEST et CONTRÔLE

10.2.3.1 Instanciation des modules d'évaluation des jugements et d'application

L'instanciation des deux systèmes était relativement aisée grâce à la modularité de l'architecture. Seuls deux modules ont fait l'objet de raffinements : le module d'évaluation et le module applicatif (voir l'architecture donnée en figure 8.35). Afin de distinguer les deux systèmes nous avons procédé à des *restrictions* sur les jugements de type SS ou SO que les systèmes pouvaient construire. Ces restrictions sont simplement effectuées en supprimant certains jugements de la sortie du module d'évaluation. Les différentes restrictions de jugements que l'on peut adopter correspondent à différentes conditions d'évaluation. On peut choisir par exemple que les deux partenaires peuvent abandonner ou qu'ils peuvent tous les deux poser des questions de clarifications. Nous avons préféré effectuer une instanciation dans laquelle seul le CONTRÔLE avait le choix de l'abandon en donnant plusieurs chances au TEST de manifester sa compréhension. Afin de ne pas perturber le processus d'ancrage, nous avons interdit le CONTRÔLE de manifester sa non-compréhension. Le TEST n'a quant à lui pas la possibilité de manifester la mauvaise compréhension du CONTRÔLE mais seulement celle de poser des questions de clarification que nous restreignons à l'ambiguïté. Les pré-tests nous ont indiqués en effet des difficultés à correctement répondre aux questions de vide ou d'incertitude. Le tableau 10.1 résume les jugements que les participants sont autorisés d'avoir dans l'évaluation.

Le module applicatif a été instancié de deux manières différentes résumées dans le tableau 10.2. Pour le TEST, la méthode *executeRequest* est entièrement dédiée à l'exécution des requêtes d'interprétation. La méthode *integrateAnswer* est interdite

Point de vue	**Jugements SS et SO autorisés**
Test	$\overline{U}$:Ambiguous, UR, URA, $UR\overline{A}$
Contrôle	Ab, $U\overline{R}$, UR, URA

Tab. 10.1 – Restriction des jugements pour l'évaluation

étant donné que le Test ne peut initier de nouveaux échanges au niveau dialogue. La méthode *nextAction* est limitée à produire un URA (le reste du contenu provenant directement de l'ancrage des énoncés du Contrôle). Pour le Contrôle, c'est la méthode *executeRequest* qui est interdite car le Test ne produit pas de nouvelles requêtes. La méthode *integrateAnswer* est toujours URA, puisque cette méthode n'est appelée qu'en cas d'énoncé du Test jugé UR. Enfin, la méthode *nextAction* effectue la projection inverse à partir d'un nouvel énoncé du corpus et produit une requête d'interprétation correspondante.

Point de vue	*executeRequest*	*integrateAnswer*	*nextAction*
Test	exécute la requête d'interprétation	interdit	toujours URA
Contrôle	interdit	toujours URA	produit une nouvelle requête d'interprétation

Tab. 10.2 – Instanciation des méthodes de l'application

Un dernier module très simple coordonne le Test et le Contrôle en branchant l'entrée de l'un sur la sortie de l'autre. Il correspond à l'agent coordinateur dans (Araki, Watanabe, and Doshita 1997) mais contrairement à ce dernier n'introduit pas de bruit linguistique.

10.2.3.2 Production des paraphrases

La production du contenu sémantique à verbaliser a fait l'objet d'une limitation. Nous n'avons pas considéré les descriptions relationnelles (par exemple « la chambre à l'hôtel Ibis ») en raison de la difficulté à les verbaliser et à les interpréter correctement. En effet, les référents décrits par d'autres référents perturbent le processus d'ancrage : si le Contrôle dit « non, la chambre à l'hôtel Ibis », le Test va rechercher à tort une expression susceptible d'avoir référé à l'hôtel Ibis dans l'énoncé initial et provoquer une erreur *Missed* (section 8.4.2.4). Pour traiter correctement ces cas, il est nécessaire d'interpréter l'expression « la chambre à l'hôtel Ibis » en prenant en compte clairement la relation de dépendance syntaxique dans la forme sémantique[1]. Nous avons cependant pu observer, lors de l'évaluation en contexte, les problèmes de notre système d'interprétation à bien prendre en compte au niveau référentiel cette caractéristique (les modifieurs relationnels sont la seconde source

[1]Ce problème est lié à la nature de la relation, syntaxiquement elle est orientée mais sémantiquement on pourrait la représenter dans les deux sens. On ne peut donc pas s'appuyer de manière générique sur le sens de la relation sémantique pour déterminer le modifieur.

d'erreur, voir section 7.3 p. 115). Nous avons alors préféré éluder ces difficultés de l'évaluation miroir. Ce cas est symptomatique de l'influence de la robustesse *interne* dans la capacité de robustesse *externe*. Ici, notre système n'était pas suffisamment robuste de manière interne pour pouvoir considérer les descriptions relationnelles dans une démarche de robustesse externe. Cela ne signifie toutefois pas notre incapacité à résoudre les modifieurs relationnels par l'ancrage, en particulier ceux qui contraignent à tort la résolution, mais seulement notre incapacité à effectuer des dialogues de clarification à l'aide de descriptions relationnelles.

Etant donné le caractère critique de la génération pour l'ancrage, nous avons adopté plusieurs stratégies de robustesse interne. D'abord on essaie de générer un composant sémantique entier, et si cette génération échoue (la chaîne de caractères retournée par GenI est vide), on essaie de générer chaque participant en les coordonnant par une virgule ou par un « ou » dans le cas de l'ambiguïté. Si un participant ne parvient pas à être généré, on essaie de retirer progressivement des traits de sa description et on s'arrête dès qu'on obtient une chaîne de caractères non vide. Si on ne parvient toujours pas à obtenir une verbalisation, on teste si le participant possède un attribut MMIL *lex* auquel cas on l'utilise. Sinon on essaie de générer le participant à l'aide d'un générateur spécifique en fonction du type de participant. Nous n'avons toutefois implémenté qu'un seul générateur spécifique dédié aux hôtels.

10.3 Résultats

10.3.1 Quantitatifs

Nous comparons ici les résultats de IREF et de DREF avec et sans ancrage. Ils ont été calculés *avecHC* sur le corpus de test de l'évaluation en contexte (174 dialogues) pour un total de 455 lienRef.

Type	**F-mesure**	**Précision**	**Rappel**
IREF sans ancrage	54.8%	66.6%	46.5%
IREF avec ancrage	56.1%	71.1%	46.3%
DREF sans ancrage	59.4%	65.6%	54.3%
DREF avec ancrage	57.9%	60.4%	55.6%

TAB. 10.3 – Résultats comparatifs sans et avec ancrage

Nous pouvons observer un gain minime en IREF (+1.3) issu d'un gain en précision (+4.5) mais d'une baisse en rappel (-0.2). La f-mesure en DREF diminue même de 1.5. Le processus d'ancrage ne semble pas fonctionner de manière significative. Cette mesure globale ne nous dit toutefois rien des circonstances dans lesquelles l'ancrage réussit ou échoue. Nous avons donc procédé à l'examen manuel des situations problématiques afin d'examiner d'une part les causes d'échec du processus d'ancrage et d'autre part de déterminer quels étaient les types de divergence susceptibles d'être le mieux résolu.

10.3.2 Problèmes d'ancrage

Avant de détailler les problèmes d'ancrages, nous montrons un exemple tiré du dialogue 1192 auquel on a appliqué le protocole miroir, dans lequel l'ancrage fonctionne bien (figure 10.4). L'expression référentielle « ces deux hôtels » en C_{18} n'est pas résolue (en l'occurrence à cause du non-traitement des groupements extra-énoncés) et le Test manifeste sa compréhension par un seul « OK ». Le Contrôle jugeant l'énoncé C_{18} mal compris effectue un $U\overline{R}$ sous la forme d'un Reject+Inform. Cet énoncé est compris par le Test qui peut réinterpréter C_{18} et manifester sa nouvelle compréhension en T_{21}. Le Contrôle jugeant cette compréhension satisfaisante peut poursuivre.

C_{18} : OK, interprète "je voudrais confirmation concernant *ces deux hôtels*
: si c' est deux hôtels de bon standing"
T_{19} : OK
C_{20} : non, l'hôtel du lac, l'hôtel du rocher
T_{21} : OK donc l'hôtel du lac, l'hôtel du rocher
C_{22} : OK, interprète ...

Fig. 10.4 – Exemple d'ancrage satisfait (miroir 1192)

Le dépouillement des erreurs a été réalisé grâce à un outil permettant de parcourir facilement l'ensemble des résultats afin de les classer manuellement. Pour catégoriser les problèmes d'ancrage nous distinguons les situations suivantes :

- le Test ne parvient pas à corriger son interprétation malgré une preuve donnée par le Contrôle
- le Contrôle ancre un énoncé alors qu'il n'aurait pas dû le faire (jugement propre UR au lieu d'$U\overline{R}$)
- le Contrôle n'ancre pas un énoncé alors qu'il aurait dû le faire (jugement propre $U\overline{R}$ au lieu d'UR)
- le Contrôle répond mal à une question du Test
- l'ancrage est réalisé avec succès mais la projection finale échoue, comptant alors pour une erreur

Nous donnons ici quelques exemples de situations d'ancrage problématiques issues de l'évaluation.

10.3.2.1 Mauvaise réinterprétation du Test

Le cas le plus naturel semble être celui où le Contrôle manifeste la mauvaise compréhension du Test sans que celui-ci ne parvienne à réinterpréter correctement l'énoncé initial à cause d'une non-compréhension. Dans l'exemple de la figure 10.5, bien que le Contrôle parvienne d'une part à manifester au Test sa mauvaise compréhension et à correctement répondre à sa question de clarification, le Test échoue à effectuer la réinterprétation de C_{24} en raison de la présence d'une seconde erreur de non-compréhension. L'absence de stratégie alternative lors de la manifestation de la non-compréhension du Test conduit finalement à l'abandon.

C_{24} : OK, interprète "*le premier*"
T_{25} : OK
C_{26} : non, l'hôtel boulogne
T_{27} : je ne comprends pas "l'hôtel". C'est ambigu.
: L'esplanade ou un hôtel, lequel hôtel
C_{28} : l'hôtel boulogne
T_{29} : je ne comprends pas "l'hôtel". C'est ambigu.
: L'esplanade ou un hôtel, lequel hôtel
C_{30} : laisse tomber
T_{31} : OK
C_{32} : OK, interprète ...

FIG. 10.5 – Exemple de mauvaise réinterprétation du TEST (miroir 933)

10.3.2.2 Mauvaise réponse du CONTRÔLE à une question du TEST

Cependant, le processus d'ancrage requiert la participation des deux individus et le CONTRÔLE peut également être à l'origine des problèmes d'ancrage. En particulier lorsqu'il ne parvient pas à répondre correctement à une question de non-compréhension du TEST. Dans l'exemple de la figure 10.6, le CONTRÔLE ne fournit que deux hôtels sur les trois qu'il attend. Sa mauvaise interprétation de la question ne peut permettre au TEST de correctement réinterpréter C_{34} et grâce à lui C_{32}.

C_{32} : OK, interprète "euh je voudrais que *l'hôtel* soit proche d'un casino"
T_{33} : OK, l'hôtel Ibis saint gratien
C_{34} : non, l'hôtel Mercure, l'hôtel Ibis, l'hôtel meridien
T_{35} : je ne comprends pas "l'hôtel". C'est ambigu.
: L'hôtel Ibis ou l'hôtel Mercure, lequel hôtel
C_{36} : l'hôtel Mercure, l'hôtel Ibis
T_{37} : OK OK donc l'hôtel Mercure, l'hôtel Ibis
C_{38} : non, l'hôtel Mercure, l'hôtel Ibis, l'hôtel meridien
T_{39} : je ne comprends pas "l'hôtel". C'est ambigu.
: L'hôtel Ibis ou l'hôtel Mercure, lequel hôtel
C_{40} : laisse tomber
T_{41} : OK
C_{42} : OK, interprète ...

FIG. 10.6 – Exemple de mauvaise réponse à une question du TEST (miroir 735)

10.3.2.3 Jugement UR au lieu d'$U\overline{R}$

Les deux cas suivants correspondent à une mauvaise estimation de la compréhension d'autrui, en l'occurrence provenant du CONTRÔLE. Le premier de ces cas est celui où le CONTRÔLE construit un jugement UR au lieu d'un jugement $U\overline{R}$. Cette erreur peut provenir d'une mauvaise projection inverse, ou d'une non-compréhension

du Contrôle non manifestée en raison de la restriction des jugements. Dans les deux cas, il ne repère pas la divergence et manifeste à tort la bonne compréhension du Test. Dans l'exemple de la figure 10.7, le Contrôle n'a pas détecté le problème de compréhension de l'expression « l'hôtel » et a manifesté à tort un UR. Aucun gain de compréhension n'est donc à attendre d'un tel comportement.

C_{14} : OK, interprète "euh je veux d() plus de détails euh je veux que euh
: *l'hôtel* propose un parking surveillé et un restaurant et
: que chaque chambre soit à moins de cent euros"
T_{15} : OK
C_{16} : OK, interprète ...

Fig. 10.7 – Exemple d'UR au lieu d'$\overline{UR}$ (miroir 318)

10.3.2.4 Jugement $\overline{UR}$ au lieu d'UR

Le Contrôle peut au contraire produire un $\overline{UR}$ alors que la compréhension du Test est manifestement bonne. A l'instar de la situation précédente, ce cas peut provenir d'une mauvaise projection inverse ou d'une mauvaise interprétation de la preuve donnée par le Test. Mais contrairement à la situation précédente, ce cas peut entraîner un gain de compréhension étant donné que le Test peut tout de même réviser son interprétation. L'exemple de la figure 10.8 illustre un comportement erroné du Contrôle : bien que le Test ait atteint la bonne interprétation de « l'hôtel », le Contrôle lui manifeste sa mauvaise compréhension. Le Test explicite son problème de réinterprétation de C_{10} mais l'échange se conclut en abandon[2].

C_{10} : OK, interprète "*l'hôtel* à la porte de Versailles pour une chambre
: avec un couple pour un couple avec enfant"
T_{11} : OK
C_{12} : non, l'hôtel Mercure
T_{13} : OK donc l'hôtel Mercure
C_{14} : non, l'hôtel Mercure
T_{15} : je ne comprends pas "l'hôtel". Pour moi, "l'hôtel" réfère à l'hôtel Mercure,
: "une chambre" réfère à ..., "un couple" réfère à ..., et "un couple"
: réfère à ...
C_{16} : laisse tomber
T_{17} : OK
C_{18} : OK, interprète ...

Fig. 10.8 – Exemple d'$\overline{UR}$ au lieu d'UR (miroir 308)

[2]Les « ... » correspondent à des problèmes de génération.

10.3.2.5 Mauvaise projection

Enfin, le contexte de notre évaluation introduit une autre catégorie d'erreur, causée ni par le TEST ni par le CONTRÔLE mais par la projection qui autorise la mesure finale. Celle-ci peut conduire à évaluer négativement la bonne compréhension du TEST, que cette bonne compréhension ait été acquise à l'issue d'une réinterprétation ou qu'elle soit directement issue de son interprétation initiale. L'exemple de la figure 10.9 montre un des cas où l'ancrage est efficace alors que la projection a conduit à une erreur (ici causée par une mauvaise projection de l'ambiguïté).

C_{44} : OK, interprète "y a-t-il une piscine dans *l'hôtel*"
T_{45} : OK
C_{46} : non, l'hôtel copacabana, l'hôtel pinos
T_{47} : OK donc l'hôtel copacabana, l'hôtel pinos
C_{48} : OK, interprète ...

FIG. 10.9 – Exemple d'ancrage correct mais ayant conduit à une erreur de projection (miroir 1146)

10.3.3 Résultats par catégorie de problèmes

Un total de 198 problèmes d'ancrage a été relevé en vérifiant manuellement chaque tour de parole. Les problèmes d'ancrage ont été annotés en corrélant l'absence de gain en rappel ou sa dégradation avec le dialogue de clarification correspondant. Certains dialogues de clarification comprenaient plusieurs problèmes et dans ce cas nous les avons tous annotés. La proportion de chaque problème par catégorie est donnée dans le tableau 10.4.

Type	**Proportion**	**Occurrences**
projection	34.8%	69
UR au lieu d'$U\overline{R}$	29.3%	58
$U\overline{R}$ au lieu d'UR	20.7%	41
réinterprétation	10.1%	20
réponse	5.1%	10

TAB. 10.4 – Problèmes d'ancrage par catégorie

Le premier résultat important est que la projection est fautive en majorité des cas (34.8%). Ce problème peut être expliqué par la mauvaise coordination de la réinterprétation et de la projection. Afin d'associer aux lienRef les bons référents réinterprétés, il est nécessaire de reconstituer correctement les relations référentielles avant la projection (voir section 6.2.2 p. 108). Il est d'autant plus nécessaire de reconstruire les relations que l'annotation des entités non nommées incluent l'annotation des modifieurs relationnels (par exemple une chambre est décrite par un hôtel). Nous avons en effet intégré les problèmes liés à l'établissement de la relation associative dans les problèmes de projection. D'un certain point de vue on aurait pu

inclure ce type de problème dans les problèmes d'UR au lieu d'$\overline{UR}$ étant donné que le Contrôle ne vérifiait pas la compréhension de ces modifieurs. Nous avons préféré toutefois considérer qu'ils relevaient de la projection puisque ces relations sont construites *a posteriori* à partir de l'espace référentiel et dans le but de projection. Dans tous les cas, l'importance de ces relations fut largement sous-estimée lors de l'évaluation miroir.

Les deux sources suivantes de problèmes d'ancrage correspondent à une mauvaise estimation de la compréhension du Test par le Contrôle. La non-détection des divergences correspond à 29.3% des problèmes d'ancrage et dans chaque cas celle-ci provoque l'absence de gain de compréhension. Le Contrôle a également de nombreuses difficultés à percevoir que le Test a effectivement compris (20.7%), mais comme nous l'avons noté ces dernières n'influencent pas nécessairement le gain de compréhension final. Enfin, les deux dernières catégories posent relativement moins de problèmes. La responsabilité du Test dans la convergence s'élève à 10.1%. Ce faible score traduit le fait que la capacité du Test à atteindre l'interprétation du Contrôle soulève moins de problèmes d'ancrage que la capacité du Contrôle à la valider. La faculté du Contrôle à répondre à une question de clarification est enfin très minoritaire dans les causes de problèmes (5.1%).

Afin de mieux estimer les résultats réels du gain de compréhension en faisant abstraction des erreurs de projection, nous avons recalculé les scores d'IREF et de DREF en enlevant tous les tours qui présentaient un problème de projection[3]. Le score final est présenté dans le tableau 10.5.

Type	**F-mesure**	**Précision**	**Rappel**
IREF sans ancrage	55.9%	64.0%	49.7%
IREF avec ancrage	65.2%	73.2%	58.7%
DREF sans ancrage	61.5%	64.4%	58.9%
DREF avec ancrage	59.0%	60.2%	57.9%

Tab. 10.5 – Résultats comparatifs en enlevant les erreurs de projection

Les résultat traduisent bien l'impact des erreurs de projection puisqu'on obtient un gain de 9.3 points en f-mesure d'IREF (+9.2 en précision et +9 en rappel). A contrario le DREF diminue légérement (-2.5) en raison du plus grand nombre de référents disponibles (voir l'évaluation DREF p.112).

Nous avons enfin conduit une dernière évaluation en reconsidérant entièrement le module de projection. L'impact de la projection étant tellement important, nous avons décidé d'implémenter à nouveau le module. En particulier, nous avons abandonné une construction explicite des relations référentielles (voir p. 108) qui autorise une description incrémentale des référents mais soulève des problèmes relatifs à la réinterprétation. La projection en contexte consiste alors plus simplement à produire la représentation courante des référents, au fur et à mesure de l'interprétation ou

[3]Pour être plus précis, il serait nécessaire de n'enlever que les lienRef correspondant aux erreurs de projection, mais étant donné que nous avons procédé à l'annotation des tours de parole, il était plus simple de procéder ainsi.

après réinterprétation. Les résultats sont donnés dans le tableau 10.6. Ils confortent l'impact de la projection, puisque l'on obtient avec le nouveau module, plus du double du gain précédent d'IREF en rappel (18.6 points au lieu de 9 points) et en précision de 6.3 points (on perd cependant mécaniquement en précision de DREF). Nous avons également effectué des mesures restreintes aux hôtels seuls, et le gain d'IREF en rappel est encore plus important et s'élève à 23.7 points.

Type	**F-mesure**	**Précision**	**Rappel**
IREF sans ancrage	52.2%	63.7%	44.3%
IREF avec ancrage	66.3%	70.0%	62.9%
DREF sans ancrage	48.8%	83.0%	34.6%
DREF avec ancrage	48.3%	75.7%	35.5%

TAB. 10.6 – Résultats comparatifs avec le nouveau module de projection

10.3.4 Impact des problèmes d'ancrage sur l'abandon

Nous avons constaté que 100% des problèmes d'ancrage $U\overline{R}$ au lieu d'UR, réinterprétation et mauvaise réponse entraînaient l'abandon. L'abandon systématique est causé par l'absence de capacité des deux systèmes à adopter des stratégies alternatives. Il devrait être possible pour le CONTRÔLE d'effectuer de nouvelles paraphrases s'il réalise que le TEST ne parvient pas à comprendre deux fois de suite. Le TEST devrait également pouvoir poser une question de clarification différente s'il aperçoit que sa question précédente n'apporte pas de réponse. Nous indiquons en perspective une manière d'aborder le problème en représentant explicitement un jugement sur l'*évolution de la compréhension*.

10.3.5 Améliorations par type d'erreur

Nous n'avons procédé qu'à une analyse partielle des types de divergence les mieux résolus en analysant environ 40% du corpus de test. La projection cause toutefois de nombreux faux-négatifs, entraînant des difficultés pour estimer précisemment la quantité et la qualité de problèmes corrigés par l'ancrage. Les résultats montrent cependant des améliorations sur les erreurs de modifieurs relationnels et sur les groupements extra-énoncés. Les problèmes de restriction de sélection, de double altérité ou de genre semblent quant à eux partiellement résolus. En revanche nous n'avons trouvé aucun cas dans les dialogues analysés de bonne résolution d'anaphore associative multiple ni d'erreur référentielle provenant des niveaux lexical, syntaxique ou sémantique qui soit correctement résolue.

Bien que partiels, ces résultats montrent l'influence du contexte dans la résolution. Le problème des modifieurs relationnels était causé par la recherche d'un référent avec des contraintes relationnelles qui n'avaient pas lieu d'être (typiquement rechercher « un hôtel avec plus de détails » à partir de l'expression « plus de détails sur l'hôtel »). La paraphrase qui ne fait pas mention de ces modifieurs permet

d'identifier avec succès le référent grâce à l'absence de ces contraintes. Le problème des groupements extra-énoncés se posait lors de la construction d'un groupe de référents introduits dans des énoncés différents. La paraphrase, en énumérant les référents dans un même énoncé permet de constituer un nouveau groupe. Les autres cas d'erreurs partiellement résolus peuvent être en partie expliqués par la projection. Par exemple, les phénomènes d'altérités requièrent la construction d'une relation de codomanialité. En l'absence de cette relation, l'annotation correspondante n'est pas produite malgré une possible bonne résolution. La compréhension de l'anaphore associative multiple n'est pas améliorée, pas tant à cause de la contradiction logique de la lecture distributive (car nous avions relâché cette dernière à l'issue de l'évaluation en contexte) mais à l'absence de paraphrase relationnelle des entités de type chambre. Un examen plus précis des causes d'erreur est toutefois nécessaire afin de déterminer si l'absence de gain des autres types d'erreur est effectivement dûe à la projection ou non.

10.3.6 Conclusions

Nous avons présenté et conduit une évaluation du processus d'ancrage dans le dialogue machine-machine en langue naturelle. Cette évaluation a permis d'exhiber d'une part l'existence d'un gain de compréhension à l'issue du processus d'ancrage et d'autre part de dégager les principaux problèmes d'ancrage qui pouvaient se poser aux systèmes. Cependant, le principal problème qui a été soulevé est justement celui de l'inadéquation de la mesure du gain de compréhension. Cette mesure sous-estime en effet le gain réel de compréhension en incluant ce qui relève des problèmes de projection dans le formalisme de comparaison et ce qui relève de la compréhension effective. Nous discutons dans le chapitre suivant la méthodologie d'évaluation et certains moyens de circonvenir aux difficultés qu'elle soulève.

11 Discussion

11.1 Elimination de la projection dans l'évaluation

Nous avons proposé un modèle d'ancrage complètement symétrique dans lequel la non-compréhension et la mauvaise compréhension peuvent être détectées et manifestées par les deux participants. L'évaluation du modèle s'est alors naturellement reposée sur une estimation des capacités de convergence interprétative pour un couple de systèmes. Cette évaluation soulève néanmoins de nombreuses difficultés.

La difficulté la plus importante provient de l'évaluation MEDIA sur laquelle nous avons bâti l'évaluation du processus d'ancrage. Le paradigme de l'évaluation MEDIA fournit en effet un large corpus annoté en sémantique et en référence ainsi que des outils permettant d'effectuer une comparaison entre la forme d'interprétation d'un énoncé et l'annotation correspondante du corpus. L'intérêt de s'appuyer sur cette évaluation était de pouvoir réutiliser le corpus ainsi que les mesures développées. L'évaluation miroir consistait à faire converser deux systèmes à propos de l'interprétation d'un énoncé du corpus MEDIA et à vérifier s'ils étaient capables de s'accorder sur cette interprétation. L'un des systèmes, le CONTRÔLE avait accès à l'interprétation annotée des énoncés. Il communiquait une requête d'interprétation au TEST et ce dernier devait prouver qu'il avait effectivement bien interprété l'énoncé ou soulever des questions.

En comparant l'interprétation du TEST sans l'ancrage (évaluation MEDIA), avec celle réinterprétée à l'issue du processus d'ancrage, on peut évaluer leur capacité à ancrer un énoncé, c'est-à-dire leur capacité de détecter la mauvaise compréhension de leur partenaire, leur faculté de poser des questions de clarification ou d'y répondre. Toutefois l'évaluation MEDIA requiert des systèmes d'interprétation qu'ils puissent *projeter* leur forme d'interprétation propre dans le formalisme de comparaison. Loin d'être négligeable, nous avons constaté que la projection était fautive en majorité des cas d'absence de gain de compréhension. Autrement dit, la mesure fournie ne traduit pas correctement la réussite du processus d'ancrage.

On ne peut s'abstraire qu'avec difficulté de la projection. La première option est d'étudier *manuellement* chaque cas problématique afin d'estimer s'il est causé par la projection, de retirer ces cas et de fournir une nouvelle mesure plus à même de traduire la capacité à converger. La seconde option est de développer de nouvelles mesures fondées uniquement sur la forme d'interprétation propre, mais ces mesures risquent de soulever des problèmes de généricité. La troisième option est de déléguer entièrement la vérification du gain de compréhension au système CONTRÔLE. Cette

troisième option semble meilleure que les deux premières puisqu'elle ne requiert ni de projection, et à l'avantage d'être générique grâce à la langue naturelle. Elle nécessite toutefois que le Contrôle soit le plus robuste possible afin que son estimation de la compréhension de son partenaire soit la plus juste.

Cette nécessité n'a toutefois pas été satisfaite lors de notre évaluation. En effet, hormis la projection, la cause principale d'absence de gain de compréhension peut être imputée au Contrôle. Dans de nombreux cas, celui-ci perçoit mal la mauvaise compréhension du Test en « oubliant » de vérifier la compréhension de certains référents. A l'inverse il peut également croire à tort à la mauvaise compréhension du Test. L'évaluation a montré que finalement, la responsabilité du Test dans l'absence de gain était plus faible que celle du Contrôle. La robustesse interne du Contrôle n'était donc pas assez importante pour exclure sa responsabilité dans l'échec de l'ancrage.

Toutefois, nous ne croyons pas que les mauvaises performances de notre Contrôle soient suffisantes pour remettre en cause le paradigme d'évaluation. En effet, à l'instar de SIMDIAL, ou de DQR, l'utilisation de la langue naturelle pour évaluer la capacité de compréhension de la langue naturelle satisfait la contrainte de généricité de l'évaluation. Ces approches, comme la nôtre, requièrent néanmoins une capacité *minimale*, capacité minimale d'interaction de l'utilisateur simulé pour SIMDIAL et capacité de compréhension de la question pour DQR. L'évaluation MEDIA requiert également une capacité minimale, celle de la projection. Mais contrairement aux deux autres approches, cette capacité n'a rien à voir avec la compréhension[1]. Toutefois, dans notre cas, disposer d'une capacité minimale d'ancrage afin de permettre l'évaluation est problématique, puisque c'est précisemment celle qu'on cherche à évaluer.

Nous pouvons néanmoins faire remarquer que nous avons procédé à l'évaluation en attribuant des jugements de compréhension *différents* aux deux systèmes. Par exemple seul le Contrôle était susceptible de manifester la mauvaise compréhension du Test, et à l'inverse seul le Test était autorisé à manifester sa non-compréhension. Bien que la capacité d'ancrage requiert ces deux capacités, on peut faire l'hypothèse que l'amélioration du Contrôle passe avant tout par l'amélioration de la faculté de juger la mauvaise compréhension d'autrui, c'est-à-dire la construction du jugement *SO* (et dans un second temps par sa faculté à *répondre* aux questions de clarification). Dès lors, on peut constituer une évaluation préalable de cette seule capacité afin de procéder à son amélioration. Il est possible par exemple de fournir des couples d'énoncés au Contrôle, le premier fournissant un énoncé de référence, et le second fournissant une preuve de compréhension du premier énoncé. Chaque couple est annoté en fonction du fait que le second énoncé apporte ou non une preuve suffisante de la compréhension du premier. On peut alors évaluer la capa-

[1] On peut modérer tout de même cette remarque étant donné que nous avons proposé une méthode de projection sémantique qui justement établit un parallèle entre la faculté de projection et celle de produire une forme d'interprétation *liée à l'application* à partir d'une forme d'interprétation *liée à la langue*. Le formalisme MEDIA peut alors dans cette perspective être considéré comme une représentation applicative, et la projection comme capacité nécessaire dans un système.

cité du CONTRÔLE à correctement estimer si le second énoncé apporte une preuve de compréhension suffisante du premier. On pourrait de même procéder à l'évaluation séparée de la constitution et de la gestion des autres jugements de compréhension. De telles techniques sont possibles pour améliorer dans un premier temps le CONTRÔLE. Dans un second temps, lorsque le CONTRÔLE est suffisamment fiable on peut l'utiliser pour évaluer de manière générique d'autres systèmes, dans l'esprit du paradigme SIMDIAL. Ce faisant, on peut parvenir à l'objectif d'évaluation de la robustesse externe d'un système donné en faisant abstraction des problèmes d'ancrage causés par le CONTRÔLE et surtout, en éliminant toute nécessité de projection.

11.2 Analogies avec le dialogue homme-machine

L'évaluation de l'ancrage a été réalisée en contexte de dialogue machine-machine en langue naturelle. Que peut-elle nous dire des capacités d'ancrage dans le contexte du dialogue homme-machine ? Dans le dialogue homme-machine, les facultés d'ancrage que nous avons données séparément au CONTRÔLE et au TEST sont réunies dans un même système (on ne procède à aucune restriction de jugements).

La première remarque concerne la capacité à juger de la mauvaise compréhension d'autrui. Comme nous l'avons montré, celle-ci constitue une des sources principales des problèmes d'ancrage. Ces faibles performances semblent alors suggérer la difficulté du système à juger de la compréhension de l'utilisateur. Toutefois, il faut rappeler que la construction du jugement SO a été extrêmement contrainte pour l'évaluation : un énoncé O_n est jugé non-pertinent vis à vis d'un énoncé S_k ($k < n$) s'il existe un des référents de S_k auquel ne fait pas référence O_n (p. 161). Le jugement R requiert l'explicitation de chaque référent présent dans un énoncé. On ne peut évidemment pas faire cette hypothèse dans le dialogue homme-machine car cela obligerait l'utilisateur à prouver de manière beaucoup trop forte sa compréhension. En fait, le jugement SO qui pose le plus de problèmes dans l'évaluation miroir, est peut-être celui le moins important dans le dialogue homme-machine. Bien qu'il soit nécessaire de les prendre en compte, particulièrement en conditions bruitées, on peut supposer que les problèmes de mauvaise compréhension de l'utilisateur ne sont pas majoritaires.

Le jugement OS du CONTRÔLE est dans tous les cas bien détecté par le TEST en raison de la forme des énoncés produits. Pourtant de nombreuses formes d'$U\overline{R}$ sont possibles dans le dialogue homme-machine sans que le Reject de l'utilisateur soit explicitement verbalisé par un « non ». L'évaluation ne prend pas en compte la capacité à détecter ce type de jugement. Toutefois, il est possible d'évaluer précisemment la capacité à détecter la manifestation de la mauvaise compréhension en annotant des énoncés selon qu'ils manifestent la présence ou non d'une preuve de mauvaise compréhension. L'évaluation consiste alors en un simple test booléen, évitant tout problème de projection.

Le jugement SS est peut-être plus important dans le dialogue homme-machine. Il n'a été toutefois que partiellement évalué grâce à la prise en compte des questions de clarification portant sur l'ambiguïté. Nous avons en effet restreint les questions

de clarification à l'ambiguïté, en raison de l'impact négatif des questions de vide ou d'incertitude sur le processus d'ancrage. Il arrivait fréquemment en effet lors des pré-tests que le Contrôle ait du mal à comprendre ces questions et nous les avons enlevées de l'évaluation finale. Les questions portant sur l'ambiguïté n'ont toutefois jamais été observées en tant que clarification de l'énoncé initial mais uniquement en tant que clarification sur la preuve de mauvaise compréhension manifestée par le Contrôle. Il semble probable que cette erreur provienne d'une mauvaise construction du jugement de compréhension de la requête d'interprétation et qu'à l'issue de la réinterprétation le jugement correct soit construit.

En détaillant les causes d'ambiguïté, nous avons observé qu'elle était dans tous les cas causée par un problème d'ordre lexical ou syntaxique. L'exemple de la figure 11.1 illustre typiquement le problème d'analyse de « l'hôtel EtapHotel » en C_{82} où seul « l'hôtel » est analysé, provoquant une ambiguïté. Ici le Contrôle parvient à produire « l'hôtel EtapHotel » alors que le Test ne parvient pas à l'interpréter en raison du générateur spécifique employé pour les noms d'hôtels (p. 192). Afin de prendre en compte ces cas, il serait d'abord nécessaire d'établir des jugements à *d'autres niveaux* que la référence, au niveau lexical, syntaxique ou sémantique, pour déterminer que probablement « EtapHotel » est un nom propre désignant un hôtel. Ensuite il faudrait être capable d'acquérir le lexème « EtapHotel » comme nom propre. Enfin il serait indispensable d'identifier « l'hôtel EtapHotel » et « l'hôtel Etap » au niveau référentiel en adoptant des contraintes de résolution plus souples sur les noms propres.

C_{80}	: OK, interprète “et le troisième”
T_{81}	: OK, l'hôtel des lilas
C_{82}	: non, l'hôtel EtapHotel
T_{83}	: je ne comprends pas “l'hôtel”. C'est ambigu.
	: Le lilas ou la chapelle ou l'hôtel Etap, lequel hôtel
C_{84}	: l'hôtel EtapHotel
T_{85}	: je ne comprends pas “l'hôtel”. C'est ambigu.
	: Le lilas ou la chapelle ou l'hôtel Etap, lequel hôtel
C_{86}	: laisse tomber

Fig. 11.1 – Exemple d'ambiguïté issue d'un problème lexical (miroir 712)

L'exemple de la figure 11.1 illustre également certaines difficultés pour répondre aux questions de clarification (jugement *OO*). Nous avons comptabilisé ce type de réponse comme bonne réponse mais on peut s'interroger sur la pertinence de ce choix. En effet, bien que la question fasse l'objet d'une véritable résolution par recherche dans l'espace référentiel, elle fournit au final en C_{84} une réponse identique à celle donnée en C_{82}. Le problème est que, bien que le Contrôle perçoive la non-compréhension du Test, il ne propose pas de paraphrase alternative en réponse à sa question. Pourtant le Test mentionne explicitement « l'hôtel Etap ». Si le Contrôle était plus collaboratif, il devrait réutiliser le nom de l'hôtel mentionné par le Test au lieu de s'appuyer sur sa conceptualisation propre. L'adaptation des

deux systèmes l'un à l'autre, par exemple avec la prise en compte de conceptualisations partagées, est nécessaire : chaque système devrait effectuer un pas vers son partenaire mais ni la simple réponse à une question de clarification, ni la production de deux requêtes identiques ne sont des pas suffisants vers cette adaptation.

11.3 Conclusions de l'évaluation

L'évaluation que nous avons proposée ne permet pas d'affirmer avec certitude qu'un système sera robuste dans le dialogue homme-machine. Elle permet d'affirmer en revanche qu'un système ne le sera pas. En effet, étant donné qu'elle s'appuie sur des capacités minimales pour l'ancrage, un système qui échoue à mettre en oeuvre le protocole miroir ne sera pas à même de gérer correctement les problèmes de compréhension dans le dialogue.

Elle permet avant tout de générer automatiquement des situations assez complexes pour faire naître des problèmes d'ancrage, qu'ils proviennent du système en tant que locuteur ou en tant qu'interlocuteur. L'examen de ces problèmes requièrent toutefois la présence humaine. Grâce à l'évaluation nous avons surtout pu constater que la difficulté de notre système à estimer la compréhension de son partenaire était une cause importante de problème dans la conduite du processus d'ancrage. Plus important encore, nous avons pu également constater la nécessité de l'adaptation des partenaires, par exemple grâce à la prise en compte de jugements à d'autres niveaux, en particulier au niveau conceptuel. Cette adaptation devrait viser à dépasser le simple cadre des requêtes de clarification, en autorisant un alignement entre les conceptualisations sans que la présence de requête de clarification soit nécessaire.

D'autre part, à l'instar de Allemandou et al. (2007), nous estimons que l'évaluation miroir ne peut remplacer une évaluation de type satisfaction avec de vrais utilisateurs. Les deux types d'évaluation sont complémentaires : l'évaluation par simulation permet de générer un grand nombre de situations problématiques automatiquement mais présente l'inconvénient de ne pas être nécessairement fidèle à une situation réelle, et au contraire l'évaluation avec de vrais utilisateurs fournit un plus petit nombre de situations mais celles-ci correspondent au comportement du système en conditions réelles. Pour conduire une évaluation avec de véritables utilisateurs, il nous serait néanmoins indispensable de considérer non pas l'ancrage en tant que tel mais l'ancrage au sein d'une tâche donnée.

Un aspect en particulier n'a pas fait l'objet d'évaluation miroir mais nécessiterait une évaluation avec des utilisateurs. Il s'agit de toutes les preuves de compréhension produites par le TEST qui n'ont pas été interprétées par le CONTRÔLE. En particulier, les coordinations « donc » et « mais », ou la manifestation du type de problème (« c'est ambigu »), n'ayant pas été interprétées par le CONTRÔLE ne pouvaient pas être évaluées. Ces preuves ont été produites dans la perspective du dialogue homme-machine et l'évaluation miroir que nous avons conduite est incapable d'en mesurer les effets. Grâce à une évaluation avec des utilisateurs, l'intérêt de ces marqueurs pourrait être mieux appréhendé.

Conclusions et perspectives

12 Conclusion

Les systèmes de dialogue sont des interfaces complexes, susceptibles d'être confrontées à des problèmes d'interprétation multiples. Nous avons proposé d'étendre le paradigme détection/correction pour considérer en quoi la gestion du terrain commun permettait de résoudre ces problèmes. Dans cette approche, la compréhension doit être recherchée de manière proactive par les deux participants, chacun d'eux manifestant la convergence ou la divergence de l'interprétation afin d'ancrer les énoncés dans le terrain commun. Cet ancrage a plusieurs effets : au niveau local, il permet aux participants d'atteindre l'interprétation désirée par leur partenaire et au niveau global, il permet d'affirmer la compétence interprétative des deux agents, facilitant par la suite la communication.

12.1 Critère d'ancrage

Ancrer un énoncé revient à déterminer les conditions nécessaires à cet ancrage. Les premiers travaux en ce sens (Clark and Wilkes-Gibbs 1986; Clark and Schaefer 1989) suggèrent qu'afin d'ancrer un énoncé, il est d'abord nécessaire qu'il soit correctement compris. Ce but d'ancrage peut motiver la manifestation de la compréhension d'un énoncé dans d'autres énoncés, en particulier la manifestation *positive* de la compréhension. Cependant, cette définition entraîne un paradoxe : pour savoir si un énoncé est correctement compris, il est nécessaire de comprendre correctement les énoncés qui en manifestent la compréhension, et alors on ne peut jamais savoir si un énoncé est correctement compris. Ce problème, soulevé sous le nom d'acceptation récursive par Traum (Traum 1999), a entraîné d'autres définitions des conditions de l'ancrage. Tous les modèles que nous avons étudiés effectuent une certaine forme de troncature dans la récursion. Par exemple le modèle de Traum suppose que les preuves de compréhension sont comprises par défaut, le modèle de Larsson suppose que tous les énoncés rejoignent le terrain commun dès leur production et que ce terrain commun peut être révisé, ou encore le modèle de Bunt suppose que deux niveaux de preuves peuvent suffire.

Notre première contribution correspond à des clarifications sur le critère d'ancrage. Quel que soit le critère d'ancrage que l'on définira, celui-ci ne pourra jamais être fiable. En effet, comme toute communication, la manifestation de la compréhension peut être mal comprise, entraînant alors un ancrage erroné. Le terrain commun n'existe pas en dehors des individus, mais chaque participant construit une représentation plus ou moins explicite du terrain commun qui existe avec ses partenaires

à partir de la manifestation de la compréhension. Comme cette manifestation peut être mal comprise, les participants peuvent entretenir un terrain commun divergent. Les différents modèles d'ancrage proposent en fait différents degrés de confiance avec lequel l'ancrage peut être déterminé mais aucun d'entre eux, pas plus que le modèle que nous avons proposé, ne pourra garantir l'ancrage.

Afin de résoudre le problème de l'acceptation récursive nous n'avons pas considéré, comme le modèle de Traum, que les preuves de compréhension étaient automatiquement comprises mais plutôt que les preuves de compréhension pouvaient avoir des effets relativement à l'ancrage sans le satisfecit du partenaire. Ce faisant, nous avons laissé complètement ouvert la possibilité d'interrompre le processus d'ancrage à cause de la non-compréhension d'une preuve ou d'effectuer un ancrage erroné résultant d'une mauvaise compréhension de la preuve. Tous les énoncés peuvent jouer un effet dès qu'ils sont crûs compris par le partenaire et cet effet peut être celui désiré par le locuteur ou pas. Etant donné que les participants manifestent leur compréhension, ils peuvent déduire de cette manifestation, si les effets désirés de leurs preuves ont été joués ou non, et si ce n'est pas le cas, peuvent enclencher des stratégies visant à la convergence. Nous n'avons fait que préciser que ces effets s'appliquaient également à l'ancrage.

Pour modéliser ce principe, nous nous sommes d'abord appuyés sur la représentation des exigences de compréhension en distinguant les exigences propres d'un participant, des exigences qu'il attribue à autrui. L'ancrage d'un énoncé requiert dans notre modèle la satisfaction des exigences propres et des exigences attribuées. Cependant la connaissance des exigences attribuées, c'est-à-dire manifestées dans un énoncé, peut se fonder seulement sur les exigences de compréhension propres de cet énoncé. Autrement dit, on peut s'appuyer avec suffisamment de confiance sur la compréhension des preuves de compréhension pour qu'elles puissent jouer des effets.

Toutefois, les exigences attribuées ne sont pas uniquement construites sur la base des énoncés d'autrui. Comme le note Clark, la quantité de preuves de compréhension attendue par le locuteur influence *a priori* la manifestation de la compréhension. La constitution des exigences attribuées est un problème difficile étant donné la multiplicité des facteurs qui interviennent. Le médium utilisé, ou le rapport social entre les partenaires peuvent par exemple influencer la production des preuves. Nous avons donc procédé à plusieurs simplifications dans la construction de ces exigences. Nous avons en effet limité la prise en compte des exigences d'autrui à deux facteurs : d'une part les exigences *a priori* sont supposées corrélées à la quantité de problèmes, autrement dit plus il existe de problèmes d'interprétation, plus le système doit manifester sa compréhension et d'autre part les exigences *a posteriori* sont directement issues de la compréhension des preuves de compréhension fournies par le partenaire.

Nous avons également limité les problèmes de compréhension que ces dernières peuvent soulever. Nous n'avons pas modélisé la mauvaise compréhension des preuves mais seulement leur non-compréhension. La mauvaise compréhension des preuves est en effet un autre problème difficile puisqu'elle peut entraîner la réinterprétation de pans complets de dialogue à l'issue de la détection d'un quiproquo. La principale difficulté est qu'elle requiert de pouvoir adopter le point de vue d'autrui pour es-

timer ce qu'il signifiait en produisant un énoncé, et ce plusieurs énoncés après la production de cet énoncé. Ce problème est certainement très intéressant puisque nous supposons que les systèmes peuvent fréquemment faire face à des situations de ce type en situations réelles. Au contraire la non-compréhension des preuves est plus facile puisqu'elle interrompt le processus d'ancrage en dédiant le dialogue à leur compréhension.

12.2 Processus d'ancrage

La seconde contribution importante concerne l'implémentation du processus d'ancrage et son adjonction à un système d'interprétation. Le choix du modèle de départ s'est porté sur le modèle des contributions de Clark et Schaefer (Clark and Schaefer 1989) et en particulier le modèle des échanges (Cahn and Brennan 1999), la plus fidèle version computationnelle du modèle des contributions. Le modèle des échanges souffrait toutefois de nombreux défauts dont la compréhension systématique des preuves de compréhension, l'absence du résultat de l'interprétation dans la modélisation du processus d'ancrage ou même l'absence de définition explicite du critère d'ancrage. Nous sommes tout d'abord partis d'une définition *épistémique* du critère d'ancrage sous la forme d'une conjonction finie de croyances de compréhension. Cette définition permet de faire participer le résultat de la satisfaction des exigences de compréhension propres et attribuées, tout en autorisant la possibilité d'un ancrage plus fin. Nous n'avons cependant pas exploré outre mesure la définition épistémique de ce critère d'ancrage en raison du fait que nous nous sommes limités au niveau de la référence et un critère plus précis ne nous était pas nécessaire. Nous avons plutôt appuyé le processus d'ancrage sur un critère structurel, fondé sur la structure de dialogue du modèle des échanges. Cette structure est construite incrémentalement à partir des croyances de compréhension de l'énoncé et des croyances de compréhension qu'il manifeste.

Plutôt que d'employer le terme de croyance, nous avons préféré parler de *jugement* de compréhension dans la mesure où un jugement peut décrire aussi bien une attitude interne vis à vis d'un fait, que la manifestation de cette attitude dans le dialogue. Les jugements de compréhension permettent de résoudre les problèmes du modèle des échanges : la prise en compte de l'interprétation est effectuée en considérant le jugement d'un individu sur sa propre interprétation, et la compréhension de la preuve de la compréhension correspond simplement à un jugement de cet individu au niveau de la preuve manifestée par son partenaire.

Nous nous sommes très largement inspirés du modèle des échanges afin de modéliser les situations de non-compréhension ou de mauvaise compréhension. D'une part nos jugements de compréhension étendent les catégories de compréhension de ce modèle, et d'autre part, nous appuyons la modélisation des situations sur la représentation de la structure de dialogue (à l'instar du modèle des contributions). Les jugements de compréhension sont alors vus comme des opérateurs de mise à jour de la structure de dialogue. Notre modèle permet de considérer des révisions différées de la structure issues par exemple de la non-compréhension d'une mauvaise

compréhension ou encore certains cas particuliers de révision de preuve.

12.3 Evaluation de la gestion du terrain commun

Enfin, notre troisième contribution relève de l'évaluation du terrain commun et de sa gestion. Deux évaluations ont été conduites. Nous avons d'abord procédé à l'évaluation du terrain commun existant *au préalable* entre un utilisateur prototypique et le système d'interprétation au sein de la campagne MEDIA (projet TechnoLangue). En comparant la sortie de notre système projetée dans un formalisme pivot avec l'annotation humaine d'un corpus à l'aide de ce même formalisme, nous avons été à même de dégager les principales sources d'erreur d'interprétation, en nous focalisant toutefois sur les erreurs au niveau de la référence.

L'évaluation du processus d'ancrage lui-même nécessitait de fournir un partenaire au système. Plutôt que d'effectuer une évaluation homme-machine, peu représentative selon nous à petite échelle, nous avons préféré nous inspirer du paradigme SIMDIAL (Allemandou 2007) dans lequel la simulation déterministe d'un utilisateur permet d'évaluer automatiquement les capacités dialogiques du système. Dans notre contexte, il s'agissait d'évaluer si deux participants parviennent à ancrer un énoncé. L'un des participants, appelé Contrôle, fournit un énoncé à interpréter et son partenaire, appelé Test, doit manifester sa compréhension de l'énoncé. Conformément au principe d'ancrage, les deux partenaires sont impliqués dans l'ancrage de l'énoncé : le Test doit prouver sa compréhension et le Contrôle vérifie que cette compréhension est la bonne.

L'évaluation MEDIA nous a permis de collecter un ensemble de dialogues grâce auxquels on peut tester le processus d'ancrage. Il suffit de fournir les énoncés du corpus munis de leur annotation correcte au Contrôle, et d'en demander l'interprétation au Test. Nous avons été ensuite à même d'observer dans quelle mesure les systèmes parvenaient à converger. Les deux systèmes qui ont été utilisés sont deux instanciations du système principal qui différaient par leur capacité d'interprétation des énoncés du corpus problématique : l'interprétation du Contrôle s'appuyait sur l'annotation de référence tandis que l'interprétation du Test s'appuyait sur le module d'interprétation. Ils différaient également par leur capacité à construire les jugements, nous avons par exemple interdit au Contrôle de manifester sa non-compréhension ou au Test de décider d'un abandon.

Les résultats montrent avant tout l'impact très important de la projection dans la méthodologie d'évaluation. Afin de supprimer la nécessité de projection nous avons proposé de faire reposer l'évaluation sur la décision du Contrôle. Cette évaluation est en effet générique étant donné qu'elle s'appuie sur la langue naturelle. Cependant les performances du Contrôle ont démontré la difficulté du système à correctement estimer la compréhension de son partenaire. En effet, dans la majorité des cas problématiques d'ancrage, le Contrôle estimait à tort que le Test avait correctement compris l'énoncé. Malgré ces deux problèmes importants nous avons pu malgré tout observer un gain de compréhension des expressions référentielles à l'issue du processus d'ancrage.

L'évaluation miroir, en simulant le processus d'ancrage entre deux systèmes, fournit automatiquement un grand nombre de situations dont la variété permet surtout d'estimer les difficultés à ancrer un énoncé. Elle requiert néanmoins la présence humaine afin de pouvoir détailler précisemment les causes d'échec du processus d'ancrage. De plus, elle ne remplace pas l'évaluation avec de vrais utilisateurs dans la mesure où tous les aspects du processus n'ont pas pu être évalués. En particulier il est nécessaire pour évaluer un phénomène de pouvoir à la fois le produire mais également de l'interpréter. Les preuves de compréhension basées sur le type de problème (« c'est ambigu ») ou encore les coordinations (« donc » et « mais ») n'ont pas pu être évaluées car bien que produites elles n'étaient pas interprétées.

13 Limitations et perspectives

Un modèle de gestion du terrain commun doit définir trois aspects fondamentaux : comment établit-on un terrain commun ? que met-on dans le terrain commun ? et comment utilise-t-on ce terrain commun ? Nous avons principalement apporté des réponses à la première question de ce triptyque en vue d'améliorer la robustesse des systèmes de dialogue. Le modèle d'ancrage que nous avons défini et implémenté établit un terrain commun au niveau de l'identification d'un référent. Toutefois, ce modèle peut être amélioré dans plusieurs directions. La première direction concerne l'extension des situations prises en compte grâce à une meilleure modélisation du processus d'ancrage. La seconde direction vise à étendre le terrain commun considéré à d'autres niveaux. La troisième direction cherche à mieux utiliser le terrain commun en adoptant des conceptualisations partagées. Enfin, la quatrième piste de recherche concerne une évaluation plus fine du processus d'ancrage.

13.1 Améliorations du processus d'ancrage

13.1.1 Mauvaise compréhension des preuves

Nous avons d'abord simplifié de manière importante le processus d'ancrage. La première simplification concerne l'absence de gestion de la *mauvaise compréhension des preuves*. Nous n'avons pas considéré ce cas en raison de la complexité nécessaire à la réinterprétation. Pour parvenir à correctement réinterpréter une preuve de compréhension nous avons suggéré qu'il était indispensable de pouvoir se replacer dans le contexte de sa réception, autrement dit de faire des retours en arrière. Les retours que nous avons considérés, au niveau de la mauvaise compréhension du contenu, sont plus « faciles » à gérer puisqu'ils ne consistent qu'à réviser le contenu propositionnel ainsi que l'espace référentiel. Nous estimons que la réinterprétation de la preuve passe par le maintien des différentes évolutions du contexte de dialogue. En se replaçant dans le contexte de sa production, on pourrait alors plus facilement prendre en compte les effets qu'elle était sensée jouer pour son locuteur. Nous avons toutefois montré en quoi la réinterprétation pouvait être coûteuse en termes d'efforts cognitifs, surtout lorsque cette réinterprétation est tardive. Cependant nous ignorons dans quelle mesure ces coûts peuvent conduire à un abandon. En effet si dans le contexte de la conversation, les participants sont peut-être plus à même d'essayer de rétablir la bonne compréhension, dans le contexte du dialogue homme-machine

finalisé, il est probable que les participants cherchent plutôt à abandonner et à revenir dans la tâche. Toutefois les contraintes, les moyens et les coûts impliqués dans la réinterprétation restent mal connus et des études sur ces caractéristiques pourraient guider le choix de la réinterprétation ou non.

13.1.2 Evolution de la convergence

La seconde simplification est que nous n'avons pas représenté explicitement l'*évolution de la convergence*, c'est-à-dire la représentation explicite de l'état courant *vis à vis* de l'état antérieur. La seule évolution que nous prenons en compte est relative à l'échec ou la réussite de la réinterprétation issue de la mauvaise compréhension. Nous n'avons pas considéré l'échec de la réinterprétation issue de la non-compréhension. Par exemple, lorsque le système pose une question de clarification et reçoit une réponse satisfaisante mais que la réinterprétation ne change pas (le cas *Unchanged* pour la mauvaise compréhension), il repose actuellement la même question jusqu'à l'abandon. Il est nécessaire cependant d'adopter une stratégie alternative. La représentation explicite de l'évolution de la convergence, par exemple sous la forme d'un jugement portant sur ce niveau, permettrait d'unifier le traitement de cette évolution pour la non-compréhension ou la mauvaise compréhension. Dès lors, une fois que ces jugements sont disponibles, ils peuvent être manifestés explicitement dans le dialogue (par exemple « je ne comprends *toujours* pas »). Cependant, nous avons rencontré quelques difficultés à construire ces jugements. Il est en effet nécessaire de relâcher l'*hypothèse de prépondérance des jugements propres* afin de les considérer.

13.1.3 Remise en cause de l'hypothèse de prépondérance des jugements propres

L'hypothèse de prépondérance des jugements propres signifie que le système prend d'abord en compte son propre jugement de compréhension de l'énoncé reçu avant de considérer le jugement manifesté dans cet énoncé. En conséquence, une non-compréhension du système, même partielle, bloque totalement le processus d'ancrage pour dédier le dialogue à sa résolution. La contribution qui manifeste une preuve de compréhension est mise en attente comme contribution flottante jusqu'au moment où sa compréhension est suffisante pour considérer ses effets relatifs à l'ancrage. L'avantage de procéder ainsi est d'unifier les cas de non-compréhension du contenu et les cas de non-compréhension de la preuve de compréhension. Cependant, si le système a un problème de non-compréhension, l'énoncé problématique initie une contribution flottante et provoque une question de clarification dans sa phase d'acceptation. Toutefois si la réponse à cette question n'est pas comprise, elle initie à son tour une contribution flottante. Ce faisant, on perd le lien entre les deux contributions et la faculté de juger l'évolution de la convergence de l'interprétation de la première. Pour conserver le lien, il faut envisager que la réponse à la question puisse jouer dans une certaine mesure son rôle de réponse (et de preuve de compréhen-

sion) et non pas la mettre en attente comme contribution flottante. Ainsi, on peut considérer un jugement sur l'évolution de la compréhension de l'énoncé initial.

La seconde conséquence de cette hypothèse est qu'elle ne considère pas certains abandons. Par exemple le fait que l'utilisateur fournisse une preuve de non-compréhension d'un énoncé du système et que cette preuve soit non comprise, peut être un facteur susceptible de motiver l'abandon de l'ancrage de l'énoncé du système. Ou encore, si le système reçoit une preuve de sa mauvaise compréhension mais qu'il ne comprend que partiellement cette preuve, la prise en compte d'abord de la mauvaise compréhension peut éliminer l'intérêt de résoudre la non-compréhension de la preuve.

Bien que facilitant le processus, l'hypothèse de prépondérance des jugements propres n'est pas souhaitable dans un modèle d'ancrage. Il doit en effet être nécessaire de pouvoir prendre en compte les preuves manifestées tout en autorisant des non-compréhensions partielles sur ces énoncés. On peut alors considérer deux types de non-compréhension, celle qui est critique pour la prise en compte des effets relatifs à l'ancrage (par exemple n'avoir pas du tout entendu l'énoncé), et celle qui n'empêche pas de les prendre en compte (par exemple la compréhension contingente d'une expression référentielle). La non-compréhension critique peut être gérée de la façon que nous avons indiquée, par l'initiation d'une contribution flottante. La non-compréhension qui ne bloque pas le processus peut être gérée de manière similaire aux problèmes d'exécution ($UR\overline{A}$), en prenant en compte la preuve manifestée, puis en initiant un nouvel échange dans sa phase d'acceptation. Une autre direction possible est d'appuyer la choix de la méthode à déclencher non pas sur un arbre de décision (SS puis SO puis OO puis OS), mais sur un tableau croisé qui estimerait pour un couple de jugements propres et manifestés, l'action à déclencher. Ce genre de modélisation faciliterait en effet la prise en compte combinée du jugement propre et du jugement d'autrui.

13.1.4 Insuffisances du modèle des échanges

Nous avons effectué une autre simplification importante dans la prise en compte des preuves manifestées en considérant que les énoncés n'apportaient qu'une seule preuve, la plus forte. Cependant, le tour de parole peut être constitué de *plusieurs* actes de langage qui apportent différentes preuves à différents niveaux. La gestion de cet ensemble de preuves est relativement difficile puisqu'il faut successivement effectuer des intégrations en retrouvant à quels énoncés se rapportent les preuves. Dans le modèle des échanges, la prise en compte d'une unique preuve facilite l'intégration car il suffit de considérer l'énoncé précédent, ou un énoncé à un niveau supérieur par exemple en testant si l'énoncé peut être candidat pour la réinterprétation. Gérer plusieurs preuves de compréhension requiert de faire un appariement entre l'ensemble des preuves et les énoncés précédents. Il est peut-être possible de faire des hypothèses afin de faciliter cet appariement, par exemple que les preuves sont ordonnées et que les premières preuves visent à ancrer les énoncés plus récents. La question mériterait d'être étudiée toutefois plus en profondeur.

Le modèle des échanges n'est néanmoins pas très favorable, en tout cas dans sa formulation actuelle, à la gestion des énoncés multi-contributifs. Au moins deux aspects du modèle devraient être remis en question. Le premier est que nous nous sommes beaucoup reposés sur l'*alternance* des locuteurs pour intégrer les contributions. Cette alternance entraîne une restriction des situations possibles en termes de structure. Par exemple, on peut faire l'hypothèse que lorsqu'on reçoit un énoncé d'autrui, la contribution précédente a été initiée par soi-même. L'alternance facilite beaucoup la définition des règles, mais en contrepartie empêche de prendre en compte les tours de parole où les locuteurs effectuent plusieurs contributions à plusieurs niveaux. Par exemple il devrait être possible de pouvoir clôturer un échange incident puis d'initier une contribution au niveau supérieur dans le même tour de parole.

Le second aspect est relatif au rôle très limité des énoncés : soit ils présentent une tâche en initiant un nouvel échange, soit ils exécutent une tâche en en clôturant un. Cette limitation a plusieurs conséquences. La plus importante est qu'elle empêche la gestion des énoncés purement évaluatifs ou des *acknowledgements*. Nous avons proposé à ce sujet de rajouter une troisième contribution à l'échange (à l'instar du modèle genevois). Cependant, les *acknowledgements* peuvent se produire à n'importe quel moment, et les deux tours de parole peuvent également se recouvrir. La seconde conséquence est l'impossibilité de considérer les assertions qui viennent modifier un but ou rajouter des informations, c'est-à-dire les énoncés qui ont un autre rôle que celui d'initier ou d'exécuter une tâche.

Tous ces problèmes sont cependant résolus dans le modèle des *grounding acts*, beaucoup plus souple vis à vis de l'intégration des énoncés. L'alternance des locuteurs n'est pas requise dans ce modèle et les énoncés peuvent jouer de multiples rôles (y compris des auto-corrections). Mais ce modèle ne facilite pas la prise en compte des effets relatifs à l'ancrage d'une *discourse unit* vis à vis d'autres *discourse units*, par exemple à l'issue d'une réinterprétation. Nous estimons toutefois que le modèle des échanges devrait être rapproché du modèle des *grounding acts*. Ce rapprochement pourrait être effectué en considérant une unité dialogique à mi-chemin entre l'échange et la *discourse unit*. On pourrait par exemple modéliser des échanges possédant certaines propriétés des *discourse units* comme le fait de pouvoir ajouter du contenu des deux locuteurs tout en conservant une certaine structure interne, par exemple comme séquence de contributions, ainsi qu'une structure de dialogue externe entre les différentes *discourse units*.

13.1.5 Au niveau de la tâche

Il est nécessaire d'autre part de considérer l'ancrage dans une application réelle. Nous avons modélisé le processus d'ancrage d'une interprétation du dialogue finalisé (avec une certaine grammaire, et une certaine ontologie) mais nous n'avons pas considéré en quoi la finalité elle-même influait sur le processus. De nombreux paramètres relatifs à l'ancrage dépendent en effet de la tâche. Tout d'abord la nécessité de l'ancrage, autrement dit la quantité de preuves à produire dépend de la tâche.

Par exemple, les informations indispensables pour poursuivre la tâche doivent être ancrées avec une priorité plus importante que les informations secondaires. Leur caractère critique nécessite en effet une meilleure prise en compte de leur ancrage. A ce sujet on peut se poser la question du partage de ce caractère critique : est-ce que la nécessité d'ancrage est toujours partagée ? Ce n'est probablement pas le cas lorsque les participants ont une vision différente de la tâche, mais que faire alors pour établir une nécessité d'ancrage partagée ?

De plus, une situation problématique dans une tâche peut ne pas s'avérer problématique dans une autre. Par exemple, dans la suite d'énoncés « l'hôtel Ibis et l'hôtel Lafayette » et « à combien est la chambre ? », on peut ne pas considérer l'ambiguïté domaniale comme un problème en raison du fait que l'utilisateur désire probablement connaître le prix dans les deux chambres. La compréhension suffisante d'un énoncé est en effet relative aux buts courants des participants (Clark and Schaefer 1989), mais nous avons considéré par hypothèse que les participants cherchaient toujours à converger. La décision de l'abandon de l'ancrage dépend également de la tâche. Pour rendre compte de ces phénomènes on peut adapter la construction des jugements de compréhension au contexte applicatif. En modulant la construction des jugements, on peut prendre en compte plus facilement la nécessité de soulever un problème en fonction de la tâche.

13.2 Extensions du terrain commun

13.2.1 Niveau de la référence

Nous nous sommes limités, au niveau de la référence, à l'identification directe d'un référent au moyen de questions d'identification et de paraphrases. Cependant, l'ancrage de la référence nécessite la prise en compte de plusieurs autres aspects ou niveaux.

D'abord, nous n'avons pas considéré l'ancrage du domaine de référence au sein duquel extraire le référent, par exemple nous n'avons pas modélisé l'ambiguïté domaniale. Prendre en compte l'ancrage du domaine d'extraction permettrait de soulever des questions plus fines. Typiquement pour l'anaphore associative, la relation entre le domaine antécédent et le référent pourrait faire l'objet de questions. Ou encore dans le cas de la référence multimodale, on pourrait résoudre les problèmes de localisation en focalisant des sous-espaces (par exemple « où ça ? » comme question à « prends le stylo »). Ce point de vue correspond à dire que la convergence de la compréhension d'une expression référentielle est un processus de *restriction* du domaine dans lequel effectuer l'extraction. La restriction que nous avons effectuée est très « directe » puisqu'elle consiste à réviser la relation entre l'expression référentielle et son référent.

La prise en compte de l'ancrage d'un référent dans un domaine donné, permettrait non seulement l'identification mais également la description conceptuelle des référents. Dans l'exemple suivant, la question ne vise pas tant à identifier James qu'à le décrire.

A : j'ai vu James
B : qui est James ?
A : c'est un ami

Nous estimons possible de modéliser et d'implémenter cette description en restreignant James, initialement comme sous-domaine des Personnes au sous-domaine des Amis de A. Cette modélisation serait facilitée par les choix en logique de description que nous avons effectués mais il reste à définir clairement comment les questions ou les réponses peuvent restreindre les domaines de référence.

13.2.2 Autres niveaux

Ensuite, si on peut concevoir dans cette théorie comment aborder le problème de conceptualisation par restriction de domaines, il est plus difficile d'estimer comment étendre le terrain commun à l'expression référentielle, par exemple au niveau lexical, syntaxique, voire intentionnel. Cette instanciation serait en effet difficile dans notre implémentation à cause de la grande hétérogénéité des représentations et algorithmes utilisés à chaque niveau. Etendre le processus d'ancrage à d'autres niveaux requiert en effet de satisfaire certaines contraintes sur l'interprétation. D'abord, il doit être possible de constituer des jugements propres ou d'estimer les jugements manifestés à plusieurs niveaux. Ensuite, il est nécessaire de considérer comment concevoir un algorithme générique de réinterprétation. Mais ces deux directions requièrent une grande homogénéité dans les représentations manipulées par le processus interprétatif.

A notre avis, la première nécessité pour concevoir un ancrage à plusieurs niveaux est de définir un modèle d'interprétation homogène et non-monotone. Le modèle doit en effet satisfaire les *besoins de réinterprétation* inhérents au processus d'ancrage. Afin d'étendre le modèle à d'autres niveaux, il est nécessaire de rechercher ce qui est commun à plusieurs aspects de l'interprétation. On pourrait par exemple considérer le processus interprétatif comme une succession de *choix* parmi des alternatives. Une représentation explicite de ces choix serait relativement difficile à généraliser mais aurait de nombreux avantages. Elle faciliterait en particulier la réinterprétation en tant que remise en cause d'un choix d'interprétation effectué à un certain niveau : d'abord par la recherche du choix problématique, par le ré-examen de ce choix et par la prise en compte des effets du nouveau choix. On pourrait de plus représenter le fait que la cible ou le niveau de la réinterprétation est elle-même un choix qui peut soulever des problèmes comme nous l'avons montré.

La représentation des problèmes se trouverait également facilitée puisqu'on pourrait considérer de manière homogène les *jugements de compréhension*. Les jugements propres peuvent être définis à partir des difficultés à effectuer un *choix* d'interprétation à un moment donné, l'ambiguïté correspond par exemple à l'impossibilité de faire un choix dans une alternative, et le vide à l'absence de cette alternative. Les jugements manifestés pourraient également être traduits sous cette forme, la mauvaise compréhension ne revèle en fait qu'un choix divergent de l'interlocuteur à un

certain niveau d'interprétation. Traduire l'interprétation comme une succession de choix peut permettre de mieux cerner les problèmes qui peuvent faire sens à l'utilisateur. En particulier il est très fréquent que le système soit confronté à des choix inexistants pour l'utilisateur. Cela n'est pas tant dû à l'état dans lequel qu'il se trouve qu'à la modélisation imparfaite des choix d'interprétation. Selon nous la modélisation de la compréhension doit tendre à l'adéquation entre les types de choix du système et ceux d'un être humain, afin que les problèmes d'interprétation rencontrés puissent être analogues.

13.3 Utilisation du terrain commun

Nous n'avons cependant que trop peu abordé l'*utilisation* du terrain commun en ce qu'il a de commun pour les choix d'interprétation ou de production. Une piste très intéressante à développer est celle de l'ancrage et de la réutilisation de conceptualisations partagées des objets. Par exemple lorsque l'utilisateur emploie un certain concept pour référer à une entité, ce concept peut être réutilisé avec assurance à la condition qu'il soit effectivement ancré. La prise en compte de conceptualisations partagées permettrait de dépasser les limites de la constitution du terrain commun montrées par l'évaluation. Plusieurs difficultés existent néanmoins.

La première est l'ancrage même de la conceptualisation. Si cela ne semble pas trop difficile à réaliser dans un modèle jouet, les problèmes d'établissement d'une conceptualisation partagée nécessitent probablement de pouvoir dialoguer à propos des mots ou des concepts. Nous proposons dans cette direction de creuser la piste brièvement introduite de la métaréférence. Pour la gestion de la métaréférence nous avons ajouté dans l'ontologie le concept d'expression référentielle, mais on pourrait étendre l'approche en introduisant le concept de mot ou le concept de concept. Bien que certainement difficile, cette approche nous semble prometteuse afin de faciliter la gestion d'un terrain commun plus important. En particulier elle vise à dialoguer à propos, non plus de l'interprétation mais des ressources utilisées pour effectuer cette interprétation.

La seconde difficulté est relative à la modification dynamique des ressources. Nous disposons du modèle qui l'autoriserait, par exemple on pourrait effectuer l'apprentissage à l'issue de la clôture de l'incidence, comme dans Luzzati (1995) et Lehuen (1997b). Toutefois, en raison de la grande hétérogénéité des ressources (lexique morphologique, lexique syntaxique, grammaire, lexique sémantique, ontologies, toutes écrites dans des formats différents) il est relativement difficile de le faire. Une représentation unifiée du traitement linguistique entraînerait une unification des ressources utilisées et c'est selon nous le pré-requis indispensable pour pouvoir modifier dynamiquement différents type de ressources.

La troisième difficulté concerne l'importance de la génération dans le système de dialogue. Le choix des paraphrases n'a été effectué qu'avec le point de vue de celui qui les produit, et nous avons pu constater les problèmes que cela soulevait lors de l'évaluation. Une modélisation explicite du moindre effort collaboratif pourrait permettre de favoriser le choix des concepts déjà ancrés plutôt que des concepts

propres. On pourrait par exemple associer aux concepts une mesure indiquant la probabilité d'être ancré. Le moindre effort collaboratif entraînant alors la préférence pour les concepts possédant une plus grande probabilité d'ancrage dans la production des paraphrases.

13.4 Evaluation du processus d'ancrage

L'évaluation à laquelle nous avons procédé couvre la capacité à atteindre une interprétation de référence par le dialogue. Cependant sa mise en oeuvre a été très limitée. Tout d'abord nous n'avons évalué qu'un seul couple de stratégies Test et Contrôle. Néanmoins le paradigme permet théoriquement d'évaluer de nombreuses combinaisons : on peut par exemple autoriser le Contrôle à poser des questions de clarification ou le Test à manifester la mauvaise compréhension du Contrôle. Afin de vérifier l'impact de ces stratégies, il nous suffit de relâcher les restrictions de jugements que nous avons imposées. Ce faisant, on pourrait se rapprocher d'une situation réelle dans laquelle les deux participants peuvent manifester leur non-compréhension ou la mauvaise compréhension de leur partenaire.

Nous n'avons ensuite procédé qu'à l'évaluation du *gain* de compréhension et pas à la pertinence du *moyen* pour atteindre ce gain. Pour affiner l'évaluation il serait nécessaire de prendre en compte une mesure du respect du moindre effort collaboratif. Par exemple, on pourrait mesurer l'efficacité du processus d'ancrage en fonction du coût de traitement des preuves : moins les preuves sont coûteuses à efficacité constante et plus le gain de compréhension peut recevoir une pondération importante.

Enfin nous pouvons également effectuer une évaluation à l'aide d'utilisateurs (à la condition de situer le processus d'ancrage dans une tâche). Celle-ci pourrait en particulier permettre de mesurer la satisfaction des utilisateurs vis à vis des différentes preuves produites. Il est probable que les utilisateurs préfèrent des preuves explicites de compréhension mais dans quelle mesure ? A partir de quel moment doit-on considérer que trop ou pas assez de preuves ont été fournies ?

13.5 Synthèse

Nous avons cherché à clarifier dans cette thèse la modélisation du processus d'ancrage en vue d'améliorer la compréhension des systèmes de dialogue. Cette recherche a permis de dégager l'importance du concept de jugement de compréhension dans l'établissement d'un terrain commun conversationnel. La constitution de ces jugements s'appuie sur une estimation de sa propre compréhension ainsi que sur une estimation des preuves de compréhension manifestées par son partenaire. Ils permettent alors à un système de se représenter la divergence interprétative et de guider le dialogue vers la convergence. Toutefois, la convergence est un problème difficile que nous avons cherché à évaluer sur un large corpus. Bien que l'évaluation montre un gain de compréhension à l'issue du processus d'ancrage, elle soulève plus

de questions qu'elle n'en apporte. En particulier elle montre l'insuffisance d'une gestion *explicite* du terrain commun à l'aide de requêtes de clarification. Les requêtes de clarification représentent certes un aspect de la gestion du terrain commun mais les participants peuvent également s'adapter l'un à l'autre de manière *implicite*, par exemple en alignant leurs représentations au niveau conceptuel. La théorie de l'alignement de Pickering and Garrod (2004) constitue, à notre avis, une bonne piste de recherche dans laquelle les jugements de compréhension pourraient avoir leur place. Ils pourraient peut-être permettre d'unifier une gestion explicite et une gestion implicite du terrain commun, mais de plus amples travaux seraient nécessaires afin de le déterminer.

Corpora

Le corpus MEDIA qui a servi de base à l'évaluation a été développé par l'agence pour la distribution des ressources linguistiques et l'évaluation (ELDA - Evaluations and Language resources Distribution Agency) conjointement avec le consortium MEDIA. Il comporte 1250 dialogues oraux, constitués par la méthode du magicien d'Oz sur une tâche de réservation d'hôtel, enregistrés, transcrits et annotés hors et en contexte (principalement sémantique et référence). Le corpus et les outils d'évaluation sont disponibles auprès de ELDA (référence E0024).

Le corpus « miroir » correspond à l'application du protocole miroir sur le sous-ensemble du corpus MEDIA utilisé pour l'évaluation en contexte. Il comporte 174 dialogues et n'est pas actuellement distribué.

Bibliographie

Abney, S. (1991). Parsing by chunks. In S. A. Robert Berwick and C. Tenny (Eds.), *Principle-based parsing*, pp. 257–278. Dordrecht : Kluwer Academic Publishers.

Allemandou, J. (2007). *SIMDIAL, un paradigme d'évaluation automatique de systèmes de dialogue homme-machine par simulation déterministe d'utilisateurs*. Thèse de doctorat, Université Paris XI, Orsay.

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

Allen, J., D. Byron, M. Dzikovska, G. Ferguson, and L. Galescu (2001). Towards conversational human-computer interaction. *AI Magazine 22*(4), 27 – 37.

Allen, J., D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent (2000). An architecture for a generic dialogue shell. *NLENG : Natural Language Engineering 6*, 213–228.

Allen, J. and M. Core (1997). Draft of DAMSL : Dialog act markup in several layers. Rapport technique, University of Rochester.

Allen, J. and C. Perrault (1980). Analyzing intention in utterances. *Artificial Intelligence 15*, 143–178.

Allen, J. F., B. W. Miller, E. K. Ringger, and T. Sikorski (1996). Robust understanding in a dialogue system. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pp. 62–70. Morgan Kaufmann.

Allwood, J. (1995). An activity based approach to pragmatics. In B. Bunt, H. & Black (Ed.), *Abduction, Belief and Context in Dialogue : Studies in Computational Pragmatics*, pp. 47–80. Amsterdam : John Benjamins.

Allwood, J. and J. Abelar (1984). Lack of understanding, misunderstanding and language acquisition. In E. . Mittner (Ed.), *Proceedings of the AILA-Conference*, Brussels.

Anderson, M. L. and B. Lee (2005). Metalanguage for dialog management. In *Proceedings of the 16th Annual Winter Conference on Discourse, Text and Cognition.*

Anderson, M. L. and D. R. Perlis (2005). Logic, self-awareness and self-improvement : the metacognitive loop and the problem of brittleness. *Journal of Logic and Computation 15(1)*, 21–40.

Antoine, J.-Y. and J. Caelen (1999, Juin). Pour une évaluation objective, prédictive et générique de la compréhension en CHM orale : le paradigme DCR (demande, contrôle, résultat). *Langues 2*(2), 130–139.

Antoine, J.-Y., J. Siroux, J. Caelen, J. Villaneau, J. Goulian, and M. Ahafhaf (2000). Obtaining predictive results with an objective evaluation of spoken dialogue systems : experi-

ments with the dcr assessment paradigm. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Athens, pp. 403–409. ELRA.

Appelt, D. E. (1987). Bidirectional grammars and the design of natural language generation systems. In *Proceedings of the 1987 workshop on Theoretical issues in natural language processing*, Las Cruces, New Mexico, USA, pp. 206–212.

Araki, M., T. Watanabe, and S. Doshita (1997). Evaluating dialogue strategies for recovering from misunderstandings. In *Proceedings of the IJCAI Workshop on Collaboration Cooperation and Conflict in Dialogue Systems*, pp. 13–18.

Areces, C., A. Koller, and K. Striegnitz (2008). Referring expressions as formulas of description logic. In *Proceedings of the 5th International Natural Language Generation*, Salt Fork, OH, USA.

Asher, N. and A. Gillies (2003). Common ground, corrections and coordination. *Argumentation 17*(4), 481–512.

Austin, J. L. (1962). *How to do things with Words : The William James Lectures delivered at Harvard University in 1955*. Oxford : Clarendon.

Barr, D. J. (2004). Establishing conventional communication systems : Is common knowledge necessary ? *Cognitive Science 28*(6), 937–962.

Barwise, J. (1988). Three views of common knowledge. In M. Yardi (Ed.), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, pp. 365–379. San Francisco : Morgan Kaufman.

Bilange, E. (1992). *Dialogue personne-machine*. Hermès, Paris.

Bilange, E. and J.-Y. Magadur (1992). A robust approach for handling oral dialogues. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France, pp. 799–805.

Bonneau-Maynard, H., A. Denis, F. Béchet, L. Devillers, F. Lefèvre, M. Quignard, S. Rosset, and J. Villaneau (2008). Media : évaluation de la compréhension dans les systèmes de dialogue. In S. Chaudiron and K. Choukri (Eds.), *L'évaluation des technologies de traitement de la langue, les campagnes Technolangue*, Chapter 9, pp. 209–232. Hermès, Lavoisier.

Bonneau-Maynard, H. and F. Lefèvre (2005). A 2+1-level stochastic understanding model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding workshop (ASRU)*, pp. 256– 261.

Bonneau-Maynard, H., S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa (2005, September). Semantic annotation of the French MEDIA dialog corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Lisbon, Portugal, pp. 3457–3460.

Boufaden, N., T. L. Hoang, and P. Dumouchel (2007, juin). Détection et prédiction de la satisfaction des usagers dans les dialogues personne-machine. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse.

Brennan, S. and E. Hulteen (1995). Interaction and feedback in a spoken language system : a theoretical framework. *Knowledge-Based Systems 8*, 143–151.

Brenner, M. (1999). Etablissement de croyances mutuelles dans le dialogue homme-machine. Rapport de maîtrise, Univ. Paris XI, Orsay, France.

Bunt, H., R. Morante, and S. Keizer (2007). An empirically based computational model of grounding in dialogue. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, Anvers, Belgique, pp. 283–290.

Burnard, L. (2000). Reference guide for the british national corpus (world edition). Oxford University Computing Services.

Byron, D. and P. Heeman (1997). Discourse marker use in task-oriented spoken dialog. In *Proceedings of Eurospeech'97*, Rhodes, Greece, pp. 2223–2226.

Caelen, J. (2003, mai). Stratégies de dialogue. In C. éd (Ed.), *Actes des Secondes Journées Francophones des Modèles Formels de l'Interaction (MFI'03)*, pp. 29–39.

Caelen, J. and H. Nguyen (2006). Traitement des incompréhensions et des malentendus en dialogue homme-machine. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pp. 445–454.

Cahn, J. (1992). A computational architecture for mutual understanding in dialog. Rapport technique, Music and Cognition Group, M.I.T. Media Laboratory.

Cahn, J. and S. Brennan (1999). A psychological model of grounding and repair in dialog. In *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, North Falmouth, Massachusetts, USA, pp. 25–33.

Chaudiron, S. and K. Choukri (2008). L'évaluation : fondements, processus et résultats. In S. Chaudiron and K. Choukri (Eds.), *L'évaluation des technologies de traitement de la langue, les campagnes Technolangue*, Chapter 9, pp. 23–46. Hermès, Lavoisier.

Cherubini, M., J. V. der Pol, and P. Dillenbourg (2005). Grounding is not shared understanding : distinguishing grounding at an utterance and knowledge level. In *Proceedings of the Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'05)*, Paris, France.

Chinchor, N. and L. Hirschmann (1997). MUC-7 coreference task definition, version 3.0. In *Actes de MUC-7*, Philadelphia, PA, USA.

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

Clark, H. (1997). Dogmas of understanding. *Discourse Processes 23*, 567–598.

Clark, H. and S. Brennan (1991). Grounding in communication. In L. J. L. Resnick and S. Teasley (Eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. APA Books, Washington.

Clark, H. and E. Schaefer (1989). Contributing to discourse. *Cognitive Science 13*, 259–294.

Clark, H. H. and C. R. Marshall (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, and I. Sags (Eds.), *Elements of Discourse Understanding*, pp. 10–63. Cambridge : Cambridge University Press.

Clark, H. H. and D. Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition 22*, 1–39.

Cohen, P. R. and C. R. Perrault (1979). Elements of a plan-based theory of speech acts. *Cognitive Science 3*, 177–212.

Corblin, F. (1987). *Indéfini, Défini et Démonstratif*. Genève : Droz.

Crabbé, B., B. Gaiffe, and A. Roussalany (2003). Une plateforme de conception et d'exploitation de grammaire d'arbres adjoints lexicalisés. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Batz-sur-mer, pp. 75–84.

Danieli, M. (1996). On the use of expectations for detecting and repairing human-machine miscommunication. In *Proceedings of the AAAI-96 Workshop on Detecting, Preventing, and Repairing Human-Machine Miscommunications*, Portland, Oregon, USA, pp. 87–93.

Denis, A., F. Béchet, and M. Quignard (2007). Résolution de la référence dans des dialogues homme-machine : évaluation sur corpus de deux approches symbolique et probabiliste. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, pp. 261–270.

Denis, A., G. Pitel, and M. Quignard (2006a). A model of grouping for plural and ordinal references. In *Proceedings of ESSLLI Workshop on Ambiguity in Anaphora*, Malaga, Spain, pp. 31–39.

Denis, A., G. Pitel, and M. Quignard (2006b). Resolution of referents groupings in practical dialogues. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia.

Denis, A., G. Pitel, M. Quignard, and P. Blackburn (2007). Incorporating asymmetric and asychronous evidence of understanding in a grounding model. In *Proceedings of the 2007 Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Trento, Italy, pp. 33–40.

Denis, A., M. Quignard, and G. Pitel (2006). A deep-parsing approach to natural language understanding in dialogue system : Results of a corpus-based evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, pp. 339–344.

DeVault, D., I. Oved, and M. Stone (2006, July). Societal grounding is essential to meaningful language use. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, Massachusetts.

DeVault, D. and M. Stone (2007, May). Managing ambiguities across utterances in dialogue. In *Proceedings of the 2007 Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, University of Trento, Italy.

Devillers, L., H. Bonneau-Maynard, and P. Paroubek (2002). Méthodologies d'évaluation des systèmes de dialogue parlé : réflexion et expériences autour de la compréhension. *Traitement Automatique des Langues 43*(2), 155–184.

Devillers, L., H. Maynard, P. Paroubek, and S. Rosset (2003, April). The peace slds understanding evaluation paradigm of the french media campaign. In *Proceedings of the workshop "Evaluation Initiatives in Natural Language Processing, are evaluation methods, metrics, and resources reusable ?", at the 10th Conference of The European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Hungary, pp. 11–18.

Duff, D., B. Gates, and S. LuperFoy (1996). An architecture for spoken dialogue management. In *Proceedings of ICSLP-1996*, pp. 1025–1028.

Erickson, T. D. and M. E. Mattson (1981, Oct). From words to meaning : A semantic illusion. *Journal of Verbal Learning & Verbal Behavior 20*(5), 540–551.

Fikes, R. E. and N. J. Nilsson (1971). Strips : A new approach to the application of theorem proving. *Artificial Intelligence 2*, 189–208.

Fillmore, C. J., C. R. Johnson, and M. R. L. Petruck (2003). Background to framenet. *International Journal of Lexicography 16*(3), 235–250.

Franconi, E. (1993, November). A treatment of plurals and plural quantifications based on a theory of collections. *Minds and Machines 3*(4), 453–474.

Franconi, E. (1994). Description logics for natural language processing. In *Proceedings of the 1994 AAAI Fall Symposium on Knowledge Representation for Natural Language Processing in Implemented Systems*, New Orleans, US, pp. 37–44.

Friederici, A. D. and S. Kotz (2003). The brain basis of syntactic processes : functional imaging and lesion studies. *NeuroImage 20*, 8–17.

Friederici, A. D. and J. Weissenborn (2007). Mapping sentence form onto meaning : The syntax-semantic interface. *Brain Research 1146*, 50–58.

Funakoshi, K. and T. Tokunaga (2006). Identifying repair targets in action control dialogue. In *Proceedings of the 11th Conference of European chapter of the Association for Computational Linguistics (EACL2006)*, pp. 177–184.

Gabsdil, M. (2003). Clarification in spoken dialogue systems. In *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*, Stanford, CA, USA, pp. 28–35.

Gabsdil, M. and O. Lemon (2004). Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, pp. 343–350.

Gaiffe, B., F. Landragin, and M. Quignard (2004, mars). 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.

Gaiffe, B. and A. Reboul (1999). Référence et représentations mentales. Cargèse, France. Tutorial, TALN99.

Gardent, C. and E. Kow (2007). Geni, un réalisateur basé sur une grammaire réversible. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

Gardent, C. and K. Striegnitz (2007). Generating bridging definite descriptions. In H. Bunt and R. Muskens (Eds.), *Studies in Linguistics and Philosophy, Computing Meaning*, Volume 3, pp. 369–396. Springer.

Garrod, S. and M. J. Pickering (2004). Why is conversation so easy ? *Trends in Cognitive Sciences 8*, 8–11.

Gaudou, B., A. Herzig, and D. Longin (2006a). Grounding and the expression of belief. In P. Doherty, J. Mylopoulos, and C. A. Welty (Eds.), *Proceedings of the 10th International Conference on Principles on Principles of Knowledge Representation and Reasoning (KR 2006)*, Windermere, UK, pp. 211–229. AAAI Press.

Gaudou, B., A. Herzig, and D. Longin (2006b, may). A logical framework for grounding-based dialogue analysis. In W. van der Hoek, A. Lomuscio, E. de Vink, and M. Wooldridge (Eds.), *Proceedings of the Third International Workshop on Logic and Communication in Multi-Agent Systems (LCMAS 2005)*, Volume 157 of *Electronic Notes in Theoretical Computer Science (ENTCS)*, pp. 117–137. Edinburgh, Scotland, UK : Elsevier.

Gilovich, T. (1990). Differential construal and the false consensus effect. *Journal of Personality and Social Psychology 59*(4), 623–634.

Ginzburg, J. & Sag, I. A. (2000). Interrogative investigations : the form, meaning and use of english interrogatives. In *No. 123 in CSLI Lecture Notes*. Stanford, California, USA : CSLI Publications.

Ginzburg, J. (2007). *Semantics and Interaction in Dialogue*, Chapter 5. CSLI Publications and University of Chicago Press.

Ginzburg, J. and R. Cooper (2001). Resolving ellipsis in clarification. In *Meeting of the Association for Computational Linguistics*, pp. 236–243.

Ginzburg, J., R. Fernandez, and D. Schlangen (2007). Unifying self- and other- repair. In *The 2007 Workshop on the Semantics and Pragmatics of Dialogue (DECALOG 2007)*, Rovereto, pp. 57–63.

Goodman, B. A. (1985). Repairing reference identification failures by relaxation. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 204–217. Association for Computational Linguistics.

Grice, P. (1957). Meaning. *The Philosophical Review 66*, 377–88.

Grice, P. (1975). *Logic and Conversation*, pp. 64–75. The Logic of Grammar.

Groenendijk, J. (1999). The logic of interrogation : Classical version. In T. Matthews and D. Strolovitch (Eds.), *SALT IX : Semantics and linguistic theory*, pp. 109–126. Ithaca : Cornell University Press.

Grosz, B. J. and C. L. Sidner (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics 12*(3), 175–204.

Haarslev, V. and R. Möller (2003, October). Racer : A core inference engine for the semantic web. In *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools (EON2003) , located at the 2nd International Semantic Web Conference ISWC 2003*, Sanibel Island, Florida, USA, pp. 27–36.

Halpern, J. and Y. Moses (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM 37*(3), 549–587.

Hayes, P. J. and R. Reddy (1983). Steps toward graceful interaction in spoken and written man-machine communication. *International journal of man-machine studies 19*(3), 231–284.

Heeman, P. and G. Hirst (1995). Collaborating on referring expressions. *Computational Linguistics 21*(3), 351–382.

Heeman, P. A., M. Johnston, J. Denney, and E. Kaiser (1998). Beyond structured dialogues : Factoring out grounding. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia, pp. 863–866.

Hirst, G., S. McRoy, P. Heeman, P. Edmonds, and D. Horton (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication 15*, 213–230.

Horton, W. and B. Keysar (1996). When do speakers take into account common ground ? *Cognition 59*, 91–117.

Jakobson, R. (1963). *Linguistique et poétique« Linguistique et poétique », Essais de linguistique générale*. Paris, Minuit.

Johnson, T. (2007). You said what ? ! : Misunderstandings in im conversation among college students. Rapport de maîtrise, Swarthmore College, Pennsylvanie, Etats Unis.

Jokinen, K. (1996, July). Evaluating robustness in dialogue systems. Dialeague96 Summer session workshop, Tokyo, Japan.

Joshi, A. and Y. Schabes (1997). Tree-adjoining grammars. In G. Rozenberg and A. Salomaa (Eds.), *Handbook of Formal Languages*, Volume 3, pp. 69 – 124. Springer, Berlin, New York.

Kamp, H. and U. Reyle (1993). *From discourse to logic*. Kluwer Academic Publisher.

Keysar, B., D. J. Barr, and W. S. Horton (1998). The egocentric basis of language use : Insights from a processing approach. *Current Directions in Psychological Sciences 7*, 46–50.

Kingsbury, D. (1968). Manipulating the amount of information obtained from a person giving directions. Unpublished Honors Thesis, Department of Social Relations, Harvard University.

Koschman, T. and C. LeBaron (2003). Reconsidering common ground : Examining clark's contribution theory in the or. In P. F. P. D. . K. S. K. Kuutti, G. Karsten (Ed.), *ECSCW 2003 : Proceedings of the Eighth European Conference on Computer-Supported Cooperative Work*, pp. 81–98. Amsterdam : Kluwer Academic Publishing.

Kow, E. (2006). Geni : natural language generation in haskell. In *Proceedings of the 2006 ACM SIGPLAN workshop on Haskell*, Portland, Oregon, USA, pp. 110–119.

Kow, E. (2007). *Réalisation de surface : ambiguïté et déterminisme*. Thèse de doctorat, Université Henri Poincaré, Nancy, France.

Kumar, A., S. Salmon-Alt, and L. Romary (2003). Reference resolution as a facilitating process towards robust multimodal dialogue management : A cognitive grammar approach. In *International Symposium on Reference Resolution and Its Application to Question Answering and Summarization*, Venice, Italy.

Lamel, L., S. Rosset, J. Gauvain, S. Bennacef, M. Garnier-Rizet, and B. Prouts (2000, August). The limsi arise system. *Speech Communication 4*(31), 339–354.

Landragin, F. (2003). *Modélisation de la communication multimodale. Vers une formalisation de la pertinence*. Thèse de doctorat, Université Henri Poincaré, Nancy.

Landragin, F., A. Denis, A. Ricci, and L. Romary (2004). Multimodal meaning representation for generic dialogue systems architectures. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 521–524.

Landragin, F. and L. Romary (2004). Dialogue history modelling for multimodal human-computer interaction. In *Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue (Catalog'04)*, Barcelona, Spain, pp. 41–48.

Landragin, F., S. Salmon-Alt, and L. Romary (2002). Ancrage référentiel en situation de dialogue. *Traitement Automatique des Langues 43*(2), 99–129.

Larsson, S. (2002). *Issue-based Dialogue Management*. Thèse de doctorat, Goteborg University, Sweden.

Larsson, S. (2003). Interactive communication management in an issue-based dialogue system. In Kruijff-Korbayova and Kosny (Eds.), *Proceedings of DiaBruck, 7th Workshop on the Semantics and Pragmatics of Dialogue*, Universität des Saarlandes, pp. 75–83.

Larsson, S. and D. Traum (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering 6*, 323–340.

Lehuen, J. (1997a). Complexité et usage de la langue naturelle en dialogue homme/machine. In *Actes du congrès IAC'97*, Paris.

Lehuen, J. (1997b). *Un modèle de dialogue dynamique et générique intégrant l'acquisition de sa compétence linguistique - Le système Coala*. Thèse de doctorat, Université du Maine, Le Mans, France.

Lewis, D. (1969). *Convention : A Philosophical Study*. Harvard University Press.

Litman, D. (1985). *Plan Recognition and Discourse Analysis : an Integrated Approach for Understanding Dialogues*. Thèse de doctorat, Université de Rochester.

Litman, D. and J. Allen (1984). A plan recognition model for clarification subdialogues. In *Proceedings of Coling'84*, pp. 302–311.

Lopez, P. (1999). *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*. Thèse de doctorat, Université Henri Poincaré, Nancy, France.

Luzzati, D. (1995). De l'erreur en DHM. *Cahiers de Linguistique Française 16*, 175–192.

Matheson, C., M. Poesio, and D. Traum (2000, May). Modelling grounding and discourse obligations using update rules. In *Proceedings of the 1st Annual Meeting of the North American Association for Computational Linguistics (NAACL2000)*, Seattle, Washington, USA, pp. 1 – 8.

McRoy, S. and G. Hirst (1993). Abductive explanations of dialogue misunderstanding. In *6th Conference of the European Chapter of the Association for Computational Linguistic*, Utrecht University, Utrecht, The Netherlands, pp. 277–286.

McRoy, S. W. and G. Hirst (1995). The repair of speech act misunderstandings by abductive inference. *Computational Linguistics 21*(4), 435–478.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Paris : Éditions Gallimard, collection « Tel ».

Meurs, M.-J., F. Duvert, F. Béchet, F. Lefèvre, and R. de Mori (2008, juin). Annotation en frames sémantiques du corpus de dialogue media. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Avignon.

Moeschler, J. (1989). *Modélisation du dialogue*. Hermès, Paris.

Nakano, Y., G. Reinstein, T. Stocky, and J. Cassell (2003). Towards a model of face-to-face grounding. In *Proceedings of Association for Computational Linguistics*, Sapporo, Japan, pp. 553–561.

Nerzic, P. (1993). *Erreurs et échecs dans le dialogue oral homme-machine, détection et réparation*. Thèse de doctorat, Université de Rennes I.

Neumann, G. (1994). *A Uniform Computational Model for Natural Language Parsing and Generation*. Thèse de doctorat, Université de la Saare, Saarbruck.

O'Brien, C. (2002). Grounding strategies in dialogue systems. Unpublished.

Olson, D. R. (1970). Language and thought : Aspects of a cognitive theory of semantics. *Psychological Review 77*(4), 257–273.

Paek, T. and E. Horvitz (2000). Grounding criterion : Toward a formal theory of grounding. Rapport technique, Microsoft Research.

Pallett, D. S., J. G. Fiscus, W. M. Fisher, J. Garofolo, B. A. Lund, A. Martin, and M. A. Przybocki (1995). 1994 benchmark tests for the arpa spoken language program. In *ARPA Workshop on Spoken Language Technology*, Austin, USA, pp. 5–36.

Perlis, D. (1997). Sources of, and exploiting, inconsistency : Preliminary report. *Journal of Applied Non-Classical Logics 7*(1).

Pickering, M. J. and S. Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*, 169–226.

Pitel, G. (2003). Vers une approche fonctionnelle de la résolution de la référence dans le dialogue. In *Actes des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL2003)*, Batz-sur-Mer, France, pp. 479–488.

Poesio, M. (1996). Semantic ambiguity and perceived ambiguity. In K. van Deemter and S. Peters (Eds.), *Ambiguity and Underspecification*, pp. 159–201. CSLI Publications.

Poesio, M. and D. R. Traum (1997). Conversational actions and discourse situations. *Computational Intelligence 13*(3), 309–347.

Pollard, C. J. and I. A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago : University of Chicago Press.

Popescu-Belis, A., L. Rigouste, S. Salmon-Alt, and L. Romary (2004). Online evaluation of coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Volume 4, Lisbon, Portugal, pp. 1507–1510.

Power, R. (1979). The organisation of purposeful dialogues. *Linguistics. 17*(1-2), 107–151.

Purver, M. (2004a, July). Clarie : the clarification engine. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, Barcelona, Spain, pp. 77–84.

Purver, M. (2004b). *The Theory and Use of Clarification Requests in Dialogue*. Thèse de doctorat, King's College University of London.

Purver, M. and J. Ginzburg (2004). Clarifying noun phrase semantics. *Journal of Semantics 21*(3), 283–339.

Purver, M., F. Ratiu, and L. Cavedon (2006). Robust interpretation in dialogue by combining confidence scores with contextual features. In *Proceedings of INTERSPEECH : International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, pp. 1–4.

Reboul, A., C. Balkanski, X. Briffault, B. Gaiffe, A. Popescu-Belis, I. Robba, L. Romary, and G. S. G. (1997). Le projet cervical : Représentations mentales, référence aux objets et aux événements. Rapport technique, Loria-CNRS/Limsi, France.

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

Reiter, E. and R. Dale (1997). Building applied natural language generation systems. *Journal of Natural Language Engineering 3*(1), 57–87.

Reithinger, N., D. Fedeler, A. Kumar, C. Lauer, E. Pecourt, and L. Romary (2005). MIAMM - a multimodal dialogue system using haptics. In L. D. Jan C. J. van Kuppevelt and N. O. Bernsen (Eds.), *Advances in Natural Multimodal Dialogue Systems*, Volume 30 of *Text, Speech and Language Technology*, pp. 307–332. Springer Netherlands.

Roque, A. and D. Traum (2008). Degrees of grounding based on evidence of understanding. In *9th SIGdial Workshop on Discourse and Dialogue*, Columbus, Ohio, USA.

Ross, L., D. Greene, and P. House (1977). The 'false consensus effect' : An egocentric bias in social perception and attribution process. *Journal of Experimental Social Psychology 13*, 279–301.

Rouillard, J. (2004). *VoiceXML - Le langage d'accès à Internet par téléphone*. Editions Vuibert.

Roulet, E., A. Auchlin, J. Moeschler, C. Rubattel, and M. Schelling (1985). *L'Articulation du discours en français contemporain*. Berne.

Russell, B. (1905). On denoting. *Mind 14*, 479–493.

Sabah, G. (1991). Le traitement automatique des langues. In G. Vergnaud (Ed.), *Les sciences cognitives en débat*, pp. 107–120. Editions du CNRS.

Sacks, H., E. A. Schegloff, and G. Jefferson (1974). A simplest systematics for the organization of turn-taking in conversation. *Language 50*, 679–735.

Saget, S. and M. Guyomard (2006). Goal-oriented dialog as a collaborative subordinated activity involving collective acceptance. In *Proceedings of the 10th Workshop on the Se-*

mantics and Pragmatics of Dialogue (Brandial 2006), University of Potsdam, Potsdam, Germany, pp. 131–138.

Salmon-Alt, S. (2001). *Référence et dialogue finalisé : de la linguistique à un modèle opérationnel*. Thèse de doctorat, Université Henri Poincaré - Nancy 1.

Salmon-Alt, S. and L. Romary (2004). Towards a reference annotation framework. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 119–122.

Schegloff, E. (1987). Between macro and micro : contexts and other connections. In J. Alexander, B. Giesen, R. Munch, and N. Smelser (Eds.), *The Micro-Macro Link*, pp. 207–234. University of California Press, Berkeley.

Schegloff, E. and H. Sacks (1973). Opening up closings. *Semiotica 8*, 289–327.

Schegloff, E. A. (1992). Repair after next turn : The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology 98*, 1295–1345.

Schegloff, E. A. (1997). Third turn repair. In G. G. et al (Ed.), *Towards a social science of language : Papers in honor of William Labov*, Volume 2 : Social interaction and discourse structures, pp. 31–40. Amsterdam : John Benjamins.

Schegloff, E. A., G. Jefferson, and H. Sacks (1977). The preference for self-correction in the organization of repair in conversation. *Language 53*(2), 361–382.

Schiffer, S. R. (1972). *Meaning*. Oxford University Press.

Schlangen, D. (2004, April 30 - May 1). Causes and strategies for requesting clarification in dialogue. In M. Strube and C. Sidner (Eds.), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, Massachusetts, USA, pp. 136–143. Association for Computational Linguistics.

Searle, J. (1969). *Speech Acts*. Cambridge University Press, New York.

Searle, J. and D. Vanderveken (1985). *Foundations of illocutionary logic*. Cambridge, England : Cambridge University.

Servan, C. and F. Béchet (2006, Avril). Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Leuven.

Shannon, C. and W. Weaver (1949). *The Mathematical theory of Communication*. University of Illinois Press.

Shieber, S. (1988). A uniform architecture for parsing and generation. In A. for Computational Linguistics (Ed.), *the 12th conference on Computational linguistics*, Morristown, NJ, USA, pp. 614–619.

Skantze, G. (2003). Exploring human error handling strategies : implications for spoken dialogue systems. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau-d'Oex-Vaud, Switzerland, pp. 71–76.

Skantze, G. (2005). Exploring human error recovery strategies : implications for spoken dialogue systems. *Speech Communication 45*(3), 325–341.

Skantze, G. (2007). Making grounding decisions : Data-driven estimation of dialogue costs and confidence thresholds. In *SigDial*, Antwerp, Belgium, pp. 206–210.

Sperber, D. and D. Wilson (1989). *La pertinence*. Communication et cognition, Paris, Minuit.

Stalnaker, R. C. (1974). Pragmatic presuppositions. In M. K. Munitz and P. K. Unger (Eds.), *Semantics and Philosophy*, pp. 197–214. New York University Press.

Stalnaker, R. C. (1978). Assertion. In Cole (Ed.), *Pragmatics*. New-York : Academic Press.

Stirling, L., I. Mushin, J. Fletcher, and R. Wales (2000). The nature of common ground units : an empirical analysis using map task dialogues. In *Proceedings of the 4th Workshop on the Semantics and Pragmatics of Dialogue (Gotalog 2004)*, Gothenburg, Sweden, pp. 159–165.

Surcin, S. (1999). *Expression langagière ambiguë et modélisation cognitive symbolique, Un modèle informatique de traitement de la polysémie d'usage*. Thèse de doctorat, Université Paris 8.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research 4*, 341–376.

Traum, D. R. (1994, December). *A Computational Theory of Grounding in Natural Language Conversation*. Thèse de doctorat, Rochester.

Traum, D. R. (1999). Computational models of grounding in collaborative systems. In *Working Papers of the AAAI.*, Menlo Park, California, pp. 124–131. AAAI.

Traum, D. R. and J. F. Allen (1992, October). A speech acts approach to grounding in conversation. In *Proceedings 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pp. 137–140.

Traum, D. R. and J. F. Allen (1994). Discourse obligations in dialogue processing. In J. Pustejovsky (Ed.), *Proceedings of the Thirty-Second Meeting of the Association for Computational Linguistics*, San Francisco, pp. 1–8. Morgan Kaufmann.

Turunen, M. and J. Hakulinen (2001). Agent-based error handling in spoken dialogue systems. In *Proceedings of the Eurospeech 2001*, Aalborg, Denmark, pp. 2189–2192.

Tversky, A. and D. Kahneman (1974, Sep). Judgment under uncertainty : Heuristics and biases. *Science, New Series 185*(4157), 1124–1131.

van Deemter, K. and R. Kibble (2000). On coreferring : Coreference in MUC and related annotation schemes. *Computational Linguistics 26*(4), 629–637.

Villaneau, J., O. Ridoux, and J.-Y. Antoine (2004). Logus : compréhension de l'oral spontané. *Revue d'Intelligence Artificielle (RIA) 18*(5-6), 709–742.

Walker, M. A. (1994a). Discourse and deliberation : testing a collaborative strategy. In *the 15th conference on Computational linguistics*, Volume 2, Kyoto, Japan, pp. 1205 – 1211.

Walker, M. A. (1994b, October). Experimentally evaluating communicative strategies : the effect of the task. In *Proceedings of the twelfth national conference on Artificial intelligence*, Volume 1, Seattle, Washington, United States, pp. 86–93.

Walker, M. A., I. Langkilde-Geary, H. W. Hastie, J. Wright, and A. Gorin (2002). Automatically training a problematic dialog predictor for the HMOHY spoken dialog system. *Journal of Artificial Intelligence Research 16*, 293–319.

Wilensky, R. (1981). meta-planning : representing and using knowledge about planning in problem solving and natural language understanding. *Cognitive Science 5*, 197–233.

Zeiliger, J., J. Caelen, and J.-Y. Antoine (1997). Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme-machine. *JST-FRANCIL*, 437–446.

Table des matières

Liste des figures

Liste des tableaux

Résumé

Les systèmes de dialogue en langue naturelle sont des interfaces de communication homme-machine susceptibles de souffrir de nombreux problèmes d'incompréhension liés à la complexité de la langue. Nous appelons robustesse leur capacité à faire face aux problèmes d'interprétation. La théorie du grounding (ancrage) de Clark & Schaefer (1989) suggère que les participants à un dialogue cherchent à atteindre la compréhension mutuelle en produisant des preuves de leur compréhension et peut alors permettre d'améliorer la robustesse des systèmes. Cette théorie est confrontée toutefois au problème d'acceptation récursive : afin de savoir si une preuve de compréhension a bien été comprise il est nécessaire d'en fournir une preuve de compréhension et on ne peut au final jamais savoir si quelque chose a été correctement compris. Les modélisations informatiques de l'ancrage qui visent à résoudre ce problème font l'objet de plusieurs simplifications ou sont trop complexes à mettre en oeuvre. Nous proposons d'appuyer la modélisation du processus d'ancrage sur la croyance de compréhension des preuves de compréhension, entraînant une coupure de la récursion ainsi que la possibilité d'ancrer à tort un énoncé. Cette modélisation a été implémentée et adjointe à un système d'interprétation symbolique classique (LTAG + logique de description). L'évaluation du système a été réalisée par simulation sur corpus en générant des dialogues d'ancrage de manière artificielle entre deux instances du système. Ce type d'évaluation permet alors d'explorer automatiquement une grande diversité de problèmes d'ancrage. Les résultats obtenus à l'issue de l'évaluation montrent un gain significatif de compréhension et valident en cela l'approche générale.

Mots clés : systèmes de dialogue, dialogue, robustesse, compréhension mutuelle, ancrage, évaluation, simulation